ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

EDITED BY
DANIEL HILLEL
CYNTHIA ROSENZWEIG
DAVID POWLSON
KATE SCOW
MICHAIL SINGER
DONALD SPARKS

VOLUME FOUR

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

FOUR-VOLUME SET

by Daniel Hillel (Editor-in-Chief)

Hardcover: 2200 pages
Publisher: Academic Press; 1 edition (November 8, 2004)
Language: English
ISBN-10: 0123485304
ISBN-13: 978-0123485304

Book Description

More than ever before, a compelling need exists for an encyclopedic resource about soil the rich mix of mineral particles, organic matter, gases, and soluble compounds that foster both plant and animal growth. Civilization depends more on the soil as human populations continue to grow and increasing demands are placed upon available resources.
The Encyclopedia of Soils in the Environmentis a comprehensive and integrated consideration of a topic of vital importance to human societies in the past, present, and future.

This important work encompasses the present knowledge of the world's variegated soils, their origins, properties, classification, and roles in the biosphere. A team of outstanding, international contributors has written over 250 entries that cover a broad range of issues facing today's soil scientists, ecologists, and environmental scientists.
This four-volume set features thorough articles that survey specific aspects of soil biology, ecology, chemistry and physics. Rounding out the encyclopedia's excellent coverage, contributions cover cross-disciplinary subjects, such as the history of soil utilization for agricultural and engineering purposes and soils in relation to the remediation of pollution and the mitigation of global climate change.

This comprehensive, yet accessible source is a valuable addition to the library of scientists, researchers, students, and policy makers involved in soil science, ecology, and environmental science.

Also available online via ScienceDirect featuring extensive browsing, searching, and internal cross-referencing between articles in the work, plus dynamic linking to journal articles and abstract databases, making navigation flexible and easy. For more information, pricing options and availability visit www.info.sciencedirect.com.

* A distinguished international group of editors and contributors
* Well-organized encyclopedic format providing concise, readable entries, easy searches, and thorough cross-references
* Abundant visual resources — photographs, figures, tables, and graphs — in every entry
* Complete up-to-date coverage of many important topics — essential information for scientists, students and professionals alike

EDITORS

Jerry L Hatfield
National Soil Tilth Laboratory
Ames, IA
USA

Kate M Scow
University of California
Davis, CA
USA

David S Powlson
Rothamsted Research
Harpenden
UK

Michael J Singer
University of California
Davis, CA
USA

Cynthia Rosenzweig
NASA Goddard Institute for Space Studies
New York, NY
USA

Donald L Sparks
University of Delaware
Newark, DE
USA

FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated

as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

CONTENTS

Contents are given as follows: CHAPTER NAME Author(s) Page number

VOLUME 1

VOLUME 2

VOLUME 3

S

VOLUME 4

Table of Contents, Volume 4

SPATIAL VARIATION, SOIL PROPERTIES

R Webster, Rothamsted Research, Harpenden, UK

Introduction

The soil mantles the land more or less continuously, except where there is bare rock and ice, and in a way so complex that no description of it can follow every detail. Further, our knowledge of the soil's properties beneath the surface is fragmentary because it derives from observations on small volumes of material widely separated from one another. Any representation of the whole mantle involves simplification and inference or prediction as to what the soil is like between sampling points with the uncertainty that they entail.

Research in the last 40 years has provided us with quantitative descriptions based on samples. Two main approaches may be discerned. In the first the soil is divided into discrete classes (strata) which are sampled to give estimates of mean values and variances using classical statistics. The other sees soil as a suite of continuous variables and seeks to describe the continuity in terms of spatial dependence and specifically uses geostatistics. The two approaches are not mutually exclusive and they can be combined.

Soil Classification

Peasant cultivators and farmers have for centuries recognized different kinds of soil, and they have divided their land where the soil changes so that they can manage it. In other words, they have classified the soil and land spatially. More formal classification of soil has its roots in nineteenth-century biological taxonomy and practice in geological survey. Finite circumscribed regions are divided into parcels by boundaries, which are sharp lines across which the soil changes in some sense. For any one region, the outcome is a map, technically a choropleth map, showing the region tessellated into spatial classes, which constitute a general-purpose classification. The map may purport to show the classes of some predefined scheme of classification; alternatively the boundaries on it may be drawn where the soil changes more than elsewhere and between which the soil is relatively homogeneous. There are thousands of examples.

The map has usually been accompanied by a text describing each of the classes displayed, with data on individual soil properties from representative sites. The spatial variation for any one soil property thus appears as a stepped function, as in Figure 1a of a transect across a region. Variation within the classes may be acknowledged, but it is not evident. The reality is more like Figure 1b, which is the same transect but now with all the data from sampling at 10-m intervals shown, and for which there is a summary in Table 1. Some of the boundaries can still be recognized where there are large jumps in the data, but others are not so obvious.

By the 1960s, taxonomists were putting numerical limits on the discriminating criteria for consistency. This helped to codify description. It did nothing, however, to quantify the variation in properties that could not be assessed readily in the field; and it was unhelpful to the map-maker who wished to place boundaries where there were maxima in the rate of change in the landscape. Description needed a formal statistical basis, a need first recognized by civil engineers in the 1960s.

Sampling and Estimation

In the classical approach, a soil property, z, takes values at an infinity of points, $\mathbf{x}_i = \{x_{i1}\, x_{i2}\}$, $i = 1, 2, \ldots, \infty$, in a region $\mathcal{R}$. These values, $z(\mathbf{x}_i)$, comprise the population, which has a mean, μ, and variance, here denoted as σ_i^2 signifying the total variance in $\mathcal{R}$. The region is divided into K spatial strata or classes, $\mathcal{R}_k$, $k = 1, 2, \ldots, K$, which are mutually exclusive and exhaustive and which are what the map displays; each has its own mean and variance, denoted μ_k and σ_k^2, respectively. The region is then sampled, and the property at the N sampling points is measured to give data, $z(\mathbf{x}_1), z(\mathbf{x}_2), \ldots, z(\mathbf{x}_N)$, of which n_k belong in class $\mathcal{R}_k$.

If sampling is unbiased, then the mean for the kth class is estimated simply by:

$$\hat{\mu}_k = \bar{z}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} z(\mathbf{x}_i) \quad \text{for} \quad \mathbf{x}_i \in \mathcal{R}_k \qquad [1]$$

The variance within $\mathcal{R}_k$ is estimated from the same sample by

$$\hat{\sigma}_k^2 = s_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} \{z(\mathbf{x}_i) - \bar{z}_k\}^2 \quad \text{for} \quad \mathbf{x}_i \in \mathcal{R}_k \quad [2]$$

In the classical approach, the soil map plus the class means and variances summarize the available information on the spatial variation of z in $\mathcal{R}$. The analysis may be elaborated by adding skewness coefficients, computed from the third moments about the means

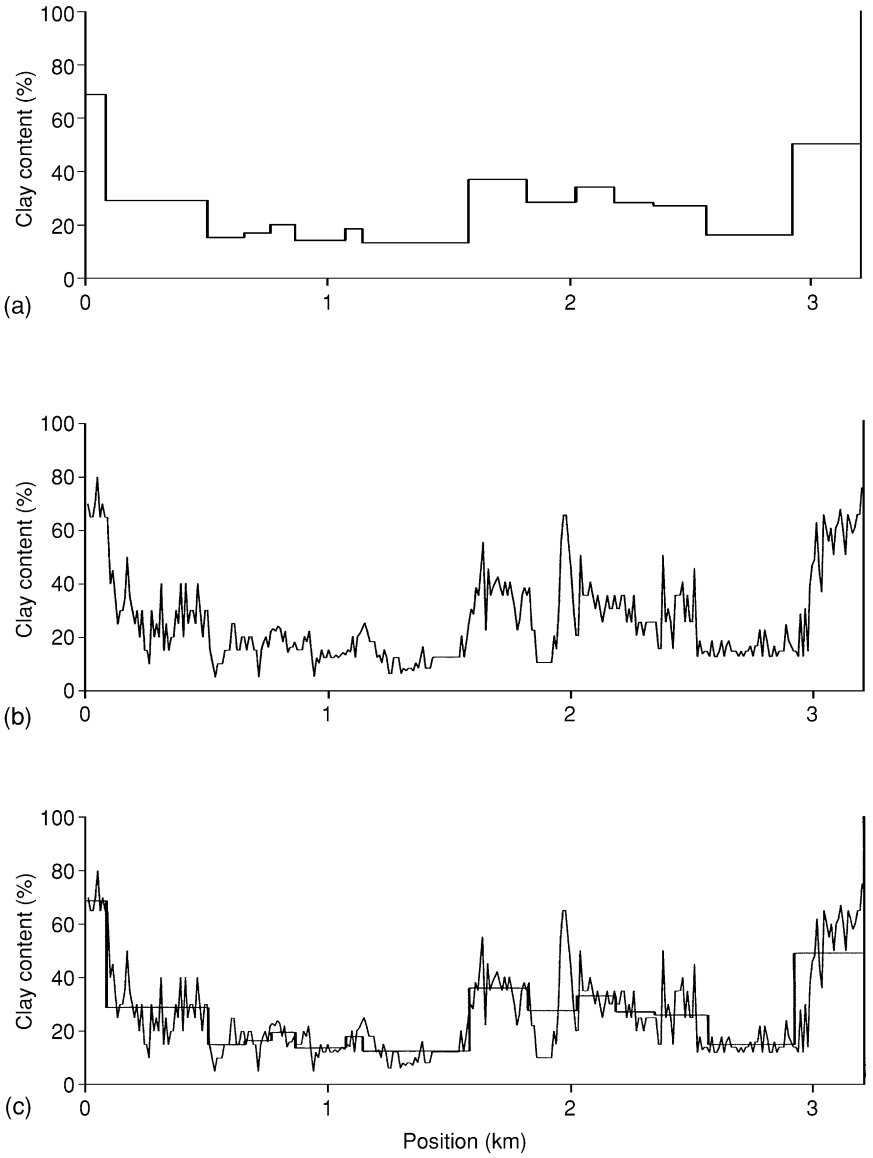


Figure 1 Variation of clay content along a transect: (a) as represented by a classification; (b) actual values measured in topsoil at 10-m intervals; and (c) superimposition of the classification on the reality.

Table 1 Summary of 321 observations of clay content (recorded as percentage by weight) in topsoil and subsoil

	Topsoil	*Subsoil*
Mean (%)	25.6	39.0
Median (%)	20.0	36.0
Variance $(\%)^2$	255.49	936.81
Standard deviation (%)	16.0	30.6
Skewness	1.2	0.2

for the classes, and perhaps higher-order moments. The data might be transformed so that they approximate a normal distribution and so that the means and variances are sufficient summaries.

More importantly, the statistics can be embraced in a single summary. When a soil surveyor subdivides a particular region to display the spatial distribution of the soil, he or she usually tries to create classes of the same categorical level, for example, all soil series or all soil families. Ideally the variances within these are equal, i.e., there is a common within-class variance:

$$\sigma_{\mathrm{W}}^2 = \sigma_k^2 \quad \text{for all} \quad k \qquad [3]$$

The differences between classes can be represented by the between-class variance, σ_{B}^2. This is essentially the variance among the means of the classes, and can be estimated from the data as follows. A quantity B is calculated as the sum of the squares of the differences

Table 2 One-way analysis of variance

Source	*Degrees of freedom*	*Mean square*	*Parameter estimated*
Between classes	$K-1$	$\frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{z}_k - \bar{\bar{z}})^2$	$n^*\sigma_B^2 + \sigma_W^2$
Within classes	$N-K$	$\frac{1}{N-K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{z(\mathbf{x}_{ik}) - \bar{z}_k\}^2$	σ_W^2
Total	$N-1$	$\frac{1}{N-1}\sum_{i=1}^{N}\{z(\mathbf{x}_i) - \bar{\bar{z}}\}^2$	σ_T^2

between the class means and the mean of all the data, $\bar{z}_k - \bar{\bar{z}}$, the latter being the equivalent of $\hat{\mu}$:

$$B = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{z}_k - \bar{\bar{z}})^2 \quad [4]$$

If all the classes are sampled equally, so that $n = n_k$ for all k, then s_B^2 is computed from B simply by:

$$s_B^2 = (B - s_W^2)/n \quad [5]$$

If the n_k are not equal, n in Eqn [5] is replaced by:

$$n^* = \frac{1}{K-1}\left(N - \frac{\sum_{k=1}^{K} n_k^2}{N}\right) \quad [6]$$

and

$$s_B^2 = (B - s_W^2)/n^* \quad [7]$$

The whole can be set out in a one-way analysis of variance, as in Table 2.

For the whole population in $\mathcal{R}$, the variances σ_W^2 and σ_B^2 sum to the total variance, σ_T^2:

$$\sigma_T^2 = \sigma_W^2 + \sigma_B^2 \quad [8]$$

The ratios of these variances describe the relative effects of the spatial classification. One ratio is the intraclass correlation:

$$\rho_i = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2} \quad [9]$$

estimated by:

$$r_i = \frac{s_B^2}{\sigma_W^2 + \sigma_B^2} = \frac{B - s_W^2}{B + (n^* - 1)s_W^2} \quad [10]$$

The last expression enables the intraclass correlation to be calculated directly from the table of analysis of variance. The term derives its name from the fact that it expresses the 'correlation' among individuals within the same class.

The theoretical maximum of ρ_i is 1 when every class is uniform ($\sigma_W^2 = 0$). In practice there is always some variation within the classes, and so $\rho_i < 1$. Its theoretical minimum is zero when all the μ_k are equal, so that $\sigma_B^2 = 0$. Its estimate, r_i, and the estimate s_B^2 are often negative. The usual cause is sampling fluctuation, where the differences between means are small in relation to the variation within the classes, and one can take negative values of r_i as estimates of $\rho_i = 0$.

Another ratio expressing the effectiveness of the classification is simply s_W^2/s_T^2, sometimes called the 'relative variance.' Its complement, $1 - s_W^2/s_T^2$, can be regarded as the proportion of the variance accounted for by the classification, and in this respect it is like the coefficient of determination, R^2, in regression analysis. Like the intraclass correlation, it varies between 1 (uniformity within classes) and 0 (no differences between them), and for large N and K the two have very similar values.

This analysis is applied to the classification of the clay content of the topsoil at Sandford. Table 3 lists the means and variances of the 15 classes, and Table 4 summarizes the analysis of variance.

Table 3 Means and variances of clay content in the topsoil, recorded as percentage by weight, for 15 classes

Class	*Mean*	*Variance*
1	68.8	26.79
2	28.8	110.74
3	15.0	32.14
4	16.5	21.07
5	19.6	12.93
6	13.7	12.33
7	17.9	17.14
8	12.4	21.96
9	35.9	54.69
10	27.6	380.36
11	33.1	42.92
12	27.2	23.23
13	26.0	121.85
14	15.0	9.71
15	49.1	341.78

Table 4 Analysis of variance of clay content of topsoil

Source	*Degrees of freedom*	*Mean square*	*F-ratio*
Between classes	14	3786.67	40.3
Within classes	306	93.94	
Total	320	255.5	

The within-class component of variance (in Table 4) is $s^2_W = 93.94$, and the between-class component, s^2_B, is 176.1. This leads to an intraclass correlation for the classification as 0.65. The complement of the relative variance is 0.63.

There are substantial differences between the variances within the individual classes, so it is something of a liberty to treat them as estimates of the same quantity to arrive at sensible values for r_i and $1 - s^2_W/s^2_T$.

The Geostatistical Approach

The artificiality of imposing boundaries between classes to describe variation that is patently continuous worried quantitatively minded soil scientists, and soil physicists in a particular, for many years. A practicable alternative eluded them, however. They toyed with polynomials, but any such function would have to be of a very high order and could have no generality. The variation was too complex, perhaps chaotic, as Figure 1b shows. Such variation looks as though it might be random. It was this last idea that provided the breakthrough: if the variation appears random then why not treat it as if it were random? This is the basis of modern geostatistics and its approach to describing soil variation.

Random Variables and Random Functions

As in the classical approach, a region $\mathcal{R}$ is regarded as comprising an infinite number of points $\mathbf{x}_i$, $i = 1, 2, \ldots, \infty$. Whereas in the classical approach the values of z at these points constitute the population, in the geostatistical approach this population is assumed to be just one realization of a random process or random function that could generate any number of such populations. At each place $\mathbf{x}$ the soil property is a random variable, $Z(\mathbf{x})$ – notice the capital 'Z' – of many values. For a continuous variable such as hydraulic conductivity or pH, this number is infinite, and the whole process may be regarded as a doubly infinite superpopulation. The random variable at $\mathbf{x}$ has a distribution with a mean and variance and higher-order moments, and the actual value there, $z(\mathbf{x})$, is just one drawn at random from that distribution.

In these circumstances the quantitative description of the variation involves estimating the characteristics of what are assumed to be the underlying random processes. The characteristics include the means and variances, and perhaps higher-order moments, but most importantly the spatial covariances.

The spatial covariance between the variables at any two places $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by:

$$C(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{E}\left[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\}\{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}\right] \quad [11]$$

where $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ are the means at $\mathbf{x}_1$ and $\mathbf{x}_2$, and E denotes the expected value. In practice $C(\mathbf{x}_1, \mathbf{x}_2)$ cannot be estimated, because there is only ever the one realization, and to overcome this apparent impasse assumptions of stationarity must be invoked.

Stationarity

Starting with the first moment, we assume that the mean, $\mu = \mathrm{E}[Z(\mathbf{x})]$, is constant for all $\mathbf{x}$, and so $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ can be replaced by the single value μ, which is estimated by repetitive sampling.

Next, when $\mathbf{x}_1$ and $\mathbf{x}_2$ coincide, Eqn [11] defines the variance, $\sigma^2 = \mathrm{E}[\{Z(\mathbf{x}) - \mu\}^2]$. This is assumed to be finite and, like the mean, to be the same everywhere. Equation [11] is then generalized so that it applies to any pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$ separated by a vector, or lag $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$, so that:

$$\begin{aligned} C(\mathbf{x}_i, \mathbf{x}_j) &= \mathrm{E}\left[\{Z(\mathbf{x}_i) - \mu\}\{Z(\mathbf{x}_j) - \mu\}\right] \\ &= \mathrm{E}\left[\{Z(\mathbf{x})\}\{Z(\mathbf{x} + \mathbf{h})\} - \mu^2\right] \\ &= C(\mathbf{h}) \end{aligned} \quad [12]$$

and this is also constant for any given $\mathbf{h}$. This constancy of the mean and variance and of a covariance that depends only on separation and not on absolute position constitutes second-order stationarity.

Equation [12] shows that the covariance is a function of the lag and only of the lag; it describes quantitatively the dependence between values of Z with changing lag. It is readily converted to the dimensionless autocorrelation by:

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0}) \quad [13]$$

where $C(\mathbf{0}) = \sigma^2$ is the covariance at lag $\mathbf{0}$.

Intrinsic Variation and the Variogram

In many instances the assumption of constant mean throughout a region is untenable, and if the mean changes the variance will appear to increase indefinitely with increasing area. The covariance cannot be defined then, because there is no value for μ to insert in Eqn [12]. Faced with this situation, geostatisticians consider the differences from place to place, and their squares, as follows. For small lag distances, the expected differences are zero:

$$\mathrm{E}\left[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\right] = 0 \quad [14]$$

and the expected squared differences define the variances for those lags:

$$\begin{aligned} \mathrm{E}\left[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2\right] &= \mathrm{var}\left[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\right] \\ &= 2\gamma(\mathbf{h}) \end{aligned} \quad [15]$$

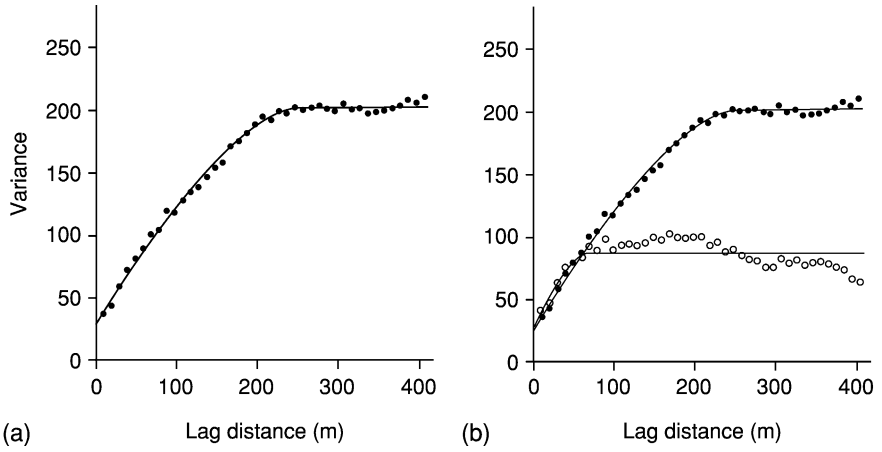


Figure 2 Experimental variograms of clay content of topsoil and the spherical models fitted to them: (a) global variogram with black circles for the experimental semivariances; and (b) the global variogram plus the within-class variogram with open circles for the experimental values.

Eqn [15] gives the variance of the difference at lag $\mathbf{h}$; and the quantity $\gamma(\mathbf{h})$, known as the semivariance, is the variance per point. Equations [14] and [15] constitute the intrinsic hypothesis of geostatistics. Like the covariance, the semivariance depends only on the lag and not on the absolute positions $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$. As a function, $\gamma(\mathbf{h})$ is the variogram, often still called the 'semivariogram.'

If the process $Z(\mathbf{x})$ is second-order stationary then the semivariance and the covariance are equivalent:

$$\begin{aligned}\gamma(\mathbf{h}) &= C(\mathbf{0}) - C(\mathbf{h}) \\ &= \sigma^2\{1 - \rho(\mathbf{h})\}\end{aligned} \qquad [16]$$

If it is intrinsic only, the covariance does not exist, but the semivariance remains valid, and it is this validity in a wide range of circumstances that makes the variogram so useful in summarizing spatial variation.

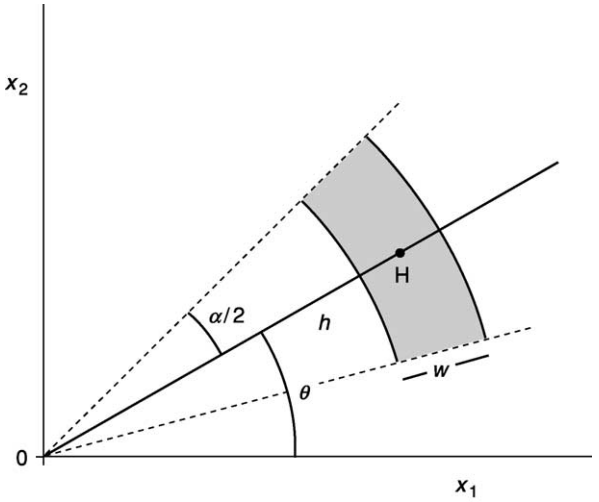


Figure 3 Discretization of the lag in two dimensions for irregularly scattered data. All separations within the gray sector are assigned to lag distance h and direction θ.

Estimating the Variogram

Semivariances are readily estimated from data, $z(\mathbf{x}_1)$, $z(\mathbf{x}_2), \ldots$, by the method of moments:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})}\sum_{i=1}^{m(\mathbf{h})}\{z(\mathbf{x}_i) - z(\mathbf{x}_i+\mathbf{h})\}^2 \qquad [17]$$

in which $m(\mathbf{h})$ is the number of paired comparisons at lag $\mathbf{h}$. By changing $\mathbf{h}$ we obtain a sample or experimental variogram, which can be displayed as a graph of $\hat{\gamma}$ against $\mathbf{h}$. Figure 2a is an example in which the experimental semivariances for the data in Figure 1b are plotted as points against the lag distance, $h = |\mathbf{h}|$, for the one-dimensional transect.

The values of $\mathbf{h}$ define discrete points on the variogram, and so sampling is best planned with regular intervals along a line in one dimension or on a grid in two or three. Otherwise the actual separations have to be placed into 'bins,' with limits in separating distance and also in direction if there is more than one dimension (Figure 3).

Models for Variograms

The underlying variogram, Eqn [15], is a continuous function in as many dimensions as the variable $Z(\mathbf{x})$. The experimental variogram estimates it at a set of points with more or less error and point-to-point fluctuation arising from the sampling. To obtain a variogram to describe the spatial variation in $\mathcal{R}$, a plausible function is fitted to the experimental values. The usual approach is to fit the simplest model that makes sense.

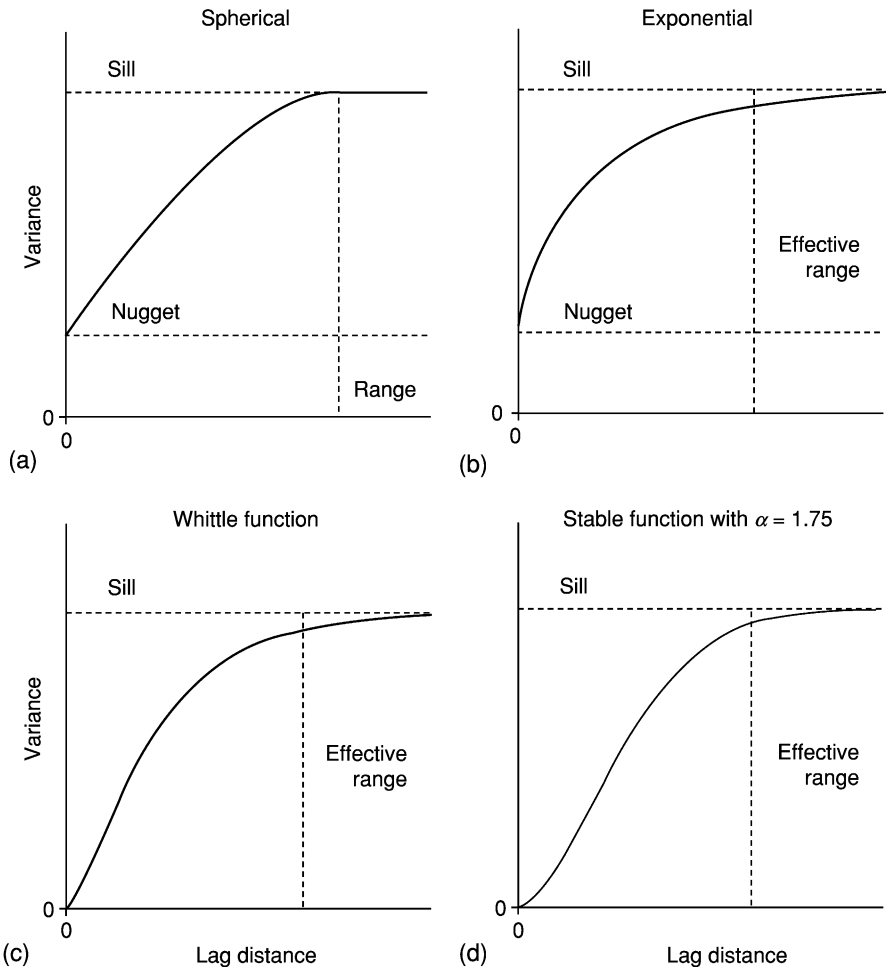


Figure 4 Four kinds of bounded variogram.

Figure 4a shows the principal features of many, if not most, experimental variograms. They are as follows:

1. The variance increases from near the ordinate with increasing lag distance;
2. The variance reaches a maximum at which it remains thereafter;
3. Any simple smooth line or surface placed through the points and projected to the ordinate cuts the ordinate at some value greater than zero.

The model must also be mathematically acceptable in that it cannot give rise to 'negative variances' when random variables are combined. Let $z(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$, be a realization of the random variable $Z(\mathbf{x})$ with covariance function $C(\mathbf{h})$ and variogram $\gamma(\mathbf{h})$, and consider the linear sum:

$$y = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i) \qquad [18]$$

where the λ_i are any arbitrary weights. The variable Y from which y derives is also a random variable, and its variance is given by:

$$\text{var}[Y] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j) \qquad [19]$$

where $C(\mathbf{x}_i - \mathbf{x}_j)$ is the covariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. This variance must be positive or zero; it may not be negative. If $Z(\mathbf{x})$ is intrinsic only then the covariances do not exist, and we must rewrite [19] as:

$$\text{var}[Y] = C(\mathbf{0}) \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \qquad [20]$$

where $\gamma(\mathbf{x}_i - \mathbf{x}_j)$ is the semivariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. The first term on the right-hand side of this equation is eliminated if the weights sum to zero, so that:

$$\text{var}[Y] = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{x}_i - \mathbf{x}_j) \qquad [21]$$

This too must guarantee non-negative variances, and only functions that do that are admissible. They are said to be 'conditional negative semidefinite,' the condition being that the weights in Eqn [21] sum to zero.

There are only a few families of simple functions that satisfy the above criteria. They can be divided into those that are bounded and those that are not. In the first group are the popular spherical and exponential models. Their formulae in their isotropic forms, i.e., for $h = |\mathbf{h}|$, are as follows.

Spherical The spherical function has the following equation:

$$\begin{aligned}\gamma(h) &= c_0 + c\left\{\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \quad \text{for} \quad 0 < h \le a \\ &= c \quad \text{for} \quad h > a \\ &= 0 \quad \text{for} \quad h = 0\end{aligned} \qquad [22]$$

Here $\gamma(h)$ is the semivariance at lag h, and c is the *a priori* variance of the autocorrelated process. The quantity c_0 is the intercept on the ordinate and is known as the 'nugget variance,' a term derived from gold-mining. The combined $c_0 + c$ is known as the 'sill' of the model, and c is the sill of the correlated variance. These quantities are illustrated in **Figure 4a**, and **Figure 2a** shows the function fitted to the experimental variogram of clay content, **Figure 1b**. The values of the parameters, c_0, c, and a, are listed in **Table 5**.

The function has a distance parameter, a; this is its range, also know as its 'correlation range.' It marks the limit of spatial dependence; values at places closer to one another than a are more or less correlated, whereas those further apart are not. It implies that all the variance in $\mathcal{R}$ is encountered within that distance, and in this sense it corresponds to the concept of the representative elementary volume (REV). The spherical function gets its name from the formula for the volume of two intersecting spheres, which are of diameter a.

Table 5 Parameters of spherical models, Eqn [22], fitted to the experimental global and within-class variograms of clay in topsoil and subsoil

		Nugget, c_0	*Sill*, c	*Range*, a (m)
Topsoil	Global	28.2	172.8	265
	Within-class	28.1	59.3	70.9
Subsoil	Global	117.5	551.4	191
	Within-class	108.4	257.2	77.1
Subsoil × *topsoil*	Global	−1.1	210.8	229.6

The semivariance at lag zero is itself zero, and for continuous processes such as most physical properties of the soil, $\gamma(h)$ should increase gradually as h increases from zero. In practice, there are usually insufficient estimates of $\gamma(h)$ near the ordinate to fit a model through the origin, and therefore the conservative approach (described above) is taken. The nugget variance is therefore best regarded as embodying variation within the shortest sampling interval plus any measurement error.

Other functions with the same general form and finite ranges are the bounded linear function (valid in one dimension only), the circular (valid in one and two dimensions, but not in three), and the pentaspherical.

Exponential The equation for the exponential function is:

$$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h}{r}\right)\right\} \qquad [23]$$

in which c_0 and c have the same meanings as before, but now with a distance parameter, r. The exponential model approaches its sill asymptotically and has no definite range therefore. A working range is often taken as $a' = 3r$, at which point the function has reached 95% of c. This model is shown in **Figure 4b**.

Models with reverse curvature at the origin Some variograms appear to approach the origin with decreasing gradients. These may be represented by the general equation:

$$\gamma(h) = c\left\{1 - \exp\left(-\frac{h^\alpha}{r^\alpha}\right)\right\} \qquad [24]$$

in which $0 < \alpha \le 2$. If $\alpha = 2$ we have the Gaussian function. This is at the limit of acceptability and gives rise to unstable prediction. It is best replaced by stable models with $\alpha < 2$; **Figure 4d** is an example. Another recommended function to describe such variation is the Whittle elementary correlation:

$$\gamma(h) = c\left\{1 - \frac{h}{r}K_1\left(\frac{h}{r}\right)\right\} \qquad [25]$$

in which r is again a distance parameter, and K_1 is the modified Bessel function of the second kind (**Figure 4c**). It has the added attraction in that it derives theoretically from diffusion in two dimensions. To all can be added a nugget variance, c_0, if desired.

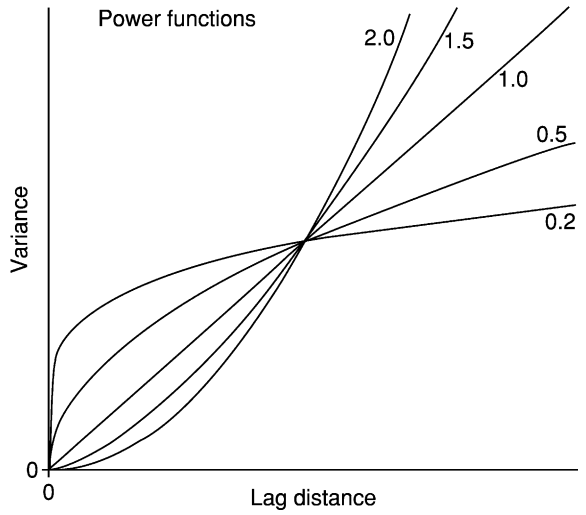


Figure 5 Four valid power functions, Eqn [26] with exponents, α, 0.2, 0.5, 1.0, and 1.5, plus the inadmissible limiting function with $\alpha = 2$.

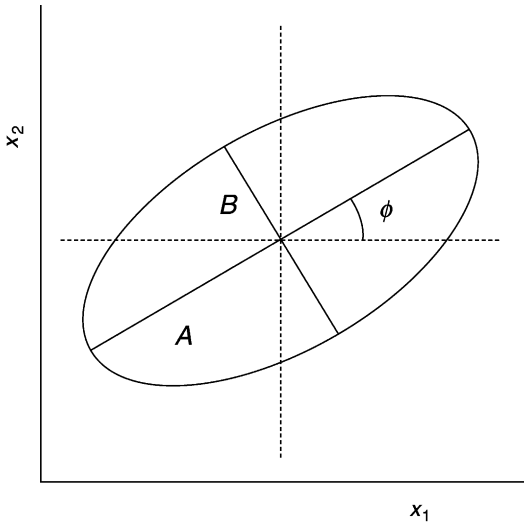


Figure 6 Ellipse showing parameters of anisotropy.

Unbounded models Variograms of processes that are intrinsic but not second-order stationary increase without bound as the lag distance increases. These can usually be fitted by power functions, for which the general equation including a nugget is:

$$\gamma(h) = c_0 + wh^{\alpha} \qquad [26]$$

The parameter w describes the intensity of the process, and the exponent, which must lie strictly between 0 and 2 (these limits are excluded), describes the curvature. If $\alpha < 1$ the curve is convex upward; if it is 1 we have a straight line; and if $\alpha > 1$ the curve is concave upward. The curve with $\alpha = 2$ is a parabola and describes a smoothly continuous process that is not random. Figure 5 shows the curves for several values of α.

Anisotropy The variogram of a two-dimensional process is itself two-dimensional, and if the process is anisotropic so is its variogram, which is then a function of both distance h and direction θ. In the simplest cases, the anisotropy is geometric, meaning that it can be made isotropic by a linear transformation of the coordinates. The transformation is defined by reference to an ellipse:

$$\Omega(\theta) = \sqrt{A^2\cos^2(\theta - \phi) + B^2\sin(\theta - \phi)} \qquad [27]$$

where A and B are the long and short diameters, respectively, of the ellipse, and ϕ is its orientation, i.e., the direction of the long axis (Figure 6). Equation [27] is embodied into the models as follows: For the bounded models, Ω replaces the distance parameter

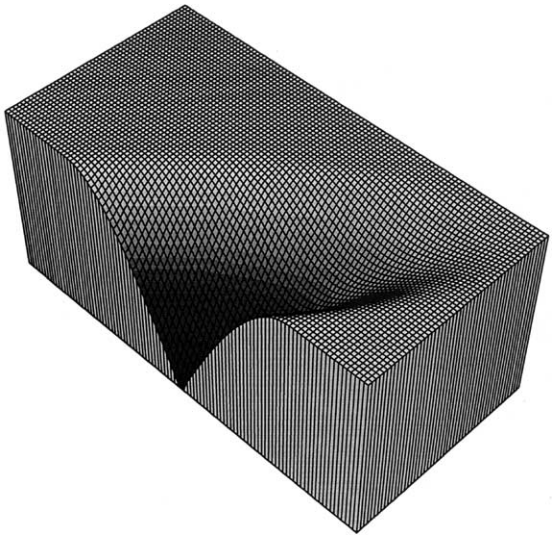


Figure 7 Perspective diagram of the variogram surface of an anisotropic spherical function.

of the isotropic variogram. So, for example, in the exponential:

$$\gamma(h,\theta) = c_0 + c\left\{1 - \exp\left(-\frac{h}{\Omega(\theta)}\right)\right\} \qquad [28]$$

and in the linear function:

$$\gamma(h,\theta) = c_0 + \Omega(\theta)h^{\alpha} \qquad [29]$$

in which $\alpha = 1$. Figure 7 shows an anisotropic spherical function. Notice how the range of the model changes with changing direction.

Combining Trend and Random Fluctuation

The above functions describe processes that are entirely random though correlated. We can represent the processes by the general model:

$$Z(\mathbf{x}) = \mu_V + \epsilon(\mathbf{x}) \quad [30]$$

in which μ_V is the mean, i.e., constant, in some neighbourhood V, and $\epsilon(\mathbf{x})$ is the autocorrelated variance as defined in Eqn [15]. It often happens that such models are unacceptable, either because there is an evident long-range trend across a region or because over short distances the variation appears smooth. In these circumstances, μ_V cannot be treated as constant but must be replaced by a deterministic term, say $u(\mathbf{x})$, that depends on the position $\mathbf{x}$. The model becomes:

$$Z(\mathbf{x}) = u(\mathbf{x}) + \epsilon(\mathbf{x}) \quad [31]$$

If $u(\mathbf{x})$ can describe the variation over the whole of $\mathcal{R}$ it is called 'trend.' If it is local only then is it known as 'drift.' In either event it is usually represented by a low-order polynomial, so Eqn [31] becomes:

$$Z(\mathbf{x}) = \sum_{j=0}^{J} a_j f_j(\mathbf{x}) + \epsilon(\mathbf{x}) \quad [32]$$

in which the a_j are unknown coefficients and the $f_j(\mathbf{x})$ known functions of our choosing.

It is fairly easy, even if somewhat arbitrary, to separate any long-range trend from the short-range, apparently random fluctuation and to estimate the parameters of the two components separately. It is not at all easy to do it where there is short-range drift. In these circumstances it involves a full structural analysis, effectively a process of trial and error.

Combining Classification with Geostatistics

In some instances neither a classification nor a variogram alone can serve to represent spatial variation in soil properties. The choropleth map implies abrupt changes, whereas the variogram is based on a model of random but continuous fluctuation. If there appears to be both kinds of variation then the two approaches may be combined. By recognizing the class boundaries, that is, by combining the information in **Figure 1a** and **1b**, and analyzing the variance (**Table 4**), residuals can be obtained from the class means. Their variance is the residual mean square, and a portion of this is likely to be autocorrelated and have its own variogram. **Figure 2b** shows by the circles an example of a within-class variogram obtained by superimposing the classification on the data. The curve through the points is again that of a spherical model with parameter values as given in **Table 5**.

The variogram of the residuals differs from the variogram of the original data in two important respects:

1. The sill of the fitted model is less by an amount approximately equal to the between-class variance, as expected;
2. The range of the model is much less. This is because the class-to-class variation, which evidently dominated the variation over the whole transect, has been removed to leave only the short-range correlation.

Coregionalization – Simultaneous Variation in Two or More Variables

Any two variables, say z_u and z_v, may be correlated and in particular linearly correlated. That relation is conventionally expressed by the product–moment correlation coefficient:

$$\rho = \frac{\mathrm{cov}[uv]}{\sqrt{\mathrm{var}[u] \times \mathrm{var}[v]}} \quad [33]$$

i.e., the covariance of z_u and z_v divided by the product of their standard deviations.

The two spatial random variables, $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$, may also be spatially intercorrelated in that each is spatially correlated both with itself, i.e., autocorrelated, and with the other. The two variables are then said to be cross-correlated. In these circumstances, the two variables have autovariograms, one each, as defined by Eqn [15] and for present purposes denoted $\gamma_{uu}(\mathbf{h})$ and $\gamma_{vv}(\mathbf{h})$. They also have a cross-variogram, $\gamma_{uv}(\mathbf{h})$, defined by:

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2}\mathrm{E}\left[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x}+\mathbf{h})\}\{Z_v(\mathbf{x}) - Z_v(\mathbf{x}+\mathbf{h})\}\right] \quad [34]$$

If both variables are second-order stationary with means μ_v and μ_v then will both have covariance functions, C_{uu} and C_{vv} as defined in Eqn [12], and a cross-covariance:

$$C_{uv}(\mathbf{h}) = \mathrm{E}[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_v\}] \quad [35]$$

There is also a cross-correlation coefficient, ρ_{uv}, given by:

$$\rho_{uv}(\mathbf{h}) = \frac{C_{uv}}{\sqrt{C_{uu}(0)C_{vv}(0)}} \quad [36]$$

This is effectively the extension of the Pearson product–moment correlation coefficient of Eqn [33] into the spatial domain, and when $\mathbf{h} = 0$ it is the Pearson coefficient.

The cross-covariance is in general not symmetric, i.e.:

$$\mathrm{E}[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_v\}] \neq \mathrm{E}[\{Z_v(\mathbf{x}) - \mu_v\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_u\}] \quad [37]$$

In words, the cross-covariance between $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$ in one direction is different from that in the other, or, expressed another way:

$$C_{uv}(\mathbf{h}) \neq C_{uv}(-\mathbf{h}) \quad \text{or equivalently} \quad C_{uv}(\mathbf{h}) \neq C_{vu}(\mathbf{h}) \quad [38]$$

since:

$$C_{uv}(\mathbf{h}) = C_{vu}(-\mathbf{h})$$

Asymmetry can be envisaged between two soil properties at different depths on a slope as a result of creep or solifluction. The subsoil would tend to lag behind the topsoil. Similarly, irrigation by flooding always from the same end of a field might distribute salts differentially in the direction of flow, but asymmetric covariances have not been reported in the literature as of 2003, as far as I know.

The cross-variogram and the cross-covariance function (if it exists) are related by:

$$\gamma_{uv}(\mathbf{h}) = C_{uv}(\mathbf{0}) - \frac{1}{2}\{C_{uv}(\mathbf{h}) + C_{uv}(-\mathbf{h})\} \quad [39]$$

This quantity contains both $C_{uv}(\mathbf{h})$ and $C_{uv}(-\mathbf{h})$ and in consequence loses any information on asymmetry; it is an even function, i.e., symmetric:

$$\gamma_{uv}(\mathbf{h}) = \gamma_{vu}(\mathbf{h}) \quad \text{for all} \quad \mathbf{h}$$

Cross-semivariances can be estimated in a way similar to that of the autosemivariances by:

$$\hat{\gamma}_{uv}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})}\sum_{i=1}^{m(\mathbf{h})}\{z_u(\mathbf{x}_i) - z_u(\mathbf{x}_i+\mathbf{h})\}\{z_v(\mathbf{x}_i) - z_v(\mathbf{x}_i+\mathbf{h})\} \quad [40]$$

and the sample cross-variogram is formed by simple incrementation of $\mathbf{h}$. There is an equivalent formula for computing the cross-covariances. Notice that there must be numerous places where both z_u and z_v have been measured.

Modeling the Coregionalization

The cross-variogram can be modeled in the same way as the autovariogram, and the same restricted set of functions is available. There is one additional constraint. Any linear combination of the variables is itself a regionalized variable, and its variance cannot be negative. This is assured by adopting the linear model of coregionalization. In it the variable $Z_u(\mathbf{x})$ is assumed to be the sum of independent (orthogonal) random variables, $Y_j^k(\mathbf{x})$:

$$Z_u(\mathbf{x}) = \sum_{k=1}^{K}\sum_{j=1}^{k} a_{uj}^k Y_j^k(\mathbf{x}) + \mu_u \quad [41]$$

in which the superscript k is an index, not a power. There is a similar assumption for $Z_v(\mathbf{x})$. If the assumptions hold then the pair of variables has a cross-variogram:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^{K}\sum_{j=1}^{k} a_{uj}^k a_{vj}^k g^k(\mathbf{h}) \quad [42]$$

The products in the second summation can be replaced by b_{uv}^k to give

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^{K} b_{uv}^k g^k(\mathbf{h}) \quad [43]$$

The quantities are the variances and covariances, e.g., nugget and sill variances, for the independent components of a spherical model. For two variables there are the three nugget variances, b_{uu}^1, b_{vv}^1, and b_{uv}^1, and similarly three for the sills of the correlated variances. The coefficients $b_{uv}^k = b_{vu}^k$ for all k, and, for each k, the matrix of coefficients:

$$\begin{bmatrix} b_{uu}^k & b_{uv}^k \\ b_{vu}^k & b_{vv}^k \end{bmatrix}$$

must be positive definite. Since the matrix is symmetric, it is sufficient that $b_{uu}^k \geq 0$ and $b_{vv}^k \geq 0$ and that its determinant is positive or zero:

$$\left|b_{uv}^k\right| = \left|b_{vu}^k\right| \leq \sqrt{b_{uu}^k b_{vv}^k} \quad [44]$$

This is Schwarz's inequality.

Any number of regionalized variables may be embodied in the linear model of coregionalization. If there are V of them the full matrix of coefficients, $[b_{ij}]$, will be of order V, and its determinant and all its principal minors must be positive or zero.

Schwarz's inequality has the following consequences:

1. Every basic structure, $g^k(\mathbf{h})$, present in the cross-variogram must also appear in the two autovariograms, i.e., $b_{uu}^k \neq 0$ and $b_{vv}^k \neq 0$ if $b_{vv}^k \neq 0$. As a corollary, if a basic structure $g^k(\mathbf{h})$ is absent from either autovariogram it may not be included in the cross-variogram;

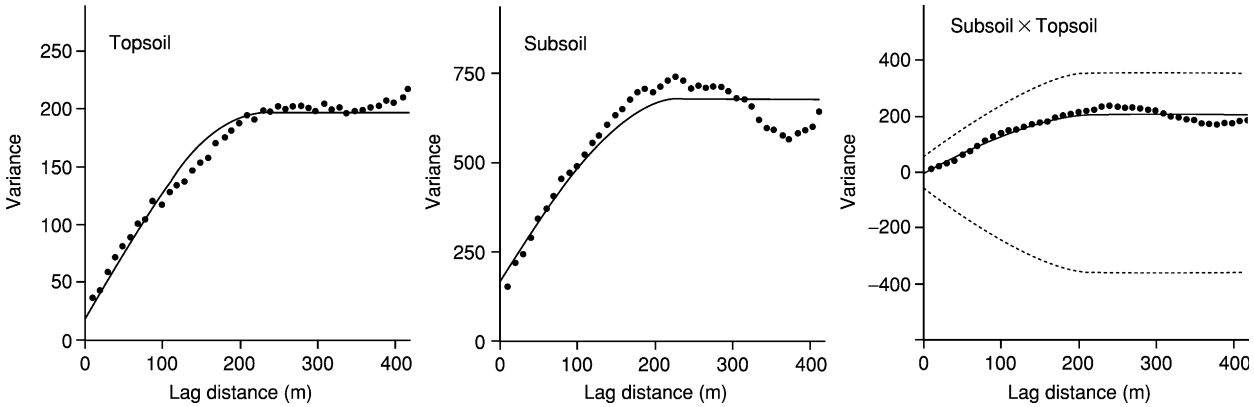


Figure 8 Auto- and cross-variograms of clay content in topsoil and subsoil with the linear model of coregionalization fitted. The dashed lines depict the hull of perfect correlation.

2. Structures may be present in the autovariograms without their appearing in the cross-variogram, i.e., b^k_{uv} may be zero when $b^k_{uu} > 0$ and $b^k_{vv} > 0$.

Parameters of the linear model of the coregionalization with the above constraints can be fitted by iteration. The distance parameters are usually first approximated by fitting models independently to the experimental variograms, and good compromise values are chosen from these. Then with those values fixed the values of the b^k_{uv} are found to minimize the sums of the squares of the residuals, subject to the condition that the solution guarantees non-negative variances, i.e., is conditionally negative semi-definite (CNSD). The validity of the resulting model may be checked by plotting it on a graph of the experimental cross-semivariances plus the limiting values that would hold if the correlation between the variables were perfect. These limits constitute the hull of perfect correlation, which is obtained from the coefficients b^k_{uu} and b^k_{vv} by:

$$\text{hull}\,[\gamma(\mathbf{h})] = \pm \sum_{k=1}^{K} \sqrt{b^k_{uu} b^k_{vv}}\, g^k(\mathbf{h})$$

The line should fit close to the experimental values for the model. It must also fall within the hull to be acceptable. If it lies close to the hull, the cross-correlation is strong; if, in contrast, it is far from the bounds, then the cross-correlation is weak.

Example The clay content of the subsoil was recorded at the same sampling points as those for topsoil in Figure 1b and, in combination with the topsoil data, illustrate the coregionalization. Table 1 contains a summary that includes the Pearson correlation coefficient, which is 0.59, indicating a modest correlation overall. The experimental autovariograms for the two depths together with their cross-variogram are shown in Figure 8 by the black circles. Spherical models can be fitted to all of them, and their parameters are listed in Table 5. Their ranges vary between 191 and 265 m, with a mean of 228 m. If this value is fixed for the coregionalization model the nugget and sill component components of the model are obtained as described above (Table 6). They are somewhat different from those fitted independently, but only somewhat. The final result is shown in Figure 8 by the solid lines through the plotted points. The model evidently fits well.

Table 6 Fitted nugget variance and sill variances of the correlated structure, i.e., the coefficients b^k_{uv} of Eqn [43], of the model of coregionalization of clay in topsoil and subsoil

		Topsoil	*Subsoil*
Nugget variance	Topsoil	17.841	
	Subsoil	−0.799	167.986
Sill variance	Topsoil	178.659	
	Subsoil	211.436	509.723

The dashed lines on the graph for subsoil × topsoil define the hull of perfect correlation. The cross-variogram falls within the hull, as it should, and the moderate distance it keeps from the upper bound is a measure of moderate cross-correlation over the lag distances computed.

Spatial Prediction – Kriging

The variogram and covariance functions are not only elegant mathematical descriptions of the real world of the soil, they are crucial for local estimation, or spatial prediction as it might better be called, by kriging.

'Kriging' is a general term for processes of weighted averaging of data to provide unbiased local estimates of unknown values of a variable with minimum

variance. It is named after D.G. Krige, who developed it for estimating the gold content of ore bodies in South Africa. In the simplest case, for a place $\mathbf{x}_0$ where the mean is unknown (the usual situation), the estimate is formed as:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) \qquad [45]$$

where the $z(\mathbf{x}_i), i = 1, 2, \ldots, N$ are sample data at places $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, and the λ_i are weights. The weights sum to 1 to assure unbiasedness, i.e., $\sum_{i=1}^{N} \lambda_i = 1$, and the variance of the estimate is given by:

$$\begin{aligned}\sigma^2(\mathbf{x}_0) &= \mathrm{E}\left[\left\{\hat{Z}(\mathbf{x}_0) - z(\mathbf{x}_0)\right\}^2\right] \\ &= 2\sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) \\ &\quad - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) \qquad [46]\end{aligned}$$

Here $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ is the semivariance between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\gamma(\mathbf{x}_i, \mathbf{x}_0)$ is the semivariance between the data point $\mathbf{x}_i$ and the target point $\mathbf{x}_0$.

This variance is minimized by solution of the $N + 1$ equations:

$$\begin{aligned}\sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(\mathbf{x}_0) &= \gamma(\mathbf{x}_i, \mathbf{x}_0) \,\forall\, j \\ \sum_{i=1}^{N} &= 1 \qquad [47]\end{aligned}$$

in which the $\psi(\mathbf{x}_0)$ is a Lagrange multiplier introduced for the minimization. The solution yields the optimum weights, and these are inserted into Eqn [45] to give the required estimate at $\mathbf{x}_0$.

The minimized variance is also obtained from the solution as

$$\sigma^2_{\mathrm{OK}}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) + \psi(\mathbf{x}_0) \qquad [48]$$

This particular form of the technique is ordinary punctual kriging – 'ordinary' because it is the most used, and 'punctual' because the estimates are for points of the same size and shape, i.e., the same supports, as the bodies of soil or other material on which the measurements were made. It is readily generalized for estimating blocks, B, larger than the supports of the data. Eqn [45] holds for the averaging, though with B replacing $\mathbf{x}_0$. In Eqn [47] the individual semivariances on the right-hand sides are replaced by the means of the semivariances between the data points and the target block, B, $\bar{\gamma}(\mathbf{x}_i, B)$. Finally, the kriging variance is given by:

$$\sigma^2_{\mathrm{OK}}(B) = \sum_{i=1}^{N} \lambda_i \bar{\gamma}(\mathbf{x}_i, B) + \psi(B) - \bar{\gamma}(B, B) \qquad [49]$$

where $\bar{\gamma}(B, B)$ is the mean semivariance within B, i.e., the within-block variance.

When the kriging equations are solved, it usually turns out that only the few points nearest to the target point or block carry any appreciable weight; the weights of the others are so close to zero that they can be ignored. Kriging is thus a local weighted average. It has two other intuitively attractive features:

1. Where data points are clustered the weight of the cluster is divided among its members so that the individual weights are small compared with those of isolated points;
2. Where data points lie approximately in a line between the target and more distant points they screen the latter, which tend as a result to have virtually no weight, however close they are to the target.

This also has practical implications. The kriging systems, Eqn [47], need never be large; they are swiftly solved, and instabilities with matrix inversion are rare.

It is now evident why the variogram, or the equivalent covariance function, is so important; the kriging systems need values drawn from it. As above, these must not give rise to negative kriging variances, and so a valid function must be fitted to the experimental variogram.

Ordinary kriging will serve in some 90% of cases; it is the 'work horse' of practical geostatistics. It requires the fewest assumptions and the least knowledge. Kriging has been much used to map soil properties, including concentrations of plant nutrients, salinity, trace element contents, and nematode infestation. The kriging systems are solved at close intervals on a grid, from which isarithms, 'contours,' of the estimates can be drawn by other graphics programs. This has led to the application of kriging in land reclamation and precision agriculture. The kriging variances can be mapped similarly; patches of large variance coincide with sparse sampling, and so the maps can show where denser sampling is necessary or desirable to achieve more reliable estimates.

Kriging can be elaborated to embody other knowledge. Universal kriging takes into account known or estimated trend in the target variable. The underlying model is that of Eqn [31], usually with a simple

low-order polynomial for the trend, as in Eqn [32]. Another way of dealing with trend is to regard the target variable as an intrinsic random function of order k, $k > 0$ (IRF-k), and working with generalized covariances. Measurements on related subsidiary variables can be combined with those of the target variable by co-kriging, in which the semivariances are drawn from the model of coregionalization, Eqn [43], with all combinations of u and v in the model. Kriging with external drift embodies knowledge of the trend in related subsidiary variables and is an extension of universal kriging. Indicator kriging and disjunctive kriging form linear combinations of nonlinear transforms of data to estimate the probabilities of variables' exceeding specified threshold values, and these techniques are of potential value in statutory controls of pollution and restoration of contaminated land.

See also: **Spatial Patterns**; **Statistics in Soil Science**

Further Reading

Chilès J-P and Delfiner P (1999) *Geostatistics: Modeling Spatial Uncertainty*. New York: John Wiley.

Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Olea RA (1999) *Geostatistics for Engineers and Earth Scientists*. Boston, MA: Kluwer Academic Publishers.

Webster R (1985) Quantitative Spatial Analysis of Soil in the Field. *Advances in Soil Science* 3: 1–70.

Webster R and Oliver MA (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

SPECIFIC SURFACE AREA

K D Pennell, School of Civil & Environmental Engineering, Atlanta, GA, USA

Introduction

The surfaces of soil particles play a critical role in many processes, including chemical reactions, contaminant adsorption, colloid filtration, and water imbibition and drainage. The specific surface area of soils and soil constituents can range from less than $0.1\,m^2\,g^{-1}$ or $1 \times 10^2\,m^2\,kg^{-1}$ up to $800\,m^2\,g^{-1}$ or $8 \times 10^5\,m^2\,kg^{-1}$. Soils consisting primarily of sands (i.e., particle diameters of 0.05–2.0 mm or 5×10^{-5} m to 2×10^{-3} m) possess relatively small specific surface areas, usually less than $0.5\,m^2\,g^{-1}$ or $5 \times 10^2\,m^2\,kg^{-1}$. In contrast, soils containing appreciable amounts of clay minerals and organic matter tend to have much larger specific surface areas (Table 1). Quantification of specific surface area may involve direct physical measurement of particle size and shape, adsorption of probe molecules from either the gas or aqueous phases, or the retention of polar liquids. The adsorption of nitrogen (N_2) gas, in conjunction with the Brenauer–Emmett–Teller (BET) equation, is the most common method of surface area determination. However, it is widely recognized that N_2 does not access the interlayer surfaces of expandable clay minerals upon drying. To overcome this limitation, the retention of polar compounds such as ethylene glycol monoethyl ether (EMGE) has been utilized to measure the total specific surface area (i.e., internal + external) of soils and expandable clay minerals. However, specific surface area values obtained by these two methods (i.e., N_2/BET and EGME) may be similar or divergent, depending upon sample composition and pretreatments. Therefore, it is important to recognize the dependence of specific surface area data on the measurement technique, and to select one or more methods that are appropriate for the system of interest.

Direct Physical Measurement

Direct physical measurement of specific surface area typically involves the use of light or electron microscopy to determine the shape and dimensions of individual soil particles. Such observations are often supplemented with X-ray diffraction measurements to assess crystallographic structure and interlayer spacing of clay minerals. Provided that a characteristic particle shape and size can be determined, the specific surface area can be obtained from mass–volume relationships. For example, the specific surface area (A_s) of a spherical particle may be calculated in the following manner:

$$A_s = \frac{4\pi r^2}{\rho_s V_s} = \frac{4\pi r^2}{\rho_s \frac{4}{3}\pi r^3} = \frac{3}{\rho_s r} \qquad [1]$$

where r is the radius of the solid particle, ρ_s is the density of the solid, and V_s is the volume of the solid. Using this approach, the specific surface area of quartz sand with a particle diameter of 1.0 mm

Table 1 Comparison of specific surface area values obtained by N_2 gas adsorption and ethylene glycol (EG) or ethylene glycol monoethyl ether (EGME) retention

Sample	Organic carbon content ($g\,kg^{-1}$)	Specific surface area ($\times 10^3\ m^2\ kg^{-1}$)	
		N_2/BET method	EG/EGME method
Kaolinite (KGa-1)[a]	0.0	10.05	16.0
Montmorillonite (SWy-1)[a]	0.0	31.82	662.0
Montmorillonite (SAz-1)[a]	0.0	97.42	820.0
Wyoming bentonite[b]	0.0	65.0	372.0
Lula aquifer sand[c]	0.1	7.7	10.5
Boston silt[b]	26.6	28.6	46.0
Webster soil[d]	33.2	8.2	168.4
Ashurst soil[b]	45.5	6.3	25.8
Houghton muck[d]	445.7	0.8	162.9

[a]van Olphen H and Fripiat JJ (1979) *Data Handbook for Clay Minerals and Other Non-metallic Minerals*, pp. 203–211. New York: Pergamon Press.
[b]Call F (1957) The mechanism of sorption of ethylene dibromide on moist soils. *Journal of the Science of Food and Agriculture* 8: 630–639.
[c]Rhue RD, Rao PSC, and Smith RE (1988) Vapor-phase adsorption of alkylbenzenes and water on soils and clays. *Chemosphere* 17: 727–741.
[d]Pennell KD, Boyd SA, and Abriola LM (1995) Surface area of soil organic matter reexamined. *Soil Science Society of America Journal* 59: 1012–1018.
BET, Brunauer–Emmett–Teller.

or 1.0×10^{-3} m and a particle density of $2.65\,g\,cm^{-3}$ or $2.65 \times 10^3\,kg\,m^{-3}$ would be $2.26 \times 10^{-3}\,m^2\,g^{-1}$ or $2.26\,m^2\,kg^{-1}$. In practice, specific surface area values of sands determined by direct observation are often several orders-of-magnitude smaller than measured values due to presence of nonspherical particles, surface roughness, and fine particles (Table 1).

An analogous mass–volume approach can be used to estimate the specific surface area of clay minerals, provided that the structural formula and unit cell dimensions are known. For example, consider a montmorillonite with a nominal structural formula of $K_{0.66}Si_{8.0}(Al_{3.34}Mg_{0.66})O_{20}(OH)_4$ and unit cell dimensions of $a = 0.5$ nm or 5×10^{-10} m, $b = 0.9$ nm or 9×10^{-10} m, and $c = 9.5$ nm or 9.5×10^{-10} m. Assuming that the particle density is approximately $2.8\,g\,cm^{-3}$ or $2.8 \times 10^3\,kg\,m^{-3}$, and that the edge area (i.e., c-dimension) is negligible compared with the area of the basal surfaces (i.e., a- and b-dimensions), the specific surface area of the montmorillonite can be estimated in the following manner:

$$A_s = \frac{2ab}{\rho_s V_s}$$

$$= \frac{2(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)}{(2.8 \times 10^3 kg\,m^{-3})(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)(9.5 \times 10^{-10}\,m)}$$

$$= 7.52 \times 10^5 m^2\,kg^{-1} \quad [2]$$

Using a slight variation of eqn [2], the specific surface area of the montmorillonite can also be estimated from the molecular weight of the unit cell and Avogadro's number (N_A):

$$A_s = \frac{(2ab)(N_A)}{MW_{mont}}$$

$$= \frac{2(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)(6.022 \times 10^{23}\,mol^{-1})}{7.447 \times 10^{-1}\,kg\,mol^{-1}}$$

$$= 7.52 \times 10^5 m^2\,kg^{-1} \quad [3]$$

The relationships shown in eqns [2] and [3] are applicable to clay minerals existing as flat, plate-like structures with a thickness corresponding to that of the unit cell. When water is removed from expandable clay minerals, the interlayers collapse. The resulting clay particle will then consist of several unit cells stacked on top of one another. If the number of unit cells contained in a collapsed montmorillonite particle is 10, the resulting surface area would be approximately $75\,m^2\,g^{-1}$ or $7.5 \times 10^4\,m^2\,kg^{-1}$, which is consistent with specific surface area values reported for dry montmorillonite samples based on N_2/BET analysis (Table 1).

The mass–volume approaches described above are generally limited in applicability to clean sands or pure clay mineral samples. Many soil constituents, including metal oxides and organic matter, exist as irregular or poorly defined amorphous structures which are virtually impossible to characterize. Furthermore, the specific surface area of natural soils cannot be treated as a strictly additive property due to surface coatings and mineral-organic matter associations. Except in a few limited cases (e.g., clean sands), the specific surface area of a whole soil should not be estimated using a summation procedure based on the surface area contributions of individual soil constituents.

Adsorption from Solution

The adsorption of dissolved molecules from solution can be used to estimate the specific surface area of a solid, provided that the resulting adsorption isotherm exhibits a limiting or maximum value. Such adsorption isotherms are classified as Type I, and can be described by the Langmuir adsorption model. The Langmuir model can be derived by assuming that, once equilibrium is reached, the rate of solute adsorption on to open surface sites is equal to the rate of solute desorption from occupied surface sites. Several assumptions are inherent to the Langmuir model, including: (1) adsorption is localized or site-specific; (2) no interactions occur between adsorbed molecules; (3) the energy of adsorption is constant for all adsorption sites; (4) the adsorption capacity of the solid is limited; and (5) the maximum adsorption capacity corresponds to monolayer coverage. Although the model was originally developed to describe the adsorption of gases on solids, for solid–liquid systems the Langmuir equation may be written as:

$$C_s = \frac{C_{s,max}\,\beta\,C_a}{1+\beta\,C_a} \qquad [4]$$

where C_s is the solid-phase concentration of solute at equilibrium, $C_{s,max}$ is the maximum solid-phase concentration, β is the ratio of the adsorption and desorption rates, and C_a is the aqueous-phase concentration of solute at equilibrium. A series of Langmuir isotherms is presented in **Figure 1** to illustrate the effect of increasing the value of β from 0.01 to $0.25\,l\,mg^{-1}$ with $C_{s,max}$ fixed at $1.0\,g\,kg^{-1}$. As the value of β increases, the rate at which the adsorption isotherm approaches the maximum sorption capacity of the solid ($C_{s,max}$) increases. However, the shape or steepness of the isotherm has no bearing on the maximum sorption capacity, which is used to calculate specific surface area.

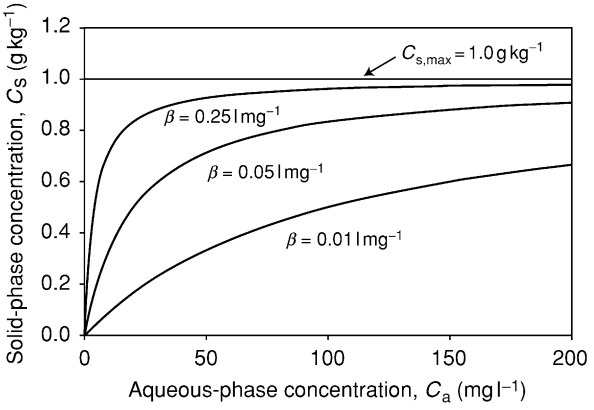


Figure 1 Effect of changes in the value of the β parameter (0.025, 0.05, and $0.01\,l\,mg^{-1}$) on Langmuir adsorption isotherms when $C_{s,max} = 1.0\,g\,kg^{-1}$.

In practice, the parameters β and $C_{s,max}$ can be obtained by directly fitting eqn [4] to experimental adsorption isotherm data (C_s versus C_a) using a nonlinear, least-squares regression procedure. Alternatively, eqn [4] can be rearranged to yield the linear form of the Langmuir equation:

$$\frac{C_w}{C_s} = \frac{1}{\beta\,C_{s,max}} + \frac{C_w}{C_{s,max}} \qquad [5]$$

Here, C_w/C_s (y-axis) is plotted against C_w (x-axis) and a linear regression procedure is then used to obtain a slope equal to $1/C_{s,max}$ and an intercept value equal to $1/\beta C_{s,max}$. With minor manipulation, the desired parameters are obtained as follows: $C_{s,max} = 1/\text{slope}$ and $\beta = (1/C_{s,max})(1/\text{intercept})$.

Several organic molecules have been used in conjunction with the Langmuir equation to determine specific surface area. In the past, organic dyes such as methylene blue were utilized because their concentration in solution could be determined by colorimetric analysis. More recently, cationic surfactants exhibiting visible or ultraviolet (UV) light absorbance have been employed for surface area determination. The cationic surfactant most widely used for this purpose is cetyl pyridinium bromide (CBP), which has a strong absorbance peak at wavelength of 259 nm or 2.59×10^{-7} m. On most mineral surfaces, adsorbed CPB molecules form a bilayer (i.e., double layer), yielding an effective molecular area (A_m) of $0.27\,nm^2$ or $2.7 \times 10^{-19}\,m^2$. If the measured value of $C_{s,max}$ for a nonexpanding clay mineral sample was $1.0\,g$ CPB kg^{-1} solid, the specific surface area would be calculated as:

$$\begin{aligned} A_s &= \frac{(C_{s,max})(N_A)(A_m)}{MW_{CBP}} \\ &= \frac{(1.0\,g\,kg^{-1})(6.22\times10^{23}\,mol^{-1})(2.7\times10^{-19}\,m^2)}{384.45\,g\,mol^{-1}} \\ &= 4.23\times10^{2}\,m^2\,kg^{-1} \end{aligned} \qquad [6]$$

Iron and aluminum oxides possess relatively low surface charge densities, and as a result CPB may not form complete bilayers on these surfaces. Therefore, soil samples are often treated to remove oxides prior to surface area analysis by CPB adsorption. In the case of expandable clay minerals such as montmorillonite, the adsorbed CPB bilayer on the interlayer surfaces is shared and therefore yields an effective molecular area of $0.54\,nm^2$ or $5.4 \times 10^{-19}\,m^2$. In addition, the external surface area must be obtained independently using the N_2/BET method in order to compute the contribution of internal surfaces to the overall adsorption of CPB.

Adsorption from the Gas Phase

The adsorption of gases is frequently used in conjunction with the BET equation to measure the specific surface area of soils and soil constituents. Gas adsorption on dry solid surfaces typically conforms to a Type II isotherm, characterized by the formation of multiple layers of adsorbed molecules. Derivation of the BET equation is based on the Langmuir model, modified to account for multilayer formation. The underlying assumptions of the BET equation are: (1) the heat of adsorption for the first layer is constant; (2) the heat of adsorption for the second and all succeeding layers is constant and equal to heat of condensation; (3) adsorption and desorption can only occur from exposed layers; and (4) the assumptions of the Langmuir model apply to each layer. Although originally derived on a molar basis, the BET equation can be expressed in terms of the solid-phase concentration (C_s) as:

$$C_s = \frac{(C_{s,mon})[(\alpha\ P/P_0)/(1-P/P_0)]\left[1-(n+1)(P/P_0)^n+n(P/P_0)^{n+1}\right]}{1+(\alpha-1)(P/P_0)-\alpha(P/P_0)^{n+1}} \quad [7]$$

where $C_{s,mon}$ is the solid-phase concentration at monolayer coverage, P is the vapor pressure, P_0 is the saturated vapor pressure, and n is the total number of adsorbed layers. The dimensionless parameter α is related to the heat of adsorption and is defined as:

$$\alpha = e^{[(Q_a-Q_c)/RT]} \quad [8]$$

where Q_a is the heat of adsorption on the exposed surface, Q_c is the heat of condensation of the liquid adsorbate, R is the ideal gas constant, and T is temperature. The BET eqn [7] reduces to the Langmuir eqn [4] when number of adsorbed layers is limited to one ($n=1$). If the number of adsorbed layers approaches infinity ($n=\infty$), eqn [7] reduces to the simplified form of BET equation commonly used for surface area determination:

$$C_s = \frac{(C_{s,mon})(\alpha)(P/P_0)}{(1-P/P_0)[1-P/P_0+\alpha(P/P_0)]} \quad [9]$$

A series of BET isotherms is shown in Figure 2 for α equal to 1, 10, and 100, with $C_{s,mon}$ fixed at $1\,g\,kg^{-1}$ and n equal to infinity. As the value of α increases, the inflection point or 'knee' of the BET adsorption isotherm becomes more evident. The inflection in the BET adsorption isotherm corresponds approximately to the point of monolayer coverage. Below the inflection point, gas adsorption occurs on exposed surfaces. The gradual increase in adsorption above the inflection point corresponds to multilayer formation on the surface, while the steep asymptotic rise in adsorption at relative vapor pressures approaching unity ($P/P_0=1$) corresponds to liquid condensation.

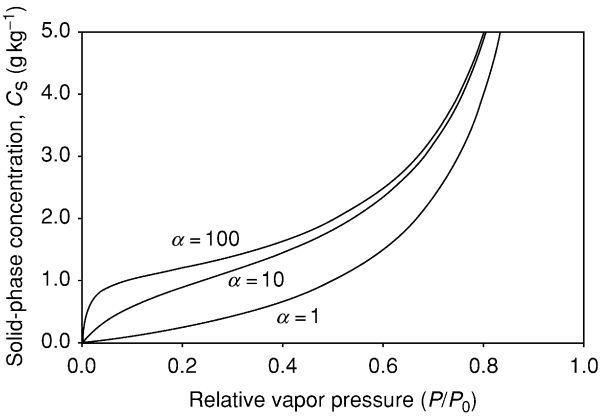


Figure 2 Effect of changes in the value of the α parameter (1, 10, and 100) on Brunauer–Emmett–Teller adsorption isotherms when $C_{s,man} = 1\ g\ kg^{-1}$ and $n =$ infinity.

The most common gas used for BET surface area analysis is nitrogen (N_2), although noble gases such as krypton (Kr) and argon (Ar) are used occasionally for solids possessing very small surface areas (less than approx. $0.1\,m^2\,g^{-1}$ or $1.0\times10^2\,m^2\,kg^{-1}$). Several volatile organic compounds (VOCs), including benzene, *p*-xylene, EGME, and water vapor have also been used for BET surface area analysis. Regardless of the gas or vapor used, the configuration of the adsorbed molecules on the surface must be known or estimated. The most common approach used to estimate the cross-sectional area of an adsorbed molecule (A_m) is related to the liquid density (ρ_l):

$$A_m = 1.091\left(\frac{MW}{\rho_l N_A}\right)^{2/3} \quad [10]$$

The coefficient of 1.091 in eqn [10] is based on the assumption of an ideal hexagonal packing of adsorbed molecules on the surface. For N_2, the value of A_m obtained using eqn [10] is $0.162\,nm^2$ or $1.62\times10^{-19}\,m^2$.

The adsorption of gases on solids can be measured experimentally using several methods. The most common method is based on the change in vapor pressure following the introduction of gas into a small glass bulb containing a dry soil sample. To obtain the adsorption isotherm, the volume of gas adsorbed is computed for each incremental gas dosage based on the change in vapor pressure at equilibrium. Upon reaching the saturated vapor pressure ($P/P_0=1$), the process may be reversed (the vapor pressure is incrementally reduced) to obtain a desorption isotherm, from which pore size analysis

can be performed. Several automated instruments based on this principle are available from commercial vendors. In general, automated surface area instruments report gas adsorption data as the volume of gas adsorbed per gram of soil (e.g., $1\,kg^{-1}$) at standard temperature and pressure (STP). Volumetric gas adsorption data are converted to a mass basis (grams per kilogram) using the molar volume of an ideal gas at STP ($22.414\,l\,mol^{-1}$) and molecular weight of the gas (e.g., $N_2 = 28.02\,g\,mol^{-1}$). The second experimental approach used to measure gas or vapor adsorption is based on the continuous introduction of gas stream at constant vapor pressure. Once equilibrium is attained, the soil sample is either weighed or extracted to determine the amount of adsorbed gas. The vapor pressure of the gas flow stream is then incrementally increased over the desired vapor pressure range to obtain an adsorption isotherm.

To obtain values for the two unknown parameters in the BET equation, α and $C_{s,mon}$, eqn [9] can be fit directly to the experimental adsorption data using a nonlinear, least-squares regression procedure. Alternatively, the experimental data can be expressed using the linear form of the BET equation:

$$\frac{P/P_0}{C_s(1-P/P_0)} = \frac{1}{C_{s,mon}\ \alpha} + \frac{(\alpha-1)P/P_0}{C_{s,mon}\ \alpha} \qquad [11]$$

Here, $(P/P_0)/[C_s(1 - P/P_0)]$ (y-axis) is plotted against P/P_0 (x-axis), and a linear regression procedure is used to obtain values for the slope, which is equal to $(\alpha - 1)/(\alpha\ C_{s,mon})$ and the intercept, which is equal to $1/(\alpha\ C_{s,mon})$. The two fitting parameters can then be obtained as follows: $\alpha = [(\text{slope})(1/\text{intercept}) + 1]$; and $C_{s,mon} = (1/\alpha)(1/\text{intercept})$. As a general rule, the value of α should be greater than 20 and the amount of sample should yield a total surface area between 40 and 120 m^2. In addition, it is often recommended that eqn [11] be applied to adsorption data over a relative vapor pressures range of 0.05–0.35. The specific surface area is then computed using the fitted value of $C_{s,mon}$ obtained from the experimental adsorption data and the cross-sectional area of the adsorbed molecule (A_m) from eqn [10]. For example, if $C_{s,mon}$ was determined to be $0.5\,g\,kg^{-1}$, the N_2/BET specific surface area (A_s) would be calculated in the following manner:

$$A_s = \frac{(C_{s,mon})(N_A)(A_m)}{MW_{N_2}}$$
$$= \frac{(0.5\,g\,kg^{-1})(6.022\times10^{23}\,mol^{-1})(1.62\times10^{-19}\,m^2)}{28.02\,g\,mol^{-1}}$$
$$= 1.74\times10^{3}\,m^2\,kg^{-1} \qquad [12]$$

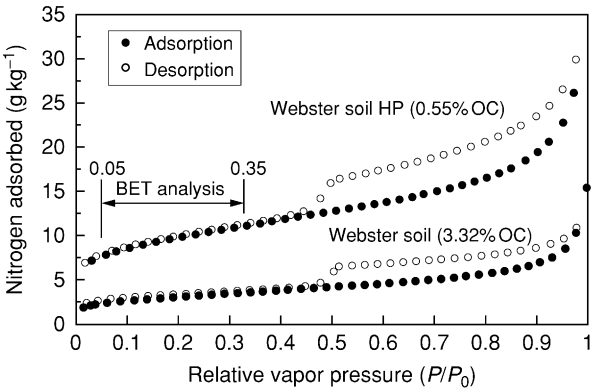


Figure 3 Adsorption and desorption of N_2 on hydrogen peroxide (HP)-treated and untreated Webster soil.

Treatment and preparation of soil samples prior to N_2 adsorption can strongly influence the measured specific surface area. For example, the removal of soil organic matter will often lead to increased specific surface area values; this has been attributed to the exposure of mineral surfaces covered by organic matter and the division of mineral particles held together by organic matter bridging. Nitrogen adsorption and desorption isotherms for hydrogen peroxide (HP)-treated and untreated Webster soil are shown in Figure 3. The resulting surface area values for the untreated and HP-treated samples were 0.79 and $7.38\,m^2\,g^{-1}$ or 7.9×10^2 and $7.38\times10^3\,m^2\,kg^{-1}$, respectively.

As noted previously, water must be removed from soil samples prior to the measurement of gas adsorption. The drying or dehydration process is known to collapse the interlayer space of expandable clay minerals, which are then not accessible to inert gases such as N_2. In addition, electron microscopy suggests that air-drying of soil samples results in the collapse and shrinkage of soil humic acid, whereas freeze-drying maintains a complex structural network characteristic of soil organic matter. Thus, freeze-drying of soils to remove water prior to N_2/BET analysis will tend to result in larger surface area values, which may be more representative of natural conditions.

Retention of Polar Liquids

The use of polar liquids such as ethylene glycol (EG) and EGME for surface area determination was based on the need to develop a relatively simple experimental technique that could be used to quantify the total surface area (i.e., internal + external) of expandable clay minerals. Due to the attractive forces between polar molecules and exchangeable cations, EG and

EGME are able to penetrate the interlayer space of expandable clay minerals. In practice, a dry soil or clay sample is placed in a vacuum desiccator containing EG or EGME as a free liquid or mixed with calcium chloride to maintain a constant vapor pressure of EG or EGME. Liquid EG or EGME is then added as drops to the solid sample until complete wetting is achieved. A suction of approximately 13.332 Pa or 0.1 mmHg is applied, and the sample is weighed over time until a constant weight is obtained. The specific surface area of the solid is then calculated as follows:

$$A_s = \frac{W_{EG/EGME}}{(W_{OD})(\text{EG/EGME conversion factor})} \quad [13]$$

where $W_{EG/EGME}$ is the weight of EG or EGME retained by the soil at the applied suction and W_{OD} is the oven-dry weight of the solid. The method is predicated on the assumption that EG and EGME are retained on solid surfaces at monolayer coverage. The mass-surface area conversion factors for EG and EGME are $3.1 \times 10^{-7}\,\text{kg}\,\text{m}^{-2}$ and $2.86 \times 10^{-7}\,\text{kg}\,\text{m}^{-2}$, respectively, and were derived from retention measurements performed on reference clay minerals with known unit cell dimensions and structural formula. It is usually recommended that the specific surface area of a reference expandable clay mineral, such as Wyoming montmorillonite (SWy-1), be measured to confirm the value of the conversion factor and to ensure that experimental procedure is operating properly. Although EG was used during the initial development of the retention method, more recent studies have employed EGME due to the shorter times required to attain a constant retention value.

Based on the success of the EG/EGME retention method for determining the total surface area of expandable clay minerals, the technique has subsequently been applied to natural soils to obtain a measure of total specific surface area. However, use of the EG/EGME retention method is complicated by the fact that polar molecules tend to form multilayers in the vicinity of cation exchange sites prior to complete monolayer coverage, retention may be influenced by the species of exchangeable cation, and the EG/EGME may partition into soil organic matter. As a result, the retention of EG or EGME by natural soils represents a measure of the uptake capacity of soils for a polar adsorbate, rather than a strict measure of total specific surface area. Despite this shortcoming, the EG/EGME retention method is an appropriate method for measuring the total specific surface area of pure clay minerals and soils that have been treated to remove organic carbon.

Selection of Surface Area Measurement Technique

Although the methods used to determine experimentally the specific surface area of soils and soil constituents are relatively well established, interpretation and appropriate use of the resulting data can be problematic. Complications arise primarily from two factors: (1) sample pretreatments including drying, organic-matter removal, and metal-oxide removal can markedly alter measured values of specific surface area; and (2) the use of different probe molecules and measurement techniques can result in similar or very divergent specific surface area values for the same soil. For example, soils that contain only trace amounts of organic matter and no expandable clay minerals will yield relatively consistent specific surface area values regardless of the method (see kaolinite and Lula aquifer sand in Table 1). In addition, the external and internal specific surface area of pure expandable clay mineral samples can be determined using a combination of N_2/BET (external surface area) and EG/EGME retention (total surface area; see SAz-1 and SWy-1 montmorillonites in Table 1). However, for natural soils containing appreciable amounts of organic matter and expandable clay minerals, the selection of an appropriate measurement technique, as well as the interpretation of specific surface area data, is far more difficult. This is particularly relevant if the intent is to quantify the total or internal surface area of a soil available under natural conditions using the EG/EGME retention method. Due to the potential for partitioning or dissolution of polar molecules into soil organic matter and interactions with exchangeable cations, the EG/EGME retention method is not suitable for most natural soils. In effect, EG/EGME retention provides an indication of the capacity of the soil to take up polar molecules, in contrast to a measure of total specific surface area. Despite the known limitations of N_2/BET surface area analysis, this method provides a relatively simple and reproducible technique for assessing the specific surface area of natural soils and soil constituents.

List of Technical Nomenclature

α	BET isotherm parameter
β	Langmuir isotherm parameter
ρ_l	Liquid density ($\text{g}\,\text{cm}^{-3}$)
ρ_s	Solid density ($\text{g}\,\text{cm}^{-3}$)
A_m	Area of adsorbed molecule (m^2)
A_s	Specific surface area ($\text{m}^2\,\text{kg}^{-1}$)

C_a	Aqueous-phase concentration ($mg\,l^{-1}$)
C_s	Solid-phase concentration ($kg\,kg^{-1}$)
$C_{s,max}$	Maximum solid-phase concentration ($kg\,kg^{-1}$)
$C_{s,mon}$	Monolayer solid-phase concentration ($kg\,kg^{-1}$)
N_A	Avogadro's number ($molecules\,mol^{-1}$)
n	Number of adsorbed layers
P	Vapor pressure (Pa)
P_0	Saturated vapor pressure (Pa)
P/P_0	Relative vapor pressure
Q_A	Heat of adsorption ($J\,mol^{-1}$)
Q_C	Heat of condensation ($J\,mol^{-1}$)
R	Ideal gas constant ($J\,K^{-1}\,mol^{-1}$)
r	Radius of particle (m)
V_s	Volume of solid (m^3)
W_{OD}	Weight of oven-dry solid (kg)

See also: **Cation Exchange**; **Clay Minerals**; **Organic Matter:** Principles and Processes; **Sorption:** Metals; Organic Chemicals

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn, pp. 591–681. New York: John Wiley.

Bower CA and Gschwend FB (1952) Ethylene glycol retention by soils as a measure of surface area and interlayer swelling. *Soil Science Society of America Proceedings* 16: 342–345.

Brunauer S, Emmett PH, and Teller E (1938) Adsorption of gases in multimolecular layers. *Journal of the American Chemical Society* 60: 309–319.

Carter DL, Mortland MM, and Kemper KD (1986) Specific surface. In: Klute A (ed.) *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, 2nd edn, pp. 413–423. Monograph No. 9. Madison, WI: American Society of Agronomy.

Greenland DJ and Mott CJB (1978) Surfaces of soil particles. In: Greenland DJ and Hayes MHB (eds) *The Chemistry of Soil Constituents*, pp. 321–353. New York: John Wiley.

Gregg SJ and Sing KSW (1982) *Adsorption, Surface Area and Porosity*, 2nd edn. New York: Academic Press.

Lowell S and Shields JE (1991) *Powder Surface Area and Porosity*, 3rd edn. London, UK: Champan & Hall.

McClellan AL and Harnsberger HF (1967) Cross-sectional areas of molecules adsorbed on solid surfaces. *Journal of Colloid and Interface Science* 23: 577–599.

Pennell KD (2003) Specific surface area. Topp GC and Dane JH (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, sect. 2.5, pp. 295–315. Madison, WI: American Society of Agronomy.

Pennell KD, Boyd SA, and Abriola LM (1995) Surface area of soil organic matter reexamined. *Soil Science Society of America Journal* 59: 1012–1018.

Sposito G (1984) *The Chemistry of Soils*, pp. 23–35. Oxford, UK: Oxford University Press.

van Olphen H and Fripiat JJ (1979) *Data Handbook for Clay Minerals and Other Non-Metallic Minerals*, pp. 203–211. New York: Pergamon Press.

STATISTICS IN SOIL SCIENCE

R Webster, Rothamsted Research, Harpenden, UK

Why Statistics?

Data on soil embody variation from many sources. Much of the variation is natural – from place to place at all scales from the global to the infinitessimal, and from time to time as the soil responds to weather, plant growth, and processes in the rhizosphere engendered by that growth. Farmers have added to the variation by their enclosing, reclaiming, clearing, and fertilizing of the land, though within fields they have removed some variation by cultivation and drainage. Further sources of variation are mineral extraction and subsequent reclamation, dumping of waste, and pollution of many kinds. Investigators design experiments and surveys in such a way that they can estimate the variation from particular sources such as imposed treatments or strata in a population and the differences between them. Variation in data also arises from the way observations are made; from the people who make the observations, from the imprecision instruments, from imperfect

laboratory technique, and from sampling fluctuation. One may like to think that the contributions from the laboratory are negligible, though 'ring' tests have often revealed them not to be so. In general, however, sampling fluctuation in the field, arising from the spatial variation there, is much larger.

Any one measurement of a soil property is influenced by contributions from at least some of these sources. It cannot be taken as an absolutely accurate representation of the truth therefore; rather it must be seen in relation to the probable error.

Statistics are needed in soil science to estimate and express this error and to apportion it to the different sources. In this way sense (signal) can be separated from meaningless or uninteresting variation (noise) in comparisons between classes of soil, in expressing relations, in assessing the effects of treatments in experiments, and in prediction. The statistical repertoire is huge, so here is presented the fairly elementary techniques that soil scientists most often need.

The techniques can be divided into two groups, namely, descriptive and analytical. They also have two fairly distinct fields of application: survey and experiment. In the first, investigators observe and record the soil as it is on samples, and descriptive techniques tend to dominate. In the second, they deliberately control some of the variation so that they can assess the effects of changes in one or a few factors that are of specific interest by analysis.

Population, Units, and Samples

The soil is regarded for statistical purposes as a population comprising elements or units. The units are the bodies of soil on which measurements are made. They are more or less arbitrary and determined largely by convenience and practicality. They may be individual cores of soil, pits, or pedons; each may be the volume of soil occupied by the roots of a single plant or that deformed under the wheel of a tractor; they may be pots in a greenhouse experiment, samples in lysimeters, plots in a field experiment, or whole fields or farms. The variation among them depends very much on the size of the units; the larger they are the more variation they encompass within them and the less there is between them to be revealed in data. The size, shape, and orientation of the units, known as the 'support' in survey, must be defined and adhered to throughout an investigation. The population is then all such units in a specified region or falling within some other definition for the purposes of the investigation. It is an operational definition, often known as the 'target population.' In a more restricted sense, the population and the units comprising it may be the values of a particular soil property in the defined supports.

Populations in surveys are typically very large, in many instances infinite, or hypothetical, and in some they are poorly defined. Measurement is feasible only on small subsets, i.e., on samples, and these subsets must properly represent their populations for the measurements to apply to the larger populations. The units in an experiment, in contrast, are typically a few dozen, though ideally the experimental material should be representative of some larger population.

Replication and Randomization

Estimating the mean in a population and its associated error from measurements on a sample requires both replication and randomization. The need for replication is evident: a single value can contain no information on variation. Randomization is needed to avoid bias. At its simplest it means choosing units such that all have the same chance of being selected.

In a survey this takes the form of simple random sampling. To apply it requires (1) that each unit can be identified uniquely, and (2) a rule for the selection. Simple random sampling, as in Figure 1a, tends to be inefficient in that it takes no account of anything known about the population beforehand, such as its spatial dependence, a soil map of the region, the relations between soil and vegetation or physiography, or the farming history. To take advantage of such knowledge, the population is first stratified either into small grid cells, as in Figure 1b, or by type of soil, vegetation, physiography, or farming, as in Figure 1c.

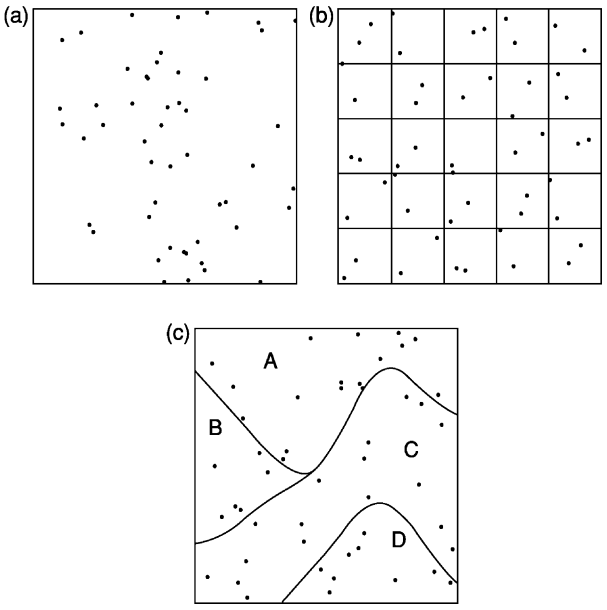


Figure 1 Probabilistic sampling of a region: (a) simple random sampling; (b) stratified random sampling with the region divided into 25 square cells (strata) and two points per cell; (c) stratified random sampling with four mapped classes (A–D) of soil as strata.

(a)

W	O	F	W	D
F	D	W	F	O
W	O	W	O	D
D	F	F	D	O

(b)

W	D	F	W	F
F	O	D	F	O
O	W	W	O	D
D	F	O	D	W

Figure 2 Layout of a randomized field experiment with four treatments; O (control), D (dung), W (industrial waste), and F (NPK fertilizer), and five-fold replication. (a) Completely randomized; (b) with the replicates arranged in five blocks.

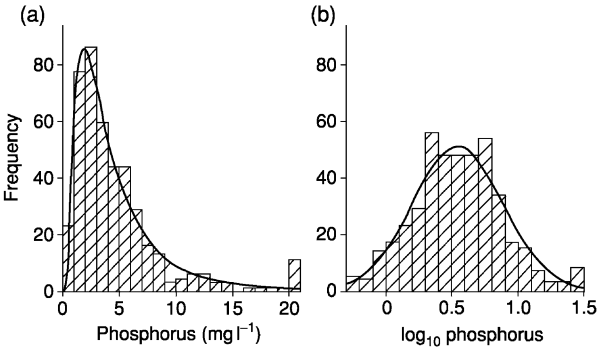


Figure 3 Histograms of 434 data on available phosphorus (a) in milligrams per liter and (b) transformed to common logarithms. The curve in b is that of the fitted normal distribution, and that in a shows the lognormal distribution.

The soil is then sampled at random within each stratum independently; this is stratified random sampling, and it enables the variation due to the strata to be distinguished from variation within them.

In experiments treatments are allocated to the units at random. The units may be arranged completely randomly, as in Figure 2a. In the field and greenhouse, they are usually grouped into blocks such that each block contains one unit of each treatment, as in Figure 2b; long-range variation then appears in the variation among blocks. There are many more elaborate designs.

Descriptive Statistics

The Mean

In almost all investigations, mean values are of prime interest. Provided sampling has been properly randomized, the mean of a set of N data, denoted z_1, $z_2, \ldots, z_N$,

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \qquad [1]$$

estimates without bias the mean, μ, in the population from which the sample is drawn. How well it does so depends on the variation it embodies.

Characteristics of Variation

Histogram The variation in a set of measurements, if there are sufficient of them, can be displayed in a histogram. The scale of measurement is divided into segments of equal width, or 'bins,' the values falling in each bin are counted, and bars of height proportional to the counts are drawn. Figure 3 is an example; it summarizes graphically the way in which the frequency is distributed over the range of the data.

Variance Variation is best expressed quantitatively as variance. For a set of N data, it is the average squared difference between the observations and their mean:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 \qquad [2]$$

Its square root, S, is the standard deviation, which is often preferred because it is in the same units as the measurements and is therefore more intelligible.

Ideally this variance should estimate the variance of the larger population, of which the N observations are a sample. This variance of a population is by definition:

$$\sigma^2 = \mathrm{E}\left[(z-\mu)^2\right] \qquad [3]$$

where μ is mean of z in the population and E denotes expectation. Equation [2] gives a biased result, however: S^2 is a biased estimator, because $\bar{z}$ in the equation is itself more or less in error as the estimate of μ. To remove the bias, N must be replaced by $N-1$ in the denominator:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{z})^2 \qquad [4]$$

The result, s^2, is now unbiased, provided the sampling was unbiased in the first place.

Estimation variance, standard error, and confidence Both S and s measure the dispersion in the observations; neither expresses the reliability of the estimate of μ. The reliance one can place in an estimate of the mean can be expressed in terms of the expected squared deviation of it from the true mean, i.e., $\mathrm{E}[(\bar{z}-\mu)^2]$. Its estimate is derived simply from s^2 by:

$$s^2(\bar{z}) = s^2/N \qquad [5]$$

This is the estimation variance, and its square root is the standard error, which is the standard deviation of means of samples of size N. The larger the sample is, the smaller this error is, other things being equal, and the more confidence there can be had in the estimate. So, the standard deviation, which describes the variation in a sample, is different from the standard error, which expresses the confidence associated with a mean. Standard errors accompany means in tables compiled from replicated measurements, and they can be shown by error bars on graphs.

Equation [5] gives the estimation variance for a simple random sample of size N. If the population has been divided into strata, then s^2, the population variance on the right-hand side of the equation, can be replaced by s^2_W, the variance within strata. The latter is generally less, often much less, than s^2, and so stratified sampling is more precise than simple random sampling by the factor $s^2_{\text{random}}(\bar{z})/s^2_{\text{stratified}}(\bar{z})$. It also means that one can achieve the same precision with a smaller sample, and in this sense stratified sampling is more efficient. This efficiency can be expressed as $N_{\text{random}}/N_{\text{stratified}}$. The within-stratum variance, s^2_W, can be estimated by the analysis of variance.

Standard errors can be converted to confidence intervals for known kinds of distribution. If the data are drawn from a normal (Gaussian) distribution and there are many of them, then an interval of $2s/\sqrt{N}$ centered on $\bar{z}$ spans a symmetric confidence interval of approximately 68%. Wider confidence intervals are readily calculated by multiplying by factors, η, such as in Table 1.

When data are few ($N < 30$) the above factors need to be replaced by Student's t for the number of degrees of freedom, f. The lower and upper symmetric confidence limits about a mean $\bar{z}$ are:

$$\bar{z} - t_f s/\sqrt{N} \quad \text{and} \quad \bar{z} + t_f s/\sqrt{N} \qquad [6]$$

where t_f is Student's t for f degrees of freedom. Values of t are available in tables and in most statistical computer packages. For $30 < N \leq 60$, t converges to η, and for $N > 60$ the values in Table 1 can be used.

There are formulae for calculating the confidence limits for other theoretical distributions. In many instances, however, it is easier to transform data to approximate a normal distribution and subsequently analyze the transformed values.

Table 1

Confidence (%)	80	90	95	99
η	1.28	1.64	1.96	2.58

Coefficient of Variation

The coefficient of variation (CV) is the standard deviation divided by the mean, i.e., $s/\bar{z}$. It is often multiplied by 100 and quoted as a percentage. The merit of the CV is that it expresses variation as pure numbers independent of the scales of measurement. It enables investigators and those reading their reports to appreciate quickly the degree of variation present and to compare one region with another and one experiment with another. The CV should not be used to compare variation in different variables, especially ones having different dimensions.

The CV is sensible only for variables measured on scales with an absolute zero. Otherwise the arbitrary choice of the zero affects it. Examples for which it should not be used are soil temperature in degrees Celsius (arbitrary zero at 273 K), color hue (which is approximately circular), and soil acidity on the pH scale (arbitrary zero equivalent to $-\log_{10}[H^+]$, with H^+ expressed in moles per liter).

For some soil properties, physics sets limits on the utility of the CV. For example, the minimum bulk density of the soil is determined by the physical structures that keep particles apart. Particles must touch one another, otherwise the soil collapses. For most mineral soils on dry land, a working minimum bulk density is approximately 1 g cm^{-3}. At the other end of the scale, the bulk density cannot exceed the average density of the mineral particles, approximately 2.7 g cm^{-3}. So the CV of bulk density is fairly tightly constrained.

The sensible use of the coefficient of variation for comparing two variables y and z relies on the assumption that they are the same apart from some multiplying factor, b, thus:

$$y = bz \qquad [7]$$

Then the mean of y is $\bar{y} = b\bar{z}$, its variance $s^2_y = b^2 s^2_z$, and its standard deviation is $s_y = bs_z$. From there, simple arithmetic shows that their CVs are the same. This principle offers a means of comparing variation by taking logarithms of the observations. Equation [7] becomes:

$$\log y = \log b + \log z \qquad [8]$$

The logarithm $\log b$ is a constant, and so the variances, $s^2_{\log y}$ and $s^2_{\log z}$, are equal, as are their standard deviations. The resulting measure of variation is therefore independent of the original scale of measurement.

The measure can be used to compare variation in two groups of observations. Consider again soil acidity. To compare the variation in acidity of a class A with that in another, class B, we treat the hydrogen ion concentration as the original variable, transform

it to pH, and compute the variances of pH. Whichever has the larger variance is the more variable, regardless of the mean. Further, we can make a formal significance test by computing $F = s^2_{\text{logy}}/s^2_{\text{logz}}$ and compare the result with F for N_y-1 and N_z-1 degrees of freedom.

Additivity of variances Variances are additive; those from two or more independent sources in an investigation sum to the total in the data. Their square roots, the corresponding standard deviations, are not. To obtain an average variation on the original scale of measurement from several sets of data, the arithmetic mean of their variances is computed, weighted as appropriate by the numbers of degrees of freedom, and then the square root of it is taken to give an 'average' or pooled standard deviation. This, divided by the mean, gives an average CV. More generally, the additive nature of variances confers great flexibility in analysis, enabling investigators to distinguish variation from two or more sources and estimate their components according to the design, as by the analysis of variance.

Statistical Significance

Significance in a statistical context means distinguishing a signal or the effects of some imposed treatment, or detecting differences between strata against a background of 'noise.' It is a matter of separating the variance due to the signal, treatments, or strata from that from other sources that are of no interest. The question being addressed is as follows: Given the magnitudes of the several components of variance, is the signal so strong or are the differences observed so large that they are highly unlikely to have occurred by chance? If the answer is 'yes,' then the result is said to be significant.

A significance test is prefaced by a hypothesis. This is usually that there is no real difference between populations or treatments and that any differences among the means of observations are due to sampling fluctuation. That is the 'null hypothesis,' often designated H_0, and the test is designed to reject it (not confirm it). The alternative, that there is a difference, is denoted either H_1 or H_A.

To judge, for example, whether two means differ, one computes from the sampling error, the probability, P, of obtaining the observed difference if the true means were identical, assuming that the form of the distribution is known. If P is small (conventionally <0.05), the null hypothesis is rejected, and the difference is judged significant. If P is large, then the null hypothesis is likely to be correct, but we have no measure of the probability that the two means are indeed identical; instead we take the view that we have too-little evidence to conclude that the difference observed applies to the population from which the sample has been drawn.

Mistaken conclusions can still be drawn as the result of significance tests. Mistakes can be of two kinds, denoted type I and type II. The first occurs when the null hypothesis is rejected on the basis of the sample evidence, i.e., a difference is declared significant when the populations are not different. The second is when the null hypothesis is accepted, i.e., there is insufficient evidence for a difference when the populations do differ.

The likelihood of drawing wrong conclusions can be diminished by increasing the sensitivity of the test, and that depends on the precision with which the means have been estimated, i.e., on their estimation variances or on the estimation variance of their difference. The latter is given by:

$$s^2_{\text{diff}} = \frac{s^2_{\text{W}}}{n_1} + \frac{s^2_{\text{W}}}{n_2} \qquad [9]$$

where n_1 and n_2 are the numbers of observations from which the means in classes 1 and 2 derive, and s^2_{W} is the variance within the classes, assumed to be common. If $n_1 = n_2$ then the variance of the difference is simply twice the estimation variance of the individual means.

Equation [9] shows two features. One is that the larger the variance is within the populations, the larger the variance between the means is and the less likely can a difference be established as significant. The second is that larger samples result in smaller variances and hence more sensitive comparisons. If the samples are large enough any soil can be established as different from almost any other for whatever property of interest.

The significance test is valuable in preventing false claims on inadequate evidence. Thus, a result might be summarized as:

> The mean measured pH of the topsoil was 5.7 compared with 6.7 in the subsoil; but because the samples were small (or because the variances were large) the difference was not statistically significant.

However, when a difference is deemed statistically significant because the null hypothesis is rejected, that does not mean that it is important or physically or biologically meaningful. For example, the difference between an observed mean pH of 5.7 in the topsoil and 5.9 in the subsoil would be of little consequence, whatever the probability of rejecting the null hypothesis. Also, while an investigator might regard a difference as significant only if $P \leq 0.05$, a reader may be willing to recognize one for which $0.05 \leq P < 0.1$

or, more stringently, only if $P \leq 0.01$. It is to some extent a matter of personal choice, and if the standard errors are reported then readers can reach their own judgments.

Note finally that the null hypothesis is highly implausible when horizons and different types of soil are being compared; they are different.

Transformations

It is often desirable to transform data to their square roots, or logarithms, or by other more elaborate functions. One reason for doing so is to obtain a new variate that approximates some known distribution, preferably normal (Gaussian) so that the usual parametric tests of significance can be applied.

The most serious departure from normality usually encountered with soil data is skewness, i.e., asymmetry, as in Figure 3a. The normal distribution is symmetric, its mean is at its center (its mode), and the mean of the data estimates this central value without ambiguity. The mean of data from a skewed distribution does not estimate the mode, nor does the median (the central value in the data). The meaning of the statistics can be uncertain therefore. A second feature of skewed data is that the variances of subsets depend on their means. If the data are positively skewed (again the usual situation) then the variances increase with increasing mean. This is undesirable when making comparisons. Third, estimation is 'inefficient' where data are skewed; that is, the errors are greater than they need be or, put another way, more data are required to achieve a given precision than would be if the distribution were normal. Transforming data to approximate normality overcomes these disadvantages. We achieve symmetry and hence remove ambiguity concerning the center. We stabilize the variances, and we make estimation efficient. The second of these is perhaps the most important.

No real data are exactly normal; all deviate more or less from normality. We have therefore to decide whether to transform them. This is best done by judicious exploration of the data aided by graphic display.

If a histogram looks symmetrical it can have superimposed on it a normal curve computed from the mean and variance of the data. The normal curve has the formula:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} \qquad [10]$$

where y is the probability density. Scale it to fit the histogram by multiplying by the number of observations and by the width of the bins. If the curve fits well then there is no need to transform the data.

If the histogram is skewed then the skewness coefficient, γ_1, can be computed in addition to the mean and variance. This dimensionless quantity can be obtained via the third moment of the data about their mean:

$$\gamma_1 = \frac{1}{NS^3} \sum_{i=1}^{N} (z_i - \bar{z})^3 \qquad [11]$$

A symmetric histogram has $\gamma_1 = 0$. Values of γ_1 greater than 0 signify positive skewness, i.e., long upper tails to the distribution and a mean that exceeds the median, which is common. Negative values of γ_1 signify negative skewness and are unusual.

If γ_1 is positive and less than 0.5 then there should be no need to transform the data. If $0.5 < \gamma_1 \leq 1$ then it might be desirable to convert the data to their square roots; and if $\gamma_1 > 1$ a transformation to logarithms is likely to give approximate normality.

The following example illustrates the situation. The data, which are summarized in Table 2, are 433 measurements of available phosphorus, P, in the topsoil. Their skewness coefficient is 3.95, i.e., they are strongly positively skewed, and this is apparent in their histogram (Figure 3a). Transforming to logarithms makes the histogram (Figure 3b) more-nearly symmetric, and, as the skewness in the logarithms is now only 0.34, the transformation seems satisfactory. Further, the normal curve appears to fit well.

Figure 4 shows how the transformation stabilizes the variances. In Figure 4a the variances of subsets of 44 from the full set of data on phosphorus are plotted against the means. Evidently, the variance increases strongly with increasing mean. Converting the data to their logarithms produces a result in which there is virtually no relation (Figure 4b).

These simple functions change only the general form of the distribution; they do not change the detail.

Table 2 Summary statistics of 433 values of available phosphorus measured in a survey of topsoil. Values of χ^2, with 18 df are added for the hypothesis that the data or their logarithms are from a normal distribution

	Scale	
Statistic	*P (mg l^{-1})*	*Log$_{10}$P*
Mean	4.86	0.546
Variance	26.52	0.1142
Standard deviation	5.15	0.338
Skewness	3.95	0.23
χ^2_{df18}	368.2	26.7

Normalizing the detail requires a more elaborate, normal score transform.

Analysis of Variance

The analysis of variance is at once one of the most powerful and elegant techniques in statistics. Its basis is that variances are additive and that the total variance in a population is the sum of the variances contributed by two or more sources. Working from the design of an investigation, it analyzes the data by separating the contributions from those sources and estimating the variances in them. Designs vary from the simple to highly complex, but all embody the same principle.

Here, two of the simplest designs are considered, such as appear in Figure 2. An investigator wants to know how manuring improves crop yield in the field. He or she has several (k) treatments, e.g., nothing (O), dung (D), industrial waste (W), and a complete artificial (NPK) fertilizer (F). The researcher replicates each m times by assigning them completely at random to plots of equal size. Figure 2a shows how the experiment might be layed out with $m = 5$ replicates. The investigator applies the treatments, grows the crop, and measures the yield at harvest, designated z.

The total variance in the yields in the experiment, s_T^2, comprises variance from two sources, namely that between the treatments, s_B^2, and that within them, s_W^2, and $s_T^2 = s_B^2 + s_W^2$. The total variance is estimated by Eqn [4]. The variance within any one treatment is estimated by the same formula but for only those data in that treatment. Pooling estimates for all treatments gives s_W^2. To complete the analysis, a quantity B can also be computed:

$$B = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2 \qquad [12]$$

where n_i is the number of plots in the ith treatment, $\bar{z}_i$ is the mean of the ith treatment, and $\bar{\bar{z}}$ is the general mean of the data. The computations are set out in Table 3. Finally, the analysis leads to a test of significance. The ratio $F = B/W$ is computed, the distribution of which has been worked out and tabulated for degrees of freedom $k-1$ in the numerator and $N-k$ in the denominator. If F exceeds the tabulated value at probability $P = 0.05$ (or $P = 0.01$ or $P - 0.001$, according to choice), the treatments are judged to have produced significant differences.

In the simple experiment illustrated in Figure 2a, all n_i are equal to 5, so that n_i can be replaced by $n = 5$, and $N = mk = 5 \times 4 = 20$. Things do not always go as planned, however, and if some of the plot yields are lost then the n_i can vary from treatment to treatment; and the more general formulae in Table 3 will take care of that.

The soil might vary across the experiment systematically, so that there is trend, or in an apparently random way at a coarse scale. This variation could swell the residual variation and so mask that due to the treatments. It can be taken into account by blocking. The m replicates are now arranged in m blocks such that in each block every treatment appears once and once only. Figure 2b shows an example in which five blocks are laid out side by side. The analysis follows the same procedure as in Table 3 except that there is an additional line for the blocks in which the sum of squares is that of deviations from the block means (Table 4). The residual sum of squares is diminished by this quantity, and, although the residual degrees of freedom are also diminished, the residual mean square, i.e., the residual variance, is usually less, and the experiment more sensitive therefore.

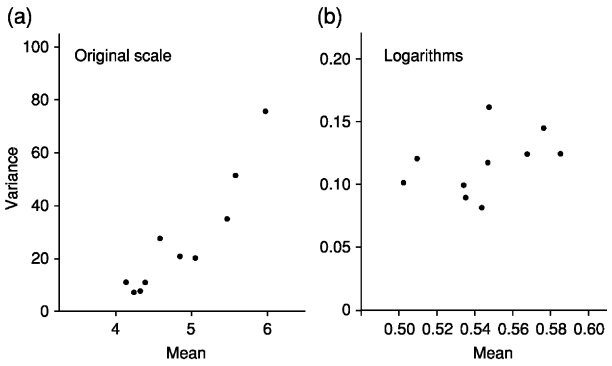


Figure 4 Graphs of variance against mean for 10 subsets of 44 phosphorus data (a) on the original scale in milligrams per liter and (b) after transformation to common logarithms.

Table 3 Table for the analysis of variance for a completely randomized design with a single classification

Source	*Degrees of freedom*	*Sums of squares*	*Mean square*	*F*
Treatments	$k-1$	$\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2 = B$	B/W
Residual	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2$	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2 = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{\bar{z}})^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{\bar{z}})^2 = T$	

Table 4 Table for the analysis of variance for a balanced randomized block design with a single set of treatments

Source	*Degrees of freedom*	*Sums of squares*	*Mean square*	*F*
Treatments	$k-1$	$\sum_{i=1}^{k} n_i(\bar{z}_i-\bar{\bar{z}})^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i-\bar{\bar{z}})^2 = B$	B/W
Blocks	$m-1$	$\sum_{l=1}^{m} n_l(\bar{z}_l-\bar{\bar{z}})^2$	$\frac{1}{m-1}\sum_{l=1}^{m} n_l(\bar{z}_l-\bar{\bar{z}})^2 = M$	
Residual	$(k-1)\times(m-1)$	$T(N-1)-B(k-1)-M(m-1)$	$\frac{T(N-1)-B(k-1)-M(m-1)}{(k-1)(m-1)} = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij}-\bar{\bar{z}})^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij}-\bar{\bar{z}})^2 = T$	

In soil survey, different classes of soil replace treatments. In the simplest cases, each class is sampled at random, and the n_i are rarely equal, either because it is difficult to obtain equal representation or because sampling is deliberately in proportion to area so as to maintain a fairly constant density. Effectively the classes are weighted in proportion to the areas they cover. $F = B/W$ can still be computed, and tested for significance, but, as above, this is less interesting than the differences between the means.

The variance between treatments or classes, s_B^2 can be obtained from B. The latter combines variation both from between treatments or classes and within them:

$$B = ns_B^2 + s_W^2 \qquad [13]$$

if $n_i = n$ for all i. Rearranging then gives:

$$s_B^2 = (B - s_W^2)/n \qquad [14]$$

If the n_i are unequal then n is replaced by n^*, given by:

$$n^* = \frac{1}{k-1}\left(N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right) \qquad [15]$$

and:

$$s_B^2 = (B - s_W^2)/n^* \qquad [16]$$

Fixed Effects, Random Effects, and Intraclass Correlation

The value s_B^2 obtained as above estimates $\sum_{i=1}^{k}(\mu_i - \mu)^2/(k-1)$, where μ_i is the expected value of treatment or class i, and μ is the expected value in the whole population. In designed experiments, in which the effects are fixed by the experimenter, s_B^2 is of little interest. In soil survey, however, where it is often a matter of chance which classes are actually sampled, the differences $\mu_i - \mu$ are subject to random fluctuation. In this event, s_B^2 estimates the variance, σ_B^2, among a larger population of means and is termed a 'component of variance,' which is of considerable interest.

The between-class variance expressed as a proportion of the total variance, $\sigma_B^2 + \sigma_W^2$, is the intraclass correlation, ρ_i:

$$\rho_i = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \qquad [17]$$

which is estimated from the analysis of variance table by:

$$r_i = \frac{s_B^2}{s_B^2 + s_W^2} = \frac{B - W}{B + (n^* - 1)W} \qquad [18]$$

The intraclass correlation has a theoretical maximum of 1 when every class is uniform. In practice there is always some variation within classes, and so ρ_i never attains 1. The minimum of ρ_i is zero, when $s_W^2 = 0$. The calculated estimate of ρ_i can be negative, because $B < W$, and is usually best explained by sampling fluctuation.

Covariance, Correlation, and Regression

The relations between two variables can be expressed by correlation and regression.

Covariance

The covariance of a pair of variables, y and z, is estimated from data in a way analogous to the estimation of the variance, Eqn [4], by:

$$\hat{c}_{y,z} = \frac{1}{N-1}\sum_{i=1}^{N}\{y_i - \bar{y}\}\{z_i - \bar{z}\} \qquad [19]$$

where $\bar{y}$ and $\bar{z}$ are the means of y and z, respectively. Covariance is not easy to envisage, especially if y and z have different dimensions. The relation may be standardized by converting it to correlation, below.

Correlation

The correlation between two variables, strictly the linear correlation, or the product–moment coefficient of linear correlation, is a dimensionless quantity, usually denoted by ρ for the population parameter. Its estimate, r, is obtained from the covariance by:

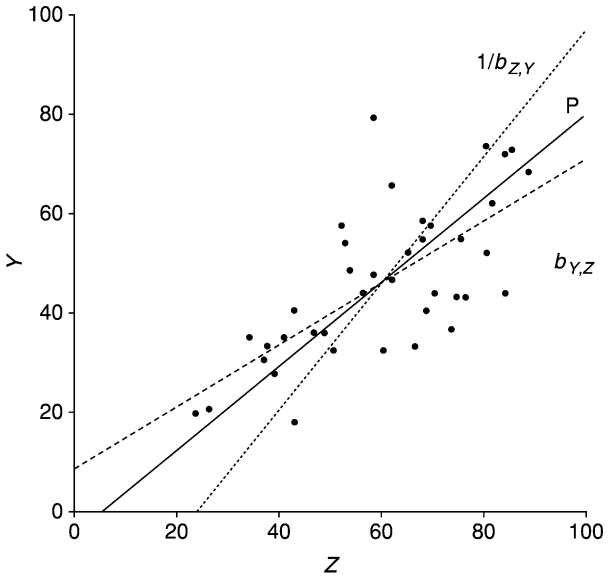


Figure 5 Scatter graph showing the relations between two variables Y and Z for which the correlation coefficient, r, is 0.699. The symbols are the observed values, the solid line, labeled 'P,' is the principal axis, with gradient 0.844, equivalent to an angle of 40.1° to the horizontal. The dashed line shows the regression of Y on Z, with gradient $b_{Y,Z} = 0.621$; and the dotted line shows that of Z on Y, for which $b_{Y,Z} = 0.788$, giving it a gradient of approximately $1/b_{Z,Y} = 1.269$.

$$r_{y,z} = \frac{\hat{c}_{y,z}}{\sqrt{s_y^2 s_z^2}} \qquad [20]$$

where s_y^2 and s_z^2 are the estimated variances of y and z. The coefficient was proposed by Karl Pearson, and for that reason it is often called the 'Pearson correlation coefficient.'

The coefficient is effectively a standardized version of the covariance. It measures the extent to which the data, when plotted as one variable against the other in a scatter graph, depart from a straight line (see Figure 5). It may vary between +1, signifying perfect positive correlation, and −1, for perfect negative correlation. Intermediate values indicate departures from the straight line, as in Figure 5, for which $r = 0.699$. In general, positive values of r indicate the tendency of y and z to increase together, whereas negative values arise when y increases as z decreases. A value of zero represents no linear relation.

Notice that the statistic refers specifically to linear correlation. The relation between two variables might be curved; the absolute value of r would then be necessarily less than 1 regardless of any scatter about the curve.

When the data, y_i, z_i, $i = 1, 2, \ldots, N$, are from a sample, then $c_{y,z}$ and $r_{y,z}$ estimate corresponding population parameters, $\mathrm{cov}_{y,z}$ and $\rho_{y,z}$. If the data can be assumed to be drawn from a bivariate normal distribution then one can test r for significance. One computes Student's t with $N - 2$ degrees of freedom:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \qquad [21]$$

The probability of this value's occurring on the null hypothesis that $\rho = 0$ can then be computed or looked up in a table of t.

Spearman rank correlation Where the distributions of the underlying variables are far from normal, the Pearson coefficient can be replaced by the Spearman rank correlation coefficient, usually denoted r_S. The values of each variable are ranked from smallest to largest and given new values 1, 2, ..., N. The correlation coefficient is then computed by applying eqns [19] and [20]. Alternatively, one may take the differences, d_i, $i = 1, 2, \ldots, N$, between the ranks and compute:

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}. \qquad [22]$$

Many soil variables observed in the field, such as grade of structure and frequency of mottles, are recorded as rankings rather than measured. In these circumstances the correlations between them can be expressed by the Spearman coefficient, whereas the Pearson coefficient would be inappropriate. Also, in these circumstances, tied ranks in any large set of data are inevitable, and Eqn [22] must be elaborated. The coefficient can be calculated in various ways, but from Eqns [19], [20], and [22] can be derived:

$$r_s = \frac{\sum_{i=1}^{N}(y_i-\bar{y})^2 - \sum_{i=1}^{N} T_{i,y} + \sum_{i=1}^{N}(z_i-\bar{z})^2 - \sum_{i=1}^{N} T_{i,z} - \sum_{i=1}^{N} d_i^2}{2\sqrt{\left\{(y_i-\bar{y})^2 - \sum_{i=1}^{N} T_{i,y}\right\}\left\{(z_i-\bar{z})^2 - \sum_{i=1}^{N} T_{i,z}\right\}}} \qquad [23]$$

in which:

$$T_i = \frac{t_i(t_i^2-1)}{12} \qquad [24]$$

where t_i is the number of observations tied at rank i.

For small samples, r_S is somewhat less sensitive than the Pearson coefficient in that larger values are necessary to establish statistical significance.

Regression

Regression treats the relation between two variables in a somewhat different way by designating one of them, y, as depending on the other, z, represented by the equation:

$$y = \beta_0 + \beta_1 z \quad [25]$$

The underlying rationale is often physical. For example, How is the soil's strength changed by additions of gypsum? Known amounts of gypsum can be added to the soil, the resultant changes in strength measured, and from the data how much on average the strength is changed by each increment in gypsum added can be estimated. The Gauss linear model can be adopted for this purpose:

$$y_i = \beta_0 + \beta_1 z_i + \epsilon_i \quad [26]$$

where y_i is the value of the random dependent variable, Y, here strength, in unit i, z_i is that of the independent variable, added gypsum, and ϵ_i is random error term that is uniformly and independently distributed with variance σ_ϵ^2. The quantities β_0 and β_1 are parameters of the model and are estimated as follows:

$$\hat{\beta}_1 = \frac{\hat{c}_{y,z}}{s_z^2} \quad [27]$$

and:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z} \quad [28]$$

Equation [27] gives the rate at which strength changes in response to increments in gypsum. The quantity $\hat{\beta}_0$ in Eqn [28] is the intercept at $y = 0$ and is likely to be of subsidiary interest. Together $\hat{\beta}_0$ and $\hat{\beta}_1$ may be inserted into Eqn [25] for predicting unknown values of Y if we know those of z:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 z \quad [29]$$

The procedure minimizes the sum of the squares of the differences between the measured values y_i, $i = 1, 2, \ldots, N$, and those expected from Eqn [25], $\hat{y}_i$:

$$s_{Y,z}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad [30]$$

A somewhat different situation is common in soil survey. In it we sample the soil without knowing in advance what values we shall obtain for the variables of interest. For example, we may measure both cation exchange capacity, CEC, and clay content, and we may regard CEC as depending in a physical sense on clay content. The data on both variables now contain random components, and for this reason we designate them with the capital letters, Y and Z, respectively. We may express the relation as the regression of CEC on clay and estimate the parameters in the same way as for the Gauss linear model. In doing so we assign all the error to CEC and minimize the sum of squares, $s_{Y,Z}^2$, as in Eqn [30]. The purpose is now prediction, i.e., prediction of CEC, knowing the clay content. We could equally well compute the regression of clay on CEC. The roles of Y and Z are reversed, and we minimize:

$$s_{Z,Y}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (z_i - \hat{z}_i)^2 \quad [31]$$

where:

$$\hat{z} = \hat{\beta}_0' + \hat{\beta}_1' y \quad [32]$$

The primes attached to $\hat{\beta}_0'$ and $\hat{\beta}_1'$ signify that these quantities refer to the regression of Z on Y, and that they are different from those for the regression of Y on Z. In other words, the line defined by Eqn [29] differs from that defined by [32], as **Figure 5** shows. The correct line to choose depends on which variable is to be predicted. To predict y from z, Eqn [29] is used; to predict z from y, Eqn [32] is used.

Further Reading

Cochran WG (1977) *Sampling Techniques*, 3rd edn. New York: John Wiley.

Draper NR and Smith H (1981) *Applied Regression Analysis*, 2nd edn. New York: John Wiley.

Mardia KV, Kent JT, and Bibby JM (1979) *Multivariate Analysis*. London, UK: Academic Press.

Snedecor GW and Cochran WG (1989) *Statistical Methods*, 8th edn. Ames, IA: Iowa State University Press.

Sokal RR and Rohlf FJ (1981) *Biometry*, 2nd edn. San Francisco, CA: WH Freeman.

Townend J (2002) *Practical Statistics for Environmental and Biological Scientists*. Chichester, UK: John Wiley.

Webster R (1997) Regression and functional relations. *European Journal of Soil Science* 48: 557–566.

Webster R (2001) Statistics in soil research and their presentation. *European Journal of Soil Science* 52: 331–340.

Webster R and Oliver MA (1990) *Statistical Methods in Soil and Land Resource Survey*. Oxford, UK: Oxford University Press.

Sterilization *See* **Disinfestation**

STOCHASTIC ANALYSIS OF SOIL PROCESSES

D Russo, Volcani Center, Bet Dagan, Israel

Introduction

Quantitative field-scale descriptions of water flow and chemical transport in the unsaturated (vadose) zone are essential for improving the basic understanding of the transport process in near-surface geological environments, and for providing predictive tools that, in turn, will be used to predict the future spread of pollutants in these environments. One of the distinctive features of a natural formation at the field scale is the spatial heterogeneity of its properties that affect flow and transport. This spatial heterogeneity is generally irregular; it occurs on a scale beyond the scope of laboratory samples and has a distinct effect on the spatial distribution of solutes, which results from transport through the formation. A fundamental question is that of how to develop predictive models that may incorporate the impact of field-scale spatial variability of the formation properties on vadose-zone flow and transport.

Modeling Flow and Transport in Unsaturated, Heterogeneous Soils

The starting point in modeling soil water flow is the uncertainty in the soil properties that affect flow and transport (because of their inherently erratic nature and the paucity of measurements). This uncertainty generally precludes the use of the traditional, deterministic modeling approach for the prediction of flow and transport on the field scale. In an alternative approach, uncertainty is set in a mathematical framework by modeling the relevant soil properties as random space functions (RSFs). As a consequence, the flow and transport equations are of a stochastic nature, and the dependent variables are also RSFs. The aim of the stochastic approach, therefore, is to evaluate the statistical moments of the variables of interest, given the statistical moments of the formation properties. In general, this is a formidable task and usually its scope is restricted to finding the first two moments. To simplify matters, it is common to treat each of the soil properties, as well as the various flow-controlled attributes, denoted by $p'(\mathbf{x})$, where $\mathbf{x}=(x_1,x_2,x_3)$ is the spatial coordinate vector, as stationary, characterized at second order by a spatially invariant mean, $P=<p'(\mathbf{x})>$, and a covariance, $C_{pp}(\mathbf{x}',\mathbf{x}'')=<p(\mathbf{x}')p(\mathbf{x}'')>$, in which $p(\mathbf{x})=p'(\mathbf{x})-P$, that depends on the separation vector $\boldsymbol{\xi}=\mathbf{x}'-\mathbf{x}''$, and not on $\mathbf{x}'$ and $\mathbf{x}''$ individually.

Under unsaturated flow conditions, the analysis is further complicated by the fact that the relevant flow parameters – hydraulic conductivity, K, and water capacity, C, which depend on formation properties – depend also on flow-controlled attributes in a highly nonlinear fashion. Consequently, under variably saturated conditions, the evaluation of the effects on flow and transport of spatial variations in the formation properties is extremely complex; it requires several simplifying assumptions regarding the constitutive relationships for unsaturated flow, the flow regime, and the spatial structure of the formation heterogeneity.

The focus here is on transport of conservative, nonreactive, nonvolatile solutes, under steady flow and in the absence of plant roots. Furthermore, the emphasis is on formations, for which the spatial distribution of their properties can be modeled by a unimodal distribution with a spatial correlation structure characterized by a covariance with a single, finite-length scale. It should be emphasized, however, that the theoretical framework described here can serve as a basis for more complicated situations, including transient water flow, transport of reactive solutes, and multiple-length scale, heterogeneous formations.

Quantification of solute transport in partially saturated, heterogeneous porous formations may be accomplished with a two-stage approach, which combines a stochastic continuum description of the steady-state flow with a general Lagrangian description of the motion of an indivisible particle of a passive solute that is carried by a steady-state flow. The first stage involves relating the statistical moments of the probability density function (PDF) of the velocity to those of the properties of the formation, while the second stage involves relating the statistical moments of the particle displacement PDF to those of the velocity PDF.

Stochastic Analysis of the Flow

A first-order perturbation approach is used here in order to obtain the first two statistical moments of the PDF of flow-controlled attributes, i.e., the pressure head, ψ, the log-unsaturated conductivity, $\log K$, the water flux vector, $\mathbf{q}'$, the water saturation, Θ, and, concurrently, those of the Eulerian velocity vector, $\mathbf{u}'$, for given statistics of the formation properties. The following assumptions are made:

1. The heterogeneous formation has a three-dimensional structure with axisymmetric anisotropy, and the flow domain is variably saturated and unbounded;
2. The local steady-state unsaturated flow obeys the Darcy law and continuity, which, if local isotropy is assumed, reads:

$$q_i'(\mathbf{x}) = -K(\psi,\mathbf{x})\frac{\partial\varphi(\mathbf{x})}{\partial x_i}$$
$$\frac{\partial q_i'(\mathbf{x})}{\partial x_i} = 0 \quad i = 1,2,3 \qquad [1]$$

where q_i' $(i=1,2,3)$ is the water flux vector, $\varphi = -x_1 - \psi$ is the hydraulic head, x_1 is directed vertically downward, and the Einstein summation convention is used in eqn [1] and elsewhere;
3. The local relationships between K and ψ (considered here as a positive quantity) and Θ and ψ are nonhysteretic and are given by the Gardner–Russo model, i.e.,

$$K(\psi,\mathbf{x}) = K_s(\mathbf{x})\exp[-\alpha(\mathbf{x})\psi] \qquad [2a]$$

$$\Theta(\psi,\mathbf{x}) = \left\{\exp\left[-\frac{1}{2}\alpha(\mathbf{x})\psi\right]\left[1+\frac{1}{2}\alpha(\mathbf{x})\psi\right]\right\}^{2/(m'+2)} \qquad [2b]$$

where K_s is saturated conductivity, $\alpha = \lambda^{-1}$ is a parameter of the formation, λ is the macroscopic capillary length scale and m′ is a parameter, selected here as $m' = 0$;
4. Both $\log K_s$ and $\log\alpha$ are multivariate normal (MVN) RSFs, ergodic over the region of interest, characterized by constant means, $F = E[\log K_s]$ and $A = E[\log\alpha]$, and by covariances $C_{ff}(\boldsymbol{\xi})$ and $C_{aa}(\boldsymbol{\xi})$, respectively, and cross-covariance, $C_{fa}(\boldsymbol{\xi})$ given by:

$$C_{pp}(\xi) = \sigma_{p*}^2\exp(-\xi') \qquad [3]$$

where f and a are the perturbations of $\log K_s$ and $\log\alpha$, respectively, $p = f$ or a, $p^* = f$, a, or fa; $\boldsymbol{\xi}' = (\mathbf{x}-\mathbf{x}'')/\mathbf{I}_p$, is the scaled separation vector, $\xi' = |\boldsymbol{\xi}'|$; σ_f^2 and σ_a^2, and $\mathbf{I}_f = (I_{f1},I_{f2},I_{f3})$ and $\mathbf{I}_a = (I_{a1},I_{a2},I_{a3})$ are the respective variances and correlation scales of $\log K_s$ and $\log\alpha$, σ_{fa}^2 is the cross-variance between perturbations of $\log K_s$ and $\log\alpha$, and $\mathbf{I}_{fa} = 2\mathbf{I}_f\mathbf{I}_a/(\mathbf{I}_f+\mathbf{I}_a)$;
5. The porosity, ϕ is constant and uniform throughout the formation.

Use of the aforementioned assumptions under ergodic conditions, by elimination of $\mathbf{q}'$ from eqn [1], expression of the various parameters and variables on the right-hand side (RHS) of eqn [2a] in terms of means and perturbations, i.e., $\log K_s = F+f$, $\log\alpha = A+a$, $\psi = H+h$, and use of the Taylor expansion with first-order terms retained, gives the first-order perturbation approximation of the steady flow equation as:

$$\frac{\partial^2 h}{\partial x_i^2} - \Gamma(2J_i - \delta_{1i})\frac{\partial h}{\partial x_i} - J_i(J_i - \delta_{1i})\Gamma a + J_i\left(\frac{\partial f}{\partial x_i} - \Gamma H\frac{\partial a}{\partial x_i}\right) = 0, \quad i = 1,2,3 \qquad [4]$$

where $\Gamma = \exp(A)$ is the geometric mean of α, $J_i = \partial(-H - x_1)/\partial x_i$ $(i=1,2,3)$ is the mean of the head-gradient vector and δ_{1i} $(i=1,2,3)$ is the Kronecker delta.

On the assumption that the flow is gravity-dominated, i.e., the mean pressure-head gradient is zero so that the mean head-gradient, J_i, is given by $J_i = \delta_{i1}$ $(i=1,2,3)$, the exact solutions of eqn [4] in terms of the (cross-)spectral relationships between the RSFs of f, a, and h, which are obtained by using Fourier–Stieljes integral representations, are:

$$\hat{C}_{hh}(\mathbf{k}) = \frac{k_1^2[\hat{C}_{ff}(\mathbf{k}) + \Gamma^2H^2\hat{C}_{aa}(\mathbf{k}) - 2\Gamma H\hat{C}_{fa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [5]$$

$$\hat{C}_{hf}(\mathbf{k}) = \frac{k_1(\Gamma k_1 - j'k^2)[\hat{C}_{ff}(\mathbf{k}) - \Gamma H\hat{C}_{fa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [6]$$

$$\hat{C}_{ha}(\mathbf{k}) = \frac{k_1(\Gamma k_1 - j'k^2)[\hat{C}_{fa}(k) - \Gamma H\hat{C}_{aa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [7]$$

where j' is the imaginary unit, $\mathbf{k} = (k_1,k_2,k_3)$ is the wave number vector, $k = |\mathbf{k}|$, and the inverse Fourier transform of the covariance of the formation properties, $C_{pp}(\boldsymbol{\xi})$ (eqn [3] with $p = \log K_s$ or $\log\alpha$), is given by:

$$\hat{C}_{pp}(\mathbf{k}) = \frac{\sigma_p^2 I_{p1}I_{p2}I_{p3}}{\pi^2(1 + I_{p1}^2k_1^2 + I_{p2}^2k_2^2 + I_{p3}^2k_3^2)^2} \qquad [8]$$

In a similar way, if the variables on the left-hand-side (LHS) of eqns [1] and [2b] are expressed in terms of means and perturbations, i.e., $\mathbf{q}' = \mathbf{Q}+\mathbf{q}$ and $\Theta = S+s$, respectively, and the Taylor expansion is used with first-order terms retained, the (cross-) spectral relationships between the RSFs of f, a, h, and s, and f, a, h, and q_i $(i=1,2,3)$, which are obtained by using Fourier–Stieljes integral representations, are:

$$\hat{C}_{ss}(\mathbf{k}) = B^2\{\hat{C}_{hh}(\mathbf{k}) + H^2\hat{C}_{aa}(\mathbf{k}) + H[\hat{C}_{ha}(\mathbf{k}) + \hat{C}_{ah}(\mathbf{k})]\} \qquad [9]$$

$$\hat{C}_{sh}(\mathbf{k}) = B[\hat{C}_{hh}(\mathbf{k}) + H\hat{C}_{ha}(\mathbf{k})] \quad [10]$$

$$\hat{C}_{sf}(\mathbf{k}) = B[\hat{C}_{hf}(\mathbf{k}) - H\hat{C}_{fa}(\mathbf{k})] \quad [11]$$

$$\hat{C}_{sa}(\mathbf{k}) = B[\hat{C}_{ha}(\mathbf{k}) - H\hat{C}_{aa}(\mathbf{k})] \quad [12]$$

$$\begin{aligned}\hat{C}_{q_{ij}}(\mathbf{k}) = K_g^2\{&\delta_{1i}\delta_{1j}[\hat{C}_{ff}(\mathbf{K}) + (\Gamma H)^2\hat{C}_{aa}(\mathbf{k}) + 2\Gamma H\hat{C}_{fa}(\mathbf{k})] \\ &+ \delta_{1i}[(\Gamma + j'k_i)(\hat{C}_{fh}(\mathbf{k}) + \Gamma H\hat{C}_{ah}(\mathbf{k}))] \\ &+ \delta_{1j}[\Gamma - j'k_j)(\hat{C}_{hf}(\mathbf{k}) + \Gamma H\hat{C}_{ha}(\mathbf{k})] \\ &+ (\delta_{1i}\delta_{1j}\Gamma^2 + k_i^2)\hat{C}_{hh}(\mathbf{k})\} \quad i,j = 1,2,3 \end{aligned} \quad [13]$$

where $B = (1/4)\ \Gamma^2 H\ \exp[-(1/2)\Gamma H]$, $K_g = \exp(Y)$ is the geometric mean conductivity, and $Y = F - \Gamma H$ is the mean of log$K$. Note that because $\psi = H + h$ is a function of water saturation, the (cross-)spectra $\hat{C}_{mn}$ $(\boldsymbol{k})$ (m, $n = f$, a, h, s; eqns [5–7] and [9–13]) depend on saturation. This dependence, however, is omitted for simplicity of notation. Furthermore, for $H \to 0$ (and $\Gamma \to 0$), the RHS of eqns [9–12] vanish, while eqns [5], [6], [7], and [13] reduce to their counterparts associated with steady flow in saturated formations.

The (cross-)covariances, $C_{mn}(\xi)$ associated with the (cross-)spectra $\hat{C}_{mn}(\mathbf{k})$ (m, $n = f, a, h, s$; eqns [5–13]), are then calculated by taking the Fourier transform of the respective $\hat{C}_{mn}(\mathbf{k})$, i.e.:

$$C_{mn}(\boldsymbol{\xi}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(j'\,\mathbf{k}\cdot\boldsymbol{\xi})\hat{C}_{mn}(\mathbf{k})d\mathbf{k} \quad [14]$$

Scaled forms of the pressure-head covariance, $C'_{hh}(\boldsymbol{\xi}) = C_{hh}(\boldsymbol{\xi})/I_{yv}^2\sigma_y^2$, independent of water saturation, Θ, are illustrated graphically in **Figure 1** as functions of the scaled separation distance in the direction parallel to the mean flow, $\xi'_1 = \xi_1/I_{yv}$ (**Figure 1a, c**), and in the direction perpendicular to the mean flow, $\xi'_2 = \xi_2/I_{yv}$ (**Figure 1b, d**), for selected values of the length-scale ratios, $\eta = \lambda/I_v$ and $\rho = I_h/I_v$, where $I_v = I_{f1} = I_{a1}$, and $I_h = I_{f2} = I_{f3} = I_{a2} = I_{a3}$. Here σ_y^2 and $I_{yv} = I_{y1}$ are the Θ-dependent, log$K$ variance and correlation length-scale in the vertical direction, respectively, given by:

$$\sigma_y^2 = \sigma_f^2 + \Gamma^2H^2\sigma_a^2 - 2\Gamma H\sigma_{fa}^2 - \Gamma^2\sigma_h^2 \quad [15a]$$

$$I_{yv} = \frac{\sigma_f^2 I_{f1} + \Gamma^2H^2\sigma_a^2 I_{a1} - 2\Gamma H\sigma_{fa}^2 I_{fa1} - \Gamma^2\int_0^\infty C_{hh}(\xi_1,0,0)d\xi_1}{\sigma_f^2 + \Gamma^2H^2\sigma_a^2 - 2\Gamma H\sigma_{fa}^2 - \Gamma^2\sigma_h^2} \quad [15b]$$

and $\sigma_h^2 = C_{hh}(0)$ is the variance of the pressure-head perturbations.

Note that the correlation length-scales, I_v and I_h, determine approximately the distances, parallel and perpendicular to the mean flow, respectively, at which property variations cease to be correlated. On the other hand, the macroscopic capillary length-scale,

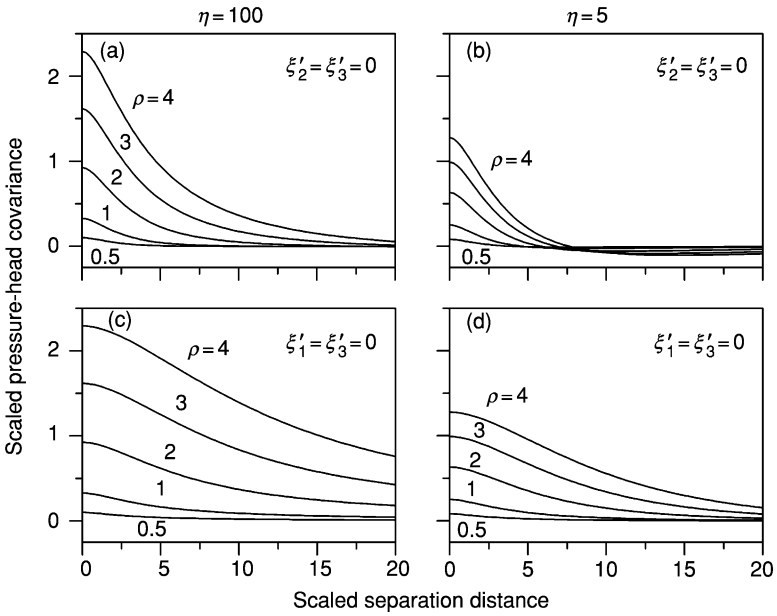


Figure 1 Scaled pressure-head covariance, C'_{hh}, independent of mean water saturation, S, as a function of the scaled separation distance parallel, ξ'_1 (a and b), and perpendicular, ξ'_2 (c and d) to the direction of the mean flow, for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$ m, $Q/\phi S = 1$m day^{-1}, and $J = (1,0,0)$.

λ, determines the relative magnitude of the capillary forces in unsaturated flow; it may be regarded as a natural length-scale of the unsaturated soil. The results depicted in Figure 1 show that, for given variability in the soil properties, both the magnitude and the persistence of the pressure-head variability increase with increasing ρ and η. Furthermore, the persistence of the pressure-head variability in the direction perpendicular to the mean flow is larger than that in the direction of the mean flow, particularly when both ρ and η are relatively large. This can be explained as follows: when the mean flow is vertical and I_v is kept constant, an increasing ρ expresses an increase in the size of the typical flow barriers in a direction normal to the mean flow; on the other hand, an increasing η expresses an increase in the macroscopic capillary length-scale, λ, i.e., a transition from a coarse-textured soil material, associated with negligible capillary forces to a fine-textured soil material, associated with significant capillary forces. Consequently, as ρ increases, the streamlines are deflected less easily; similarly, as η increases, the lateral head perturbation gradients increase. In other words, increasing stratification and capillary forces are expected to enhance the lateral dissipatation of water and, concurrently, to enhance the variability in the pressure head.

In Figure 2, principal components of the scaled flux covariance tensor, independent of S, i.e., $q'_{ii}(\boldsymbol{\xi}') = Cq_{ii}(\boldsymbol{\xi}')/\sigma_y^2 Q^2$ $(i = 1,2,3)$, where $Q = |\mathbf{Q}|$, are depicted as functions of $\xi'_1 = \xi_1/I_{yv}$, for selected values of the length-scale ratios, $\eta = \lambda/I_v$ and $\rho = I_h/I_v$. Note that because of the assumption of axisymmetric anisotropy, and for $J_2 = J_3 = 0$, $q'_{22}(\boldsymbol{\xi}) = q'_{33}(\boldsymbol{\xi})$. Figure 2 shows that the longitudinal component of the scaled flux covariance tensor is much larger than its transverse components, and also that both the magnitude and the persistence of the transverse components of the flux covariance tensor may be greater in fine-textured soil material. However, in such soil, the magnitude of the longitudinal component of the flux covariance tensor increases and its persistence decreases. On the other hand, soil stratification may increase the persistence of the longitudinal and the transverse components of the flux covariance tensor, may decrease the magnitude of the longitudinal component, and may either increase (for $\rho < \pi/2$) or decrease (for $\rho > \pi/2$) the magnitude of the transverse components of the flux covariance tensor.

The behavior of the principal components of the flux covariance tensor depicted in Figure 2 is explained on the same basis as above, i.e., when the mean flow is vertical and I_v is kept constant, an increasing ρ expresses increasing size of the typical flow barriers in a direction normal to the mean flow, while an increasing η expresses increasing capillary forces. Consequently, $q'_{11}(0)$ approaches unity at the small ρ limit, $\rho \to 0$ (which implies heterogeneity in the horizontal directions only) and decreases with increasing ρ, vanishing at the large ρ limit, $\rho \to \infty$ (which implies a perfectly stratified formation), while $q'_{ii}(i = 2, 3)$ are nonmonotonic functions of ρ with a maximum at

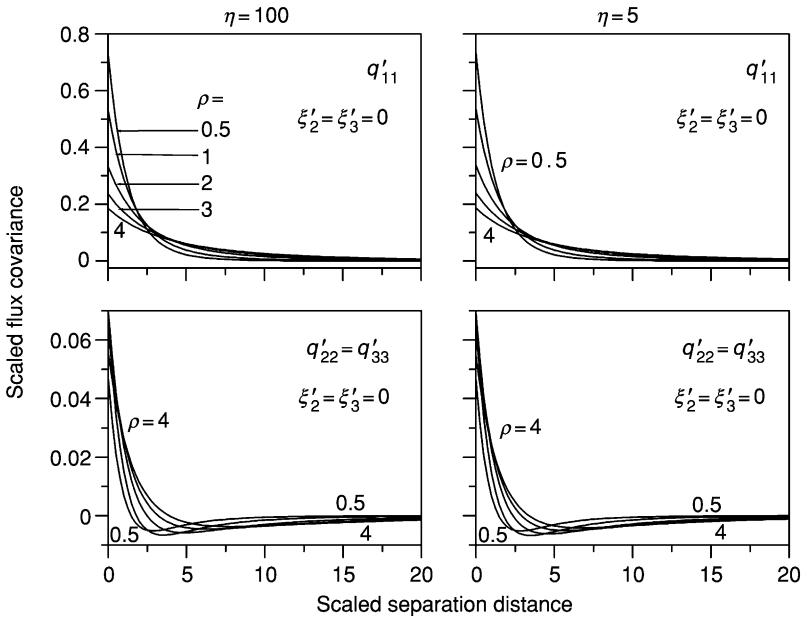


Figure 2 Principal components of the scaled flux covariance tensor, q'_{ii}, $i = (1,2,3)$, independent of mean water saturation, S, as a function of the scaled separation distance parallel to the direction of the mean flow, ξ'_1, for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$m, $Q/\phi S = 1$m day^{-1}, and $J = (1,0,0)$.

$\rho = \pi/2$, that vanishes at the small ρ limit, $\rho \to 0$ and at the large ρ limit, $\rho \to \infty$. On the other hand, $q'_{ii}(0)$ $(i = 1,2,3)$ increases with increasing η.

When the mean flow is aligned with the x_1-axis only, the first term on the RHS of the Fourier transform of eqn [13] vanishes for $i,j = 2,3$, while both its second and third terms vanish at $\xi_1 = 0$ and $\xi_1 \to \infty$. Furthermore, the last term is antisymmetric in ξ_1. Consequently, integration of the principal components of the flux covariance tensor along the ξ_1-axis, when performed up to $\xi_1 \to \infty$, yields $\int_0^\infty Cq_{11}(\boldsymbol{\xi})d\xi_1 = Q^2\sigma_y^2 I_{yv}$ and $\int_0^\infty Cq_{ii}(\boldsymbol{\xi})d\xi_1 = 0$, $(i = 2$ to $3)$, irrespective of the value of ρ or η. Hence, the decrease in $q'_{11}(0)$ with increasing ρ and decreasing η must be compensated for by an increase in the separation distance over which $q_1(x_1)$ and $q_1(x_1 + \xi_1)$ are positively correlated. Consequently, the persistence of $q'_{11}(\xi_1)$ increases with increasing ρ and decreasing η. On the other hand, for $i = 2$ or $i = 3$, the increase in the separation distance over which $q_i(x_1)$ and $q_i(x_1 + \xi_1)$ are positively correlated with increasing ρ and η must be compensated by an increase in the separation distance over which $q_i(x_1)$ and $q_i(x_1 + \xi_1)$ are negatively correlated. Hence, the persistence of $q'_{ii}(\xi_1)$ $(i = 2,3)$ increases with increasing ρ and η.

Stochastic Analysis of the Transport

A first-order perturbation approach is used here in order to obtain the first two statistical moments of the PDF of the one-particle-trajectory for given statistics of the Eulerian velocity vector, $\mathbf{u}'(\mathbf{x})$, and, concurrently, for given statistics of the formation properties. In a general Lagrangian framework, the transport is described in terms of the motion of indivisible solute particles which are convected by the fluid. If pore-scale dispersion is neglected, the trajectory of a solute particle is related to $\mathbf{u}'(\mathbf{x})$ by the fundamental kinematic relationship:

$$\frac{d\mathbf{X}'}{dt} = \mathbf{u}'(\mathbf{X}') \quad \text{for} \quad t > 0; \mathbf{X}' = \mathbf{a} \quad \text{for} \quad t = 0 \qquad [16]$$

where t is time, and $\mathbf{X}' = \mathbf{X}'(t,\mathbf{a})$ is the trajectory of a particle which is at $\mathbf{X}' = \mathbf{a}$ when $t = 0$.

In order to relate the statistical moments of the PDF of the one-particle trajectory to those of the velocity PDF, the following assumptions are employed: (1) Lagrangian and Eulerian stationarity and homogeneity of the velocity field; (2) given statistics of the velocity field; (3) small fluctuations of the particle displacement about the mean trajectory; and (4) large Peclet numbers, i.e., local dispersion is omitted.

Under the aforementioned assumptions, for ergodic conditions and for a particle of a solute injected into the flow field at time $t = 0$ and location $\mathbf{x} = \mathbf{a}$, the solution of eqn [16] is:

$$\mathbf{X}'(t,\mathbf{a}) = \mathbf{a} + \mathbf{U}t + \int_0^t \mathbf{u}'(\mathbf{a} + \mathbf{U}\tau)d\tau \qquad [17]$$

where $\mathbf{U} = <\mathbf{u}'(\mathbf{x})>$ is the mean Eulerian velocity vector.

For fixed $\mathbf{a}$, the first two moments of $\mathbf{X}'$ (eqn [17]), the mean, $<\mathbf{X}'(t)>$ and, by a first-order approximation in the velocity variance, the covariance tensor, $X_{ij}(t) = <X_i(t)X_j(t)>$, $(i,j = 1,2,3)$, of the one-particle displacement at time t, where $\mathbf{X} = \mathbf{X}' - <\mathbf{X}'>$ is the fluctuation, are given by:

$$<\mathbf{X}'(t)> = \mathbf{U}t \qquad [18a]$$

$$X_{ij}(t) = 2\int_0^t (t - \tau)C_{u_{ij}}(\mathbf{U}\tau)d\tau \qquad [18b]$$

Here $C_{uij}(\boldsymbol{\xi}) = <u_i(\mathbf{x})u_j(\mathbf{x} + \boldsymbol{\xi})>$, $(i,j = 1,2,3)$, the velocity covariance tensor, where $u_i(\mathbf{x}) = u_i'(\mathbf{x}) - <u_i'(\mathbf{x})> = q_i(\mathbf{x})/\phi S - Q_i s(\mathbf{x})/\phi S^2$ is the velocity fluctuation, is given by:

$$C_{u_{ij}}(\boldsymbol{\xi}) = \frac{K_g^2 \delta_{1i}\delta_{1j}}{\phi^2 S^4} C_{ss}(\boldsymbol{\xi}) + \frac{C_{q_{ij}}(\boldsymbol{\xi})}{\phi^2 S^2} - \frac{K_g^2}{n^2\phi^3}\left[- \delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})] - \delta_{1j}\frac{\partial C_{sh}(\boldsymbol{\xi})}{\partial \xi_i} + \delta_{1i}\frac{\partial C_{sh}(-\boldsymbol{\xi})}{\partial \xi_j}\right] \qquad [19]$$

where $C_{sy}(\boldsymbol{\xi})$ is the cross-covariance between perturbations of $\log K$ and Θ given by:

$$C_{sy}(\boldsymbol{\xi}) = C_{sf}(\boldsymbol{\xi}) - \Gamma C_{sh}(\boldsymbol{\xi}) - \Gamma H C_{sa}(\boldsymbol{\xi}) \qquad [20]$$

Eqns [18a] and [18b] are of a general nature, consistent with the linearization of the flow equation. Furthermore, because the soil properties, $\log K_s$ and $\log\alpha$ are MVN, $\log K$, ψ, and Θ are also MVN. Consequently, the velocity is MVN, and, by eqn [17], the one-particle trajectory, $\mathbf{X}'(t)$, is also MVN at any time. This means that, generally, the PDF of $\mathbf{X}'(t)$ satisfies the Focker–Planck equation and the expected concentration satisfies the classical convection dispersion equation. Furthermore, under ergodic conditions, the one-particle trajectory statistical moments eqn [18], can be equated with those of the spatial moments of the resident concentration, $c(\mathbf{x};t)$ (defined as the mass of the solute per unit volume of aqueous solution) of a plume. For a passive solute which lies at $t = 0$, within a finite volume V_0, the ith coordinate of the centroid of the solute plume, $X_i^e(t)$, and the

moment of inertia of the plume with respect to the x_i-axis, $X^e_{ii}(t)$ $(i=1,2,3)$, are:

$$X^e_i(t) = M^1_i(t) = <X_i(t)> + \frac{C_s X'_i(t)}{S} \qquad [21a]$$

$$X^e_{ii}(t) = \left\{M^2_i(t) - [M^1_i(t)]^2\right\} = X_{ii}(t) - \left[\frac{C_{sX'_i}(t)}{S}\right]^2 \qquad [21b]$$

where $M^N_1(t)$, $N=1$, 2, are the first and the second normalized spatial moments of the distribution of $c(\mathbf{x};t)$ in the ith coordinate, $<X_i(t)>$ and $X_{ii}(t)$ are the principal components of eqns [18a] and [18b], respectively, and $C_{sXi}(t) = <X_i(t)s(0)>$ is the cross-covariance between perturbations of the ith component of $\mathbf{X}'$ and water saturation, Θ.

It is clear from eqn [21] that, because Θ is spatially variable, the velocity $\mathbf{u}'(\mathbf{x}) = \mathbf{q}'(\mathbf{x})/\phi\Theta(\mathbf{x})$ is not divergence-free, i.e., $\mathrm{div}[\mathbf{u}'(\mathbf{x})] \neq 0$. Consequently, $X^e_i(t)$ differs from $<X_i(t)>$ by the correction factor $C_{sXi}(t)/S$. This implies time dependence of the effective solute velocity, $v^e_i = \mathrm{d}(X^e_i)/\mathrm{d}t = U_i + C_{sui}(\mathbf{U}t)/S$, where $C_{sui}(\boldsymbol{\xi}) = <u_i(\mathbf{U}\tau)s(0)>$ $(i=1,2,3)$ is the cross-covariance between perturbations of the ith component of the Eulerian velocity vector, $\mathbf{u}'$ and water saturation Θ, given by:

$$C_{su_i}(\boldsymbol{\xi}) = \frac{K_g}{\phi S^2}\left[-\delta_{1i}C_{ss}(\boldsymbol{\xi}) + \delta_{1i}SC_{sy}(\boldsymbol{\xi}) - S\frac{\partial C_{sh}(\boldsymbol{\xi})}{\partial \xi_i}\right] \qquad [22]$$

On the other hand, if the second term on the RHS of eqn [21*b*] is regarded as a second-order term which can be neglected as a part of the first-order approximation, then $X^e_{ii}(t)$ and the effective macrodispersion, $D^e_{ii}(t) = (1/2)\mathrm{d}[X^e_{ii}(t)]/\mathrm{d}t$ are not affected by the fact that $\mathrm{div}[\mathbf{u}'(\mathbf{x})] \neq 0$, and are identical to their counterparts, $X_{ii}(t)$ and $D_{ii}(t)$, the principal components of eqns [18b] and [23] below, respectively.

Note that, from a physical point of view, the ratios $X_{ij}(t)/2t$ $(i,j=1,2,3)$ may be regarded as apparent dispersion coefficients, that, for $t\to\infty$, tend to the macrodispersion tensor, $D_{ij}(t)$, $(i,j=1,2,3)$ given by:

$$D_{ij}(t) = \frac{1}{2}\frac{\mathrm{d}X_{ij}(t)}{\mathrm{d}t} = \int_0^t C_{u_{ij}}(\mathbf{U}\tau)\mathrm{d}\tau \quad (i,j=1,2,3) \qquad [23]$$

In the case of the velocity covariance tensor eqn [19], the first term on the RHS of eqn [19] depends only on the variability of the water saturation, Θ, while the second term on the RHS of eqn [19] depends only on the variability of the water flux, $\mathbf{q}$. The third term on the RHS of eqn [19] represents the effect on the velocity of the interaction between the Θ and the $\mathbf{q}$ heterogeneities. Inasmuch as the term $\delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})]$ is negative and $|\delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})]| > -\delta_{1j}[\partial C_{sh}(\boldsymbol{\xi})/\partial\xi_i] + \delta_{1i}[\partial C_{sh}(-\boldsymbol{\xi})/\partial\xi_j]$, all three terms on the RHS of eqn [19] contribute to the variability in velocity, and, concurrently, to macrodispersion (eqn [23]). Note that when the mean flow is aligned with the x_1-axis only, i.e., $\mathbf{J} = (J_1,0,0)$, the first and third terms on the RHS of eqn [19] vanish for $i=j=2$ or 3. In other words, in this case the transverse components of the velocity covariance tensor (eqn [19]) and, concurrently, those of the macrodispersion tensor (eqn [23]), are independent of the spatial variability of Θ and are only influenced by its mean value, $S = <\Theta(\mathbf{x})>$. Note also that when the formation is saturated (i.e., when $H\to 0$ and $S\to 1$), and when Γ approaches its small limit, $\Gamma\to 0$, the first and third terms on the RHS of eqn [19] vanish; consequently, eqn [19] and, concurrently, eqn [23] reduce to their counterparts associated with steady flow in saturated, three-dimensional heterogeneous formations.

Scaled forms of the principal components of eqn [23], $D'_{ii} = \phi S D_{ii}/Q\sigma^2_y I_{yv}$, $(i=1$ to $3)$, are presented graphically in **Figure 3**, as functions of the scaled travel time, $t' = tQ/\phi S I_{yv}$, for selected values of mean water saturation, S, and the length-scales ratios, $\rho = I_h/I_v$ and $\eta = \lambda/I_v$. Note that because of the assumption of axisymmetric anisotropy, and for $J_2 = J_3 = 0$, $D'_{22}(t) = D'_{33}(t)$. **Figure 3** suggests that the longitudinal component of the scaled dispersion tensor is much larger than its transverse components, particularly in formations in which $\rho \leq 1$. Note that for $\mathbf{J} = (J_1,0,0)$, because of the first and the third terms on the RHS of eqn [19], D'_{11} is saturation-dependent, while $D'_{ii}(i=2,3)$ are not.

The behavior of the principal components of eqn [23] demonstrates the combined influence of the formation heterogeneity, the capillary forces and water saturation on solute spreading under unsaturated flow. This arises directly from their effects on the velocity covariance (eqn [19]). For given $\nu = \sigma^2_a/\sigma^2_f$, S, ρ, and η, and when $\mathbf{J} = (J_1,0,0)$, the time behavior of $D'_{ii}(i=1,2,3)$ describes a continuous transition from a non-Ficksian to a Ficksian regime, i.e., they grow monotonously and linearly with t' at small t' $(t' \ll 1)$. As t' increases, however, the behavior of D'_{ii} $(i=2,3)$ is completely different from that of D'_{11}. The latter is a monotonically increasing function of t', approaching a constant value at the large t' limit, $t'\to\infty$, while $D'_{ii}(i=2,3)$ are nonmonotonic functions of t', vanishing at the large t' limit.

Figure 3 suggests that in unsaturated flow, solute spreading is expected to increase with diminishing

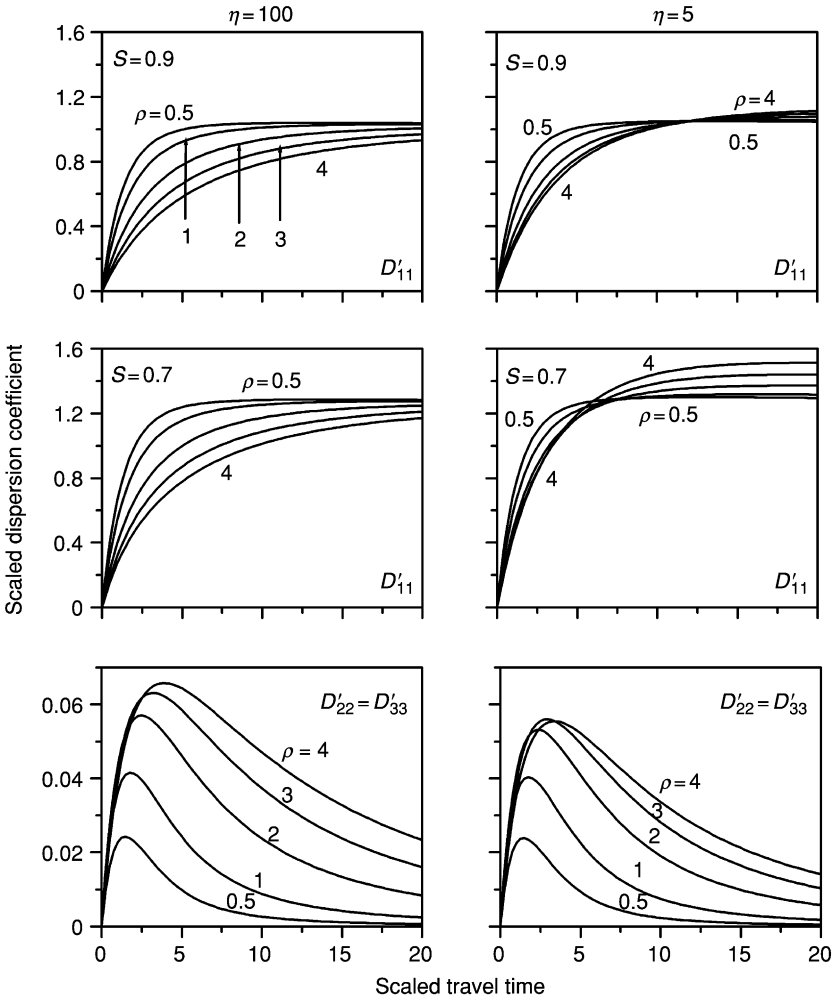


Figure 3 Principal components of the scaled macrodispersion coefficient tensor, D'_{i1}, $i=(1,2,3)$, as a function of the scaled travel time, t', for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of mean water saturation, S and $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$m, $Q/\phi S = 1$m day^{-1}, and $J=(1,0,0)$. Note that D'_{ii} ($i=(2,3)$) are independent of S.

water saturation. This is particularly so when the variances ratio, $\nu = \sigma_a^2/\sigma_f^2$, increases. Increasing stratification is expected to enhance solute spreading in the transverse directions in formations of both fine-textured and, especially, coarse-textured soil materials. On the other hand, increasing stratification is expected to reduce solute spreading in the longitudinal direction in formations of relatively fine-textured soil materials. In formations of relatively coarse-textured soil material, however, at a relatively small distance traveled, increasing stratification is expected to diminish solute spreading in the longitudinal direction while the converse is true when the travel distance exceeds a few log$K$ correlation length-scales. Furthermore, the travel distance required for the principal components of the macrodispersion tensor to approach their asymptotic values may be exceedingly large, particularly in relatively wet formations with significant stratification and with coarse-textured soil material that is associated with small capillary forces. Hence, in many practical situations of vadose-zone transport, the typical travel distance may be small compared with the travel distance required for the solute plume to reach its asymptotic, Ficksian behavior.

Summary

Before concluding, it is worthwhile emphasizing the limitations of the approach presented here. One limitation is due to the small-perturbation, first-order approximations of the velocity and the displacement covariance tensors, as a consequence of which the results are formally limited to formations with σ_y^2

smaller than unity. Other limitations stem from the fact that the calculations, which rely upon the assumption of ergodicity, are also restricted to conditions of constant mean-head gradient. The assumptions regarding the spatial structure of the formation (exponential covariance with axisymmetric anisotropy), the statistics of the relevant formation properties, and the flow-controlled attributes (statistically homogeneous), the local flow regime (steady-state flow), the local constitutive relationships for unsaturated flow (Gardner–Russo model), and the transport (Lagrangian stationarity, the neglect of both the fluctuations of the particle displacement about the mean trajectory and pore-scale dispersion), might also limit the applicability of the results of this approach. Nevertheless, the results of the first-order perturbation analysis are sufficiently reliable to indicate appropriate trends.

Most of the stringent assumptions of the first-order perturbation approach may be relaxed by using a numerical approach, which, in general, combines a statistical generation method that produces realizations of the heterogeneous formation properties in sufficient resolution, with an efficient numerical method to solve the partial differential equations that govern flow and transport in heterogeneous, variably saturated formations. However, the complexity and high degree of nonlinearity of such flow, compounded by the serious numerical problems which may be encountered when dealing with steep, spatial-head gradients, and the requirement that grid blocks must be much smaller than the length scale of the heterogeneity, all lead to a numerical problem of extreme difficulty which might demand a formidable computational capacity. This might substantially limit the capability of the numerical approach to solve flow and transport problems on a large, field scale.

List of Technical Nomenclature

$\boldsymbol{\alpha}$	Gardner–Russo parameter (m^{-1})
$\boldsymbol{\Gamma}$	Geometric mean of α (m^{-1})
$\boldsymbol{\delta}$	Kronecker delta (dimensionless)
$\boldsymbol{\eta} = \boldsymbol{\lambda}/I_{yv}$	Scale ratio (dimensionless)
$\boldsymbol{\Upsilon}$	Water saturation (dimensionless)
$\boldsymbol{\lambda}$	Macroscopic capillary length scale (m)
$\boldsymbol{\nu} = \boldsymbol{\sigma}_a^2/\boldsymbol{\sigma}_f^2$	Variance ratio (dimensionless)
$\boldsymbol{\xi}$	Separation distance vector (m)
$\boldsymbol{\xi}'$	Scaled separation distance vector (dimensionless)
$\boldsymbol{\rho} = I_h/I_v$	Length-scale ratio (dimensionless)
$\boldsymbol{\sigma}_p^2$	Variance of p' (units of p'^2)
$\boldsymbol{\tau}$	Dummy variable of integration (s)
$\boldsymbol{v}^e$	Effective solute velocity vector (ms^{-1})
$\boldsymbol{\phi}$	Porosity ($m^3\,m^{-3}$)
$\boldsymbol{\varphi}$	Hydraulic head (m)
$\boldsymbol{\psi}$	Pressure head (m)
$\mathbf{a}$	Coordinate vector of a point in V_0 (m)
a	A perturbation and mean of $\log\alpha$ (dimensionless)
$B = (1/4)\Gamma^2 H$	$\exp[-(1/2)\Gamma H]$ (m^{-1})
c	Resident concentration ($kg\,m^{-3}$)
$\mathbf{C}_{mn}$	(Cross-)covariance between perturbations m and n (units of $m \times n$)
$\hat{\mathbf{C}}_{mn}$	(Cross-)spectra between perturbations m and n (units of $m \times n$)
$\mathbf{C}_{q_{ij}}$	Flux covariance tensor ($m^2\,s^{-2}$)
$\mathbf{C}_{u_{ij}}$	Velocity covariance tensor ($m^2\,s^{-2}$)
$\mathbf{D}$	Macrodispersion tensor ($m^2\,s^{-2}$)
$\mathbf{D}'$	Scaled macrodispersion tensor (dimensionless)
D_{ii}^e	ith component of effective macrodispersion ($m^2\,s^{-1}$)
f, F	Perturbation and mean of $\log K_s$ (dimensionless)
h, H	Perturbation and mean of ψ (m)
$\mathbf{I}_p$	Correlation scale of p'(m)
I_v, I_h	Vertical and horizontal components of $\mathbf{I}_p$(m)
I_{yv}	Vertical correlation scale of $\log K$(m)
$\mathbf{J}$	Mean head-gradient vector (mm^{-1})
j'	The imaginary unit (dimensionless)
K	Unsaturated conductivity (ms^{-1})
K_g	Geometric mean of K (ms^{-1})
K_s	Saturated conductivity (ms^{-1})
$\mathbf{k}$	Wave number vector (m^{-1})
M_i^N	Nth normalized spatial moment in the ith coordinate (m^N)
$m = a, f, h, s$	(Units of a, f, h, or s)
m'	Parameter (eqn [2b]) (dimensionless)

$n = a, f, h, s$	(Units of a, f, h, or s)
$p' = f$ or a	(Dimensionless)
$p^* = f$, a, or fa	(Dimensionless)
$\mathbf{q}$, $\mathbf{Q}$	Perturbation and mean of $\mathbf{q}'$ (ms^{-1})
$\mathbf{q}'$	Water flux vector (ms^{-1})
q'_{ij}	Scaled flux covariance tensor (dimensionless)
s, S	Perturbation and mean of water saturation (dimensionless)
t	Time (s)
t'	Scaled value of t (dimensionless)
$\mathbf{u}$, $\mathbf{U}$	Perturbation and mean of $\mathbf{u}'$ (ms^{-1})
$\mathbf{u}'$	Eulerian velocity vector (ms^{-1})
V_0	Initial volume of the solute body (m^3)
$\mathbf{X}'$	Particle displacement vector (m)
X_{ij}	Particle displacement covariance tensor (m^2)
X_i^e	ith coordinate of the centroid of the solute plume (m)
X_{ii}^e	Moment of inertia of the plume with respect to the x_i-axis (m^2)
$\mathbf{x}$	Spatial coordinate vector (m)
$<\mathbf{X}>$ $\mathbf{X}$	Mean and perturbation of $\mathbf{X}'$ (m)
y, Y	Perturbation and mean of log$K$ (dimensionless)

Further Reading

Beckett PHT and Webster R (1971) Soil variability, a review. *Soils and Fertility* 34: 1–15.

Bras RL and Rodriguez-Iturbe I (1985) *Random Functions and Hydrology*. Reading, MA: Addison-Wesley Professional.

Dagan G (1989) *Flow and Transport in Porous Formations*. Berlin, Germany: Springer-Verlag.

Mantoglou A and Gelhar LW (1987) Stochastic modeling of large-scale transient unsaturated flow systems. *Water Resources Research* 23: 37–46.

Nielsen DR, Biggar JW, and Erh KH (1973) Spatial variability of field-measured soil-water properties. *Hilgardia* 42: 215–260.

Polmann DJ, Mclaughlin D, Luis S, Gelhar LW, and Ababou R (1991) Stochastic modeling of large-scale flow in heterogeneous unsaturated soils. *Water Resources Research* 27: 1447–1458.

Russo D (1998) Stochastic modeling of scale-dependent macrodispersion in the vadose zone. In: Sposito G (ed.) *Scale Dependence and Scale Invariance in Hydrology*, pp. 266–290. Stanford, UK: Cambridge University Press.

Russo D (1998) Stochastic analysis of flow and transport in unsaturated heterogeneous porous formation: effects of variability in water saturation. *Water Resources Research* 34: 569–581.

Russo D, Zaidel J, and Laufer A (1994) Stochastic analysis of solute transport in partially saturated heterogeneous soil. I. Numerical experiments. *Water Resources Research* 30: 769–779.

Russo D, Russo I, and Laufer A (1997) On the spatial variability of parameters of the unsaturated hydraulic conductivity. *Water Resources Research* 33: 947–956.

Russo D, Zaidel J, and Laufer A (1998) Numerical analysis of flow and transport in a three-dimensional partially saturated heterogeneous soil. *Water Resources Research* 34: 1451–1468.

Taylor GI (1921) Diffusion by continuous movements. *Proceedings of the London Mathematics Society* A20: 196–211.

Yeh T-CJ (1998) Scales issues of heterogeneity in vadose-zone hydrology. In: Sposito G (ed.) *Scale Dependence and Scale Invariance in Hydrology*, pp. 224–265. Stanford, UK: Cambridge University Press.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985a) Stochastic analysis of unsaturated flow in heterogeneous soils. 1. Statistically isotropic media. *Water Resources Research* 21: 447–456.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985b) Stochastic analysis of unsaturated flow in heterogeneous soils. 2. Statistically anisotropic media with variable α. *Water Resources Research* 21: 457–464.

STRESS–STRAIN AND SOIL STRENGTH

S K Upadhyaya, University of California–Davis, Davis, CA, USA

Concept of Stress

The concept of stress is the most fundamental concept of continuum mechanics and elasticity theory. When a body is subjected to external forces, it experiences deformation (strain) and stress. Consider a rectangular, parallelepiped-shaped volume element, ABCDEFGH, oriented such that its face BCGF is perpendicular to the X_1 axis; the face CDHG is perpendicular to the X_2 axis; EFGH is perpendicular to the X_3 axis; and where $X_1X_2X_3$ represents a right-handed Cartesian coordinate system such that the X_3 axis is vertical (Figure 1). Let ΔF_1, ΔF_2, and ΔF_3 respectively represent forces acting on these surfaces. Then the force per unit area acting on these surfaces (i.e., T_1, T_2, and T_3, respectively) as the size of the surface shrinks to zero can be represented by:

$$T_1 = \Delta F_1/(\Delta_2\Delta_3) \qquad [1a]$$

$$T_2 = \Delta F_2/(\Delta_1\Delta_3) \qquad [1b]$$

$$T_3 = \Delta F_3/(\Delta_1\Delta_2) \qquad [1c]$$

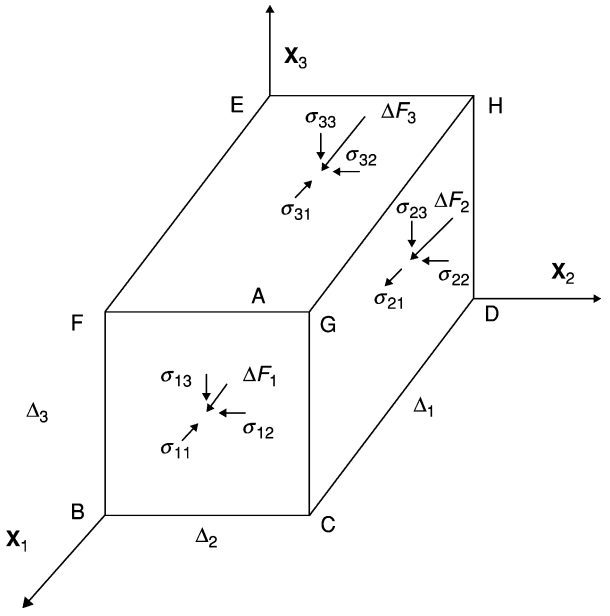


Figure 1 Components of a stress tensor.

in the limit Δ_1, Δ_2, and Δ_3 approach zero. In Eqns [1a]–[1c], Δ_1, Δ_2, Δ_3 are the linear dimensions of the volume element in X_1, X_2, and X_3 directions, respectively.

These three vectors can be decomposed along the coordinate axis as follows:

$$T_1 = \sigma_{11}\mathbf{i}_1 + \sigma_{12}\mathbf{i}_2 + \sigma_{13}\mathbf{i}_3 \qquad [2a]$$

$$T_2 = \sigma_{21}\mathbf{i}_1 + \sigma_{22}\mathbf{i}_2 + \sigma_{23}\mathbf{i}_3 \qquad [2b]$$

$$T_3 = \sigma_{31}\mathbf{i}_1 + \sigma_{32}\mathbf{i}_2 + \sigma_{33}\mathbf{i}_3 \qquad [2c]$$

where $\mathbf{i}_1$, $\mathbf{i}_2$, and $\mathbf{i}_3$ are unit vectors along the coordinate axes X_1, X_2, and X_3, respectively. The components of these stress vectors T_1, T_2, and T_3, form the elements of the stress tensor $[\boldsymbol{\sigma}]$, which is written as:

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \qquad [3]$$

The diagonal terms of this stress tensor are known as 'normal stresses' and the off-diagonal terms are known as 'shear stresses.' Figure 1 shows the various components of the stress tensor. As is usual in soil mechanics, compression is considered to be positive. Moreover, the subscripts of stress tensor component $\boldsymbol{\sigma}_{ij}$ are chosen so that the first subscript, 'i,' indicates the direction of the surface normal on which the stress component acts and the second subscript, 'j,' indicates the direction along which the stress component is directed. Thus σ_{13} indicates that this stress component acts on the surface whose normal is along the X_1 axis (i.e., BCGF) and σ_{13} is directed in the X_3 direction. In the absence of body couples, $\sigma_{ij} = \sigma_{ji}$, thus making the stress tensor symmetric. (Body couples, or distributed moments within the body, arise from the action of an electric field on polarized matter or the action of a magnetic field on magnetized particles. Such couples are not a concern in most mechanics problems and are ignored here.)

The first invariant of the stress tensor (the sum of the diagonal terms or the trace of the stress tensor, which does not change under a coordinate transformation) is related to the hydrostatic, spherical, or octahedral normal stress. It is given by:

$$J_1 = \mathrm{tr}(\sigma) = \sigma_{11} + \sigma_{22} + \sigma_{33} = 3\sigma_h = 3\sigma_{oct} \qquad [4]$$

where J_1 is the first invariant of the stress tensor, σ_h is the hydrostatic stress, and σ_{oct} is the mean octahedral normal stress. Mean pressure or bulk stress, p, is given by the mean of the diagonal elements, i.e.:

$$p = \sigma_h = \sigma_{oct} \quad [5]$$

The stress tensor can be decomposed into two parts in the following manner:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} p & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & p \end{bmatrix} + \begin{bmatrix} \sigma_{11}-p & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22}-p & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}-p \end{bmatrix} \quad [6]$$

The first tensor on the right-hand side is known as the 'hydrostatic' or 'spherical stress tensor,' and the second tensor is known as the 'deviatoric stress tensor' [S]. The second invariant of the stress tensor is related to the octahedral shear stress, τ_{oct}, and is given by:

$$J_{2D} = 1/6[(\sigma_{11}-\sigma_{22})^2 + (\sigma_{22}-\sigma_{33})^2 + (\sigma_{33}-\sigma_{11})^2] + \sigma_{12}^2 + \sigma_{23}^2 + \sigma_{13}^2 \quad [7]$$

where we have utilized the fact that the stress tensor is symmetric. The relation between octahedral shear stress, τ_{oct}, and J_{2D} is given by:

$$\tau_{oct} = [2/3\, J_{2D}]^{1/2} \quad [8]$$

This second invariant of the deviatoric stress tensor or the related quantity τ_{oct} plays an important role in describing the yield behavior in soil mechanics. In some cases the third invariant of the deviatoric stress tensor is also used to represent the yield behavior, such as in the extended Drucker–Prager yield criteria.

Principal Stress

For the symmetric version of the stress tensor $[\boldsymbol{\sigma}]$ given in Eqn [3], which is associated with the coordinate system shown in Figure 1, it is always possible to locate a set of mutually orthogonal planes along which all shear stress components will be zero, leaving only the normal components of the stress tensor. These normal components of stress tensor are known as principal stresses, the planes are called principal planes, and the normal vectors associated with these planes are known as principal axes or directions. If the principal axes are taken as the reference axes, then the stress tensor given in Eqn [3] becomes:

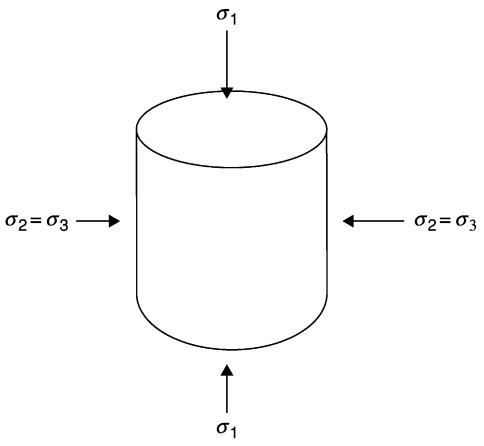


Figure 2 Cylindrical state of stress in soil mass.

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \quad [9]$$

where σ_1, σ_2, and σ_3 are known as 'principal stresses.' The largest (in magnitude) of the principal stresses is known as the 'major principal stress,' the smallest one is known as the 'minor principal stress,' and the other one is known as the 'intermediate principal stress.' In terms of principal stresses, the second invariant of stress tensor and τ_{oct} become:

$$J_{2D} = 1/6[(\sigma_1-\sigma_2)^2 + (\sigma_2-\sigma_3)^2 + (\sigma_3-\sigma_1)^2] \quad [10]$$

For example, consider a cylindrical body of soil under a triaxial loading situation, as shown in Figure 2. For this case, σ_1, and $\sigma_2 = \sigma_3$ are the principal stresses. This example will be used throughout this discussion. The mean hydrostatic stress and the octahedral shear stress are given by:

$$p = (\sigma_1 + 2\sigma_3)/3 \quad [11]$$

$$\tau_{oct} = \left(\frac{2}{3}J_{2D}\right)^{1/2} = \frac{\sqrt{2}}{3}(\sigma_1 - \sigma_3) \quad [12]$$

In soil plasticity theory, often the deviator stress $q = (\sigma_1 - \sigma_3) = (3J_{2D})^{1/2}$ is used to describe shear behavior, and mean hydrostatic pressure, p, is used to describe volumetric compression.

Effective Stress

In soil mechanics usually effective stress rather than the total stress is used in constitutive relationships. For saturated soils, effective hydrostatic pressure, p_{eff}, is defined as the difference between the total

hydrostatic stress, p_t, minus the pore-water pressure, u_w (i.e., $p = p_{eff} = p_t - u_w$). For unsaturated soils the effective pressure is a function of soil suction or soil-moisture tension. The effective stress in unsaturated soils may be represented by:

$$p_{eff} = p_t - u_a + \chi(u_a - u_w) \qquad [13]$$

where u_a is pore air pressure and χ depends on water content. Often it is taken as a fraction of unit cross-sectional area of soil occupied by water. However, this is problematic from a continuum mechanics and thermodynamic point of view. There is a need to treat unsaturated soil as a four-phase medium consisting of air, soil particles, water, and 'contractile skin' which separates air from water. Instead of using a single effective stress as given in Eqn [13] to define the stress state of soil, it is possible to use two different stress-state variables: net stress $(p_t - u_a)$ and matric suction $(u_a - u_w)$. The additional stress-state variable, matrix suction, is related to soil-moisture content through the soil-water release characteristic, which is regarded as an important constitutive function.

A total stress approach has been used to determine critical state parameters for unsaturated, agricultural soils. This method is acceptable as long as it is recognized that the material parameters are treated as a function of soil-moisture content. Unsaturated agricultural soils can be modeled in terms of total stress as long as the effect of unsaturated conditions is duly recognized. Thus the effective hydrostatic pressure, p_{eff} (i.e., $p = p_{eff}$), is used for saturated soil, and total pressure, p_t (i.e., $p = p_t$), is used for unsaturated soil.

Concept of Strain

When a body is subjected to external forces, it experiences deformation. The ratios of these deformations to the respective original dimensions are known as 'strain.' Thus, if a slender rod of length L is subjected to a uniaxial compressive load, its length will decrease, leading to a compressive strain of $\epsilon_a = u/L$, where u is the decrease in length (Figure 3). This is known as the 'Cauchy strain.'

This concept can be easily extended to three dimensions. Just as in the case of stress, it is possible to represent the strain experienced by any body under the generalized external forces in terms of a strain tensor, $[\boldsymbol{\epsilon}]$ as follows:

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \qquad [14]$$

where each ϵ_{ij} is known as a normal strain if $i = j$ and known as a shear strain if $i \neq j$. If there are no body

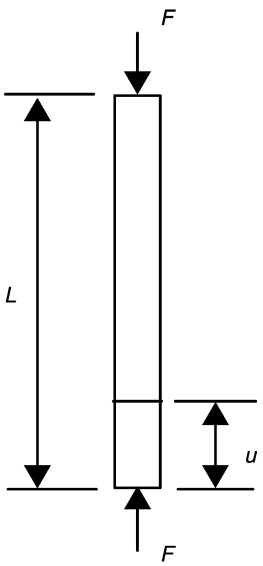


Figure 3 A slender rod under compressive axial strain.

moments, then $\epsilon_{ij} = \epsilon_{ji}$. This gives six strain components for a three-dimensional case. The subscripts 'i' and 'j' have the same implications as they did in the case of the stress tensor described above. If strain levels are small, these strain components can be expressed as:

$$\epsilon_{11} = \frac{\partial u_1}{\partial x_1}; \quad \epsilon_{22} = \frac{\partial u_2}{\partial x_2}; \quad \epsilon_{33} = \frac{\partial u_3}{\partial x_3}$$
$$\epsilon_{12} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right); \quad \epsilon_{13} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1}\right);$$
$$\epsilon_{23} = \frac{1}{2}\left(\frac{\partial u_3}{\partial x_2} + \frac{\partial u_2}{\partial x_3}\right) \qquad [15]$$

where u_1, u_2, and u_3 and deformations of the body along three mutually orthogonal, rectangular coordinate system axes such as those in Figure 1. It is a common practice in engineering to use engineering shear strains, γ_{ij}, which are related to ϵ_{ij} as follows:

$$\gamma_{ij} = 2\epsilon_{ij} \qquad [16]$$

The first invariant of the strain tensor, $I_1 = \text{tr}\ (\epsilon) = \epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon_v$, is known as the 'volumetric strain.' Just as in the case of stress tensor, it is possible to decompose the strain tensor into two parts as follows:

$$\begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} = \begin{bmatrix} \epsilon_v/3 & 0 & 0 \\ 0 & \epsilon_v/3 & 0 \\ 0 & 0 & \epsilon_v/3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} - \epsilon_v/3 & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} - \epsilon_v/3 & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} - \epsilon_v/3 \end{bmatrix} \qquad [17]$$

The first tensor on the right-hand side is the volumetric strain tensor and the second one is the deviatoric strain tensor. The second invariant of the strain tensor, I_{2D}, is of particular interest and is given by:

$$I_{2D} = 1/6[(\epsilon_{11} - \epsilon_{22})^2 + (\epsilon_{22} - \epsilon_{33})^2 + (\epsilon_{33} - \epsilon_{11})^2] + \epsilon_{12}^2 + \epsilon_{23}^2 + \epsilon_{13}^2 \quad [18]$$

In engineering practice, a closely related term, 'octahedral shear strain,' is used and is given by:

$$\gamma_{oct} = \left[\frac{2}{3} I_{2D}\right]^{\frac{1}{2}} \quad [19]$$

Principal Strain

For the symmetric version of the strain tensor given in Eqn [14], which is represented in the $X_1X_2X_3$ coordinate system, it is always possible to find a set of three mutually perpendicular planes along which all shear-strain components will be zero. The directions normal to these planes are known as principal strain axes. In terms of the principal strains, the strain tensor reduces to:

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{bmatrix} \quad [20]$$

where ϵ_1, ϵ_2, and ϵ_3 are principal strains. In terms of these principal strains, the second invariant of the deviatoric strain reduces to:

$$I_{2D} = 1/6[(\epsilon_1 - \epsilon_2)^2 + (\epsilon_2 - \epsilon_3)^2 + (\epsilon_3 - \epsilon_1)^2] \quad [21]$$

For the example problem, shown in Figure 2, $\epsilon_2 = \epsilon_3$. Under these conditions, the volumetric strain becomes:

$$\epsilon_v = \epsilon_1 + 2\epsilon_3 \quad [22]$$

and the octahedral shear strain becomes:

$$\gamma_{oct} = \left[\frac{2}{3} I_{2D}\right]^{\frac{1}{2}} = \frac{\sqrt{2}}{3}(\epsilon_1 - \epsilon_3) \quad [23]$$

In soil plasticity, ϵ_v and ϵ_s (the latter of which is given by $2/3\ (\epsilon_1 - \epsilon_3) = \sqrt{2}\gamma_{oct}$) are often used to represent soil behavior.

Void Ratio and Volumetric Strain

Often volumetric strain is represented in terms of changes in void ratio, e, which is defined as the ratio of total pore space to the total volume of individual solid particles, i.e.:

$$e = \frac{V_0}{V_s} \quad [24a]$$

where V_0 is the total volume of voids and V_s is the total volume of solid particles. The volumetric strain ϵ_v is related to the changes in the void ratio by the following equation:

$$d\epsilon_v = -\frac{de}{1 + e_i} \quad [24b]$$

where e_i is the initial void ratio. Note that the negative sign in Eqn [24b] is owing to the sign convention that treats compressive strain as positive.

Constitutive Laws

The constitutive law has been defined as a general functional relationship between stress, strain, and their rates, i.e.:

$$f(\sigma, \dot{\sigma}, \epsilon, \dot{\epsilon}) = 0 \quad [25]$$

where $\dot{\sigma}$ and $\dot{\epsilon}$ represent stress and strain rates, respectively. For static equilibrium conditions, the most general linear relationship between stress and strain tensors takes the following form:

$$[\boldsymbol{\sigma}] = [\mathrm{C}][\boldsymbol{\epsilon}] \quad [26]$$

where [C] is a fourth-order tensor, known as 'material property tensor,' since it relates two second-order tensors. For homogeneous, elastic material, the material property tensor consists of 81 elements. However, symmetry of stress and strain tensors, and strain energy considerations reduce the number of independent constants required to 21. If this discussion is limited to isotropic elastic materials, only two independent material constants are required to relate stress and strain. The Young modulus, E, and Poisson ratio, ν, are usually selected as material properties to represent elastic isotropic behavior. With reference to Figure 4a, which shows a rod under uniaxial loading, the following are definitions for the Young modulus and Poisson ratio:

$$E = \frac{\sigma_1}{\epsilon_1} \quad [27a]$$

$$\nu = -\frac{\epsilon_3}{\epsilon_1} \quad [27b]$$

Now consider a component under a general state of stress. If we consider the principal stress and strain as shown in Figure 4b, then the strain component ϵ_1 is caused by the direct stress, σ_1, as well as two transverse stress σ_2 and σ_3 due to the Poisson effect. Similar arguments hold for strain components ϵ_2 and ϵ_3.

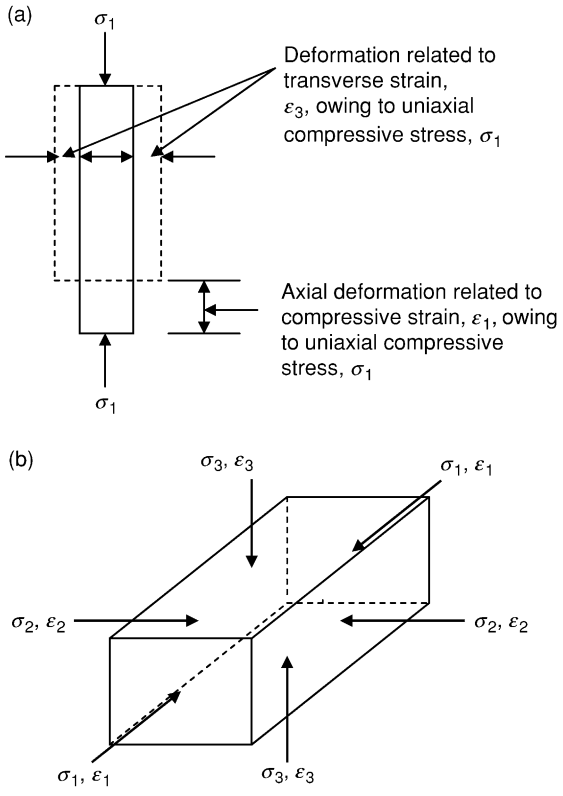


Figure 4 Behavior of (a) a rod under uniaxial loading and (b) a prismatic member under general loading.

Therefore:

$$\epsilon_1 = \frac{1}{E}\sigma_1 - \frac{\nu}{E}\sigma_2 - \frac{\nu}{E}\sigma_3 \qquad [28a]$$

$$\epsilon_2 = \frac{1}{E}\sigma_2 - \frac{\nu}{E}\sigma_3 - \frac{\nu}{E}\sigma_1 \qquad [28b]$$

$$\epsilon_3 = \frac{1}{E}\sigma_3 - \frac{\nu}{E}\sigma_1 - \frac{\nu}{E}\sigma_2 \qquad [28c]$$

Moreover, summing Eqns [28a]–[28c]:

$$\epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon_v = \frac{(1-2\nu)}{E}(\sigma_1 + \sigma_2 + \sigma_3) = \frac{3(1-2\nu)}{E}p \qquad [29]$$

where $\epsilon_v = \epsilon_1 + \epsilon_2 + \epsilon_3$ is the volumetric strain and $p = (\sigma_1 + \sigma_2 + \sigma_3)/3$ has been substituted (see Eqns [4] and [5]). Alternately Eqn [29] can be written as:

$$p = \frac{E}{3(1-2\nu)}\epsilon_v = K\epsilon_v \qquad [30]$$

where $K = E/3(1-2\nu)$ is known as the 'elastic bulk modulus.'

It is always true for an axisymmetric case such as cylindrical loading that $\sigma_2 = \sigma_3$. Subtracting Eqn [28c] from [28a]:

$$\epsilon_1 - \epsilon_3 = \frac{1+\nu}{E}(\sigma_1 - \sigma_3) \qquad [31a]$$

As mentioned earlier, $\epsilon_1 - \epsilon_3 = 2/3\ \epsilon_s$ and $\sigma_1 - \sigma_3 = q$. Therefore, Eqn [31a] becomes:

$$q = 3\left\{\frac{E}{2(1+\nu)}\right\}\epsilon_s = 3G \cdot \epsilon_s \qquad [31b]$$

where $G = E/2(1+\nu)$ is the shear modulus.

In soil mechanics, elastic bulk modulus, K, and shear modulus, G, are usually chosen as material parameters. If the symmetric properties of stress and strain tensors are utilized, Eqn [26] can be rewritten in terms of material parameters, K and G, as:

$$\begin{Bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{23}\\ \sigma_{13}\end{Bmatrix} = \begin{bmatrix} K+\frac{4G}{3} & K-\frac{2G}{3} & K-\frac{2G}{3} & 0 & 0 & 0\\ K-\frac{2G}{3} & K+\frac{4G}{3} & K-\frac{2G}{3} & 0 & 0 & 0\\ K-\frac{2G}{3} & K-\frac{2G}{3} & K+\frac{4G}{3} & 0 & 0 & 0\\ 0 & 0 & 0 & 2G & 0 & 0\\ 0 & 0 & 0 & 0 & 2G & 0\\ 0 & 0 & 0 & 0 & 0 & 2G \end{bmatrix} \begin{Bmatrix}\epsilon_{11}\\ \epsilon_{22}\\ \epsilon_{33}\\ \epsilon_{12}\\ \epsilon_{23}\\ \epsilon_{13}\end{Bmatrix} \qquad [32a]$$

Sometimes it is useful to represent strain in terms of stress (i.e., inverse of Eqn [32a]). This inverse relationship is given by Eqn [32b]:

$$\begin{Bmatrix}\epsilon_{11}\\ \epsilon_{22}\\ \epsilon_{33}\\ \epsilon_{12}\\ \epsilon_{23}\\ \epsilon_{13}\end{Bmatrix} = \begin{bmatrix} \frac{1}{3G}+\frac{1}{9K} & \frac{1}{9K}-\frac{1}{6G} & \frac{1}{9K}-\frac{1}{6G} & 0 & 0 & 0\\ \frac{1}{9K}-\frac{1}{6G} & \frac{1}{3G}+\frac{1}{9K} & \frac{1}{9K}-\frac{1}{6G} & 0 & 0 & 0\\ \frac{1}{9K}-\frac{1}{6G} & \frac{1}{9K}-\frac{1}{6G} & \frac{1}{3G}+\frac{1}{9K} & 0 & 0 & 0\\ 0 & 0 & 0 & \frac{1}{2G} & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{2G} & 0\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2G} \end{bmatrix} \begin{Bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{23}\\ \sigma_{13}\end{Bmatrix} \qquad [32b]$$

For the example problem of cylindrical state of stress, Eqns [30] and [31b] apply. These equations can be written in a matrix form as follows:

$$\begin{Bmatrix} p \\ q \end{Bmatrix} = \begin{bmatrix} K & 0 \\ 0 & 3G \end{bmatrix} \begin{Bmatrix} \epsilon_v \\ \epsilon_s \end{Bmatrix} \quad [33a]$$

Alternatively:

$$\begin{Bmatrix} \epsilon_v \\ \epsilon_s \end{Bmatrix} = \begin{bmatrix} \frac{1}{K} & 0 \\ 0 & \frac{1}{3G} \end{bmatrix} \begin{Bmatrix} p \\ q \end{Bmatrix} \quad [33b]$$

From Eqns [33a] and [33b], it is clear that volumetric and shear behavior are completely decoupled.

Material Behavior under Load

Response of a material to external loading can be linear-elastic, nonlinear-elastic, rigid-perfectly plastic, elastic-perfectly plastic, or elastoplastic. Figure 5 shows these responses.

Soil rarely exhibits a linear-elastic behavior (Figure 5a). A nonlinear-elastic constitutive relationship (Figure 5b) can be used to describe soil loading behavior along a particular loading or unloading path with reasonable accuracy. However, it is not a good indicator of soil behavior under wide varieties of loading and unloading conditions (i.e., general loading–unloading conditions). Rigid-perfectly plastic (Figure 5c) and elastic-perfectly plastic (Figure 5d) models are sometimes used in soil analysis, particularly in limit equilibrium analysis. Figure 5e represents a typical metal bar under uniaxial tension. Often, soil exhibits similar behavior under isotropic compression as well as triaxial compression. In the case of metals, when a specimen is loaded in a uniaxial direction, it usually exhibits a linear elastic behavior until the yield point (point A) is reached on its stress–strain curve. After yield it will further deform if loaded to a point B. If the specimen is now unloaded, it will usually follow an unloading path such as BCD. When the load is completely removed, it will not have recovered its original length but will have a permanent deformation or plastic deformation. If this specimen is reloaded, it may follow a path such as DEB, during which the material usually exhibits elastic behavior. At point B it will yield again and follow path BF. Unloading–reloading curves are generally close to each other, and the area under the curve BCDE represents hysterisis loss. Since point B corresponds to a higher stress level than point A, this type of material is often called a 'work-hardening' material. Note that beyond point B, along path BF, both elastic and plastic deformations will occur. If the specimen is unloaded and reloaded at point F it may trace an unloading–reloading curve FGHI similar to the curve BCDE. As mentioned earlier, soil exhibits similar behavior (overconsolidated soil may show a distinct hump) except that it seldom shows linear elastic behavior even when loaded from a stress-free, undeformed state. During unloading–reloading conditions, soil usually exhibits logarithmic (nonlinear) elastic behavior until the yield point is reached.

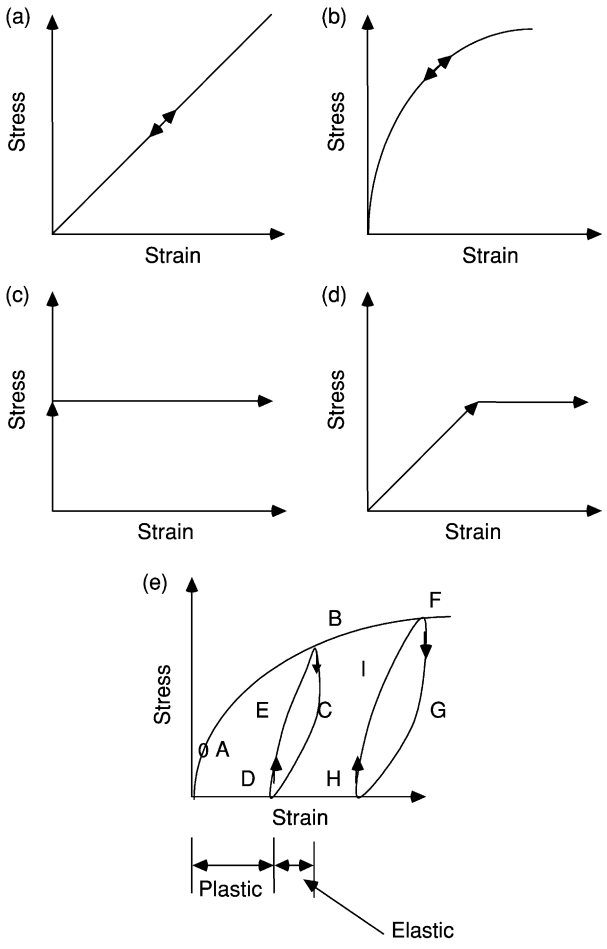


Figure 5 Idealized stress–strain curves for materials: (a) linear-elastic; (b) nonlinear elastic; (c) rigid-perfectly plastic; (d) elastic-perfectly plastic; (e) and elastoplastic behavior.

Nonlinear Elastic Behavior

In principle, it is possible to model the stress–strain behavior of soil using nonlinear elastic models for specific loading paths and conditions. Nonlinearity implies that the elastic parameters are no longer constants but depend on the states of stress and strain of the soil. For these conditions usually an incremental analysis is conducted using tangent moduli values: tangent bulk modulus, K_t, and tangent shear modulus, G_t. Figure 6 shows graphically the significance of tangent moduli. For the case of nonlinear elastic constitutive models, Eqns [32a] and [b] can be written in an incremental form to account for nonlinear behavior as follows:

$$\begin{Bmatrix} d\sigma_{11} \\ d\sigma_{22} \\ d\sigma_{33} \\ d\sigma_{12} \\ d\sigma_{23} \\ d\sigma_{13} \end{Bmatrix} = \begin{bmatrix} K_t + \frac{4G_t}{3} & K_t - \frac{2G_t}{3} & K_t - \frac{2G_t}{3} & 0 & 0 & 0 \\ K_t - \frac{2G_t}{3} & K_t + \frac{4G_t}{3} & K_t + \frac{2G_t}{3} & 0 & 0 & 0 \\ K_t - \frac{2G_t}{3} & K_t - \frac{2G_t}{3} & K_t + \frac{4G_t}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2G_t & 0 & 0 \\ 0 & 0 & 0 & 0 & 2G_t & 0 \\ 0 & 0 & 0 & 0 & 0 & 2G_t \end{bmatrix} \begin{Bmatrix} d\epsilon_{11} \\ d\epsilon_{22} \\ d\epsilon_{33} \\ d\epsilon_{12} \\ d\epsilon_{23} \\ d\epsilon_{13} \end{Bmatrix} \quad [34]$$

and the inverse relationship is given by:

$$\begin{Bmatrix} d\epsilon_{11} \\ d\epsilon_{22} \\ d\epsilon_{33} \\ d\epsilon_{12} \\ d\epsilon_{23} \\ d\epsilon_{13} \end{Bmatrix} = \begin{bmatrix} \frac{1}{3G_t} + \frac{1}{9K_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & 0 & 0 & 0 \\ \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{3G_t} + \frac{1}{9K_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & 0 & 0 & 0 \\ \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{3G_t} + \frac{1}{9K_t} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2G_t} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2G_t} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2G_t} \end{bmatrix} \begin{Bmatrix} d\sigma_{11} \\ d\sigma_{22} \\ d\sigma_{33} \\ d\sigma_{12} \\ d\sigma_{23} \\ d\sigma_{13} \end{Bmatrix} \quad [35]$$

For the triaxial, cylindrical-loading example problem, Eqn [35] takes the following form (cf. Eqn [33b]):

$$\begin{Bmatrix} d\epsilon_v \\ d\epsilon_s \end{Bmatrix} = \begin{bmatrix} 1/K_t & 0 \\ 0 & 1/(3G_t) \end{bmatrix} \begin{Bmatrix} dp \\ dq \end{Bmatrix} \quad [36]$$

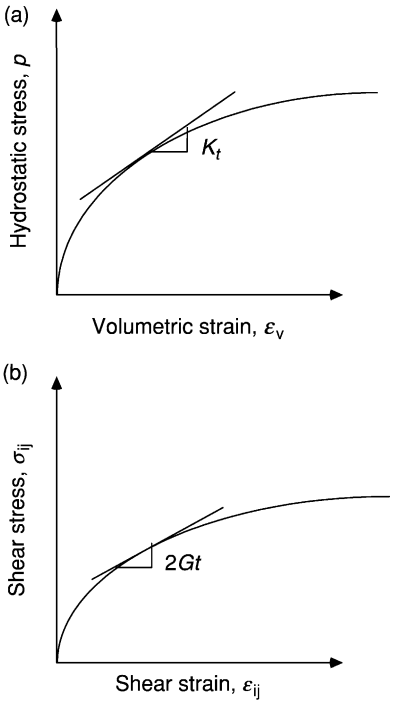


Figure 6 Stress–strain relationship for nonlinear elastic, isotropic material: (a) hydrostatic stress versus volumetric strain; (b) shear stress versus shear strain.

Variable-Moduli Models

Since both K_t and G_t are expected to be a function of stress and strain state of soil, these two tangent moduli can be represented as some explicit function of stress and strain, or stress and strain invariants. These are known as 'variable-moduli models' and many have been proposed for soil. One such model is:

$$K_t = K_0 + K_1 \cdot p \quad [37]$$

and

$$G_t = G_0 + G_1\sqrt{J_{2D}} \quad [38]$$

where K_0 and G_0 are the initial bulk and shear moduli, respectively, and K_1 and G_1 are stress-related material parameters. Here the tangent bulk modulus is assumed to depend on the mean hydrostatic stress and the tangent shear modulus is assumed to depend on the second invariant of the deviatoric stress tensor.

For the example problem using Eqns [36], [37], and [38], the following differential equations are obtained:

$$d\epsilon_v = \frac{dp}{(K_0 + K_1 p)} \quad [39]$$

$$d\epsilon_s = \frac{dq}{3(G_0 + G_1' q)} \quad [40]$$

where $G_1' = G_1/\sqrt{3}$, since $q = \sqrt{3J_{2D}}$.

Upon integration, Eqns [39] and [40] yield:

$$\epsilon_v = \frac{1}{K_1} \ln\left(\frac{K_0 + K_1 p}{K_0}\right) \quad [41]$$

$$\epsilon_s = \frac{1}{3G_1'} \ln\left(\frac{G_0 + G_1' q}{G_0}\right) \quad [42]$$

Under certain circumstances, this model allows the tangent shear modulus to become negative; but, in reality, material would have failed before this happens. From Eqn [42] it is clear that $\epsilon_s \to -\infty$, as $(G_0 + G_1' q) \to 0$. A modification of Eqn [39] is often used to describe the nonlinear elastic behavior of soil before it yields. If K_0 is zero, integration of Eqn [39] gives:

$$\epsilon_v = \frac{1}{K_1} \ln(p) + C$$

or:

$$e = e_i - \frac{(1 + e_i)}{K_1} \ln(p) = e_i - \kappa \ln(p) \quad [43]$$

where e_i is the void ratio at an initial pressure of 1 kPa, κ is the logarithmic bulk modulus which is related to K_1 and e_i (i.e., $\kappa = (1 + e_i)/K_1$). Eqn [43] is commonly used to describe the elastic behavior of soil. For normally consolidated soil (i.e., path A in Figure 7), isotropic compression is also described by a similar equation of the type:

$$e = \Gamma - \lambda \ln(p) \quad [44]$$

where Γ and λ are soil parameters. Eqns [43] and [44] are graphically represented in Figure 7.

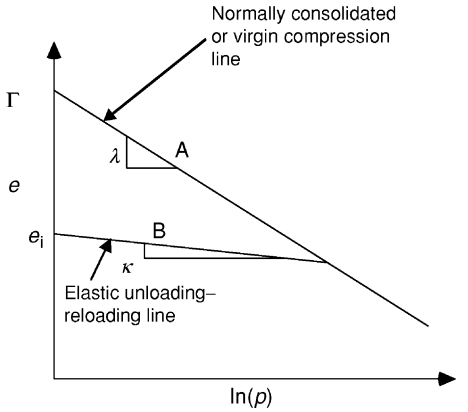


Figure 7 Isotropic consolidation for normally consolidated soil (curve A) and elastic unloading and reloading line (curve B).

The three soil parameters κ, λ, and Γ are used in describing soil behavior in the Cam-clay critical-state model (see the section Critical-State Soil Mechanics, below). Note that both κ and λ are associated with variable-moduli models that represent nonlinear behavior of soil in isotropic compression of normally consolidated soils and unloading–reloading characteristics of soils.

Duncan and Chang model

A well-known, hyperbolic-type stress–strain relationship has been proposed by Duncan and Chang. This model uses the tangent Young modulus and Poisson ratio. It assumes that the stress difference q is a hyperbolic function of axial strain. Figure 8 shows a typical stress–strain curve for both sandy and clayey soils.

This stress–strain curve can be represented by:

$$q = \sigma_1 - \sigma_3 = \frac{\epsilon}{a + b\epsilon} \quad [45]$$

where a is related to the initial tangent modulus ($E_i = 1/a$) and b is related to the asymptotic value of stress difference, $(\sigma_1 - \sigma_3)_{ult}$ ($(\sigma_1 - \sigma_3)_{ult} = 1/b$). It was found that compressive strength of soil at failure was always slightly smaller than this asymptotic stress value. The asymptotic value can be related to compressive failure stress by a factor, R_f, i.e.:

$$(\sigma_1 - \sigma_3)_f = R_f (\sigma_1 - \sigma_3)_{ult} \quad [46]$$

Combining Eqn [45] with [46]:

$$(\sigma_1 - \sigma_3) = \frac{\epsilon}{\left[\frac{1}{E_i} + \frac{\epsilon R_f}{(\sigma_1 - \sigma_3)_f}\right]} \quad [47]$$

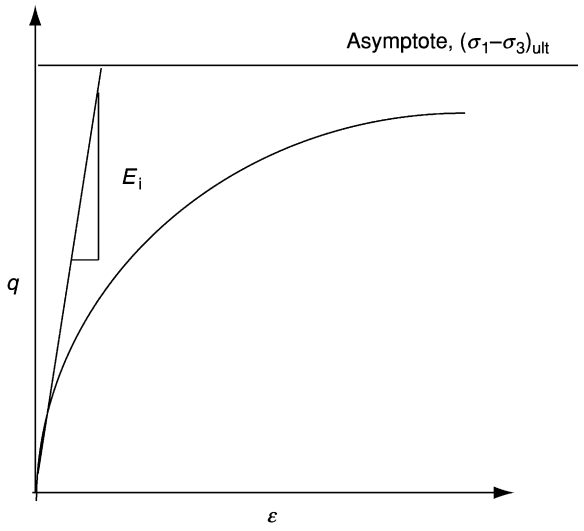


Figure 8 Variation of stress difference or deviator stress q as a function of axial strain ϵ.

The response of many soils depends on the confining pressure for a given stress path. Duncan and Chang used a relationship proposed by Janbu to represent initial tangent modulus as a function of confining stress as follows:

$$E_i = K' p_0 (\sigma_3 / p_0)^n \quad [48]$$

where K' and n are empirical constants, p_0 is the atmospheric pressure, and σ_3 is the minor principal stress. Moreover, the compressive pressure at failure in terms of the Mohr–Coulomb failure criterion is: funny initial moe

$$(\sigma_1 - \sigma_3)_f = \frac{2c\ \cos(\phi) + 2\sigma_3\ \sin(\phi)}{1 - \sin(\phi)} \quad [49]$$

where c is cohesion and ϕ is soil internal angle of friction. The tangent Young modulus can be obtained by differentiating Eqn [47] with respect to axial strain, i.e.:

$$E_t = \frac{\partial(\sigma_1 - \sigma_3)}{\partial \epsilon} \quad [50]$$

Eqns [47]–[50] show that:

$$E_t = (1 - R_f S)^2 E_i \quad [51]$$

where S is the fraction of mobilized stress strength and is given by:

$$S = \frac{(\sigma_1 - \sigma_3)}{(\sigma_1 - \sigma_3)_f} \quad [52]$$

or

$$E_t = \left[1 - \frac{R_f(1 - \sin(\phi))(\sigma_1 - \sigma_3)}{2c\ \cos(\phi) + 2\sigma_3\ \sin(\phi)}\right]^2 K' p_0 \left(\frac{\sigma_3}{p_0}\right)^n \quad [53]$$

This expression involves five parameters. It is relatively easy to implement such models into numerical analysis techniques such as the finite element technique (FEM), and a constant value of the Poisson ratio can be used. It should be noted that although these models can be readily incorporated into numerical analysis, the path dependency of the parameters used limits their widespread use. Since plastic strain dominates soil behavior following yield (point A in Figure 5), soil behavior can be better represented by elastoplastic models.

Elastoplastic Behavior of Soil

As mentioned previously, once the material yields (point A in Figure 5), there will be some irrecoverable (plastic) deformation. There are two important aspects of plastic behavior of materials: (1) a yield criterion, and (2) postyield behavior.

Yield Criteria

When the material arrives at a certain state of stress under the action of an external load, it may permanently lose the ability to regain its original dimensions for any further increase in that load. The relationship between various stress components at this limiting situation (transition from elastic to plastic region) is expressed in terms of a scalar function, f, known as the yield criterion, i.e.:

$$f = f(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}) \quad [54]$$

For isotropic materials, this criterion can be expressed in terms of principal stresses or stress invariants as follows:

$$f = f(J_1, J_2, J_3) \quad [55]$$

Hydrostatic stress seldom plays a role in metal failure, therefore the yield criterion given in Eqn [54] is usually expressed only in terms of stress invariants of the deviatoric stress tensor as:

$$f = f(J_{2D}, J_{3D}) \quad [56]$$

where J_{3D} is the third invariant of the deviatoric stress tensor. J_{3D} is not dealt with here, since most of the widely used yield criteria are represented in terms of J_{2D} only. The well-known von Mises yield criterion depends on J_{2D} as follows:

$$J_{2D} = \frac{\sigma_y^2}{3} \quad [57]$$

where σ_y is the yield stress under uniaxial load. Another widely used yield criterion is the Tresca yield criterion, which is given by:

$$\sigma_1 - \sigma_3 = \sigma_y \quad [58]$$

Although, for the case of the cylindrical triaxial loading (i.e., the example problem), both von Mises and Tresca yield criteria reduce to the same equation, since $J_{2D} = (\sigma_1 - \sigma_3)^2/3$ (cf. Eqn [10]), in general these two yield criteria differ from each other.

Unlike for metals, the yield criterion for geologic materials depends on the first invariant of the stress tensor or the hydrostatic pressure. This is because these types of materials exhibit internal friction, which results in frictional forces that increase with normal load. The well-known Mohr–Coulomb failure criterion is a yield function that accounts for internal friction within the soil mass. This yield criterion is commonly expressed as:

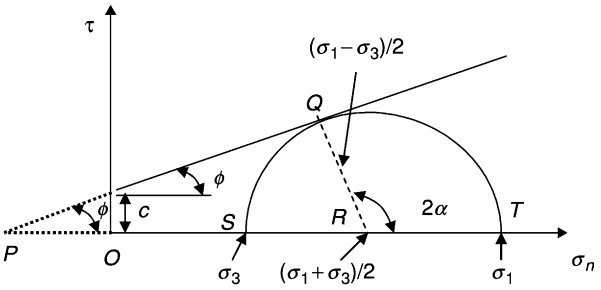


Figure 9 Mohr–Coulomb failure criterion for geologic materials. Point O is the origin of the coordinate axes, *R* is the center of the Mohr circle, *S* and *T* are intersection points of the Mohr circle with the normal stress axis, *Q* is the point of tangency between the Mohr circle and Mohr–Coulomb failure line, *P* is the point of intersection of the Mohr–Coulomb failure line with the normal stress axis.

$$\tau_{\max} = c + \sigma_{\mathrm{n}} \tan(\phi) \qquad [59]$$

where $\tau_{\max}$ is the maximum shear stress; σ_{n} is the normal stress; c is cohesion; and ϕ is the soil internal angle of friction. Figure 9 is a graphical representation of the Mohr–Coulomb failure criterion superimposed on a typical Mohr circle. Eqn [59] can also be written in terms of principal stress as:

$$\sigma_1 = \sigma_3 \tan^2\alpha + 2c \tan\alpha \qquad [60]$$

where α is $45^\circ + \phi/2$. It should be noted that, for frictionless materials (i.e., $\phi = 0$), the Mohr–Coulomb failure criterion reduces to the Tresca yield criterion. The Mohr–Coulomb yield criterion is difficult to implement in three-dimensional applications, because the yield surface is an irregular hexagonal pyramid with sharp corners (mathematical singularities). A generalization, known as the 'Drucker–Prager yield criterion,' that results in a yield surface in the form of a smooth right circular cone is commonly used in modeling granular materials. This yield criterion is expressed in terms of stress invariants as follows:

$$\sqrt{J_{2\mathrm{D}}} - \alpha J_1 - k = 0 \qquad [61]$$

where α and k are related to internal angle of friction and cohesion. Figure 10 shows a plot of the Drucker–Prager yield criterion in the $\sqrt{J_{2\mathrm{D}}}$ versus J_1 plane. For purely cohesive soils, α is zero and the Drucker–Prager yield criterion reduces to von Mises yield criterion.

Hardening Cap

Many geologic materials experience almost continuous plastic deformation from the very beginning of loading. Thus if a mass of soil previously loaded to point B in Figure 10 is unloaded along path BA and then loaded even hydrostatically (i.e., $\sigma_1 = \sigma_2 = \sigma_3 = p$) along the J_1 axis, it will deform elastically until it is loaded to point C, and any further loading leads to plastic deformation. The continuous curved surface, which is assumed to join the Drucker–Prager fixed-yield surface smoothly at point C, is called the 'hardening cap.' As the material is loaded to point E along the hydrostatic axis, the hardening cap is pushed out to location DE, thus extending the elastic region from OPBC to OPDE. In essence, the material has become harder under unloading and reloading. Thus, if the soil mass is loaded along an arbitrary path EFG after the hydrostatic loading to E, the soil will behave elastically up until it reaches a stress state defined by point F. Further loading will cause the soil to deform plastically until it reaches point G on the fixed failure surface defined by the Drucker–Prager yield condition. A state of stress above the line PG is not attainable according to this yield criterion. This idea of hardening caps has been incorporated in the development of an unified theory of soil mechanical behavior that links volume changes, stress state, and yielding into a single framework called the critical-state soil mechanics.

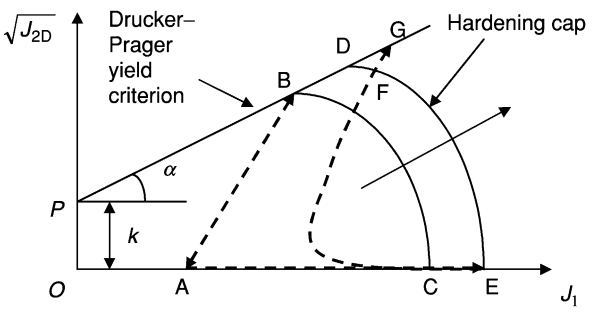


Figure 10 Drucker–Prager yield criterion with hardening caps.

Postyield Behavior

Upon yield, soil mass undergoes plastic deformation. The total strain can be decomposed into elastic and plastic strain, i.e.:

$$\epsilon_{\mathrm{ij}} = \epsilon^{\mathrm{e}}_{\mathrm{ij}} + \epsilon^{\mathrm{p}}_{\mathrm{ij}} \qquad [62]$$

where $\epsilon^{\mathrm{e}}_{\mathrm{ij}}$ is the recoverable elastic strain and $\epsilon^{\mathrm{p}}_{\mathrm{ij}}$ is the permanent plastic deformation. The elastic strain can be determined based on a constitutive relationship such as in Eqn [35]. The plastic strain is determined using the concept of plastic potential and flow rule.

Plastic potential and flow rule The plastic potential, g, is assumed to be a scalar function of the state of stress within the soil mass (i.e., $g = g(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23})$). The direction of the plastic strain is

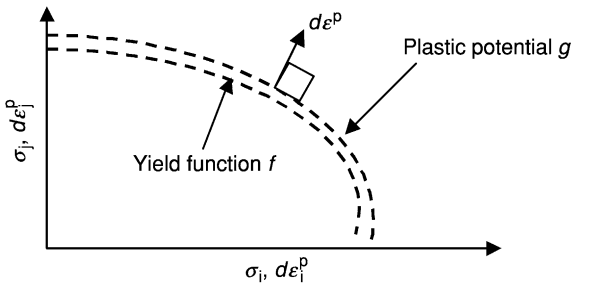


Figure 11 Plastic potential and yield function for a material governed by the associative flow rule.

defined by a flow rule. According to the flow rule, the incremental plastic strain vector is oriented in the direction of the normal to the yield potential (Figure 11). This condition is known as 'normality rule.' Using this definition, the plastic strain is given by:

$$\epsilon_{ij}^p = \chi \frac{\partial g}{\partial \sigma_{ij}} \qquad [63]$$

where χ is a positive scalar multiplier or loading index. For some materials, the plastic potential g is assumed to be the same as the yield function f. These materials are said to follow the associative flow rule; materials for which the plastic potential differs from the yield function are said to follow the nonassociative flow rule.

Hardening law An additional concept that is necessary to describe the plastic behavior completely is the hardening law. The hardening law describes the growth of yield function or hardening cap as the material undergoes plastic deformation (e.g., the rule for the growth of the hardening cap from BC to DE in Figure 10). A hardening parameter, h, is often included in the yield function to describe the changes in the yield function with plastic flow, i.e.:

$$f = f(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}, h) \qquad [64a]$$

Critical-State Soil Mechanics

The critical-state concept is based on the observation that when a soil sample is subjected to increasing shear loading, it will yield and undergo plastic deformations (both volumetric and shear) and finally arrive at a critical volume, after which its volume remains unchanged, although its shape continues to change. The Cam-clay model for the elastoplastic behavior of wet clay includes this critical-state concept while accounting for the stress and volume changes during yielding. A slightly modified version of this original model has been used widely in numerical simulation studies. Here this 'modified Cam clay model' is considered for the particular case of a triaxial loading of a cylindrical soil sample to keep the mathematical treatment relatively simple. The hydrostatic stress ($p = (\sigma_1 + 2\sigma_3)/3$), shear stress ($q = (\sigma_1 - \sigma_3)$), and void ratio ($e$) are used as the state variables in developing the model.

Figure 12 graphically represents the soil behavior in the modified Cam-clay model. Figure 12a shows the variation of void ratio as a function of the hydrostatic pressure. Curve ABCD is the normal consolidation line (NCL) under hydrostatic loading. Curves EB and FC are elastic unloading–reloading curves (URL). The NCL and URL curves are mathematically represented by Eqns [43] and [44]. Figure 12b is the representation of NCL and URL in the void ratio versus $\ln(p)$ plane (cf. Figure 7). Figure 12c shows the critical-state line (CSL) and the elliptical yield surface, OPQ, that is hypothesized by the modified Cam-clay model in the q–p plane. The major axis of the ellipse coincides with the hydrostatic stress axis and is equal to $(p_0/2)$, which is the hardening parameter, h. The critical-state line is assumed to pass through the origin with slope M in the q–p plane and intersects the yield surface at point P. When the soil arrives at the critical-state line, it undergoes no more volume change, i.e., $d\epsilon_v$ is zero. This requirement along with the normality condition can be used to show that point P is directly above the center of the elliptical yield surface. Therefore the coordinates of this point are $(p_0/2, Mp_0/2)$, and $(Mp_0/2)$ is the minor axis of the ellipse:

$$\frac{\left(p - \frac{p_0}{2}\right)^2}{\left(\frac{p_0}{2}\right)^2} + \frac{q^2}{\left(\frac{Mp_0}{2}\right)^2} = 1 \qquad [64b]$$

Multiplying by $(Mp_0/2)^2$ and simplifying:

$$M^2[p(p - p_0)] + q^2 = 0 \qquad [64c]$$

Figure 12d is a three-dimensional (3-D) representation of the modified Cam-clay model which shows that the critical-state line is a space curve, the projection of which is a straight line on the q–p plane.

Modified Cam-Clay Constitutive Relationship

The following relationships are developed using the associative flow rule (i.e., plastic potential is the same as the yield function). The elastic portion of the total strain is given by Eqn [36], i.e.:

$$\begin{Bmatrix} d\epsilon_v^e \\ d\epsilon_s^e \end{Bmatrix} = \begin{bmatrix} 1/K_t & 0 \\ 0 & 1/(3G_t) \end{bmatrix} \begin{Bmatrix} dp \\ dq \end{Bmatrix} \qquad [65]$$

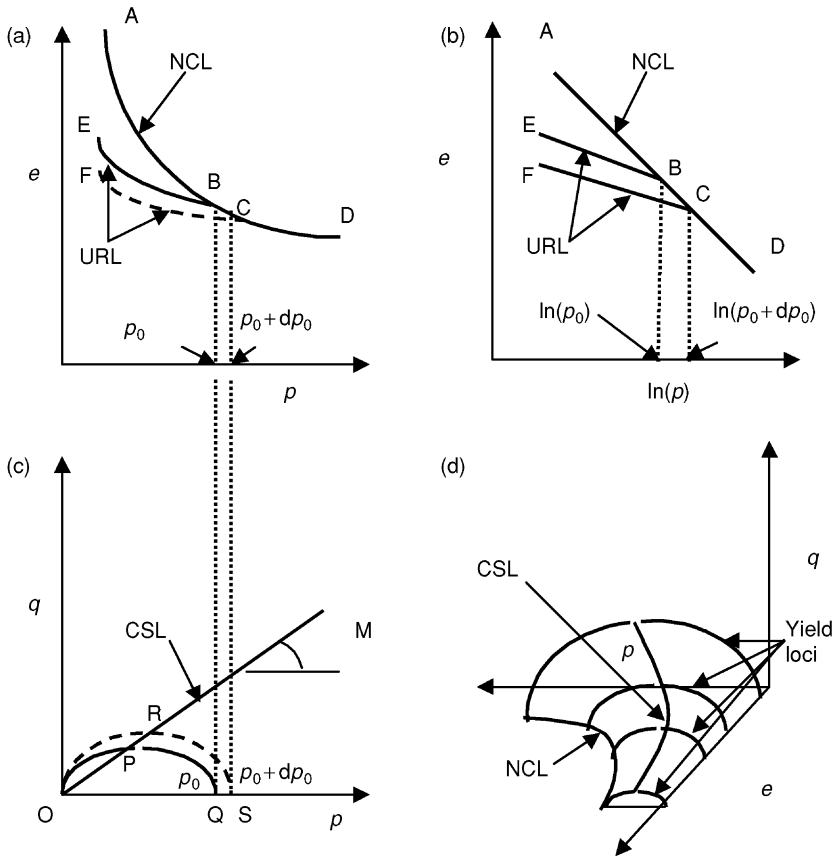


Figure 12 Modified Cam-clay model with elliptical yield surface: (a) deformation in the *e*-*p* plane; (b) deformation in the *e*-ln(*p*) plane; (c) behavior in the *q*-*p* plane; and (d) three-dimensional representation of the model. NCL, normal consolidation line; URL, unloading–reloading line; CSL, critical-state line.

where superscript 'e' represents elastic strain components. The plastic strain components are given by:

$$\mathrm{d}\epsilon_{\mathrm{v}}^{\mathrm{p}} = \chi \frac{\partial g}{\partial p}$$
$$\mathrm{d}\epsilon_{\mathrm{s}}^{\mathrm{p}} = \chi \frac{\partial g}{\partial q} \qquad [66]$$

where superscript 'p' stands for plastic strain components. Because of the assumption of associative flow, the plastic potential, g, is the same as the yield function, f. The elliptical yield function described in the previous section is given by (cf. Eqn [64c]):

$$f(p,q,p_0) = g(p,q,p_0) = M^2[p(p-p_0)] + q^2 = 0 \qquad [67]$$

Eqn [67] can be used to obtain an expression for loading index, χ, as follows. Taking the total differential of Eqn [67]:

$$\mathrm{d}f = \mathrm{d}g = \frac{\partial f}{\partial p}\mathrm{d}p + \frac{\partial f}{\partial q}\mathrm{d}q + \frac{\partial f}{\partial p_0}\mathrm{d}p_0 \qquad [68]$$

However, $\mathrm{d}p_0$, which is related to the size of the yield surface, i.e., changes in the hardening parameter, can be related to volumetric plastic strain using Eqns [43] and [44]. Referring to **Figure 12b**, from Eqns [43] and [44]:

$$e_0 = e_{\mathrm{i}} - \kappa \ln(p_0) \qquad [69a]$$

$$e_0 = \Gamma - \lambda \ln(p_0) \qquad [69b]$$

The elastic part of the volume change is related to that described in Eqn [69a], and the total volume change is related to that described in Eqn [69b]. Taking differentials of Eqns [69a] and [69b]:

$$\mathrm{d}e^{\mathrm{e}} = -\kappa \frac{\mathrm{d}p_0}{p_0} \qquad [70a]$$

$$\mathrm{d}e = -\lambda \frac{\mathrm{d}p_0}{p_0} \qquad [70b]$$

Therefore, the plastic part of the deformation is given by:

$$\mathrm{d}e^{\mathrm{p}} = \mathrm{d}e - \mathrm{d}e^{\mathrm{e}} = -(\lambda - \kappa)\frac{\mathrm{d}p_0}{p_0} \qquad [71]$$

From Eqns [24b] and [71]:

$$d\epsilon_v^p = -\frac{de^p}{1+e_i} = \left\{\frac{\lambda-\kappa}{1+e_i}\right\}\frac{dp_0}{p_0} \quad [72]$$

Eqn [72] can be solved for dp_0 as follows:

$$dp_0 = p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}d\epsilon_v^p \quad [73]$$

Therefore, from Eqns [66], [68], and [73]:

$$\frac{\partial f}{\partial p}dp + \frac{\partial f}{\partial q}dq + \frac{\partial f}{\partial p_0}p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}\chi\frac{\partial g}{\partial p} = 0 \quad [74]$$

Solving Eqn [74] for χ gives:

$$\chi = \frac{-\left(\frac{\partial f}{\partial p}dp + \frac{\partial f}{\partial q}dq\right)}{\frac{\partial f}{\partial p_0}p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}\frac{\partial g}{\partial p}} \quad [75]$$

If $\eta = q/p$ is substituted as η, then for the Cam-clay model:

$$\frac{\partial f}{\partial p} = \frac{\partial g}{\partial p} = (M^2-\eta^2)p$$
$$\frac{\partial f}{\partial q} = \frac{\partial g}{\partial q} = 2\eta p$$
$$\frac{\partial f}{\partial p_0} = -M^2 p \quad [76]$$

where $p_0 = (M^2+\eta^2)p/M^2$ is substituted for from Eqn [67]. (From Eqn [67]: $f(p, q, p_0) = M^2\,[p\,(p - p_0)] + q^2 = 0$; therefore $M^2\,p^2 - M^2\,pp_0 + \eta^2\,p^2 = 0$, since $q = \eta p$; further simplification of this equation leads to $p^2\,(M^2+\eta^2) = M\,pp_0$, or $p_0 = (M^2+\eta^2)\,p/M^2$.) Substituting these partial derivatives in Eqn [75] and simplifying:

$$\chi = \left(\frac{\lambda-\kappa}{1+e_i}\right)\frac{(M^2-\eta^2)dp + 2\eta\,dq}{(M^2+\eta^2)(M^2-\eta^2)p^2} \quad [77]$$

Finally, from Eqn [66] and [76], upon simplification:

$$d\epsilon_v^p = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\left[(M^2-\eta^2)dp + 2\eta dq\right] \quad [78a]$$

$$d\epsilon_s^p = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\left[2\eta dp + \frac{4\eta^2}{(M^2-\eta^2)}\right] \quad [78b]$$

or in matrix form,

$$\left\{\begin{matrix} d\epsilon_v^p \\ d\epsilon_s^p \end{matrix}\right\} = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\begin{bmatrix} (M^2-\eta^2) & 2\eta \\ 2\eta & \frac{4\eta^2}{(M^2-\eta^2)} \end{bmatrix} \quad [79]$$

If the soil mass is within the elastic region, Eqn [65] can be used to compute elastic strain. However, if the soil mass has yielded and is undergoing elastic as well as plastic deformation, Eqn [79] can be used to obtain plastic strain while Eqn [65] provides the magnitude of the elastic strain. This model contains five soil parameters (κ, ν or G, λ, M, Γ). The first two parameters are related to the elastic behavior of soil mass whereas the last three parameters are related to the plastic behavior. Note that the parameter κ is related to soil bulk modulus K by the relationship, $\kappa = \frac{1+e_i}{\kappa}$, as shown previously.

Although the elliptic yield function represents the behavior of normally consolidated and lightly overconsolidated soils reasonably well, it overpredicts the strength of overconsolidated soils.

NSDL–AU Model

Based on numerous tests conducted using unsaturated agricultural soils at the National Soil Dynamics Laboratory (NSDL) and Auburn University (AU) in Alabama, USA, a model for plastic behavior of soils has been formulated, the NSDL–AU model, which is based on two key developments. The first is the development of a model to describe the hydrostatic compression of unsaturated agricultural soils. This equation for NCL differs from Eqn [43] and has the following form:

$$\bar{\epsilon}_v = (A + Bp)(1 - e^{-Cp}) \quad [80]$$

where $\bar{\epsilon}_v$ is the natural volumetric strain, $\ln(V/V_0)$, and A, B, and C are material parameters. Eqn [80] has been extended to represent cylindrical triaxial test results. The extended equation is given by:

$$\bar{\epsilon}_v = (A + Bp)(1 - e^{-Cp}) + \frac{\sqrt{2}}{3}D\frac{q}{p} \quad [81]$$

Note the use of $\tau_{oct} = \sqrt{2}/3\ q$ in this formulation. Moreover, this equation contains an additional material parameter, D. Eqn [81] represents the yield surface for unsaturated agricultural soils and is quite similar to the yield surface shown in **Figure 12d**. Furthermore, the model inherently contains a critical-state line. It should be noted that Eqn [81] represents plastic behavior of soil and does not account for any unloading–reloading action (i.e., elastic behavior). Most agricultural soils tend to be slightly overconsolidated and their deformation is primarily controlled by plastic deformation.

Composite Soil-Strength Parameters

Although the foregoing discussion treats mechanical behavior of soil from the veiwpoint of continuum mechanics, it has not been used widely by field researchers and practitioners due to its complexity and the difficulty in obtaining engineering soil parameters of undisturbed *in situ* soils. Devices such as cone penetrometers, and shear and sinkage devices have been developed to provide various measures of soil strength for use in tillage, traction, and soil-compaction studies. These devices provide soil-strength parameters that tend to depend on the geometry of the test device and type of loading applied. Moreover, these soil parameters often do not represent any single soil property, but are usually functions of several fundamental engineering properties of soil. Therefore soil parameters obtained using these devices are often called 'composite soil parameters'; e.g., penetrometers, shear-vane devices, and shear graphs. There are also various techniques used for determining soil stickiness, shatter resistance, and cutting resistance. The cone penetrometer is well-known among these devices because it is a simple device and very easy to use. Shear-vane devices and shear graphs have been used to obtain soil shear and sinkage parameters. They are useful in predicting tractive ability of wheels and tracks using a semiempirical approach.

Cone Penetrometer

Perhaps the most widely used device to measure soil strength in the field is the cone penetrometer. Although the cone penetrometer was developed to determine the mobility of off-road vehicles at the Waterways Experiment Station in Vicksburg, Mississippi, USA, it has been used to predict traction, draft requirements of tillage implements, and to quantify soil strength to indicate soil-compaction level and impedance to root growth. The most common form of this device consists of a polished steel cone, which is pushed against the soil and then the force of penetration is measured. The American Society of Agricultural Engineers (ASAE) has developed a standard (S313.3) that describes the geometry of a standard cone penetrometer, whereas a second standard (EP542) outlines the proper procedure for using this device.

Since the force needed to penetrate the soil is related to the geometry of the device, a cone with a base diameter of 20.27 mm and an apex angle of 30° is selected as the standard shape (Figure 13). For harder soil conditions, a smaller cone with a base diameter of 12.83 mm is used. A second key variable that influences the force of penetration is the penetration rate. ASAE Standard EP542 recommends a quasistatic rate of 1.83 m min^{-1}. Although it is difficult to control the insertion rate with handheld devices, hydraulically or electrically operated devices can be designed to operate at this standard speed. Force is usually measured using a load-sensing mechanism such as a load cell. Newer devices often include a depth-measuring sensor such as a potentiometer so that a soil penetration resistance profile can be obtained. The penetration resistance force is expressed as cone index, which is the ratio of the force to the base area of the cone. The soil cone index value obtained using a soil cone penetrometer is a composite value that depends on soil texture, bulk density, and moisture content. In terms of engineering properties of soil, it depends on cohesion, soil internal angle of friction, soil metal friction, and adhesion.

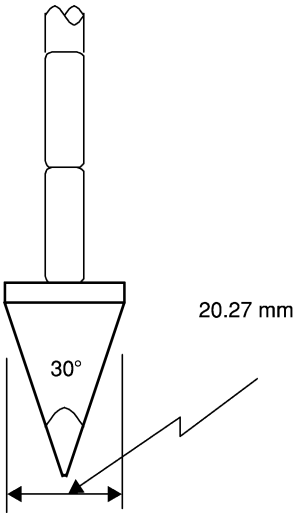


Figure 13 Standard cone penetrometer of the American Society of Agricultural Engineers.

One of the main concerns with the use of cone index to represent soil strength is its variability, especially in dry and cloddy conditions. ASAE standard EP542 recommends that at least 20 measurements must be taken near the field capacity of soil in a given location to obtain a representative measure of soil strength. With the advent of precision agriculture and the potential role of soil compaction in limiting water infiltration, drainage, and root growth, there is an increased interest in the cone penetrometer as a soil-strength mapping tool. Consequently, fully automated cone penetrometers with global positioning systems (GPS) to provide geographic position data are now commercially available.

Measurement of Soil Sinkage and Shear

Inadequacy of cone index values in representing soil characteristics relevant to tractive ability of wheels

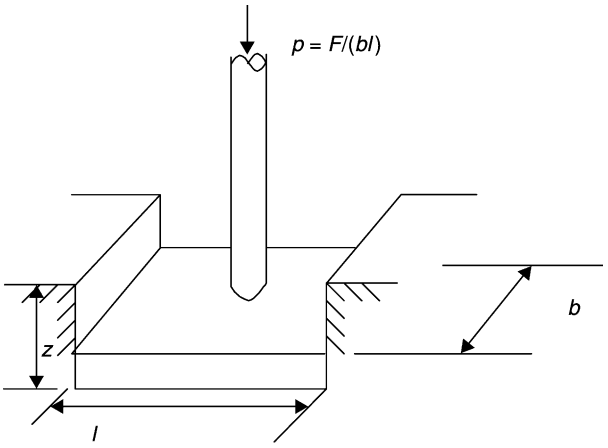


Figure 14 A rectangular sinkage plate.

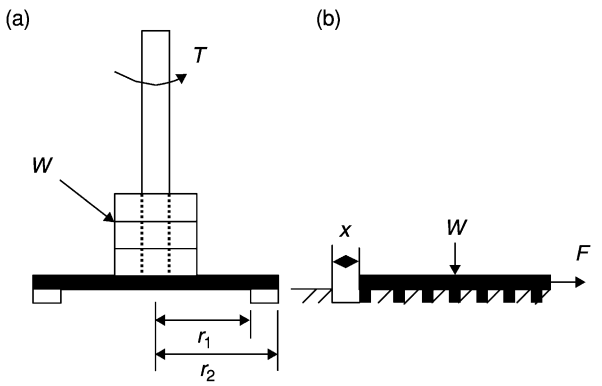


Figure 15 Soil-shear test devices: (a) shear ring; and (b) grouser plate.

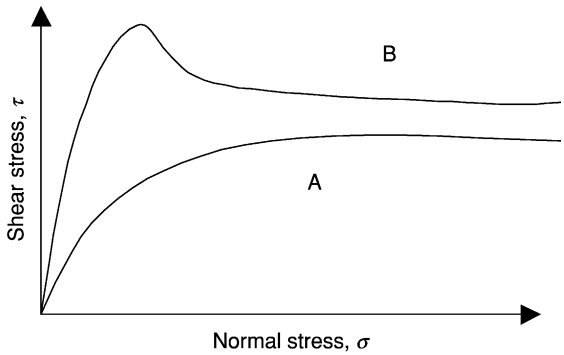


Figure 16 Shear characteristics of soil (A) normally consolidated or slightly overconsolidated and (B) overconsolidated.

and tracks has led to the development of sinkage and shear devices such as the bevameter.

Soil sinkage devices consist of either circular or rectangular plates that are pushed against soil, and their load deformation characteristics are recorded (Figure 14). The Bernstein equation is often used to relate applied load to soil deformation as follows:

$$p_s = kz^n \qquad [82]$$

where p_s is the applied compressive pressure, z is the soil sinkage, and k and n are constants. The sinkage parameter was found to depend on plate width by the following relationships:

$$k = \frac{k_c}{b} + k_\phi \qquad [83]$$

where k_c and k_ϕ are parameters related to soil cohesion and angle of internal friction, respectively, and b is the minimum plate dimension. Since the parameter n in Eqn [82] is usually not an integer, units of k are not straightforward. An alternate formulation that overcomes the problem of dimension of k is:

$$p = k_r \left(\frac{z}{b}\right)^n \qquad [84a]$$

and:

$$k_r = (k_1 + k_2\, b) \qquad [84b]$$

where k_1 and k_2 are once again parameters related to soil cohesion and angle of internal friction, respectively. Note that the unit of k_r in Eqn [84a] is the same as that of pressure. These plate sinkage relationships have been used to model rolling resistances of wheels and tracks.

Shear characteristics of soil such as cohesion and angle of internal friction have been used to model the tractive ability of wheels and tracks and soil cutting by tillage tools. Torsional shear devices (shear vane, shear cone, shear ring, shear graph) and rectangular grouser plates are often used to measure soil-shear characteristics. Figure 15 shows a circular shear head and a rectangular grouser plate. The maximum shear stress at a specific normal load, W, on a torsional device as shown in Figure 15a is given by:

$$\tau = \frac{3T}{2\pi(r_1^3 - r_2^3)} \qquad [85]$$

where T is the torque and r_1 and r_2 are the radii of the shear vane.

The original bevameter used a shear device similar to the one shown in Figure 15a. A recent development includes an instrumented soil test device that uses a grouser plate, sinkage plate, and a standard cone to obtain soil shear, sinkage, and cone index data. Figure 16 shows the shear characteristics of soil measured using a shear test device. The curve A in Figure 16 is a representation of a normally consolidated or slightly overconsolidated soil. The

curve B is a typical response of an overconsolidated soil. Most agricultural soils behave as shown in curve A, which can be represented by the empirical equation:

$$\tau = \tau_{max}(1 - e^{-x/k_s}) \quad [86]$$

where τ_{max} is the maximum shear stress, k_s is the soil-shear modulus, and x is the soil deformation in the direction of shear stress.

Moreover, the Mohr–Coulomb criterion (cf. Eqn [59]) can be used to represent τ_{max}. Note that $\tau_{max} = F/A$ and $\sigma_n = W/A$, where A is the plate area. If the shear test is repeated at two different values of the normal load, W, both c and ϕ can be determined.

Some Recent Developments

Increased interest in precision agriculture within the last decade has led to the investigation of causes of yield variability within a field. One factor believed to influence crop yield is soil compaction, since it has a direct impact on soil hydraulic conductivity. As mentioned previously, the cone penetrometer is the soil-strength measuring device that is being used increasingly to map soil compaction level. Since it is a highly variable point measurement, numerous cone index values are needed to obtain proper representation of a field. This limitation of the cone penetrometer has led to the development of alternative devices that can measure and map soil strength in a continuous manner. One such device consists of a texture–soil-compaction sensing system that consists of a simple tine that is instrumented with a load cell to measure soil-cutting force. It also incorporates a dielectric-based soil-moisture sensor, because soil-moisture content influences soil-cutting force significantly. The soil-cutting force, F, is a function of soil bulk density, ρ, texture, ξ, and moisture content, θ, when the device is operated at a constant speed and operating depth; i.e.:

$$F = f(\rho, \xi, \theta) \quad [87]$$

Based on the field measurements the soil-cutting force was determined as:

$$F = g(\rho, \xi) * e^{-c\theta} \quad [88]$$

where c is an empirical constant. The unknown function $g(\rho, \xi)$ is 'texture/soil compaction index' (TCI). Therefore, TCI is given by:

$$TCI = F/e^{-c\theta} \quad [89]$$

Note that this TCI value depends on both soil bulk density and texture. Since texture is a static property in the absence of cut-and-fill operation, TCI values can be used as an indicator of soil-compaction level. The TCI sensor has been interfaced to a differential global positioning system (DGPS) to obtain soil-strength maps of tomato fields and correlate them to tomato yield. The field test results indicate that, although the TCI sensor works reasonably well, it is not helpful in locating the compact layer within the soil mass. As of 2003, active research is currently in progress to develop a compaction profile sensor, which can measure the compaction level of soil with depth. Successful development of such a real-time soil-compaction profile sensor may contribute to the development of site-specific tillage (tilling only where there is a need to loosen soil) and limit tillage depth to the hardpan depth.

See also: **Compaction**; **Conservation Tillage**; **Cultivation and Tillage**; **Site-Specific Soil Management**; **Structure**; **Subsoiling**; **Swelling and Shrinking**

Further Reading

ASAE (2000) *Soil Cone Penetrometer.* ASAE Standard S313.3, pp. 832–833. St. Joseph, MI: ASAE.

ASAE (2000) *Procedures for Using and Reporting Data Obtained with the Cone Penetrometer.* ASAE Standard EP542, pp. 987–989. St. Joseph, MI: ASAE.

Bailey AC and Johnson CE (1989) A soil compaction model for cylindrical stress states. *Transactions of the American Society of Agricultural Engineers* 32(3): 822–825.

Bailey AC and Johnson CE (1994) *NSDL-AU Model, Soil Stress–Strain Behavior.* ASAE Paper No. 94-1074. St. Joseph, MI: ASAE.

Bailey AC, Johnson CE, and Schafer RL (1984) Hydrostatic compaction of agricultural soils. *Transactions of the American Society of Agricultural Engineers* 27(4): 925–955.

Das BJ (1983) *Advanced Soil Mechanics*. New York: McGraw-Hill.

Desai CS and Siriwardane HJ (1994) *Constitutive Laws for Engineering Materials with Emphasis on Geologic Materials*. Englewood Cliffs, NJ: Prentice Hall.

Duncan JM and Chang CY (1970) Non-linear analysis of stress and strain in soils. *Journal of Soil Mechanics and Foundations Division* 96: 1629–1653.

Hettiaratchi DRP (1987) A critical state soil mechanic model for agricultural soil. *Soil Use and Management* 3(3): 94–105.

Kirby JM (1989) Measurements of the yield surfaces and critical-state of some unsaturated agricultural soils. *Journal of Soil Science* 40: 167–182.

Roscoe KH and Burland JB (1968) On the generalized stress–strain behavior of "wet" clay. In: Heyman J and Leckie FA (eds) *Engineering Plasticity*, pp. 335–609. Cambridge, UK: Cambridge University Press.

Roscoe KH, Schofield A, and Worth CP (1958) On yielding of clay soils. *Geotechnique* 8: 22–53.
Upadhyaya SK, Chancellor WJ, Perumpral RL *et al.* (1994) *Advances in Soil Dynamics*, vol. 1. St Joseph, MI: ASAE.
Upadhyaya WJ, Chancellor JV, Perumpral RL *et al.* (2002) *Advances in Soil Dynamics*, vol. 2. St Joseph, MI: ASAE.
Wood DM (1990) *Soil Behavior and Critical State Soil Mechanics*. New York: Cambridge University Press.

STRUCTURE

V A Snyder and M A Vázquez, University of Puerto Rico Agricultural Experiment Station, San Juan, Puerto Rico 00926–1118

Introduction

The term 'structure' of a granular medium refers to the spatial arrangement of solid particles and void spaces. In materials such as coarse sands and gravels, the particles are loosely bound and tend to arrange themselves in closely packed, minimum-energy configurations (Figure 1a). Most soils, however, tend to exhibit a hierarchical structure. That is, primary mineral particles, usually in association with organic materials, form small clusters or 'first-order aggregates.' These in turn form larger clusters or 'second-order aggregates,' and so on, as illustrated schematically in Figure 1b.

Aggregate hierarchy in soils is reflected not only in increasing aggregate size with each successive level, but also in the predominant mechanisms by which particles in aggregates are bonded together; i.e., aggregates at different hierarchical levels tend to bond together by different mechanisms. Hence, the term 'structure' in soil science generally carries a connotation of bonding mechanisms in addition to geometrical configuration of particles.

Without hierarchical structure, medium- and fine-textured soils such as loams and clays would be nearly impermeable to fluids and gases, and at 'typical' moisture contents would possess a mechanical strength prohibitive to growth of plant roots and soil organisms. Thus, structure plays a crucial role in the transport of water, gases, and solutes in the environment, and in transforming soil into a suitable growth medium for plants and other biological organisms. Physical appearance of structured vs unstructured 'puddled' soil is shown in Figure 2.

This article briefly describes different hierarchical levels of soil structure, and dominant processes and mechanisms through which structural bonding occurs. Also discussed are statistical models and geometric scaling concepts used to describe the hierarchical system as a whole. A final section summarizes the dynamics of structure in tilled soils. Useful books and reviews on soil structure are listed at the end.

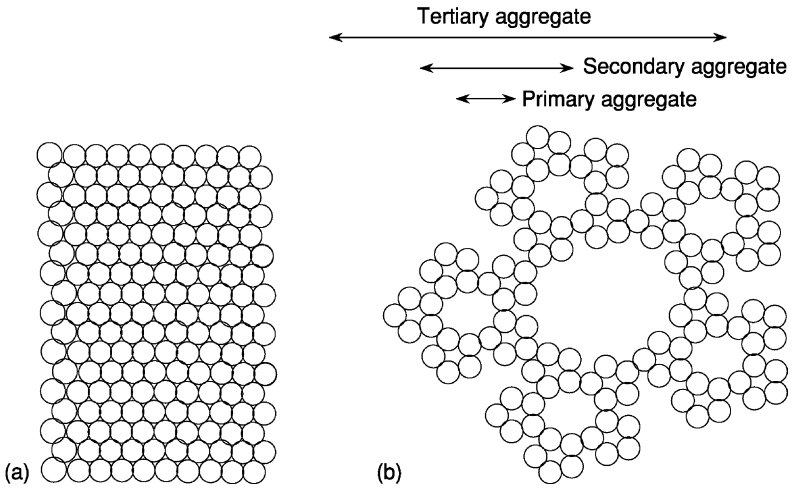


Figure 1 Idealized diagrams of (a) unstructured close-packed particles; and (b) system of hierarchically structured particles.

Hierarchical Levels of Soil Structure and Bonding Mechanisms

A fairly broad consensus exists for classifying soil aggregates into two main hierarchical categories, microaggregates (<250 μm diameter) and macroaggregates (>250 μm diameter). The microaggregates are typically subdivided into subclasses, <2 μm, 2–20 μm, and 20–250 μm. Salient properties of these aggregate categories are summarized below, and illustrated schematically in Figure 3. Comparison of aggregate size scales with other characteristic soil dimensions is presented in Figure 4.

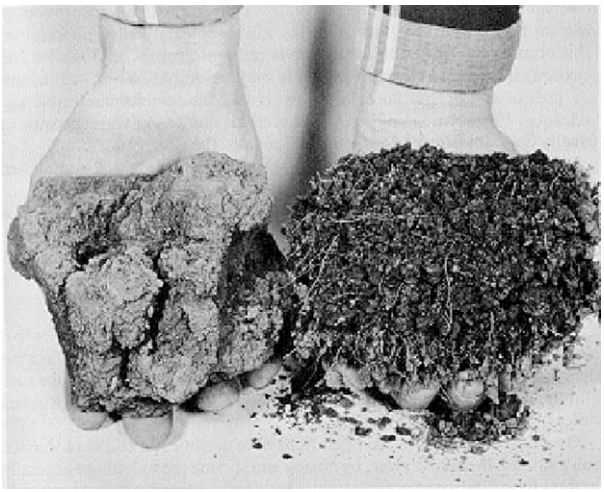


Figure 2 Photograph of structured soil and the same soil in a 'puddled' state where structure has been mostly destroyed. Courtesy of the USDA Natural Resources Conservation Service.

Microaggregates less than 2 μm in Diameter

These appear to be initially formed by flocculation of clay particles into domains or quasicrystals (Figures 3 and 5). Particularly in very small aggregates (<0.2 μm), most organic matter appears to be absorbed only onto external surfaces of the quasicrystals. Thus, bonding between particles in soil quasicrystals is governed by essentially the same van der Waals and electrical double-layer phenomena that produce quasicrystals in simple clay–water systems. Exclusion of organic matter from internal quasicrystal surfaces appears quite pronounced in montmorillonitic soils where clay particles exhibit a strongly oriented, mutually parallel structure (regions marked 'T' in Figure 5), and to a lesser extent in soils with illitic or kaolinitic mineralogy and correspondingly less-ordered domains. Very small microaggregates are highly resistant to mechanical disruption, typically requiring several minutes of ultrasonic dispersion, often with the aid of an oxidizing agent.

In a next hierarchical level, quasicrystals and other mineral particles coalesce around central bonding nuclei of highly processed organic materials of humic and polysaccharide nature. The clay domains and other mineral particles on the outside of these microaggregates protect the inner organic core against access by microorganisms. Figure 5 shows pockets of polysaccharide material (P) surrounded by clay tactoid coatings (T). The humic and polysaccharide materials typically consist of multiple strands and functional groups, which provide many

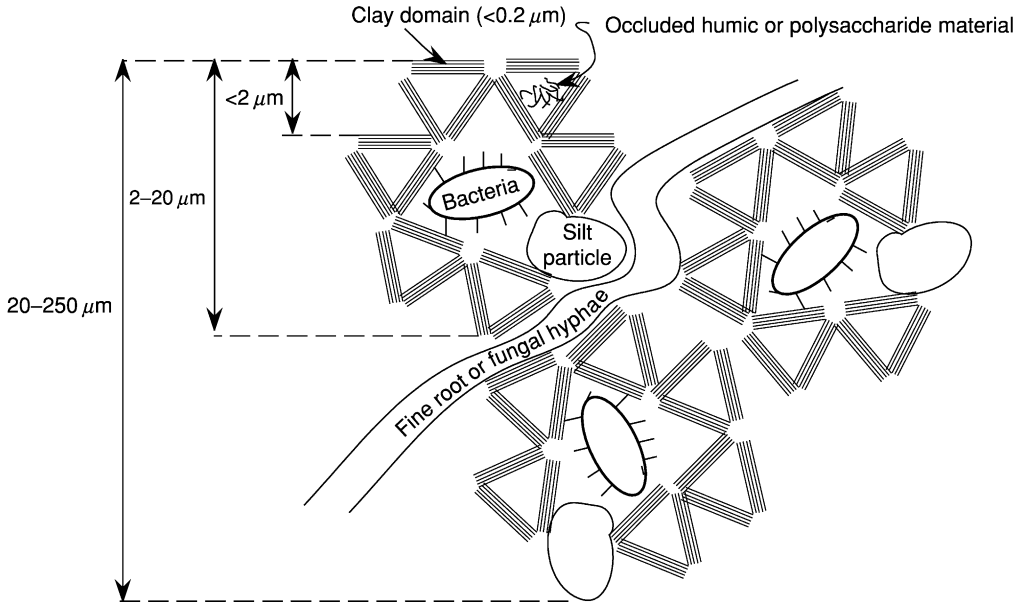


Figure 3 Schematic diagram of a hierarchical system of microaggregates illustrating special characteristics of each class.

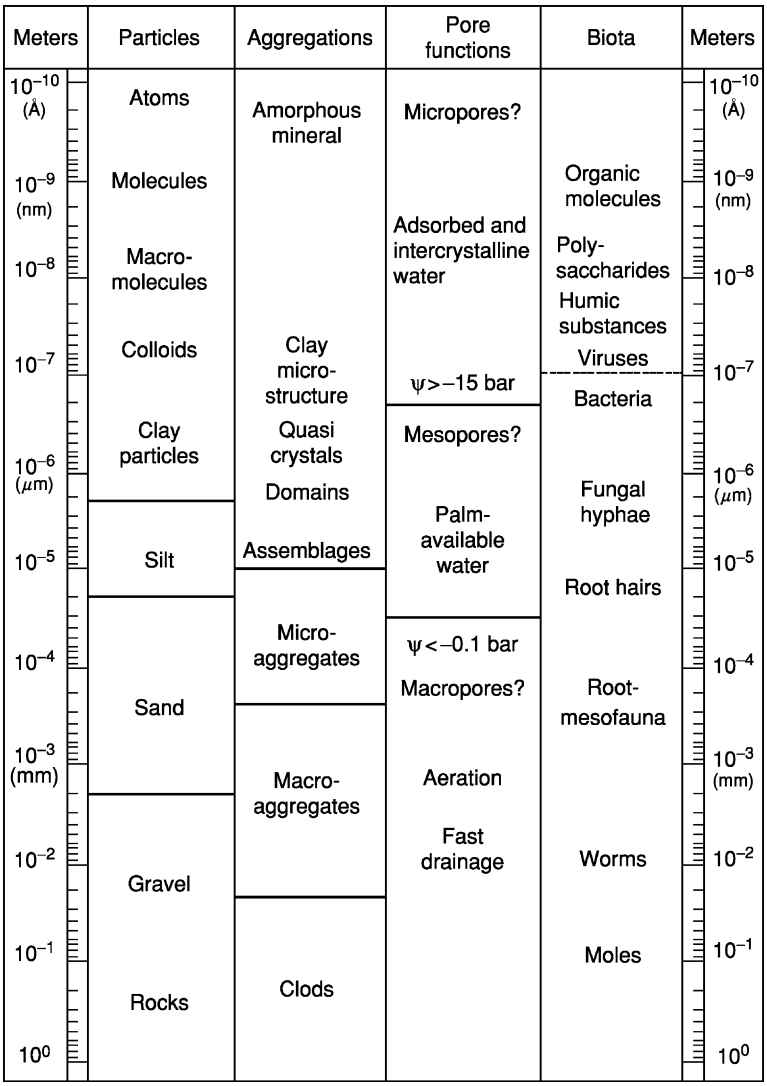


Figure 4 Comparative scales in soil structure. Reproduced with permission from Waters AG and Oades JM (1991) In: Wilson WS (ed.) *Advances in Soil Organic Research*, p. 164. Cambridge: Royal Society of Chemistry.

sites for bonding with the mineral surfaces. Since the predominant electrical charge of organic polymers at 'normal' soil pH values is usually negative, bonding with negatively charged inorganic colloids is largely achieved through 'bridging' by multivalent cations such as Ca^{2+}, Fe^{3+}, Al^{3+} and their hydrous oxides, which are able to complex with both mineral surfaces and organic functional groups (Figure 6 and Table 1). In highly weathered soils with abundant variable-charge minerals, ligand exchange between mineral surfaces and organic functional groups can produce particularly strong bonds. This is particularly notable in Oxisols, characterized by extremely strong micro-porous microaggregates resulting in a characteristic bimodal pore-size distribution (Figure 7).

In order for microaggregates to form effectively in soil, organic binding materials must be finely distributed throughout the soil, rather than deposited in isolated pockets. Particularly effective mechanisms appear to be *in situ* biosynthesis of organic materials by microorganisms associated with extensive networks of fine roots with high turnover rates, such as under grass vegetation.

Microaggregates tend to form slowly in soils, but once formed they also degrade slowly, even under unfavorable soil management systems. Organic substances in microaggregates have been observed to remain stable for hundreds and in some cases even thousands of years. Among the stabilized organic materials are enzymes that may contribute to processes such as N mineralization and herbicide degradation. The amount of carbon 'sequestered' in soil microaggregates constitutes a substantial fraction of the total amount organic carbon on Earth. This has

stirred much interest in the impact of soil structure management on atmospheric CO_2.

Microaggregates 2–20 μm in Diameter

These are formed from oriented clay domains, microaggregates $<2\,\mu$m in diameter, and/or coarse clay and fine silt particles, which coalesce around a central core of hyphal fragments and bacterial cells or colonies (Figures 3 and 5). Bonding is effected by microbial materials such as polysaccharide synthesized by the bacteria and hyphae. The outer layer of clay domains and microaggregates protects the bacteria from organisms such as nematodes and protozoa, which are too large to penetrate the outer layer. Thus, up to 40–60% of the microbial biomass in soil has been found associated with microaggregates 2–20 μm in diameter.

Like $<2\,\mu$m aggregates, 2–20 μm aggregates are very resistant to mechanical disruption, sometimes resisting up to 5 min of ultrasonic dispersion.

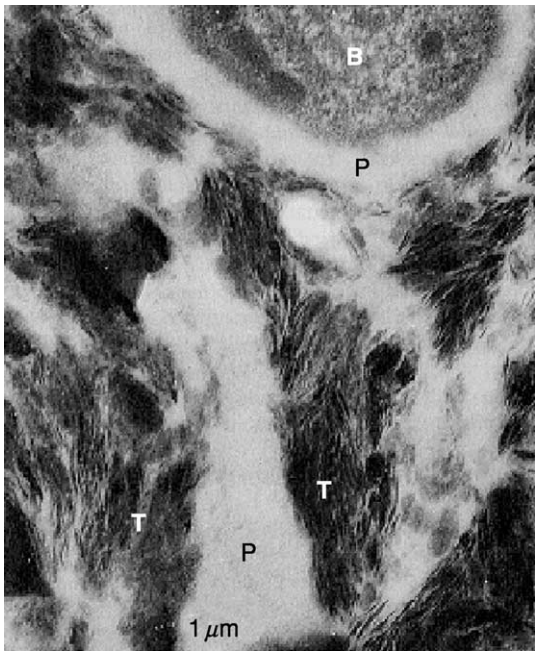


Figure 5 Micrograph of a $<2\,\mu$m microaggregate showing clay tactoids (T), occluded polysaccharide material (P), and bacterial capsule (B) surrounded by polysaccharide material. Reproduced with permission from Foster RC (1978) In: Emerson WW, Bond RD, and Dexter AR (eds) *Modification of Soil Structure*, p. 104. Chichester: Wiley.

Microaggregates 20–250 μm in Diameter

These are formed largely by particles or aggregates $<20\,\mu$m in diameter, bonded by polysaccharide material around central nuclei of fine roots and fungal hyphae, which may or may not be subsequently completely degraded by soil microorganisms. Bonding is generally strong enough that the aggregates are stable to slaking upon direct immersion of air-dry soil into water. Micrographs of two such aggregates are shown in Figure 8.

Macroaggregates (greater than 250 μm Diameter)

General properties of macroaggregates and their dynamic nature in soil management systems Due to their effect on size of the largest soil pores, these aggregates are very influential in determining macroscopic soil properties such as mechanical strength,

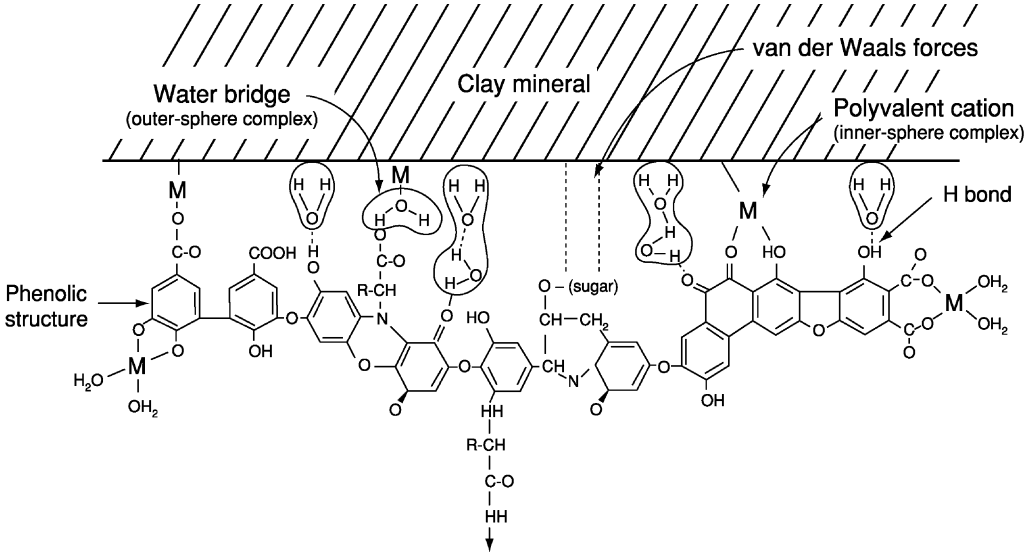


Figure 6 Illustration of bonding mechanisms between humic acid and mineral surfaces. Reproduced with permission from Haynes RJ and Beare MH (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 217. Boca Raton, FL: CRC Press.

Table 1 Possible aggregate-bonding mechanisms

I. Clay domain–clay domain

A. Domain face–domain face

Cations bridge between negative faces. Mechanism similar to that for orientation of clay platelets into domains

$$\text{Face}^{-} \text{- - - } M^{n+} \text{ - - -}^{-} \text{ Face}$$

B. Domain edge—domain face

Positive edge site to negative face

$$\text{Edge Al—}OH_2^{+} \text{- - -}^{-} \text{ Face}$$

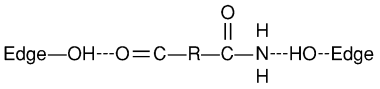
II. Clay domain—organic polymer—clay domain

A. Domain edge—organic polymer—(domain)

1. Anion exchange: Positive edge site to polymer carboxyl

$$\text{Edge Al—}OH_2^{+} \text{ - - - } {}^{-}\text{OOC—R—}$$

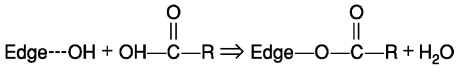
2. Hydrogen bonding between edge hydroxyl and polymer carbonyl or amide



3. Cation bridge between negative edge site and polymer carboxyl

$$\text{Edge—}O^{-}\text{—}M^{n+}\text{—}{}^{-}\text{OOC—R—COO}^{-}$$

4. van der Waals attraction between edge and polymer
5. Ligand exchange



B. Domain face—organic polymer—(domain)

$$\text{Face Si—O—HO—R—OH}$$

1. Hydrogen bonding between polymer hydroxyl and external or internal (expanding lattice minerals) face silicate oxygens

$$\text{External face}^{-} \text{- - -} M^{n+}\text{—OOC—R—COO}^{-}\text{—}$$

2. Cation bridge between domain external face and polymer carboxyl or other polarizable group
3. van der Waals attraction between face and polymer

III. Quartz—(silt, inorganic, and organic colloids)—quartz

A. Chemical bonds established between quartz surface gels of hydrated alumino silicates and active groups of other aggregate constituents

B. Quartz grains held in a matrix of silt and clay stabilized primarily by:

1. Oriented clay particles
2. Irreversibly dehydrated silicates, sesquioxides, or humic–sesquioxide complexes
3. Irreversibly dehydrated humic materials
4. Silt-size microaggregates stabilized by iron humates
5. Organic colloids and clay domains bonded by mechanisms cited under I and II

Adapted with permission from Hillel D (1998) *Environmental Soil Physics*. London: Academic Press.

hydraulic conductivity, and aeration status. They are also the aggregates most susceptible to breakdown under stresses due to tillage, compaction, raindrop impact, or wetting-induced slaking.

Macroaggregation seems primarily caused by reorientation and binding of clay particles and microaggregates by fine roots and hyphae, and further cementation by extracellular polysaccharides. Associations between extensive fine root systems and vesicular arbuscular mycorrhizal (VAM) fungi, which produce large amounts of polysaccharide, appear particularly effective in forming macroaggregates. Saprophytic fungi also stabilize aggregates, particularly when substrates such as combinations of straw and manure are added to the soil. Algal filaments, covered with slimy gels, have also been observed to be effective. It is often difficult to distinguish between the relative importance of binding by fine plant roots and fungal hyphae because, particularly in the case of VAM fungi, fungi and root growth are often positively

correlated. What does seem fairly clear is that plant species with large systems of fibrous roots, associated with VAM hyphae that secrete large amounts of polysaccharide gel, are effective stabilizers.

Bonding by polysaccharides apparently does not occur uniformly throughout macroaggregates, but rather primarily in the larger cracks or pores (15–50 μm in diameter), precisely where aggregates are most likely to rupture. Due to this efficient placement of bonds, even small increases or removals of polysaccharide can cause large increases or decreases in macroaggregate stability. Polysaccharides in large pores are more accessible to degradation by microorganisms than in microaggregates, so they are relatively labile and sensitive to changes in management.

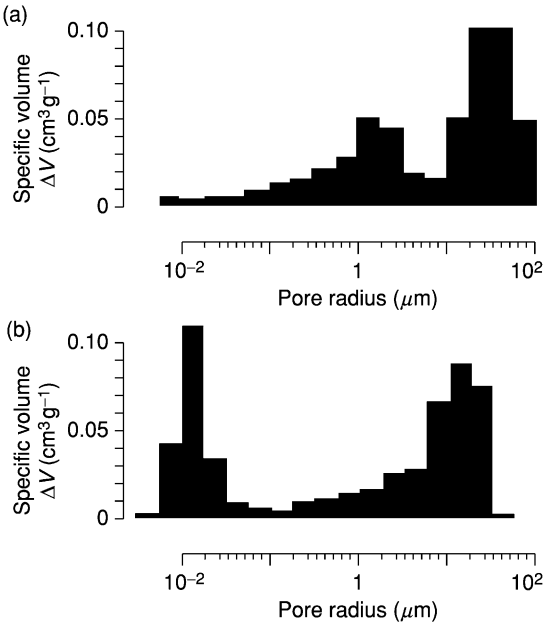


Figure 7 Pore size distribution for (a) silty soil and (b) an oxisol obtained by mercury porosimeter. Adapted from Bartoli F, Dutartre Ph, Gomendy V, *et al.* (1998) In: Baveye PJ, Parlange JV, and Stewart BA (eds) *Fractals in Soil Science*, p. 220. Boca Raton, FL: CRC Press, with permission.

Because labile polysaccharides constitute only a small fraction of the total amount of soil organic matter, it is not surprising that total soil organic matter content does not always correlate well with macroaggregate stability. This is particularly true shortly after sudden changes in management, where changes in total organic matter may be minimal but significant changes may have occurred in the network of fine roots, fungal hyphae, and associated polysaccharides, with associated rapid changes in aggregate stability. For example, Figure 9 shows that, after changing from continuous corn to an alfalfa cropping system, soil organic carbon accumulation lagged behind aggregate formation. In such situations, measurements of labile polysaccharide or related parameters are often better indicators of trends in macroaggregate stability than is total organic matter content (Table 2).

In cases where total soil organic matter content does correlate positively with macroaggregate stability, management has often been stable over long periods of time, e.g., continuous forest, pasture, or tillage. The high correlation probably reflects not only aggregate stabilization by organic matter, but also organic matter stabilization by aggregates. For a given type of soil management, correlation between organic matter and structural stability can be highly

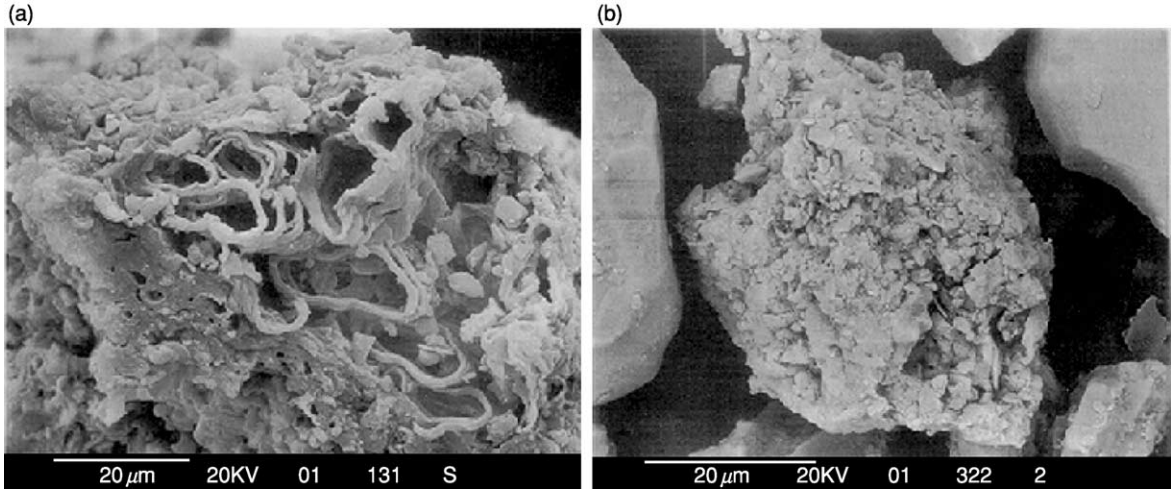


Figure 8 Micrographs of 20–250 μm aggregates showing (a) partly decayed vascular bundle surrounded by inorganics; (b) aggregates with elongated void running from top left to bottom right, with no remnants of plant anatomy evident. Reproduced with permission from Waters AG and Oades JM (1991) In: Wilson WS (ed.) *Advances in Soil Organic Research*, pp. 169–170. Cambridge: Royal Society of Chemistry.

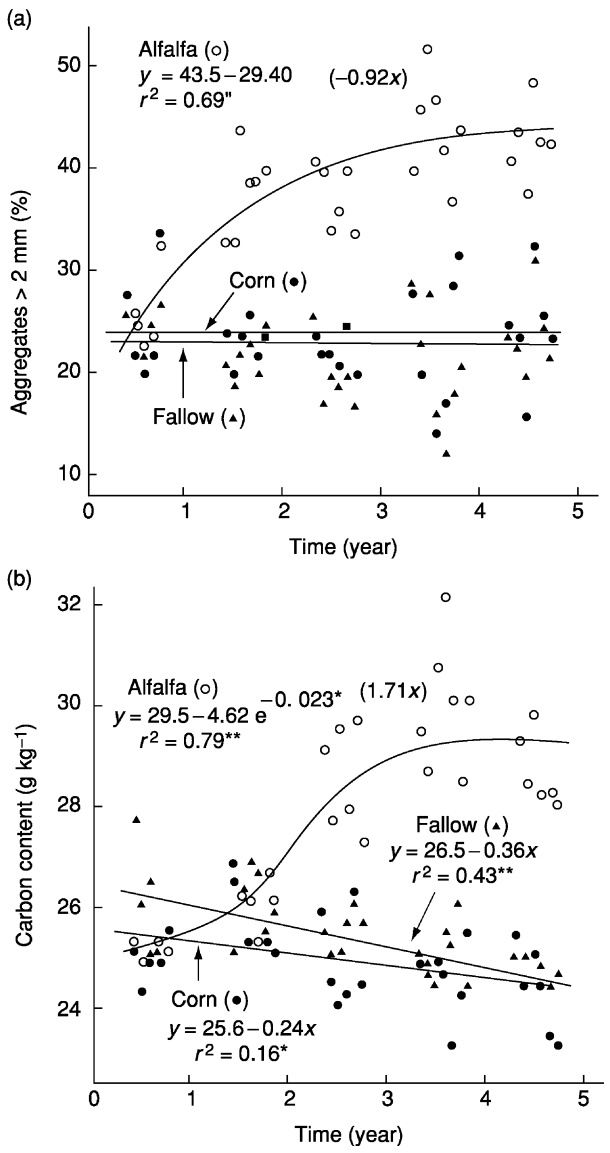


Figure 9 Data (open circles) showing how increases in soil organic carbon content (b) lag behind increase in aggregate stability (a) following changes in land management. Reproduced with permission from Angers DA and Carter MR (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 199. Boca Raton, FL: CRC Press.

soil-dependent, with texture playing a major role. In many cases a certain threshold level of organic matter is necessary before macroaggregates begin stabilizing (Figure 10). The threshold value tends to increase with increasing soil clay content, suggesting that a critical amount of organic matter per unit mineral surface area must be exceeded in order for effective aggregation to occur.

Earthworm casts Earthworm casts are the indurated soil material surrounding channels left in the soil by earthworms. Casts are initially quite unstable, since they are essentially puddled soil emitted from the earthworm's gut. Stability comes later, as a result of microbial interaction with organic materials mixed with soil in the gut. Fungal hyphae also contribute to cast stability, but probably originate from fungi outside the casts since fungi are strict aerobes that cannot survive inside the earthworm gut. When disrupted by tillage or mild shaking in water, earthworm casts generally break down into macroaggregate-sized fragments, and thus are often classified as macroaggregates.

Casts provide stability to earthworm channels, which over time under zero-tillage conditions may occur in such numbers as to contribute significantly to soil hydraulic conductivity and aeration status. An important property of these channels is their continuity, which makes them particularly effective as preferential flow conduits. Tillage disrupts continuity of earthworm channels, rendering them largely ineffective for transport.

Role of Wetting and Drying on Soil Structure Development

A key element in the development of soil structure is the cyclic pattern of wetting and drying of field soils.

When wet soils dry, particles are drawn together by negative pressures or suctions that develop in the pore water. Microscopically, this promotes clay orientation and interparticle bonding. Macroscopically, the soil as a whole shrinks or decreases in volume, causing the development of crack networks and surfaces of weakness between neighboring soil elements. These to a large extent define the boundaries of soil aggregates, particularly macroaggregates, and form continuous channels that facilitate water infiltration and gas movement.

The extent of soil cracking depends primarily on the intensity of drying, and on the amount and type of soil clay minerals. Soils with large amounts of 2:1 clay minerals, such as vertisols, are notorious for networks of large shrinkage cracks. At the other extreme, soils with predominantly nonexpandable clay minerals, such as oxisols, exhibit very little shrinkage and cracking behavior. This to a large extent explains the characteristic 'massive' soil structure of oxisols, manifested by a lack of visually distinct macroaggregate boundaries.

If crack networks extend to the soil surface, as is usually the case in untilled soil, then water infiltration during rainfall events occurs preferentially down these cracks until swelling causes the cracks to seal up again. Dispersed clay in the infiltrating water may be deposited at crack boundaries as water infiltrates the

surrounding aggregates, forming clay skins or 'cutans' on the aggregate surfaces. Also, since the soil at crack surfaces is the first to wet during infiltration events, soil displacements during wetting may result in shearing or 'smearing' of these surfaces. Such shear zones, or 'slickensides,' are particularly notorious in vertisols.

Characterization of Soil Structure Based on Visual Assessment

Visual Inspection of Soil Aggregates

A routine component of soil surveys is the description of soil macroaggregates based on visual examination. The criteria used in this classification scheme are type or shape, class or size, and grade or distinctness. Different structure types are illustrated in Figure 11, and further details are given in Table 3.

A certain amount of subjectivity exists in deciding the precise category for a given soil structure, particularly with respect to shape and grade. Nevertheless, the classification system has been used successfully for grouping soils according to structure-related properties such as permeability and preferential flow of solutes. This is largely due to the fact that aggregate morphology is related to morphology of the interaggregate void spaces where most water and solute transport takes place. For example, blocky

Table 2 Sensitivity of various chemical indices in relation to changes in aggregate stability over time

Previous cropping history	*Aggregate stability (MWD, mm)*	*Organic C (%)*	*Acid-hydrolyzable carbohydrate (%C)*	*Hot water-extractable carbohydrate ($\mu g\ C^{-1}$)*	*Microbial biomass C ($\mu g\ C\ g^{-1}$)*
18-year pasture	2.7	3.2	0.35	208	1018
4-year pasture[a]	2.5	2.5	0.26	169	890
1-year pasture	2.0	2.4	0.25	152	801
1-year arable	1.3	2.4	0.23	140	738
4-year arable	1.2	2.4	0.23	134	712
10-year arable	1.0	2.0	0.19	127	610

[a]The 1-year and 4-year pasture and 1-year and 4-year arable soils come from a cropping rotation of 4-years arable followed by 4-years pasture. (Data from Haynes RJ, Swift RS, and Stephen RC (1991) Influence of mixed cropping rotation (pasture-arable) on organic matter content, water stable aggregation and clod porosity in a group of soils.) *Soil Tillage Res* 19: 77–87.
Reproduced with permission from Haynes RJ and Beare MH (1996). In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 236. Boca Raton, FL: CRC Press.

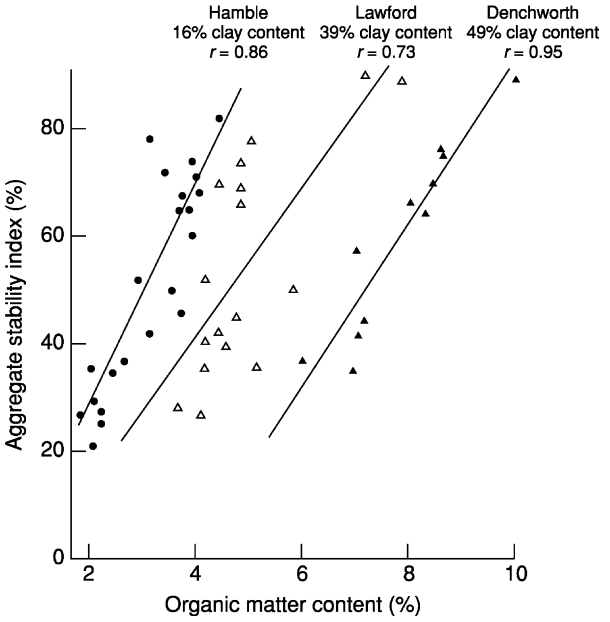


Figure 10 Relation between aggregate stability and organic matter content in soil with varying texture. Reproduced with permission from Haynes RJ and Beare MH (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 232. Boca Raton, FL: CRC Press.

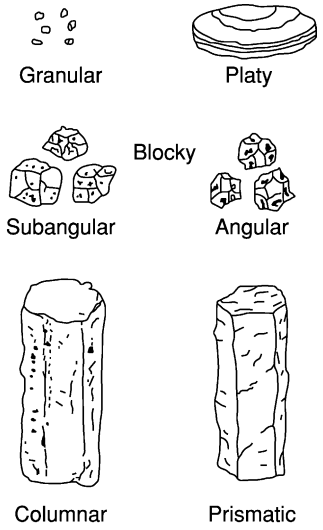


Figure 11 Classification of macroaggregates based on visual appearance. Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. London: Academic Press.

Table 3 Classification of soil structure according to US Department of Agriculture soil survey staff

A Type: shape and arrangement of peds

	Platelike	*Prismlike*		*Blocklike—polyhedral—spheroidal* Three approximately equal dimensions arranged around a point			
	Horizontal axes longer than vertical, arranged around a horizontal plane	Horizontal axes shorter than vertical. Arranged around vertical line. Vertices angular		*Blocklike—polyhedral* Plane or curved surfaces accommodated to faces of surrounding peds		*Spheroidal–polyhedral* Plane or curved surfaces not accommodated to faces of surrounding peds	
		Without rounded caps	With rounded caps	Faces flattened; vertices sharply angular	Mixed rounded, flattened faces; many rounded vertices	Relatively nonporous peds	Porous peds
B Class: size of peds	*Platy (mm)*	*Prismatic (mm)*	*Columnar (mm)*	*Blocky (mm)*	*Subangular blocky (mm)*	*Granular (mm)*	*Crumb (mm)*
1. Very fine or very thin	<1	<10	<10	<5	<5	<1	1
2. Fine or thin	1–2	10–20	10–20	5–10	5–10	1–2	1–2
3. Medium	2–5	20–50	20–50	10–20	10–20	2–5	2–5
4. Coarse or thick	5–10	50–100	50–100	20–50	20–50	5–10	
5. Very coarse (very thick)	>10	>100	>100	>50	>50	>10	
C Grade: durability of peds	0. Structureless		No aggregation or orderly arrangement				
	1. Weak		Poorly formed, nondurable, indistinct peds that break into a mixture of a few entire and many broken peds and much unaggregated material				
	2. Moderate		Well-formed, moderately durable peds, indistinct in undisturbed soil, that break into many entire and some broken peds but little unaggregated material				
	3. Strong		Well-formed, durable, distinct peds, weakly attached to each other, that break almost completely into entire peds				

Reproduced with permission from Hillel D (1998) *Environmental Soils Physics*. London: Academic Press.

and prismatic macroaggregates are often bounded by networks of large, continuous cracks, whereas interaggregate void spaces for granular or crumb structures are smaller and more finely distributed.

Image Analysis

Examination of photographs of soil crack patterns, or of thin sections, has long been used to obtain detailed information on the shape of aggregates and their spatial organization in relation to interaggregate void spaces. This has become increasingly feasible with the advent of powerful yet low-cost digital image acquisition and analysis technology. Three-dimensional image analysis is also possible using techniques such as computer-aided tomography.

Mathematical Models of Soil Structure

Aggregate Size Distributions

A common method for characterizing soil structure involves measuring the size distribution of soil fragments or 'aggregates' produced by a specified fragmentation method. The distributions are often fitted to two-parameter probability models such as in Table 4. In many cases one of the two parameters is nearly constant. For example, for 'fractal' distributions, the fractal dimension D may be relatively constant. Likewise for log-normal and Weibull distributions, the geometric standard deviation σ and exponent λ, respectively, often do not vary much. In such cases the distribution may be defined by specifying only the diameter X at some fixed probability P, or a probability-weighted diameter or 'mean weight diameter' defined over some fixed probability interval (P_1, P_2) as:

$$\mathrm{MWD} = \int_{P_1}^{P_2} X(P)\,dP \qquad [1]$$

Pore Size Distributions

Pore size distributions may be determined by microscopic examination of thin sections, or inferred indirectly from moisture release characteristics or mercury intrusion porosimetry. The latter two methods are based on the capillary relation between equivalent cylindrical pore radius (r) and gauge pressure head (h) at which liquid will just enter or recede from the pore:

$$r = 2\sigma\cos\Phi/\Delta\rho g h \qquad [2]$$

Here ρ is liquid density, σ is solid–liquid surface tension, g is the gravitational constant and Φ is the solid–liquid contact angle. Equation [2] allows inferring the total volume of voids in radius interval $r + dr$ from the measured volume of liquid drained or intruded in the pressure head range $h + dh$. The pore-size histograms in Figure 7 are an example of distributions inferred by this method.

For the common case of liquid water in non-hydrophobic soils, where $\rho \approx 1\ \mathrm{g\,cm^{-3}}$, $\sigma \approx 71\ \mathrm{dyn\,cm^{-1}}$, and $\Phi \approx 0$, eqn [2] reduces to:

$$r = 0.15/h \qquad [3]$$

where r and h are expressed in cm.

Scaling Models of Soil Structure

Soil structure often exhibits certain similitude or 'scaling' properties, which greatly simplify its mathematical representation. Two types of scaling have received wide attention in soil science, fractal scaling and Miller scaling (in honor of the brothers E.E. and R.D. Miller, who developed the original concepts). Fractal scaling concerns similarities between hierarchical levels in a given soil, whereas Miller scaling deals with structural similarities between different soils. In both cases, the underlying assumption is that the structural elements being compared are similar

Table 4 Common cumulative probability functions used to characterize aggregate size distribution in soils

Distribution	*Functional form*	*Definition of variables*
Fractal	$P(x) = \left(\frac{X}{\sigma}\right)^{3-D}$	$P(x)$ = cumulative mass fraction X = aggregate diameter σ, D = constants
Log-normal	$P(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\ln X} \exp\left\{-1/2\left[\ln\left(\frac{X}{\mu}\right)\right]^2\right\} d\ln X$	$P(x)$ = cumulative mass fraction X = aggregate diameter μ,σ = constants
Weibull	$P(x) = 1 - \exp\left\{-\left(\frac{X}{\sigma}\right)^{\lambda}\right\}$	$P(x)$ = cumulative mass fraction X = aggregate diameter σ, λ = constants

geometrically, differing only in some characteristic dimension or 'length.'

Fractal scaling relations between hierarchical levels in a given soil Fractals may be defined, somewhat simplistically, as hierarchical porous or irregular objects composed of successively nested similar elements. To illustrate, a simple fractal object is shown in Figure 12. The lowest hierarchical level (Figure 12a) plays the role of reference element or 'generator.' A second hierarchical level (Figure 12b) is established by forming a larger cluster from the primary clusters. These in turn form a third hierarchical level (Figure 12c), and so on.

The filamentous fractal object in Figure 12 is similar in many ways to particle clusters formed during diffusion-limited flocculation. Other more compact fractal objects, like the one shown earlier in Figure 1b, are perhaps more representative of soil aggregates.

A property of these and many other self-similar objects is that the number of primary particles N required to form a given hierarchical level is a non-integer power law function of the characteristic length L (aggregate diameter) at that hierarchical level, i.e.:

$$N \propto L^D = L^{E-1+f} \qquad [4]$$

Here E represents dimensionality of the problem ($E = 3$ for three-dimensional space, and 2 for two-dimensional objects in the plane) and f is a fraction $0 > f \leq 1$. The parameter $D = E - 1 + f$ is often known as the fractal dimension. The fraction f increases with the degree of 'space-filling' of the given fractal object. For example, f has a value of approximately 0.465 for the open filamentous object in Figure 12, and a value closer to unity (0.658) for the more compact 'aggregate-like' structure in Figure 1b. In the case of complete space filling (such as squares formed by placing smaller squares edge to edge, cubes formed by placing smaller cubes face to face, etc.), $f = 1$ and eqn [4] reduces to the familiar Euclidean relation $N \propto L^E$.

A consequence of scaling property eqn [4] is that the number of particles per unit 'characteristic' bulk volume, N/L^E, varies as L^{f-1}, which is an inverse relation whenever $f < 1$. Consistent with this result, aggregate bulk density often decreases in linear log-log fashion with increasing aggregate size (Figure 13).

Analysis of pore spaces in certain fractal structures has yielded theoretical power law relations between volumetric water content θ and matric suction h. At least qualitatively, this agrees with the well known empirical Brooks–Corey relation:

$$\theta(h) = \theta_{sat}(h/h_e)^b \qquad [5]$$

where θ_{sat} is saturated water content, h_e is suction at air-entry and b is a soil-dependent constant.

Fractal scaling models have proved particularly useful for describing the geometry of complicated crack networks in soils.

Miller scaling relations between different soils In Miller scaling, no *a priori* assumption is made regarding fractal hierarchy within a given soil (although fractal scaling is allowed under certain conditions). All that is required is similitude between

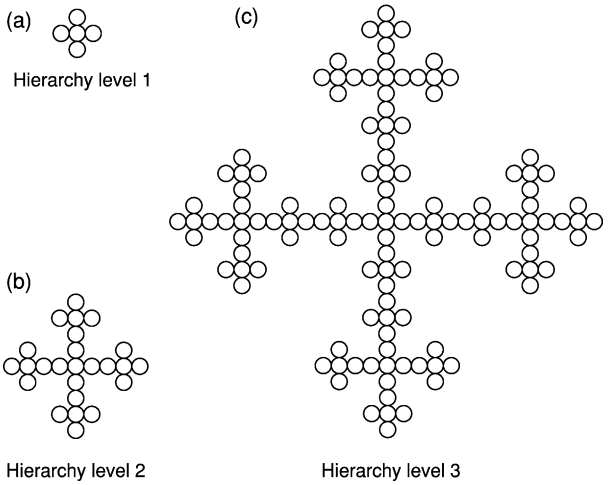


Figure 12 Self-similar filamentous objects of increasing hierarchical levels. Adapted with permission from Meakin P (1991) Fractal aggregates in geophysics. *Review on Geophysics* 29: 317–354.

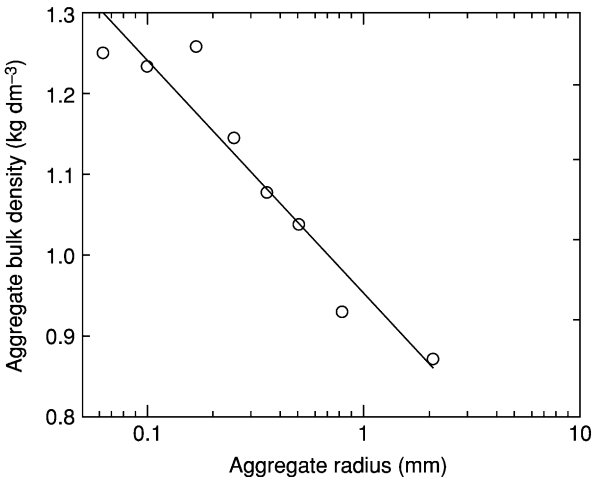


Figure 13 Bulk density of soil aggregates as a function of their radius. (Data from Chepil WS (1950) Methods of estimating apparent density of discrete soil grains and aggregates. *Soil Science* 70: 351–362, with permission.)

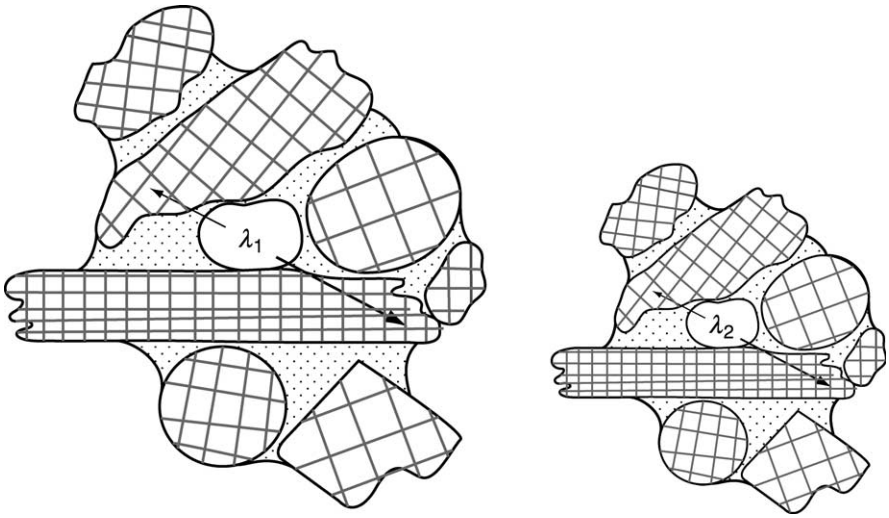


Figure 14 Illustration of two Miller similar soils with different characteristic lengths λ_1 and λ_2. After Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27: 324–332.

different soils, regardless of what the internal hierarchy may be.

In its strictest sense, Miller similitude assumes that the geometry of both elementary particles and their structural organization is identical across soils (Figure 14). Assuming dominance of capillary phenomena and creeping Newtonian fluid flow, this type of similarity imposes scale invariance on the dimensionless moisture release characteristic $\theta(h\lambda/\sigma)$ and the dimensionless hydraulic conductivity function $K(\theta)\eta/\lambda^2$, where K(θ) is hydraulic conductivity K at moisture content θ, η is fluid viscosity and σ is surface tension of the liquid–vapor interface. Miller similar soils always have identical porosity.

In practice, a weaker form of Miller similitude or 'generalized scaling' is often adopted, under recognition that it is usually the larger soil pores (i.e., those in upper hierarchical levels) that exhibit similitude and furthermore conduct most of the pore water. In generalized scaling, not only is the characteristic length λ of these large pores allowed to vary across soils, but also the soil volume fraction or 'effective' porosity P_e that they occupy. This is illustrated in Figure 15, where similarly shaped conducting or 'effective' structural units of different characteristic lengths λ_1 and λ_2 are embedded in a surrounding nonconducting soil matrix. For a given value of λ, the 'effective' porosity P_e is proportional to the number of elementary structural units per unit overall soil volume. Thus P_e may be considered a characteristic 'pore number' scaling factor, complementing λ which accounts for characteristic size. This contrasts with Miller similar soils, where λ is the only independent scale factor because P_e is always constant.

A consequence of generalized scaling behavior is invariance of the reduced moisture retention function $S_e(h\lambda/\sigma)$, where S_e is an 'effective' pore saturation defined as:

$$S_e = \theta_e/P_e = (\theta - \theta_o)/(\theta_{sat} - \theta_o) \qquad [6]$$

Here θ and θ_{sat} are the actual and saturated soil moisture contents, respectively, and θ_o is a 'residual' water content below which scaling relations no longer apply. θ_o may be considered as the water content when all the 'effective' structural units have drained and only the nonconducting matrix surrounding the 'effective' structural elements remains saturated. The parameter $\theta_e = \theta - \theta_o$ is known as the 'effective' water content and $\theta_{sat} - \theta_o$ is the effective porosity P_e.

Soils with invariant $S_e(h\lambda/\sigma)$ functions are frequently observed to have scale-invariant hydraulic conductivity functions $K(S_e)/\lambda_K^2$, where λ_K^2 is an empirical soil-dependent scale factor that may or may not correlate with the pore size scale factor λ. Only under strict Miller similarity does $\lambda = \lambda_K$.

Two fractal soils will also exhibit Miller scaling, if: (1) the shapes of the initial generating structures are the same in both soils, even though their characteristic lengths λ are different; and (2) both soils have the same number of hierarchical levels.

Structure of Tilled Agricultural Soils

Throughout history, modification of soil structure by tillage has played a central role in crop production. The tilled 'plow layer' plays a crucial role in determining plant growth and transport of gas, water, and

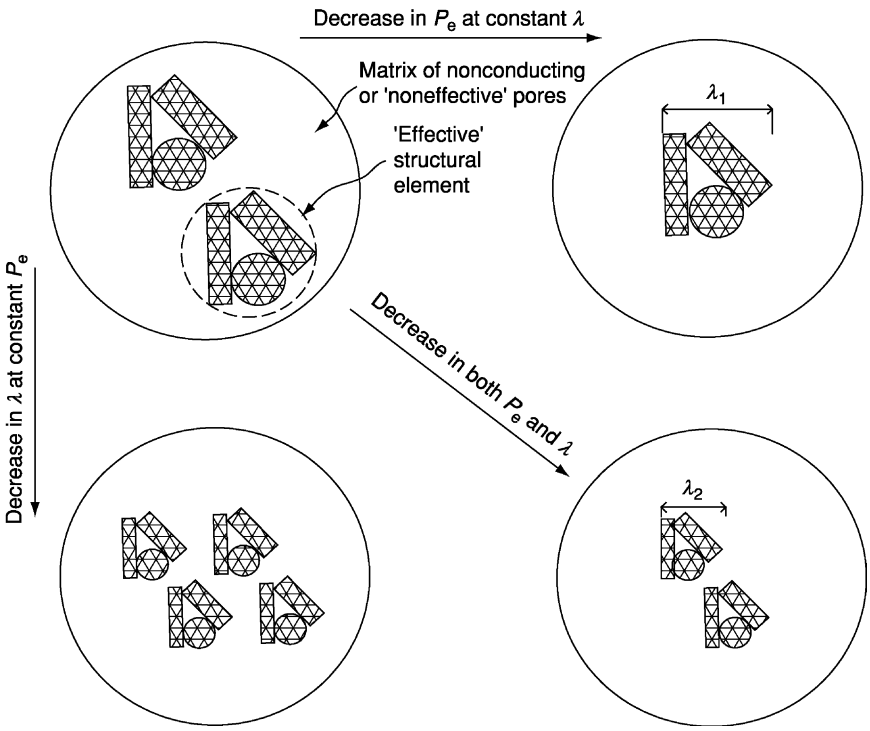


Figure 15 Illustration of four similar soils exhibiting generalized scaling.

chemicals in the environment. Thus, no review of soil structure would be complete without some discussion on structural dynamics in tilled soils.

Initial Soil Conditions Produced by Tillage: Soil Tilth

The structure or 'tilth' of freshly tilled soil is defined by the size and shape distribution, spatial arrangement, and internal structure of soil fragments produced by tillage. For a given soil and tillage implement, the tilth obtained depends primarily on soil moisture content at time of tillage.

Tillage under wet conditions, particularly in heavy-textured soils, generally results in large, plastically deformed fragments where internal structure may be seriously damaged or 'puddled.' Tillage of dry soils, on the other hand, generally results in minimal plastic deformation and structural damage, but undesirably large fragments may still be produced. The optimum soil moisture content for tillage is at some intermediate value, often close to the lower plastic limit, wherein maximum fragmentation occurs with little plastic deformation. Soils in this condition are said to exist in their most 'friable' state. Soils with predominantly large or intermediate pore sizes, associated with light texture or good structure or both, tend to be friable over fairly large ranges in moisture content. On the other hand, poorly structured fine-textured soils, particularly those dominated by high-activity clays, tend to present narrow moisture ranges for maximum friability.

Post-tillage Soil Structural Transformations

The loose structure produced by tillage tends to be highly unstable, so that under action of wetting and drying the soil resettles back toward a more stable structural state. The most important components of this resettlement process are crusting at the soil surface, and fracture and plastic deformation of aggregates deeper in the tilled layer.

Surface crusting of tilled soils The formation of thin (<2 mm) crusts at the soil surface, due to action of raindrops and sprinkler irrigation, is a common feature of cultivated soils throughout the world. Crusts are characterized by greater density, higher shear and tensile strength, finer pores and lower hydraulic conductivity than the underlying tilled soil. Important consequences are poor seedling emergence and increased runoff and erosion. A thin-section photograph taken from crusted soil is shown in Figure 16.

Crusting is thought to be a combined effect of two processes: (1) aggregate rupture caused by mechanical forces, such as raindrop impact and pore-air entrapment during rapid water infiltration; and (2) dispersion of clay particles. At low electrolyte concentrations, soils with Na^+-dominated exchange complexes tend to disperse spontaneously upon wetting. Ca^{2+}-dominated soils at low electrolyte concentrations may also disperse on wetting, but this

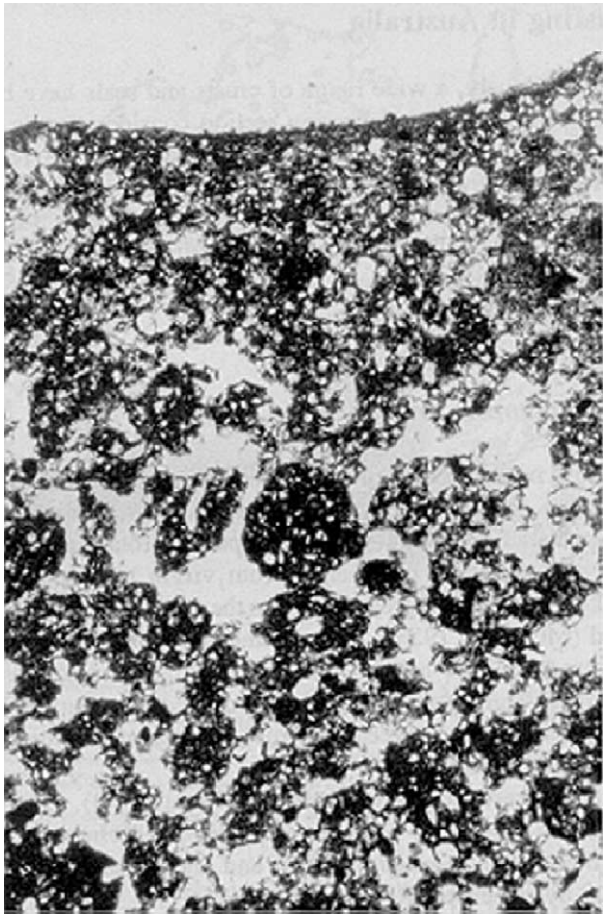


Figure 16 Photomicrograph of thin section from a crusted soil showing the denser layer near the soil surface. Frame length = 4 mm. Reproduced with permission from Chartres CJ (1992) In: Sumner ME and Stewart BA (eds) *Soil Crusting*, p. 344. Boca Raton, FL: CRC Press.

Figure 17 Fragmentation state of large clods after six wetting-drying cycles; (a) clayey vertisols; (b) clayey oxisols; and (c) clayey ultisols.

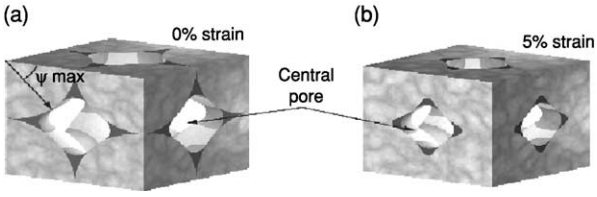


Figure 18 Schematic illustration of aggregate bed deformation resulting from plastic flattening of contact points: (a) initial undeformed aggregate bed; (b) aggregate bed after 5% volumetric strain. Courtesy of Dani Or, University of Connecticut.

frequently requires additional mechanical energy input, such as by recent tillage or raindrop impact. Dispersion in both Na^{+}- and Ca^{2+}-dominated soils is reduced at high electrolyte concentrations, but to a much greater degree in the Ca^{2+}-dominated case.

Soils with 10–30% clay content, with at least traces of smectitic clay and high amounts of silt, seem to be the most prone to crusting. Clay contents >30% tend to stabilize aggregates against disruption, whereas at <10% clay not enough fine material is present for dispersion to have significant effects.

Efforts to control crusting mainly involve ground cover to reduce raindrop impact, and chemical application to control dispersion and stabilize aggregates. Commonly used chemicals are moderately soluble Ca^{2+}-salts such as gypsum that maintain high Ca^{2+} concentrations in solution, together with polymers such as polyacrylamides (PAM) that interact with clay surfaces to promote flocculation and aggregation.

Fracture and plastic deformation of aggregates during wetting and drying cycles At depths greater than several millimeters in the plow layer, soil is not subject to direct raindrop impact, and wetting occurs more gradually and under greater tensions than in the surface layer. Thus, the mechanisms governing soil structural changes are different from the surface layer. The main mechanisms appear to be aggregate fragmentation due to differential soil swelling during water infiltration, and plastic deformation of wet aggregates due to concentrated shear stresses at interaggregate points of contact.

Wetting-induced aggregate fragmentation seems to increase with increasing soil shrink–swell potential. The most extreme and familiar manifestation is the 'self-mulching' of vertisols, wherein large soil clods completely break down to small fragments after only one or two wetting and drying cycles (Figure 17a). Other heavy-textured soils with lesser clay activity manifest the same phenomenon, but generally require more wetting and drying cycles to achieve comparable amounts of fragmentation (Figure 17b,c).

Plastic deformation of aggregates occurs when shear stresses at interaggregate contact points exceed the plastic yield strength (dependent on water content) of the soil material. The result is a gradual flattening and coalescence of interaggregate contact points (Figure 18), tending toward a state where aggregate boundaries disappear and the soil reaches its original

isotropic structure. The main causes of the concentrated stress at interaggregate contact points are believed to be overburden pressure and capillary forces due to water menisci between adjacent aggregates.

Long-Term Effects of Tillage

In the short term, tillage produces a loose soil structure conducive to plant growth and water infiltration. However, when intensively tilled soil is allowed to reconsolidate through the above mechanisms over an extended period, it often reaches a state of greater compactness than if it had never been tilled at all, typically requiring years in the undisturbed state to recover its original (pre-tillage) structure.

The Notion of Structural Quality and the Nonlimiting Water Range

An important question is: what constitutes a 'good' or 'poor' soil structure for plant growth? These attributes pertain to the general notion of structural quality. Such a notion is difficult to quantify, because plant growth in soils is strongly influenced by a number of structure-related physiological stress factors, such as soil aeration, mechanical impedance, and water availability, all of which vary in different ways with changing soil water content. For example, aeration stress is greatest at high water content, and decreases as the soil dries out, whereas stresses associated with water availability and mechanical impedance tend to increase with decreasing water content. Soil structure acts as a modifier of these relationships, by increasing or reducing the severity of each stress factor at a given soil water content.

Optimum soil physical conditions for plant growth generally reside within some intermediate range of water content where none of the stress factors is limiting. This range is known as the nonlimiting water range (NLWR). Plants growing in soils with a wide NLWR have a lesser probability of experiencing stress during their growth cycle than those growing in soils with a narrow NLWR. The NLWR has been proposed as a simple, measurable indicator of soil physical conditions or structural quality. In this framework, soils with a wide NLWR are said to have good structure, whereas those with a narrow NLWR are considered poorly structured.

Summary

The proper study of soil structure includes the definition of its essential nature, its role in the environment, methods of its characterization, and processes involved in its development and degradation. Especially important are the role and dynamics of soil structure in agricultural ecosystems, primarily because it is in these ecosystems that soil structure is most influenced by, and in turn influences, human management. Simple mathematical models of soil structure have proved useful in explaining hydrological and other macroscopic phenomena in soils.

See also: **Aggregation:** Microbial Aspects; Physical Aspects

Further Reading

Baveye P, Parlange JY, and Stewart BA (eds) (1998) *Fractals in Soil Science*. Boca Raton: CRC Press.

Carter MR and Stewart BA (eds) (1996) *Structure and Organic Matter Storage in Agricultural Soils*. Boca Raton: Lewis Publishers.

De Boodt MF, Hayes MHB, and Herbillon A (eds) (1990) *Soil Colloids and Their Associations in Aggregates*. New York: Plenum Press.

Kay BD and Angers DA (2001) Soil structure. In: Warrick AW (ed.) *Soil Physics Companion*. Boca Raton: CRC Press.

Pachepsky Y, Radcliffe DE, and Selim HM (eds) (2003) *Scaling Methods in Soil Physics*. Boca Raton: CRC Press.

Sumner ME and Stewart BA (eds) (1992) *Soil Crusting*. Boca Raton: Lewis Publishers.

SUBSOILING

R L Raper, USDA–ARS, Auburn, AL, USA

Published by Elsevier Ltd.

Introduction

'Subsoil' refers to the stratum of soil immediately below the surface soil or topsoil. Often this layer is overlooked, as most land management is focused on the topsoil, which can be altered drastically by tillage and other practices. However, the subsoil can have a large impact on a soil's potential productivity. If this layer of soil is extremely dense, roots may not penetrate, rooting volume will be decreased, nutrient uptake will be reduced, and plants may become susceptible to drought; also, water may not be able to infiltrate into the subsoil, thus limiting available water for plant growth and increasing surface runoff and potential soil erosion. Disrupting the subsoil to allow proper water infiltration and root growth may be necessary for optimum plant response.

'Subsoiling' refers to the process of soil tillage performed by a tool inserted into the soil to a depth of at least 350 mm. Tillage conducted by a narrow tool inserted to a more shallow depth is sometimes referred to as chisel plowing and is mostly used to loosen or level the soil surface and prepare a seedbed. Although tillage has been performed for several thousand years at a shallow depth, subsoiling is a relatively new operation, having only been performed since vehicles have excessively compacted the soil through their large mass and frequent traffic. Prior to the twentieth century, the ability to till deeper than just a few centimeters was not possible due to a lack of tractive force, nor was it necessary, because compaction due to repeated vehicle traffic had not yet been caused. In addition, naturally dense subsoils (e.g., hardpans, fragipans) require such treatment to allow proper root growth and drainage.

Shallow soil compaction caused by natural processes or field equipment can usually be alleviated by chisel plowing to shallow depths. However, if compaction penetrates deeper, more radical measures such as subsoiling may be necessary. This deeper compaction is often caused by repeated trafficking of the soil surface with large vehicle loads. In some locations, natural processes can lead to deep compaction, which can restrict root, water, and air movement in the soil.

Currently, subsoiling is practiced on a routine basis throughout the world. Many soils respond positively to subsoiling, resulting in yield improvements. Tillage tools used for subsoiling vary widely and result in differences in residue draft force requirements, remaining on the soil surface, and belowground soil disruption.

Measurement of Subsoiling

Determining when to subsoil requires some measurement of soil compaction. Cone index is the most accepted measure of soil compaction and has been used to determine when roots are restricted and can no longer expand into soil. This term is defined as the force required to insert a standard 30° cone into the soil. When values of cone index approach 1.5–2 MPa, root growth becomes limited and plants can start suffering the ill effects of soil compaction. After subsoiling, however, cone index values as low as 0.5 MPa are commonly found down to the depth of tillage (Figure 1).

Benefits of Subsoiling

The most obvious benefit of subsoiling is the disruption of deep, compacted subsoil layers. If soil compaction is excessive in these layers, roots cannot

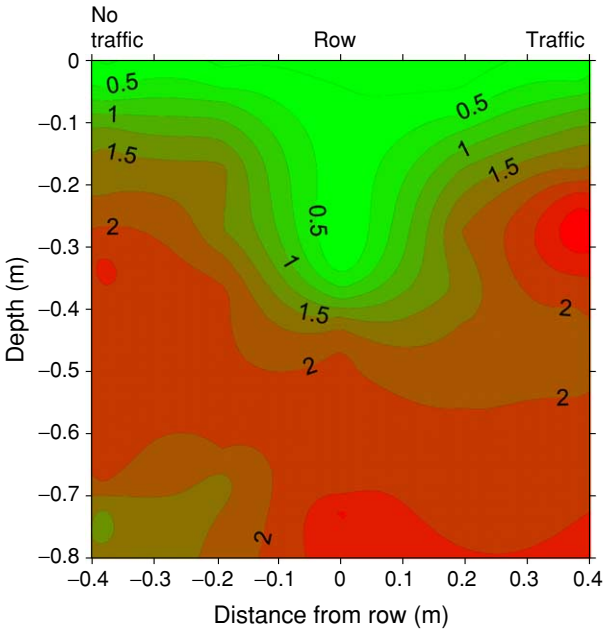


Figure 1 Cone index (mega pascals) isolines of soil showing location of nontrafficked row middle, row, and trafficked row middle. Minimal values of cone index are colored green and are shown near the center of the graph, representing root extension down to 0.3 m. Severely compacted zones are particularly evident under the trafficked row middle at the 0.3-m depth.

penetrate and are restricted to shallow depths. During times of drought, plants grown in a compacted soil are immediately susceptible, as their roots are confined to shallow zones that do not contain adequate soil moisture. Subsoiling excessively compacted soils lossens the soil for root growth, the depth of which is increased, so the plants are better able to withstand periods of drought.

Coupled with the increased root growth is the improved infiltration that usually accompanies subsoiling. Rainfall that previously exceeded infiltration capacity can be stored in the subsoil. The loosened soil provides pathways into the soil for rainfall to move quickly, instead of ponding on the soil surface and eventually evaporating or running off. Larger amounts of soil moisture may then be available to the plant during the growing season when moisture may otherwise be limited.

Increased numbers of macropores are also found after subsoiling. Even though some of these pores disappear as the soil reconsolidates, many stay open and provide increased storage of water and oxygen for plant roots. However, it is important that subsequent vehicle traffic be minimized to achieve long-lasting effects of subsoiling. Some research has reported that benefits of subsoiling are lost by the second pass of a vehicle tire. This could mean that subsoiling might not benefit a crop if traffic from a primary tillage operation and a planting operation were allowed to stray too close to the subsoiled channels. Maintaining the loosened soil profile and the increased storage capacity for water could be extremely valuable to plant roots during temporary summer droughts.

Ultimately, crop yields may improve from subsoiling, although the amount of improvement is difficult to estimate, because soil type, soil condition, plant species, and climate all have large effects. Many soils have shown benefits of being subsoiled; however, the amount of relative benefit may be offset by the expense of performing the operation. Some coarse-textured soils (sandy-to-loamy), which compact easily and require minimum tillage forces for subsoiling, show significant yield improvements when subsoiled. Some fine-textured soils are not economically subsoiled due to the lack of a yield improvement or because of the high draft forces necessary for subsoiling.

Subsoiler Design

Tillage tools used for subsoiling vary greatly in design and use. The individual vertical members that contact the soil and provide disruption are referred to as shanks. Their design varies greatly depending upon purpose, geographical location, soil type and depth of use.

Prior to 1950, most subsoiler shanks were straight, with a slight forward projection angle. However, research near the end of that decade recognized that other shapes, including curved and elliptical subsoiler shanks, can provide reduced draft forces in some soil types and soil conditions (Figure 2). Eventually, parabolic shanks became widely used and accepted as reducing draft forces. Some studies, however, have found that straight shanks mounted at an aggressive forward angle have reduced tillage draft when used in sandy soils. One negative effect of using curved shanks is that these shanks are designed to operate at a particular depth. When curved shanks are operated at depths either shallower or deeper than their intended depth, draft can increase, probably due to soil bodies that build up in front of the shanks, resulting in 'soil-on-soil' friction in contrast to the lesser 'metal-on-soil' friction (Figure 3).

Some subsoilers disrupt the soil in a symmetric manner, leaving equally disturbed soil on either side of the subsoiler as it moves in a forward direction. However, in the mid-1970s, the bentleg subsoiler was developed, which was designed to disturb the soil in a nonsymmetric manner. This shank is bent to one side by 45°, with the leading edge rotated forward by 25°.

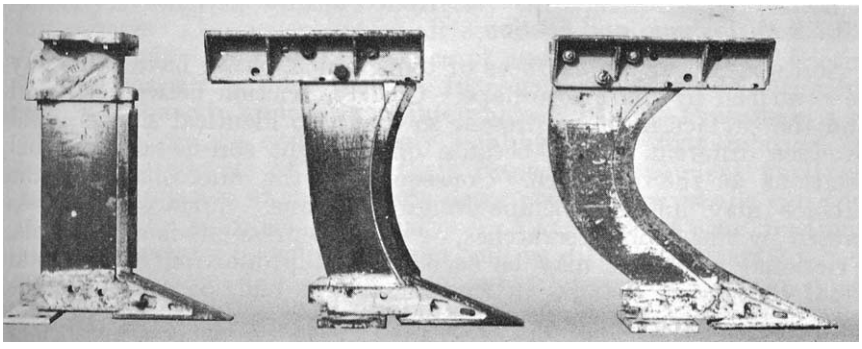


Figure 2 Straight, slightly curved, and deeply curved subsoilers, as tested in the 1950s. (Reproduced with permission from Nichols ML and Reaves CA (1958) Soil reaction to subsoiling equipment. *Agricultural Engineering* 39: 340–343.)

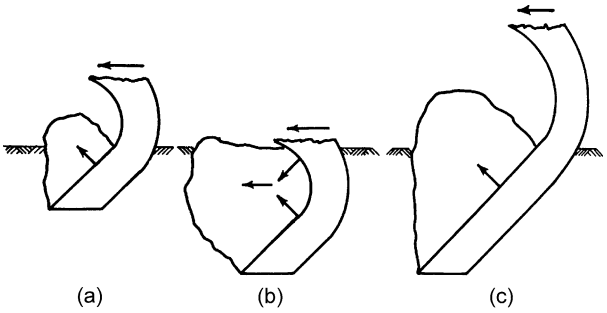


Figure 3 (a) Curved subsoiler operating at design depth; (b) curved subsoiler operating at deeper than design depth; (c) curved subsoiler scaled in size to operate at increased depth. (Reproduced from Gill WR and Vanden Berg GE (1966) *Soil Dynamics in Tillage and Traction*. Washington, DC: USDA.)

Figure 4 Side (left) and front (right) view of a bentleg subsoiler.

As the shank is traveling forward, it contacts the soil over a 216-mm width, which is substantially larger than any of the nonbentleg subsoilers. The main advantage of the bentleg subsoiler is its ability to pass through the soil leaving the surface relatively free of disturbance. For this reason, many producers have adopted this form of tillage as a method of alleviating soil compaction while maintaining large amounts of residue on the soil surface (Figure 4).

The bentleg subsoiler is commonly thought to require larger amounts of draft energy than traditional subsoilers. However, several studies have shown that comparable draft forces are generated for the bentleg subsoiler and traditional subsoilers if they are operated at similar depths.

Several other methods have been advocated for reducing draft on agricultural vehicles or increasing belowground disruption. These include vibrating the subsoiler, rotating the subsoiler shanks, placing a wing on or behind the subsoiler shank, using multiple shanks placed immediately behind each other and operating at different depths, or using a rolling coulter. Many of these experimental methods have shown promise and are undergoing further refinements, but none have been adopted by the agricultural machinery industry as of 2003.

Force Required for Subsoiling

Subsoiling requires a great amount of tillage energy. Based on experimental data obtained throughout the USA, the American Society of Agricultural Engineers' (ASAE) Standard D497.1 gives the following equation for calculation of draft force for subsoiling:

$$D = F_{\mathrm{i}}\left[A + C(S)^2\right]WT \qquad [1]$$

where D is implement draft (in newtons), F is dimensionless soil texture adjustment, i is fine (1), medium (2), or coarse-textured (3) soil, A and C are machine-specific factors, S is field speed (in kilometers per hour), W is number of rows or tools, and T is tillage depth (in centimeters).

Many factors influence draft forces of subsoiling and are represented in Eqn [1]. Soil texture certainly has a very large effect on draft force, with the factor F_{i} ranging from 0.45 for coarse-textured soils to 1.0 for fine-textured soils. (Coarse-textured soils are described as being sandy soils, medium-textured soils are described as being loamy soils, and fine-textured soils are described as being high in clay content.) Speed is also considered to be one of the most important factors. Because speed has a large effect on draft force, most subsoiling operations are conducted at relatively low speeds. The number of subsoiler shanks is also important, with the draft force increasing proportionally for each additional shank being pulled through the field.

One factor that is easily overlooked and that appears last in the equation is tillage depth. Often producers set their subsoiler depth at the deepest position that the tractor can pull. The subsoiler will only be moved upward toward the soil surface when excessive draft is sensed; after this area of excessive soil strength is passed, the original depth of subsoiling will be returned to. However, recent research reports that optimum yields are achieved when the depth of tillage is targeted to the depth of the root-impeding layer. Here, subsoiling energy is only expended to the problematic depth without wasting it by tilling too deeply. Also, excessive depths of subsoiling may increase compaction at these depths, as pressure from vehicle traffic could be propagated downward to depths below the tillage zone. As tillage is conducted more deeply every year, compaction moves downward through loosened soil horizons.

Eqn [1] only gives an approximate mean range (±50%) for subsoiler draft force, as many differences

can exist between subsoilers. One difference that is commonly found between subsoiler shanks is tip design. For Eqn [1], factors *A* and *C* have been designated as 226 and 1.8 for narrow points, but as 294 and 2.4 for 30-cm winged points. The effect of the point is to increase the draft force substantially, while increasing belowground disruption.

Management Practices

Subsoilers have been used mostly to totally disrupt entire fields or severely compacted parts of fields, especially headlands where turning and excessive trafficking has occurred. V-frame subsoilers have typically been designed with subsoiler shanks positioned closely over the width of operation (**Figure 5**). Their width is set so that the disrupted zone from a shank intersects the disrupted zone from nearby shanks. The compacted portion of the field or the entire field is loosened using this method of subsoiling with little consideration for future field activities. However, some producers subsoil at a slight angle to the previous year's rows so that excessively trafficked lanes of the field are mostly removed. Another reason for conducting the subsoiling operation at an angle is that it ensures that an entire row would not be missed if an error in subsoiler positioning occurred or if a shank were bent or broken and did not adequately disrupt the soil. Secondary tillage is normally required to prepare the soil surface for planting after a complete subsoiling disruption operation.

As many producers have begun to realize the benefits of maintaining an adequate surface-residue cover, they have become concerned that the subsoiling operation may bury excessive amounts of crop residue. To avoid this, one method that has been widely adopted is called 'strip-tillage' or 'in-row subsoiling.' This process involves subsoiling directly beneath the row. A single shank is pulled through the soil directly beneath each row to loosen the soil only in close proximity to the crop. Tillage energy is not wasted on loosening zones between rows that are not necessary for crop roots. Areas between rows are mostly left undisturbed with full residue coverage. One significant difference between strip-tillage and subsoiling as conducted by a V-frame subsoiler is the condition of the soil as it is left by the implement. Strip-tillage implements usually include a method of closing the subsoiling slot left by the shank to prepare a seedbed because planting may immediately follow the strip-tillage operation (**Figure 6**).

When to Subsoil

The frequency of subsoiling is dependent primarily upon the particular cropping system and the soil's needs. In some areas of the world, benefits of subsoiling have been found up to 10 years later. However, in the southeastern USA, most research indicates that this tillage operation could be performed every other year without limiting yields. To a large degree, it depends upon the management system that is put in place following the subsoiling operation. When

Figure 5 V-frame subsoiler used for complete disruption of soil compaction.

Figure 6 Strip-tillage subsoiler used for subsoiling beneath rows. Note closing wheels, which prepare the seedbed for planting.

controlled traffic and in-row subsoiling are both used, a soil may be able to withstand the compaction forces generated by traffic on the soil surface. The forces that tires place on the soil surface in the row middles will not propagate sideways into the row area, because the hardened soil in the row middles will be able to withstand the traffic without deformation. This process is likened to tires running on cement pads which are located between rows in the field. As long as the tires are kept on the cement pads, little compaction occurs under the rows. The loosened soil beneath the rows will be able to maintain their loose structure for longer periods of time and may not need to be subsoiled as often. Even when subsoiling is performed, the shank runs in soil that was previously loosened and not recompacted and the subsoiling forces are reduced. New automatic steering systems that enable precise control of the tractor as it passes through the field could enhance the adoption of controlled traffic as a management tool and could reduce the need for annual subsoiling (Figure 7).

Figure 7 Crops grown in a controlled-traffic system, where subsoiled zones beneath rows are kept separate from compacted zones between rows. Note deep rooting beneath rows.

Because of the large energy requirements necessary for subsoiling, some producers subsoil only a portion of their field during a year. The next year, they alternate and till other parts. This incremental approach is continued until the entire field has been subsoiled and the operation begins again. This type of rotation may provide enough overall loosening of the soil to enable good productivity. However, if deep tillage is not performed during a year that had a moisture-limiting condition, yield reductions could occur. To reduce overall risk from drought effects on their crops, many producers who must subsoil do it on an annual basis.

What time of year to subsoil is largely dependent upon the producer's schedule. Many producers subsoil in the autumn after their harvest is complete. Subsoiling at this time of the year can be efficient, because farmers have several months to prepare for planting the next season's crop. Waiting until spring to subsoil can delay planting if adverse weather conditions exist. Subsoiling can also remove compaction and rutting caused by the harvesting process, which can excessively traffic the field with very large loads (usually defined as greater than 10 Mg per axle) from

the harvesting or transport equipment. Subsoiling in the autumn can also be helpful due to the ability to have a loosened soil profile during winter months when rainfall is greater for many climatic regions. This loosened soil profile has increased surface roughness and increased infiltration, which enables more rainfall to penetrate and to be stored for future crop use.

Many producers who practice in-row subsoiling wait until spring to subsoil. They subsoil immediately before planting to give the crop the maximum benefit from the subsoiling operation. Combination planters and subsoilers have been produced for a one-pass operation. Producers who typically use this cropping system subsoil on an annual basis because their soils recompact very easily. Measurements made in the southeastern USA have shown that subsoiling conducted in the autumn does not provide adequate soil loosening necessary for optimum crop yields due to the soil's natural ability to reconsolidate over the winter months.

There is one disadvantage to subsoiling in the spring: occasionally when the subsoiling operation is closely followed by planting, a problem can develop relating to the proper emergence of seedlings. Intense rainfall events can cause the subsoiling channel to settle quickly and move the seed downward within the loosened zone. Replanting the crop may be necessary if the seeds have been excessively covered by soil.

Another consideration for deciding when to subsoil is the moisture content of the soil. Maximum disruption of the soil profile is usually provided when the soil is extremely dry, in contrast to subsoiling when the soil is almost saturated. However, subsoiling forces increase dramatically when soil is dry, and adequate tillage or tractive energy may not be available. A reasonable compromise seems to be to recommend subsoiling when soil moisture is near the permanent wilting point, but soil drying has not progressed to the point of the hygroscopic coefficient, when soil moisture is in vapor phase (i.e., soil moisture is bound tightly to soil solids and little is available for plant use). Some research indicates that little difference in soil disruption is measured between soil moisture of the hygroscopic coefficient point and that of permanent wilting point. However, the difference in tillage forces required for subsoiling between these two soil conditions can be significant.

Maintaining Surface-Residue Coverage

Consideration also has to be given to the amount of soil-surface disruption. Often producers are concerned with trying to reduce draft forces and they forget about leaving the soil in a smooth condition appropriate for planting. Also, efforts should be made to ensure that an adequate amount of crop residue is not buried by the subsoiling operation. This may be especially important to fragile residues when a crop such as soybean is followed by subsoiling. Research results also indicate that subsoiling when soil is near the hygroscopic coefficient causes maximum disturbance to the soil surface. Subsoiling near the permanent wilting point disrupts the soil surface less and probably results in less residue burial. Properly choosing a subsoiling shank that minimally disturbs crop residue and then operating it at the correct soil moisture and at a proper speed results in minimal surface disturbance and maximum subsurface disruption.

Even though most research on subsoiling has targeted force reduction, some early observations of the soil surface indicated that the overall effect of shank design had little effect on soil breakup. However, recent research indicates that bentleg subsoilers typically do a better job of maintaining surface residue than subsoilers, which are used for complete disruption. Consequently bentleg subsoilers have been readily adopted for use in many conservation tillage systems.

Subsoiling in Irrigated Fields

When irrigation is available, yield responses to subsoiling are less obvious. Increasing plant rooting depth by subsoiling may not be important, as the plants are likely to obtain all of their moisture through irrigation water. There are two advantages of using subsoiling with irrigation: the first is the ability of the soil to store additional moisture after subsoiling so less frequent irrigation is necessary; the second is that not all parts of a field respond positively to irrigation. Lower-lying areas of irrigated fields that are poorly drained show yield increases when subsoiled.

Subsoiling in Perennial Crops

Most subsoiling operations are conducted in row-cropping agriculture, although many other crops may experience potential benefits from this operation. In many forest locations a subsoiling operation is typically conducted before pine seedlings are planted. The subsoiling operation is conducted by large, single-row plows pulled either by bulldozers or log-skidders (Figure 8). Usually, this operation is conducted every 3 m and the trees are planted soon after the operation is completed. Several advantages are seen from subsoiling prior to forestry planting, for example: (1) pine tree seedlings gain from having

Figure 8 Subsoiling plow used for forest tillage to prepare soil before tree planting. Note backward sweep of subsoiler shank, which is used to allow the implement to pass over tree stumps.

compacted soil disrupted, enabling roots to penetrate to depths adequate for soil moisture; (2) the forest soil surface is smoother and more easily accessible to machinery; and (3) there is economic benefit from having the trees machine-planted rather than hand-planted.

Fields where other nonannual crops are grown such as pastures, vineyards, orchards, and fields used for sugar cane production that are prone to repeated heavy machinery traffic are frequently subsoiled prior to planting. After they are planted, the crops may be in place for many years and the opportunity for loosening is reduced. Subsoiling can be performed after the crop is established, but it is not advisable: plant growth and yields may be detrimentally affected due to excessive root pruning, potentially followed by periods of drought.

Considerations Before Subsoiling

Even though it may be possible to subsoil a field to remove compaction, care should be exercised before this potentially expensive operation is performed. Once soil is loosened by subsoiling, it will easily recompact if traffic is applied in the same area. Research indicates that two passes of a tractor in the subsoiled area will cause the soil to return to its previous state prior to subsoiling. If traffic is controlled, however, the benefits of subsoiling can be long-lasting.

Using a cover crop has been shown in some locations to replace the need for subsoiling. Winter cover crops are able to increase infiltration of winter rainfall and assist with water storage for use by the main cash crop the following summer. Evaporation of soil moisture is also hindered by the residue cover provided by the winter cover crop, even persisting several months later during summer months. The increased amount of soil moisture present under the cover crop reduces the overall soil strength and allows the plant roots to continue to grow downward. However, cover crops are not advisable under all climate and soil conditions, and subsoiling may still offer increased crop response in severely compacted soils.

The overall management of the production system should be examined to determine whether the soil compaction that is being alleviated by subsoiling is natural or whether it is traffic-induced. If it is natural, then subsoiling may have to be performed on an annual basis to give plants the maximum benefit of the operation. However, if a portion of the compaction is machine-induced, adoption of controlled traffic or a cover crop may enable the subsoiling operation to be performed less frequently.

Producers should be aware that some soils may not respond positively to subsoiling even though they may appear to be compacted. Yield improvements may not be realized in soils that are not severely

compacted or are not in need of subsoiling. Even some soils that are compacted may have adequate plant growth due to the presence of old root channels or significant earthworm activity and may not be improved significantly by subsoiling.

Summary

Subsoiling is a necessary tillage process for many fields used for crop production. Subsoiling disturbs the soil down to at least 350 mm and provides for increased rooting in soils compacted by either natural causes or by vehicle traffic. The potential success of subsoiling varies depending upon the design of the shank, the timing of the operation, the crop, the soil, and management decisions. Ongoing research is aimed at a better understanding of subsoiling so that producers can determine whether their soils might benefit from this tillage process.

List of Technical Nomenclature

draft force The horizontal force required to pull the implement through the soil

See also: **Compaction**; **Conservation Tillage**

Further Reading

ASAE (1998) *Agricultural Machinery Management Data D497.4*. ASAE Standards, pp. 360–367. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1998) *Terminology and Definitions for Agricultural Tillage Implements SAE S414.1*. ASAE Standards, pp. 261–272. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1999) *Procedures for Obtaining and Reporting Data with the Soil Cone Penetrometer EP542*. ASAE Standards, pp. 964–966. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1999) *Soil Cone Penetrometer S313.2*. ASAE Standards, pp. 808–809. St. Joseph, MI: American Society of Agricultural Engineers.

Gill WR and Vanden Berg GE (1966) Design of tillage tools. In: Gill WR and Vanden Berg GE (eds) *Soil Dynamics in Tillage and Traction*, pp. 211–297. Washington, DC: USDA.

Larson WE, Eynard A, Hadas A, and Lipiec J (1994) Control and avoidance of soil compaction. In: Soane BD and van Ouwerkerk C (eds) *Soil Compaction in Crop Production*, pp. 597–625. Amsterdam, the Netherlands: Elsevier.

Nichols ML and Reaves CA (1958) Soil reaction to subsoiling equipment. *Agricultural Engineering* 39: 340–343.

SULFUR IN SOILS

Contents

Overview

M A Tabatabai, Iowa State University, Ames, IA, USA

Introduction

The average sulfur (S) content of the earth's crust is estimated to be between 0.06 and 0.10%. It is usually ranked as the 13th most abundant element in nature. Sulfur occurs in soils in organic and inorganic forms, with organic S accounting for more than 95% of the total S in most soils from the humid and semihumid regions. The proportion of organic and inorganic S in soils samples, however, varies widely according to soil type and depth of sampling. It is usually somewhat lower in subsurface than in surface soils.

Although it is well known that S in soils is present mainly in organic combinations, very little is known about the identities of these S compounds. The inorganic S fraction in soils may occur as sulfate (SO_4^{2-}) and compounds of lower oxidation state such as sulfide (S^{2-}), thiosulfate ($S_2O_3^{2-}$), tetrathionate ($S_4O_6^{2-}$), polysulfides (S_n^{2-}, where $n > 10$), sulfite (SO_3^{2-}), and elemental S (S^0). The last four are detected in soils treated with elemental S or certain pollutants. In well-drained, aerated soils, most of the inorganic S

normally occurs as sulfate, and the concentrations of reduced S compounds are generally 1%. There are several forms of sulfate in soils. These include easily soluble sulfate, adsorbed sulfate, insoluble sulfate, and sulfate coprecipitated (cocrystallized) with $CaCO_3$. Under anaerobic conditions, particularly in tidal swamps and poorly drained or waterlogged soils, the main form of inorganic S in soils is sulfide and, often, elemental S.

Carbon–Nitrogen–Phosphorus–Sulfur Relationships

Significant information is now available on the relationships between C, N, P, and S in soils around the world. Unlike P, which can be present in significant proportions as organic and inorganic combinations, inorganic N and S values are very small relative to organic forms of these elements in soils. Therefore, often the relationships between organic C, total N (instead of organic N), organic P, and total S (instead of organic S) are reported. Significant variation can occur in the C:N:P:S ratios of individual soils, but the mean ratios for groups of soils from different regions are similar. Agricultural soils, in general, have a mean C:N:P:S ratio of approximately 130:10:1.3:1.3. Soils under native grass have ratios of the order of 200:10:1:1. Peat and organic soils have intermediate ratios of approximately 160:10:1.2:1.2. These ratios are shown in Table 1 for six Brazilian surface soils and six Iowa surface soils.

Sources of Sulfur in soils

Minerals Sources

Many S-containing minerals occur in nature. The main S-bearing minerals in rocks and soils are present in two states: (1) as sulfate, such as in gypsum ($CaSO_4 \cdot 2H_2O$), anhydrite ($CaSO_4$), epsomite ($MgSO_4 \cdot 7H_2O$), and mirabilite ($Na_2SO_4 \cdot 10H_2O$); or (2) as sulfide, such as pyrite and marcasite (FeS_2), sphalerite (ZnS), chalcopyrite ($CuFeS_2$), cobaltite (CoAsS), pyrrhotite ($Fe_{11}S_{12}$), galena (PbS), arsenopyrite ($FeS_2 \cdot FeAs_2$), and pentlandite $(Fe,Ni)_9S_8$.

Fertilizers Sources

There are many fertilizer materials, both liquid and solid, which are used to supply S to growing crops. The S source selected for any particular situation is determined by the crop to be grown, the S level of the soil, the cost of the material, and the ease with which its use can be fitted into a particular fertilizer program.

Atmospheric Sources

Rainfall and the atmosphere constitute a third important source of S in soils. It is estimated that in the USA more than 25 million tons of SO_2 or 13 million tons of S are emitted annually into the atmosphere. Most of these amounts are derived from combustion of fossil fuels, but industrial processes such as ore smelting, petroleum-refining operations, and other such sources contribute approximately 20% of

Table 1 Carbon, nitrogen, phosphorus, and sulfur relationships in some Brazilian and Iowa surface soils

	Ratio				
Soil no.	*Organic C/total N*	*Total N/organic P*	*Total N/total S*	*Organic P/total S*	*Organic C/total N/organic P/total S (organic S)*
Brazilian soils					
1	21.0	11.8	3.4	0.3	210:10:0.9:3.0 (2.5)
2	21.3	9.7	7.0	0.9	213:10:1.0:1.4 (1.1)
3	12.8	13.3	8.3	0.7	128:10:0.8:1.2 (1.1)
4	13.9	3.3	6.1	2.0	139:10:3.1:1.6 (1.6)
5	23.6	18.7	12.0	0.7	236:10:0.5:0.8 (0.8)
6	23.8	10.2	6.3	0.7	238:10:1.0:1.6 (1.5)
Mean	19.4	11.2	7.2	0.9	194:10:1.2:1.6 (1.4)
Iowa soils					
7	9.4	6.2	9.1	1.6	94:10:1.6:1.1 (1.0)
8	10.8	7.0	9.2	1.3	108:10:1.4:1.1 (1.0)
9	10.4	7.0	6.9	1.0	104:10:1.4:1.4 (1.4)
10	10.3	8.0	7.4	0.9	103:10:1.2:1.3 (1.3)
11	12.4	8.2	8.0	1.0	124:10:1.2:1.3 (1.2)
12	13.1	7.6	6.9	0.9	131:10:1.3:1.5 (1.4)
Mean	11.1	7.3	7.9	1.1	110:10:1.4:1.3 (1.2)

Reproduced with permission from Neptune AML, Tabatabai MA, and Hanway JJ (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

the total S emitted into the atmosphere. Another major source is volcanic activity around the world.

Chemical Nature of Organic Sulfur in Soils

Understanding the nature and properties of the organic S fractions in soils is important, because these compounds govern the release of plant-available S. Even though much of the organic S compounds in soils remains unidentified, three broad groups of S compounds are recognized. These groups have been classifed according to the nature of the reagents used or according to the groups of S compounds attacked by the reagents. Thus, three distinct groups of S-containing compounds have been identified (Figure 1):

1. Organic S that is not directly bonded to C and is reduced to H_2S by hydriodic acid (HI). This fraction is believed to be largely in the form of sulfate ester with C–O–S linkages. Examples of substances that contain these linkages include arylsulfate, alkylsulfates, phenolic sulfate, sulfated polysaccharides, choline sulfate, and sulfated lipids. Other organic sulfates could be present as sulfamates (C–N–S) and sulfated thioglycosides (N–O–S). On average, approximately 50% of the total organic S in humid and semihumid regions is present in this form, but it can range from 30 to 60% (Table 2). Values as high as 95% have been reported for Iowa subsoils. Unless otherwise indicated, the HI-reducible S includes the inorganic SO_4^{2-} fraction (Table 3);
2. Organic S that is directly bonded to C (C–S) and is reduced to inorganic sulfide by Raney Ni (50% each of Ni and Al powder) in an alkaline medium (NaOH). This fraction is believed to consist largely of S in the form of S-containing amino acids such as methionine and cysteine. Its concentration in soils ranges from 10 to 30% of the total organic S (Tables 2 and 3);

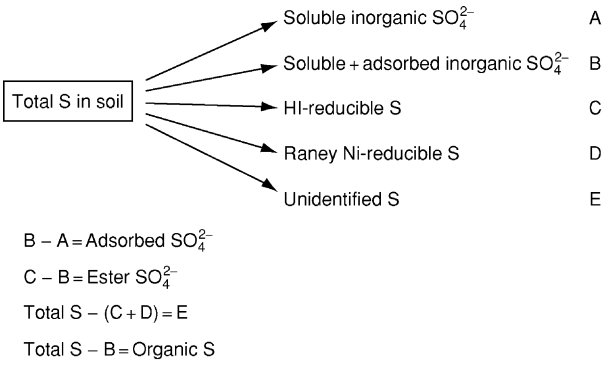


Figure 1 Organic and inorganic sulfur fractions in soils.

3. Organic S that is not reduced by either of the reagents employed in estimation of fractions 1 and 2. This unidentified fraction is inert to HI and Raney Ni. It is generally in the range of 30–40% of the total organic S (Tables 2 and 3). It is very stable, because it resists degradation by caustic chemical reagents; therefore, this fraction is of little importance as a potential source of S for plants.

Inorganic Sulfur in Soils

Inorganic S in soils may occur as sulfate and as compounds of lower oxidation states such as sulfide, polysulfides, sulfite, thiosulfate, and elemental S. In well-drained, well-aerated soils, most of the inorganic S normally occurs as sulfate, and concentrations of reduced S compounds are barely traceable, if present at all. The concentration of soluble SO_4^{2-} is estimated after extraction with 0.01% LiCl or 0.15% $CaCl_2$ solution, and the concentration of the soluble plus the insoluble fraction is estimated after extraction with a $Ca(H_2PO_4)_2$ solution containing 500 mg P l^{-1} (Table 2).

Fate of Inorganic Sulfate in Soils

The inorganic sulfate in soils may occur as water-soluble salts that can be leached from soils, adsorbed by soil colloids, or insoluble forms. Soil chemical properties such as pH, type of clay, and presence of cations are important factors in governing the leaching and adsorption of SO_4^{2-} in soils. The mechanism of SO_4^{2-} adsorption by soils involves coordination with hydrous oxides, exchange on edges of silicate clays, and molecular adsorption. Both the water-soluble and adsorbed SO_4^{2-} are available to plants. Under anaerobic (waterlogged) conditions, SO_4^{2-} is reduced to S^{2-} and precipitated as metal sulfide. This process is biological in nature.

Leaching losses

The movement of sulfate in soils determines the magnitude of losses of S in drainage water. Transport of sulfate within a soil profile is influenced by its concentration in soil solution, its reaction with the solid-phase components, and movement, velocity, and pattern of water movement within the soil. The relative magnitude of, and interactions among, these factors determines the physicochemical fate of the sulfate released from mineralization of organic S or added to soils as fertilizer, crop-residue decomposition, irrigation waters, and atmospheric deposition. From the information available on sulfate losses with percolating water, the following conclusions can be drawn:

Table 2 Total sulfur and percentage distribution of various forms of sulfur in some Brazilian and Iowa surface soils

		Percentage of total soil S in form specified						
		Inorganic sulfate S						
Soil no.	*Total S (mg kg^{-1} soil)*	*LiCl*	*Ca(H_2 PO_4)$_2$*	*Hl-reducible S*	*Ester sulfate S*	*Carbon-bonded S*	*Unidentified organic S*[a]	*Total organic S*
Brazilian soils								
1	59	5.1	15.3	35.6	20.3	5.1	59.3	84.7
2	43	9.3	23.3	51.2	27.9	4.7	44.1	76.7
3	72	6.9	12.5	52.8	40.3	11.1	36.1	87.5
4	214	3.7	5.1	70.1	65.0	5.6	24.3	94.9
5	209	1.4	4.8	53.6	48.8	12.4	34.0	95.2
6	398	2.0	6.3	43.0	36.7	5.0	52.0	93.7
Mean	166	4.7	11.2	51.1	39.8	7.3	41.7	88.8
Iowa soils								
7	55	7.3	7.5	50.9	43.4	18.2	30.9	92.5
8	174	5.8	4.9	42.9	47.0	9.2	37.9	94.1
9	331	5.4	5.5	57.7	52.2	12.1	30.2	94.5
10	338	2.4	2.5	55.3	52.8	8.9	35.8	97.5
11	438	1.8	1.8	61.6	59.8	7.3	31.1	98.2
12	580	3.5	3.5	48.6	45.1	12.1	39.3	96.5
Mean	319	4.4	4.3	52.8	50.1	11.3	34.2	95.6

[a]Unidentified organic S was calculated from total S − (Hl-reducible S + carbon-bonded S).

Reproduced from Neptune AML, Tabatabai MA, and Hanway J (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

Table 3 Fractionation of sulfur in surface soils from different regions

	Percentage of total S in form specified					
	HI-reducible		*C-bonded*[a]		*Unidentified*	
Location[b]	*Range*	*Mean*	*Range*	*Mean*	*Range*	*Mean*
Alberta, Canada (15)	25–74	49	12–32	21	7–45	30
Australia (15)	32–63	47	22–54	30	3–31	23
Brazil (6)	36–70	51	5–12	7	24–59	42
Iowa, USA (34)	36–66	52	5–20	11	21–53	37
Quebec, Canada (3)	44–78	65	12–32	24	0–44	11

[a]Determined by reduction with Raney Ni.
[b]Numbers in parentheses indicate number of samples.
Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

1. Losses of SO_4^{2-} are reduced by cropping and are less with rooting and perennial crops than with annual crops;
2. Leaching losses of SO_4^{2-} are greatest when monovalent ions such as K^+ predominate; next in order are the divalent ions such as Ca^{2+} and Mg^{2+}, and leaching losses are minimal when soils are acid and appreciable concentrations of Fe and Al hydrous oxides are present;
3. Under comparable soil and cropping conditions, the amount of SO_4^{2-} removed from the soil profile is generally directly related to the amount of leachate;
4. Sulfate adsorption would lead to more rapid and complete removal of Cl^- from acid soils;
5. Sulfate losses increase with liming or amendment with phosphate;
6. Sulfate losses are less when the S fertilizer is banded than when broadcast.

Complete separation of the influence of physical and chemical soil properties on transport of sulfate is not possible, because several factors affect the distribution of an ion such as sulfate in the soil pores. These include: (1) the electric field surrounding the individual soil particles, and (2) the magnitude of the repulsive (or attraction) forces which are dependent on the mineralogy and chemical composition of the solid phase, and the pH and salt content of the aqueous phase.

Losses of S by leaching vary widely: some drainage water contains more S than the rain supplies, even though little or none is added in fertilizer. The extra quantity may be deposited directly on plants and soils from the atmosphere or released from soil organic matter or minerals. Expressed in kilograms of sulfur per hectare, the annual losses from unfertilized fields by drainage water in the state of Illinois range from 1.5 to 65 kg S ha^{-1}, in Germany they average 33 kg S ha^{-1}, in Europe and North America they average 15 kg S ha^{-1}, in South America they average 4.5 kg S ha^{-1}, and in some areas of Australia they are less than 1 kg S ha^{-1}. It has been estimated that between 3 and 32 kg S ha^{-1} is lost by tile drainage in the state of Iowa. In general, the annual loss from soils by leaching varies from insignificant amounts to as much as 320 kg S ha^{-1} from soils treated with S fertilizers.

Sulfate Adsorption by Soils

Soils vary widely in their capacity to adsorb sulfate. Because sulfate adsorption occurs at low pH values (less than 6), its adsorption is negligible in most agricultural soils ($pH > 6$). Its adsorption in subsurface acid horizons plays an important part in contributing to S requirement of crops; conserving S from excessive leaching, and in determining S distribution in soil profiles. Sulfate adsorption is a reversible process and is influenced by a number of soil properties. These include:

1. *Clay content and type of clay mineral*. Sulfate adsorption usually increases with clay content of soils. Kaolin minerals retain more sulfate than montmorillonite clays;
2. *Hydrous oxide of Al and Fe*. Hydrous oxides of Al, and to a lesser extent of Fe, show marked tendencies to retain sulfate, especially the former in certain soils;
3. *Soil horizon or depth*. Most soils have some capacity to adsorb sulfate. The amounts of sulfate adsorption in surface horizons may be low but are often greater in lower soil horizons due to the presence of more clay and Fe and Al oxides;
4. *Soil pH*. Sulfate adsorption in soils is favored by strong acid conditions. It becomes almost negligible at $pH > 6$;
5. *Sulfate concentration and temperature*. The amount of sulfate adsorbed is concentration- and temperature-dependent. The amount of sulfate adsorbed is in kinetic equilibrium with sulfate concentration

in solution. Temperature has a relatively small effect on sulfate adsorption by soils;

6. *Effect of time*. Sulfate adsorption increases with the length of time it is in contact with soil;

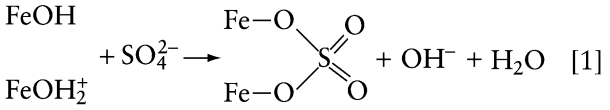
7. *Presence of other anions*. Sulfate is weakly held by soils. The retention decreases in the following order: hydroxyl > phosphate > sulfate = acetate > nitrate = chloride. Phosphate will displace or reduce the adsorption of sulfate;

8. *Effect of cations*. The amount of sulfate adsorbed is affected by the associated cation or by the exchangeable cation following the order: $H^+ > Sr^{2+} > Ba^{2+} > Ca^{2+} > Mg^{2+} > Rb^+ > K^+ > NH_4^+ > Na^+ > Li^+$. Both the cation and the sulfate from the salt are retained, but the capacity of adsorption of sulfate is different from that of the associated cation.

Mechanisms of Sulfate Adsorption by Soils

Several mechanisms have been proposed for adsorption of sulfate by soils.

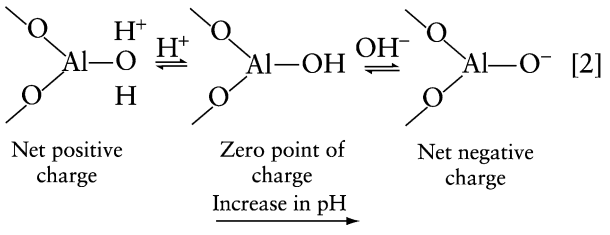
Coordination with hydrous oxides In acid soils sulfate adsorption essentially involves the chemistry of Fe and Al. The hydrous Fe and Al oxides tend to form coordination complexes due to the donor properties of oxygen. The adsorption involves the replacement of two surface OH groups (or OH_2^+) by one SO_4^{2-}. The two O atoms of the SO_4^{2-} are each bound to a different Fe^{3+}, resulting in a binuclear bridging surface complex:



[1]

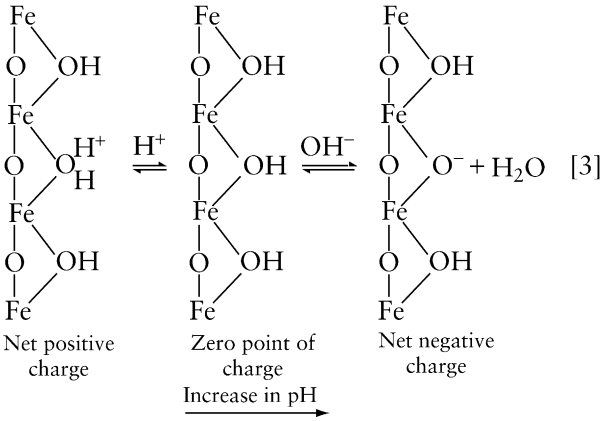
This reaction has been demonstrated to occur on the surfaces of goethite (α-FeOOH), and, because soils contain Fe oxides, it is assumed that a similar reaction takes place in soils under acid conditions.

The effect of pH on SO_4^{2-} adsorption by hydrous oxides in soil must consider the zero point of charge (the pH at which the change on the surface is zero). Deviation from this pH value involves protonation or deprotonation of Al oxides:



[2]

A similar protonation or deprotonation of Fe oxides is possible:



[3]

The positively charged Al and Fe hydrous oxides adsorb the SO_4^{2-} ions.

In addition to the mechanisms described above, homoionic Al-saturated clay coated with hydrated oxides R (Fe and Al) may adsorb sulfate ions as follows:

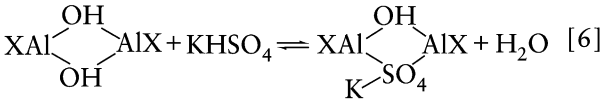
$$ {}^{y}K^+ + Al_x\text{—clay} + {}^{y}H_2O \longrightarrow Al_x(OH)_y^{Ky}\text{—clay} + {}^{y}H^+ \quad [4] $$

$$ SO_4^{2-} + R_x(OH)_y\text{—clay} \longrightarrow R_x[(OH)_{y-z}(SO_4)_z]\text{—clay} + {}^{z}OH^- \quad [5] $$

In this mechanism, it is assumed that K ion adsorption sites developed from the exchange and/or hydrolysis of Al on the clay surface. As a result of this hydrolysis, some Al ions are released into the solution. At the same time sulfate ions replace the OH ions from R(OH) coating on clay and substitutes for them. The replaced OH ions in turn react with H ions. According to this mechanism, whether the pH of the system increases or decreases depends on the relative rates of the two reactions: Al hydrolysis and OH ion exchange. It has been demonstrated that sulfate ion adsorption increases when sulfate-adsorbing soils are coated with Fe and Al oxides.

The adsorbed sulfate ions may be replaced by other anions of greater penetration (coordination) ability such as PO_4^{3-}. This has been demonstrated by displacing the adsorbed SO_4^{2-} from surface soils by application of phosphate fertilizers.

Exchange on edges of silicate clays This exchange presumably involves replacement of OH^- by SO_4^{2-} in terminal octahedral coordination with Al. The general effects and mechanism are similar to that discussed above for Al and Fe hydrous oxides. The following type of reaction can occur with both hydrated Al oxides and the Al layer kaolinite:



[6]

Molecular adsorption This mechanism is less understood than the others described above. It implies that SO_4^{2-} is adsorbed by some mechanism by which the associated cation is retained by soils. This mechanism has been referred to as 'salt' adsorption, 'molecular' adsorption, and imbibition.

Sulfur Transformations in Soils

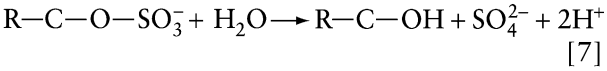
The transformations of S in soils are many and varied (i.e., oxidation, reduction, volatilization, decomposition and mineralization of plant and microbial residues), and often the changes are cyclic S changes from inorganic to organic forms (immobilization) and back again by living organisms.

Mineralization

The conversion of an element from organic form to an inorganic state as a result of microbial activity is termed 'mineralization.' As is the case with carbon and nitrogen, organic S in soils is mineralized to inorganic forms, mainly SO_4^{2-}, the form taken up by plant roots. The mechanisms involved in this transformation, however, are not clear. It appears that microorganisms are involved in this process, where they obtain their energy from the oxidation of carbonaceous materials in soils. During this process organic S is mineralized. Some of the mineralized S is used for synthesis of new microbial cell materials (immobilization), because the portion not required for synthesis is released as inorganic S. Mineralization and immobilization occur simultaneously in soils whenever organic material is undergoing microbial decomposition. The effect of temperature on the rate of N and S mineralization in 12 Iowa surface soils is shown in Table 4.

Sources of mineralizable S It is believed that ester sulfates in soils are the main sources of S mineralization in soils. However, carbon-bonded S (C–S) cannot be excluded, because this fraction contains the amino acids methionine, cystine, and cysteine, which can be converted to inorganic sulfate under aerobic conditions. The information available suggests that this form of S in soils can be a source for plant uptake.

Role of arylsulfatase in S mineralization Because a large proportion of the organic S in soils appears to be present as ester sulfate, it seems reasonable to expect that some organic S is mineralized by the action of arylsulfatase enzyme; indeed, this enzyme has been detected in soils. This enzyme catalyzes the hydrolysis of ester sulfate, releasing sulfate for plant uptake. The reaction is as follows:



[7]

Table 4 Comparison of rates of nitrogen and sulfur mineralization in Iowa surface soils; soil glass-bead columns were incubated at 20°C or 35°C and the mineral N and S produced were determined after leaching every 2 weeks with 0.01 mol l^{-1} KCl for a total of 26 weeks

	Rate of mineralization (kg ha^{-1} week^{-1})					
	20°C			*35°C*		
Soil	*N*	*S*	*N:S*	*N*	*S*	*N:S*
Lester	6.7	1.6	4.2	22.6	4.9	4.9
Ackmore	4.9	1.6	3.1	27.6	4.9	5.6
Fayette	5.6	1.8	3.1	23.5	5.2	4.5
Downs	8.1	2.7	3.0	35.9	7.0	5.1
Clarion	6.7	1.8	3.7	26.7	7.8	3.4
Muscatine	7.4	1.8	4.1	33.2	6.7	5.0
Nicollet	3.8	1.3	2.9	17.3	4.0	4.3
Tama	9.0	2.2	4.1	34.3	7.2	4.8
Webster	9.4	2.2	4.3	38.1	6.5	5.9
Canisteo	4.9	1.3	3.8	20.9	4.0	5.2
Harps	4.3	1.2	3.6	21.7	3.6	6.0
Okoboji	6.3	1.8	3.5	34.3	6.1	5.6
Mean	6.4	1.8	3.6	28.0	5.7	5.0

Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Pattern of sulfate release Because the opposing reactions of mineralization and immobilization can occur simultaneously, different patterns of SO_4^{2-} release have been observed, depending on the energy materials available for the microorganisms (Figure 2):

1. Immobilization of S during the initial stages of incubation followed by SO_4^{2-} release;
2. A steady linear release of SO_4^{2-} over the whole period of incubation;
3. A rate of release that decreases with incubation time.

The pattern of SO_4^{2-} release is not related to any specific soil properties, but is apparently caused by adjustment of the microbial populations to the incubation conditions and to the availability of the initial substrates.

Factors affecting sulfur mineralization Because S mineralization is microbiological in nature, any variable that affects microbial growth should affect S

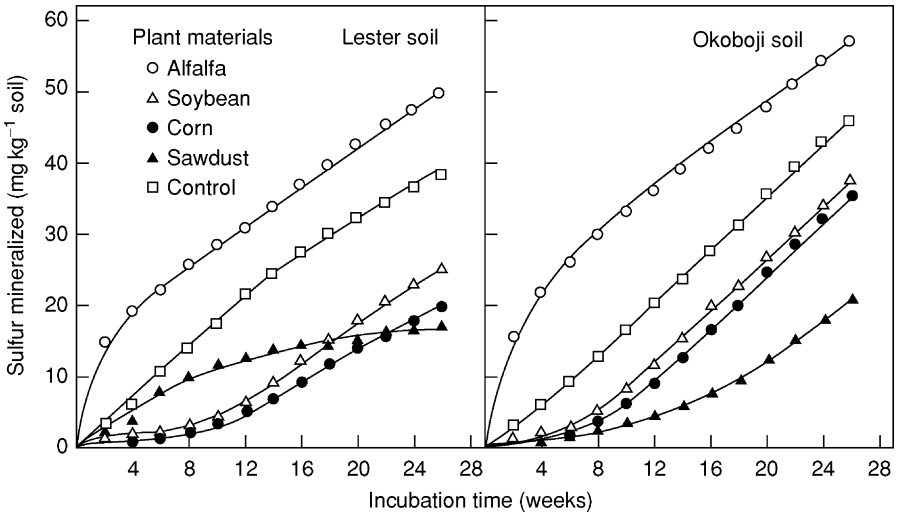


Figure 2 Effect of crop-residue treatment on sulfur mineralization in soils. (Reproduced from Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.)

mineralization. Therefore, temperature, moisture, pH, and the availability of nutrients are most important.

Oxidation of Elemental S in Soils

Elemental S is one of the main sources of S added to soils. Before it can be untilized by crops, however, elemental S has to be oxidized to sulfate. Elemental S is oxidized in soils by chemical and biochemical processes, and a number of factors affect these processes. Microbial reactions dominate the processes. The microorganisms involved in oxidation of elemental S in soils belong to three groups: (1) chemolithotrophs (e.g., members of the genus *Thiobacillus*); (2) photoautotrophs (e.g., species of purple and green S bacteria); and (3) heterotrophs, which include a wide range of bacteria and fungi. Those listed in groups (1) and (2) are mainly responsible for oxidation of reduced S compounds in aerobic soils. Phototrophic bacteria are the predominant organisms responsible for oxidizing S^{2-} at the soil–water interface in flooded soils and in the rhizosphere of rice plants. The major reduced forms of inorganic S found in elemental S-treated soils are S^0, S^{2-}, and the oxyanions $S_2O_3^{2-}$, $S_4O_6^{2-}$, and SO_3^{2-}. These anion are oxidized, ultimately to SO_4^{2-} (Figure 3). The reactions involved in oxidation of elemental S in soils seem to be as follows:

$$S^0 \longrightarrow S_2O_3^{2-} \longrightarrow S_4O_6^{2-} \longrightarrow SO_3^{2-} \longrightarrow SO_4^{2-} \quad [8]$$

It is not clear whether these reactions are biochemical, occurring as a result of microbial processes in soils, or whether some of the intermediate products formed are the results of abiotic reactions. This is

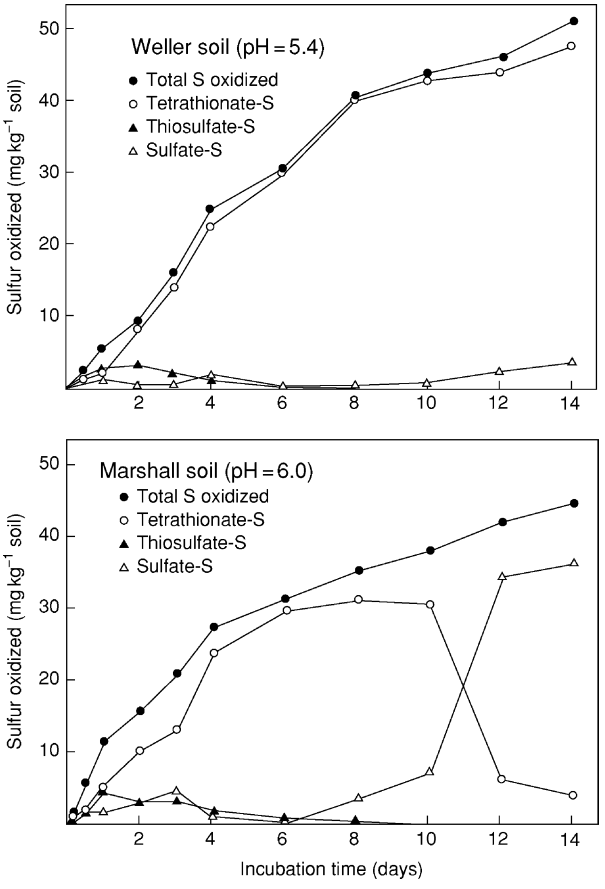


Figure 3 Amounts of thiosulfate-, tetrathionate-, and sulfate-S produced in two Iowa surface soils (Weller and Marshall soils) amended with elemental S (200 mg S kg^{-1} soil) and incubated at 30°C for various times. (Reproduced from Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.)

especially important in the case of the intermediates $S_2O_3^{2-}$ and $S_4O_6^{2-}$.

The enzyme rhodanese (thiosulfate-cyanide sulfotransferase; EC 2.8.1.1) catalyzes the conversion of the intermediate $S_2O_3^{2-}$ to SO_3^{2-} as follows:

$$S_2O_3^{2-} + CN^- \longrightarrow SO_3^{2-} + SCN^- \qquad [9]$$

Several factors affect of rate of oxidation of elemental S in soils. These include: (1) particle size: the finer the particles, the faster the reaction, which is because of the increase in surface area with decreasing particle size; (2) temperature: the higher the temperature, the greater the reaction rate (Figure 4); this is true between 10°C and 40°C; (3) time of contact with soil: the longer the reaction, the more the oxidation (Figure 5); and (4) effect of pH: the oxidation appears to be faster in alkaline than in acid soils.

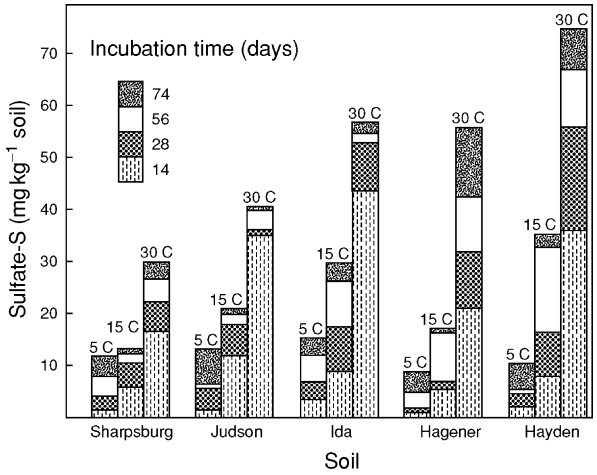


Figure 4 Effects of temperature and time of incubation on oxidation of elemental S (100 mg S kg^{-1} soil) in soils. (Reproduced from Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.)

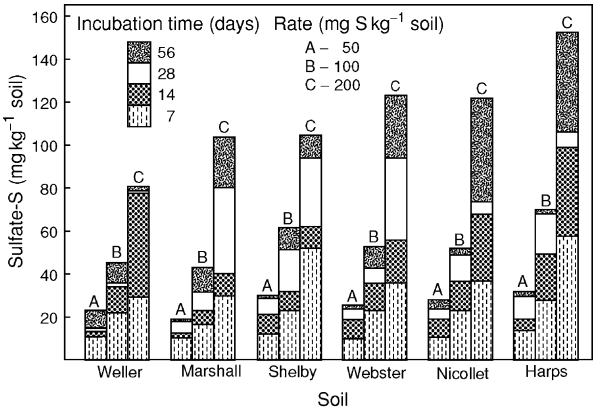


Figure 5 Effects of application rate and time of incubation on oxidation of elemental S in soils. The values for Weller and Marshall soils treated with 200 mg S kg^{-1} soil include the amounts of thiosulfate- and tetrathionate-S produced during 7 days and 14 days of incubation.

Reduction of Sulfate in Waterlogged Soils

The reduction of sulfate to H_2S is a process that occurs mainly by anaerobic bacteria; thus, it occurs only in anaerobic soils. This process is not important in aerobic agricultural soils, except perhaps in anaerobic microsites in soil aggregates. However, it is a major reaction in S cycling in waterlogged soils or periodic flooding, especially when in soils containing readily decomposable plant residues such as alfalfa (*Medicago sativa*). Bacterial reduction of sulfate involves either an assimilation or dissimilation process. In the former process, sulfate is reduced to the thiol (–SH) group of organic compounds for protein synthesis. In the latter process, the reduction leads to production of H_2S under very low redox potential (E_h) values. Under normal conditions, however, H_2S is not volatilized from soils, because it precipitates with Fe^{2+}, Mn^{2+}, Cu^{2+}, Cu^+, and/or Zn^{2+} present in soils. In the case of Fe^{2+}, it forms ferrous sulfide (FeS), and pyrite (FeS_2) is formed in severely reducing conditions by the reduction of sulfate to S^{2-} by the bacteria *Desulfovibrio desulfuricans*, which reacts with FeS to produce FeS_2.

Volatilization of S Compounds from Soils

Relative small amounts, if any, of S-containing gases, including H_2S, are released from aerobic agricultural soils, even when such soils are waterlogged. Substantial amounts of H_2S are released, however, from salt-marsh soils. Several gases are released from soils when treated with animal manures, sewage sludges, and protein-rich plant materials such as alfalfa, especially under waterlogged conditions. These include carbon disulfide (CS_2), which results from decomposition of the amino acids cystine and methionine; carbonyl sulfide (COS), released during decomposition of thiocyanate and isothiocyanate; and methyl mercaptan (CH_3SH), dimethyl sulfide (CH_3SCH_3), and dimethyl disulfide (CH_3SSCH_3), which result from decomposition of methionine-containing materials.

See also: **Acid Rain and Soil Acidification**; **Enzymes in Soils**; **Fertility**; **Greenhouse Gas Emissions**; **Minerals, Primary**; **Organic Matter:** Principles and Processes; **Organic Residues, Decomposition**; **Sorption:** Oxyanions; **Sulfur in Soils:** Biological Transformations; Nutrition

Further Reading

Freney JR (1967) Sulfur-containing organics. In: McLaren AD and Peterson GH (eds) *Soil Biochemistry*, vol. 1, pp. 229–259. New York: Marcel Dekker.

Germinda JJ and Gupta VVSR (1992) Biochemistry of sulfur cycling in soil. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 1–53. New York: Marcel Dekker.

McLachlan KD (ed.) (1975) *Sulphur in Australasian Agriculture*. Sydney, Australia: Sydney University Press.

Neptune AML, Tabatabai MA, and Hanway JJ (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.

Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Tabatabai MA (1985) Effect of acid rain on soils. *CRC Critical Reviews in Environmental Control* 15: 65–110.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Agronomy monograph no. 27. Madison, WI: American Society of Agronomy.

Tabatabai MA (1987) Physicochemical fate of sulfate in soils. *Journal of the Air Pollution Control Association* 37: 34–38.

Tabatabai MA (1994) Soil enzymes. In: Weaver RW, Angle JS, and Bottomley PS (eds) *Methods of Soil Analysis*. part 2, pp. 775–783. *Microbiological and Biochemical Properties*. Book series no. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA (1996) Sulfur. In: Sparks DL (ed.) *Methods of Soil Analysis*. part 3, *Chemical Methods*, pp. 921–960. Book series no. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.

Tisdale SL, Nelson WL, and Beaton JD (1985) *Soil Fertility and Fertilizers*, 4th edn. New York: Macmillan.

Biological Transformations

S D Siciliano and J J Germida, University of Saskatchewan, Saskatoon, SK, Canada

Introduction

In soil, sulfur exists in a variety of forms, some of which are beneficial to the environment, and other forms that are typically considered pollutants. The wide array of oxidation states and forms of sulfur results in a rich diversity of biological transformations involving this important element. These transformations are critical for many basic ecological processes such as plant growth, but also influence the movement of environmental contaminants.

The biological transformations of sulfur can be divided into four major classes: (1) oxidation; (2) reduction; (3) assimilation; and (4) mineralization (Figure 1). In addition, microbial metabolites can chemically fix or solubilize inorganic S compounds. The first two classes are related to the acquisition of energy for growth and the last two are typically related to nutrient acquisition. For energy acquisition, sulfur can act as an electron acceptor, an electron donor, and sometimes both at the same time. Sulfate (SO_4^{2-}) acts as an electron acceptor and is reduced to S^{-2} by a group of bacteria known as sulfate-reducing bacteria. These bacteria grow in the absence of oxygen. In contrast, S^{-2} can act as an electron donor and is oxidized by a group of bacteria known as sulfur oxidizers. These bacteria grow in the presence of oxygen. Interestingly, some bacteria can disproportionate sulfur, i.e., ferment sulfur. In this reaction, one of the two sulfur atoms, present in thiosulfate as S^{-1}, acts as an electron donor and the other sulfur atom, present as S^{+5}, acts as an electron acceptor.

The last two classes of sulfur transformation are directly linked to the nutritional status of the cell. Mineralization is a transformation in which organic sulfur contained in a growth substrate is released in a mineral form such as sulfate. In contrast, assimilation is a process is which sulfur present in the environment is incorporated into biomass. In the former case, the mineralized sulfur is available for other organisms to use, whereas in the latter the assimilated sulfur is no longer available to other organisms. Each of these transformations of sulfur has important side-effects on the environment. The mineralization of sulfur provides sulfate, which can potentially stimulate sulfate-reducing bacteria that play important roles in the arsenic and mercury biogeochemical cycles. The oxidation of reduced sulfur by sulfur oxidizers results in acid mine drainage, one of the primary detrimental impacts of mining.

The Global Sulfur Cycle

The global cycle of sulfur begins with sulfur volatilized largely as dimethylsulfide from marine algae, marshlands, mud flats, plants, and soil entering the atmosphere. This dimethylsulfide is converted photochemically to methanesulfonic acid, which then deposits on to the Earth and is rapidly converted by bacteria to carbon dioxide and sulfate. This sulfate

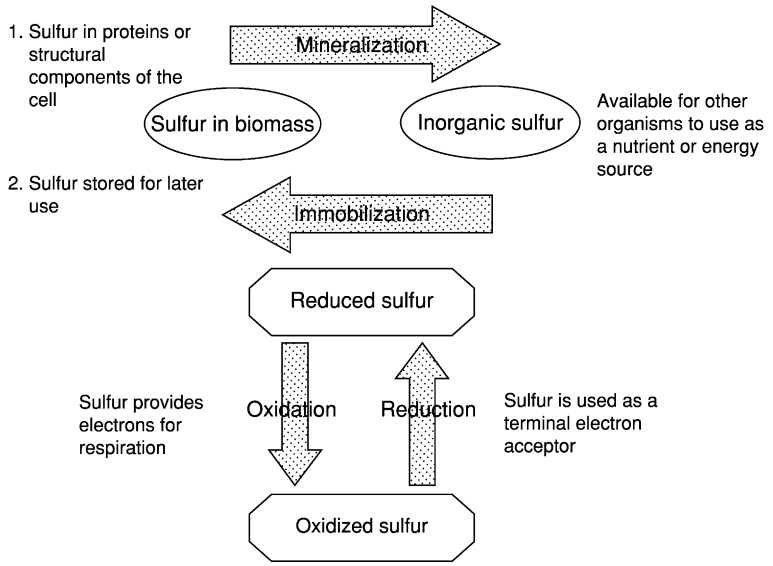


Figure 1 The four principal biological transformations of sulfur.

is assimilated by plants and immobilized into compounds like sulfoquinovosyl diacylglycerol, or alternatively the sulfur is released back to the atmosphere as dimethylsulfide. As plants die, bacteria metabolize the immobilized sulfur present in plant biomass and either immobilize this sulfur into their biomass or release it as sulfate for uptake by plants. In addition to this natural sulfur cycle, human activities have at least doubled the input of sulfur to terrestrial systems. Approximately 1.5×10^{11} kg S per year is deposited due to atmospheric pollution, largely in the form of sulfuric acid (sulfate), with small amounts of sulfite and biosulfite also being deposited. This deposited sulfate can either be used by bacteria as a terminal electron acceptor or assimilated by organisms requiring sulfur as a nutrient. Despite this recent disruption of the global sulfur cycle, the sulfur levels present in soil are largely due to pedogenic factors like climate, vegetation, and parent material. As a result, sulfur levels in soil range from 0.002 to 10%: most of this sulfur, 90%, is found in organic form.

The Biological Availability of Sulfur

Although 90% of sulfur in soil is present as organic sulfur, much of this organic S is not available for assimilation or dissimilation by microorganisms and plants. Microorganisms need to convert the organic sulfur in soil into sulfate before plants can take up the sulfate as a nutrient. This conversion of sulfate from organic sulfur is the end result of a complicated biogeochemical cycle. Sulfur enters the soil ecosystem and is converted and assimilated by microorganisms into their biomass. Predators then consume these bacteria and excess sulfur is released into the soil solution. The sulfur released into the soil solution is available for uptake by other organisms or plants. However, the predators do not mineralize all the S present in prey organisms and, as a result, a small fraction of S enters what is termed the 'resistant' pool of organic S present in soil (Figure 2). This resistant pool of organic S is found is large polymers that are in close association with clays. These large polymers resist microbial attack because it is not energetically favorable for microbes to attack these polymers. There are three distinct pools of sulfur in soil: (1) inorganic sulfate that is readily taken up by plants and microbes alike; (2) short-lived organic sulfur compounds like taurine and cysteine which are

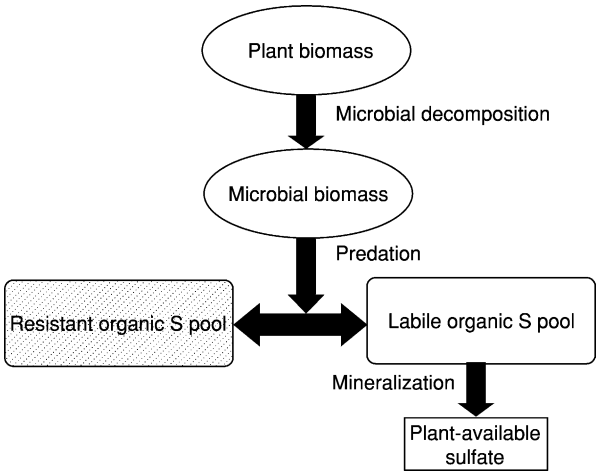


Figure 2 The different sulfur pools present in soil and their relationship to plant-available sulfur.

mineralized by microorganisms and either released as sulfate or incorporated into biomass; and (3) long-lived resistant organic sulfur polymers found in soil humus. Microbial transformations are the primary process by which sulfur is transferred between the pools of readily available, young organic sulfur and old organic sulfur. These transformations are discussed below.

Mineralization

Almost exclusively, microorganisms mediate sulfur mineralization. By definition, mineralization is the metabolic conversion of an organic form of an element to an inorganic form. Mineralization processes can be divided into cell-mediated or extracellular enzymatic processes. Cell-mediated mineralization occurs by oxidative decomposition under aerobic conditions or by desulfurization under anaerobic conditions. In cell-mediated mineralization, S present in carbon-containing compounds is mineralized as organisms consume the carbon to obtain energy. Thus, if there is more S in the substrate than the organism requires, the S will be mineralized and released to the environment. If there is not enough S, then any released S is consumed and immobilized by the organism. The break-even point for mineralization can be calculated based on the C:S ratio of the substrate, the C:S ratio of the organism, and the yield coefficient, i.e., how much of the consumed C is incorporated into biomass. As a general rule of thumb, if the C:S ratio of a substrate is 200 or less, then S is released. If the C:S ratio is greater than 400, then S is not going to be released because the microorganisms require the S for their own biomass. In this latter case, organisms rely on a second, enzymatic method of mineralization to obtain the sulfur essential for growth.

Enzymatic mineralization involves extracellular enzymes, such as arylsulfatases, released by the microorganism. These enzymes, once released from cells, will hydrolyze sulfate esters present in the soil. This process releases sulfate into the soil solution for use by the cells. The activity of these enzymes depends on a wide range of soil factors with reaction rates varying by a factor of 4 between different soils. These enzymes attack the vast amount of S present in the soil as organically bound S. Organic S in soil constitutes more than 90% of the total S present in soils. This organic S can be grouped into carbon-bonded S and organic sulfates. Organic sulfates, which comprise between 30 and 75% of sulfur in soil, typically include compounds such as sulfate esters (C-O-S), sulfamates (C-N-S), and sulfated thiogylcosides (N-O-S), and it is this form of sulfate that enzymatic mineralization processes attack.

Mineralization processes are biological transformations of sulfur that are intimately linked to cellular growth. Thus, factors that influence microbial activity, such as moisture, temperature, or plant growth, will influence mineralization. Increasing the growth of microorganisms by either adding a carbon source or plant growth typically results in reduced mineralization.

Assimilation

Assimilation is the process of converting inorganic sulfur into organic sulfur present in an organism. This assimilatory process is intimately linked to mineralization. In fact, these two processes, assimilation and mineralization, occur simultaneously. If the net effect is to release inorganic sulfate then this is termed mineralization; if instead the net effect is for the sulfur to be incorporated into the biomass, then this is termed assimilation. Thus, the first step of many assimilation processes is the conversion of sulfur into a form that can be incorporated into biomass. Taken alone, such steps might be considered mineralization processes but we term them assimilatory processes if: (1) the sum of the steps is to incorporate sulfur into biomass and (2) they are induced by sulfur limitation, suggesting that the organism is using the process to obtain sulfur. There are different biological assimilatory transformations that can occur depending on what form the available sulfur is in. For inorganic SO_4^{2-}, an enzyme called adenosine triphosphate (ATP) sulfurylase mediates the reaction between sulfate and ATP and sulfate to form adenosine-5′-phosphosulfate (APS) (Figure 3). In turn, this APS is transformed by APS kinase with another ATP molecule to form adenosine-3′-phosphate-5′-phosphosulfate (PAPS). This high-energy molecule reacts with PAPS reductase and two sulfhydryl groups to form sulfite, and sulfite is reduced by $NADPH_2$ to form sulfide which is incorporated into L-cysteine by *O*-acetylserine sulfhydrylase.

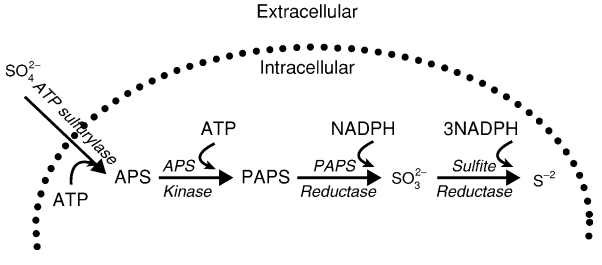


Figure 3 Assimilation of inorganic sulfur by bacteria. See text for a complete description of the reactions involved and definition of abbrevations. The sulfide produced is incorporated into cysteine by cysteine synthase.

This pathway will efficiently scavenge excess sulfate present in the soil solution that organisms require. So much so, that in the presence of an energy source, added SO_4^{2-} S is quickly incorporated into the organic fraction of the soil biomass and later found in the fulvic acid fractions. However, sulfate is not commonly found in soil solution and organisms have developed other pathways to assimilate sulfur for their metabolic needs.

Aromatic sulfonates, SO_3^{1-} groups bound to aromatic rings, can also serve as a sulfur source for assimilation. This assimilatory pathway is controlled by a genetic cluster called the 'sulfate-stimulation-induced stimulon' that encodes three distinct gene clusters, *asf*, *ssu*, and *ats*. Together these three genes control the assimilation of S by bacteria in soil. The *ats* gene cluster is responsible for the binding of aromatic sulfonates and transportation into the cell. The *ssu* genes encode for the cleavage for the C-S bond by a monooxygenase. The *asf* gene cluster provides the reducing equivalents from the oxidation of NADH necessary to assimilate aromatic sulfonates. The overall result of this pathway is the production of sulfite, which reacts with PAPS reductase to form sulfide, as described above. Alkane sulfonates are also a source of sulfur for assimilation. For most alkane sulfonates the pathway is very similar to that described above, with the exception that the transport mechanisms differ. There is a distinctly different pathway for taurine, involving an α-ketoglutarate-dependent dioxygenase which oxidizes both taurine and α-ketoglutarate to release sulfite. This pathway is regulated by *Tau*D and is specific for taurine, with little activity seen for other alkane sulfonates.

Sulfate esters are assimilated by the action of a group of enzymes termed 'arylsulfatases.' These enzymes are broadly grouped by their pH optima, with alkaline sulfatases having a pH optimum of 8.3–9.0 and acid sulfatases having a pH optimum of 6.5–7.1. In essence, these enzymes hydrate the S-O bond and release sulfate, which can then be assimilated by the pathway described above for sulfate. These assimilatory processes should be clearly differentiated from processes that use organo-sulfur compounds as a C and energy source. Organo-sulfur assimilation genes are regulated by the sulfate-stimulation-induced stimulon, which in turn is regulated by levels of cysteine, thiosulfate, and sulfite present in the cell. The presence of these compounds inside the cell will downregulate the sulfate-stimulated-induced genes and thereby repress expression of assimilatory enzymes. This is a logical regulatory control, since if there is excess cysteine and sulfate present in the cell, the cell likely does not need to assimilate sulfur.

Sulfur Oxidation

Many organisms, such as chemoautotrophs and photoautotrophs, use sulfur as a source of energy. This occurs by oxidizing S^0 to SO_3^{2-} and finally to SO_4^{2-} and in the process stripping six electrons from sulfur and using oxygen or nitrate as a terminal electron acceptor. A great variety of thiobacilli can perform these reactions. Some thiobacilli are chemoautrophs (only able to use S as an energy source), others are facultative chemoautotrophs (able to use C as an energy source if necessary) and some are mixotrophs (able to use C and S as an energy source at the same time). Most thiobacilli require the presence of oxygen but some, such as *Thiobacillus denitrificans*, can use nitrate as a terminal electron acceptor. Similarly, some thiobacilli can use Fe^0 in addition to sulfur as an electron donor.

The research on the importance of thiobacilli on sulfur oxidation comes largely from pure culture studies. The actual situation in soil involves a wide diversity of organisms. There are gliding sulfur oxidizers, including bacteria, the cells of which are arranged in trichomes, that show a gliding motion on the substrate. The most important members of this group in relation to S-oxidation in soils are species of *Beggiatoa*, bacteria that participate in sulfide oxidation in the root zone of rice. All strains of *Beggiatoa* deposit sulfur in the presence of H_2S. Phototrophic bacteria, such as *Chromatium* and *Chlorobium*, also play an important role in sulfide oxidation in rice paddy soil, but not in aerobic agricultural soils. A number of nonfilamentous, chemolithotrophic sulfur-oxidizing bacteria, such as *Sulfolobus*, *Thiospira*, or *Thiomicrospira*, have also been isolated from special habitats, but the importance of these bacteria in S oxidation in soils has yet to be determined. Some thiobacilli, such as *Thiobacillus denitrificans*, can oxidize sulfur or thiosulfate and use nitrate as an electron acceptor. Other aerobic bacteria, such as *Arthobacter*, *Bacillus*, and *Pseudomonas*, oxidize sulfur in soil during their normal metabolism of oxidizing organic compounds in soil and using oxygen as a terminal electron acceptor. These reactions have not been characterized but it is thought that this is occurring as a side reaction to the primary transformations being carried out by these organisms. In the soil ecosystem, it is widely assumed that thiobacilli are the dominant sulfur oxidizers but this view is based on the observation that if you add elemental sulfur to soil, the numbers of thiobacilli increase. However, no consistent correlation between S-oxidation rates and the prevalence of thiobacilli has been demonstrated, except in very broad terms. In general, it is assumed that initially heterotrophs oxidize elemental sulfur

until the pH is low enough that oxidation of sulfur chemoautotrophs is energetically favorable and then thiobacilli dominate. Thus, under energy-limiting conditions in an environment dominated by dissolved, reduced sulfur compounds with little organic matter, thiobacilli will dominate. In contrast, in a soil in which there are significant amounts of organic matter and less reduced sulfur, heterotrophs will dominate.

Disproportionation of Sulfur

In the latter part of the twentieth century, investigators discovered that not only can bacteria oxidize and reduce sulfur, but they can also disproportionate it. Disproportion reactions are best likened to a fermentation in which one portion of a molecule acts as an electron donor and the other, an electron acceptor. Bacteria are able to disproportionate thiosulfate by oxidizing the sulfur atom in thiosulfate that carries a +5 charge such that it now carries a charge of +6. This +6 sulfur is released as SO_4^{2-}. Simultaneously, bacteria oxidize the other sulfur atom in thiosulfate that carries a −1 charge such that it now carries a charge of −2. This −2 sulfur is released as hydrogen sulfide:

$$S_2O_3^{2-} + H_2O \rightarrow SO_4^{2-} + HS^- + H^+ \qquad [1]$$

This metabolic process has immense implications for the global sulfur cycle since typically in sediments sulfate predominates in the upper layers and sulfide in the lower layers. Hence, this metabolic reaction reveals an entirely new and, up until then, unsuspected ecological niche which is responsible for up to 60% of the thiosulfate transformations in sediment. Since then, other bacteria have been found that can disproportionate sulfonates such as taurine to sulfate and sulfide. The bacteria capable of this organic sulfur disproportation reaction form a new species called *Desulfonispora* since it appears that it is a distinct group of organisms that is capable of this reaction.

Reduction of Sulfur

The reduction of sulfur can occur for two reasons: assimilation and dissimilation. The assimilation pathway has been discussed above – bacteria convert sulfate to sulfide for inclusion in amino acids such as cysteine. Dissimilatory sulfate reductase refers to a process whereby sulfate is used as a terminal electron acceptor. Sulfate-reducing bacteria use such an enzyme, which allows them to oxidize organic substrates such as lactate, malate, and ethanol, and use sulfate as a terminal electron acceptor. Hence, these bacteria are primarily found in environments where there is little or no oxygen and live off the fermentation end products produced by other bacteria.

Sulfate-reducing bacteria catalyze the following process:

$$SO_4^{2-} + 5H_2 \rightarrow H_2S + 4H_2O \qquad [2]$$

Since hydrogen sulfide is extremely toxic, these organisms require some sort of metal to react with the hydrogen sulfide produced and precipitate as a nonsoluble metal sulfide. In soil, this metal is typically iron, with FeS being precipitated around areas of sulfate-reducing bacterial growth and activity. The reaction catalyzed by sulfate-reducing bacteria reaches an optimum at an E_h of −300mV at a pH of 7. These conditions are typically reached as one enters the highly anaerobic zone. Typically, terminal electron acceptors are depleted in the following order: oxygen, nitrate, nitrite, manganite, iron, sulfate, and finally, carbon dioxide. Thus, there are many alternative electron acceptors that will be used by a microbial community before sulfate is reduced. Despite this limitation, sulfate-reducing bacteria and activity are commonly found in soils, sediments, polluted water, oil-bearing strata, and shales.

Sulfate-reducing bacteria can implement an alternative survival strategy if sulfate levels in the water are too low to be used. The H^+ produced by organic matter oxidation is transferred directly to a methanogen which then consumes that H^+ to reduce CO_2 and produce methane. Because the methanogens keep the partial pressure of hydrogen so low, the sulfate-reducing bacteria can use this mechanism to obtain energy from organic matter degradation even when there is not enough sulfate around for them to transform to sulfide.

Sulfur Transformations and Environmental Quality

The biological transformations of sulfur can lead to significant environmental problems in the soil. The enzyme sulfate-reducing bacteria use to transfer methyl groups can also accidentally methylate mercury to form methylmercury. Sulfate-reducing bacteria have been identified as one of the primary causes of the increased mercury accumulation in many ecosystems because methylmercury is the form of mercury that most readily accumulates in food chains. Typically, this reaction occurs in anaerobic, lowland soils where the activity of sulfate-reducing bacteria is closely linked to methylmercury production. However, sulfate-reducing bacteria can also mitigate environmental pollution through the production of metal sulfides by the following reaction:

$$M^{2+} + S^{2-} \rightarrow MS \quad [3]$$

Sulfate-reducing bacteria like *Desulfovibrio* spp. can form sulfides of Sb, Co, Cd, Fe, Pb, Ni, and Zn during the reduction of sulfate. The extent of metal sulfide genesis depends on many factors, such as the amount of sulfate present and the relative toxicity of the metal ion. In nature, this toxicity is reduced when the metal ions are adsorbed on clays or complexed with organic matter.

The oxidation of metal sulfides in soil involves both chemical and microbial processes and, as a result, is a more complex process than is the oxidation of S^0 Chalcocite (Cu_2S), chalcopyrite ($CuFeS_2$), galena (PbS), pyrite (FeS_2), and nickel sulfide (NiS) are just a few examples of metal sulfides that are subject to microbial transformations. For example, the biological oxidation of pyrite (FeS_2) can follow one of two pathways. The first result follows a series of oxidation steps, described in eqns [4–7]. The second mechanism of pyrite oxidation involves thiosulfate as an intermediate but in the end requires the oxidation of elemental sulfur to sulfate by *Thiobacillus*. These biotic oxidations are responsible for the formation of acid mine drainage and acid soil formation in surface mine spoils. First, ferrous sulfate is formed as the result of an abiotic oxidation step:

$$2FeS_2 + 2H_2O + 7O_2 \rightarrow 2Fe_2SO_4 + 2H_2SO_4 \quad [4]$$

This reaction is then followed by the bacterial oxidation of ferrous sulfate, generally by *T. ferrooxidans*:

$$4FeSO_4 + O_2 + 2H_2SO_4 \rightarrow 2Fe(SO_4)_3 + 2H_2O \quad [5]$$

This reaction occurs chemically but can be accelerated 10^6–10^8 times by thiobacilli. This bacterial oxidation of ferric ion plays an important role in the bioleaching of metal sulfides in the environment because it cycles the iron between +2 and +3 and the iron is then free to oxidize sulfide minerals abiotically.

Subsequently, ferric sulfate is reduced and pyrite oxidized by a strictly chemical reaction:

$$Fe_2(SO_4)_3 + FeS_2 \rightarrow 3FeSO_4 + 2S^0 \quad [6]$$

$$2S + 6Fe_2(SO_4)_3 + 8H_2O \rightarrow 12FeSO_4 + 8H_2SO_4 \quad [7]$$

The elemental sulfur produced is finally oxidized by *T. thiooxidans*, and the acidity produced helps the whole process to continue.

$$2S^0 + 3O_2 + 2H_2O \rightarrow 2H_2SO_4 \quad [8]$$

Note the net production of 10 molecules of H_2SO_4 during the process.

Although several sulfur-oxidizing thiobacilli and heterotrophs can be isolated from acid sulfate soils in which pyrite is being oxidized, they appear not to play an important role in the process, with the exception of *T. ferrooxidans*. The biological oxidation of sulfides and other reduced S compounds can have severe consequences for the environment. For example, acid mine drainage contaminates several thousand kilometers of streams in the Appalachian coal-mining region of the USA.

Conclusion

There is a large amount of sulfur present in soil and most of it is found in a form susceptible to biological transformations. These transformations include mineralization, assimilation, oxidation, and reduction. Recently, anthropogenic activity has impacted the sulfur cycle by increasing the deposition of sulfate in the form of acid rain and by exposing more sulfur to oxidation during mining operations. The four basic transformations discussed in this article influence the fate of the movement of sulfur in terrestrial and aquatic ecosystems and the environmental impact of anthropogenic activities altering the sulfur cycle.

See also: **Organic Residues, Decomposition**; **Sulfur in Soils:** Overview; Nutrition

Further Reading

Benning C (1998) Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol. *Annual Reviews in Plant Physiology and Molecular Biology* 49: 53–75.

Ehrlich HL (2002) *Geomicrobiology*. New York: Marcel Dekker.

Fenchel T, King GM, and Blackburn TH (1998) *Bacterial Biogeochemistry: The Ecophysiology of Mineral Cycling*. San Diego, CA: Academic Press.

Howarth RW, Stewart JWB, and Ivanov MV (eds) (1992) *Sulfur Cycling on the Continents: Wetlands, Terrestrial Ecosystems, and Associated Water Bodies. SCOPE 48*. Chichester, UK: John Wiley.

Huang PM, Bollag JM, and Senesi N (eds) (2001) *Interactions Between Soil Particles and Microorganisms: Impact on the Terrestrial Ecosystem*. Chichester, UK: John Wiley.

Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*, 2nd edn. New York: Academia Press.

Sylvia DM, Furmann JJ, Hartel PG, and Zuberer DA (eds) (1998) *Principles and Applications of Soil Microbiology*. New Jersey: Prentice Hall.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Madison, WI: American Society for Agronomy.

Nutrition

M A Tabatabai, Iowa State University, Ames, IA, USA

Introduction

Sulfur(s) is considered either the fourth most important nutrient element, after N, P, and K, for plant growth, or the sixth, after C, H, O, N, and P, for protein composition. In addition to these six, there are 15 other elements that are essential for the growth of some plants. These are Ca, Mg, K, Fe, Mn, Mo, Ni, Cu, B, Zn, Cl, Na, Co, V, and Si. Not all these elements are essential for all plants, but all have been shown to be essential for some plants.

S has been recognized as an essential element for plant growth and development for more than 200 years. In 1859 Liebig was aware of the close relationship between N and S in many plants. The ancient Romans and Greeks demonstrated the beneficial use of gypsum ($CaSO_4 \cdot 2H_2O$) as a fertilizer. The history of the use of S fertilizers may be divided into three periods: (1) the first period is the reign of gypsum (1760–1845); during this period, gypsum was used widely and its beneficial effect was generally overestimated; (2) the second period is the reign of superphosphate (1845–1905), during which the need for S fertilization was ignored and the use of gypsum was discouraged by agronomists; (3) the third period is the Renaissance, or Modern period. The use of gypsum as a fertilizer in modern times can be traced back to the middle of the eighteenth century. A field trial with gypsum applied to a clover field was reported in Switzerland in 1768. Application of gypsum increased the yield of clover by more than two-fold.

Interest in fertility-related aspects of the S cycle is increasing, because S deficiencies in agronomic crops are observed with increasing frequency. Sulfur fertilization is now required for many crops around the world. Several factors are responsible for the increased need of S fertilization. These include: (1) increased use of high-analysis fertilizers that contain little or no S, (2) increased crop yields, (3) decreased use of S as a pesticide, (4) more intensive cropping, and (5) decreased gain of atmospheric S by soils and plants as a result of decreased combustion of coal and other S-containing fuels.

Sulfur Requirements of Crops

The importance of S in crop production is obvious, because plants require S for synthesis of essential amino acids and proteins, certain vitamins and coenzymes, glucoside oils, structurally and physiologically important disulfide linkages and sulfhydryl groups, and activation of certain enzymes.

Sulfur is absorbed by plant roots almost exclusively as the sulfate ion, SO_4^{2-}. Typical concentrations of S in plants range from 0.1 to 0.4%. As with N, much of the SO_4^{2-} taken up by plants is reduced in plants, and the S is found in –S–S– and –SH forms. Sulfate-S in large concentrations may occur in plant tissues and cell sap. Normally, S is present in equal or lesser concentration than phosphorus in such plants as wheat, corn, beans, and potatoes, but in greater concentrations in alfalfa, cabbage, and turnips. Generally, agronomic crops require about the same amount of S as they do for P.

Although the S content of plants varies depending on the supply available, some crops have greater S requirement than others. An average yield of forage crops removes 17–50 kg S ha^{-1}, and the cereal grains generally remove more than 30 kg ha^{-1}. Other crops such as cabbage, turnip, and alfalfa have a particularly high requirement for S. Such crops commonly need from 45 to 70 kg S ha^{-1}. One of the crops that require a very high amount of S is sugarcane. A yield of 224 tons ha^{-1} removes approximately 100 kg S (Table 1).

Another factor that affects the S requirement of plants is the available N. S and N are closely associated in protein synthesis, thus S requirements vary with the supply of N to crops. Therefore, when S becomes limiting for plant growth, addition of N does not increase the yield or protein concentration of plants.

The N:S ratios of many crops at their maximum yields have been assessed. Alfalfa requires 1 part of S for every 11–12 parts of N to ensure maximum production, while wheat, corn, beans, and sugarbeet leaf blades require 1 part of S for every 12–17 parts of N. The N:S ratios of grains such as oats and barley are 13:1 and 9:1, respectively.

Crop plants obtain their S requirements from a number of sources. These include: (1) soils, crop residues, and manures; (2) irrigation waters; (3) rainfall and the atmosphere; and (4) fertilizers and soil amendments. The order of importance of each of these sources varies with the type of crop, location, and management practices. The effects on S requirements are particularly marked in nonindustrialized areas where soil supplies are already low and additions from precipitation are being further reduced by shifts in energy sources (burning coal for energy production).

Functions of Sulfur in Plants

Sulfur has numerous functions in plant growth and metabolism. Among those are the following:

Table 1 Sulfur content of crops

Crop	Yield (tons ha^{-1})	Total S content (kg ha^{-1})
Grain and oil crops		
Barley	5.4	22
Corn	11.2	34
Grain sorghum	9.0	43
Oats	3.6	22
Rice	7.8	13
Wheat	5.4	22
Peanuts	4.5	24
Soybeans	4.0	28
Hay crops		
Alfalfa	17.9	45
Clover grass	13.4	34
Bermuda grass		
Common	9.0	17
Coastal	22.4	50
Brome grass	11.2	22
Orchard grass	13.4	39
Pangola grass	26.4	52
Timothy grass	9.0	18
Cotton and tobacco		
Cotton (lint + seed)	4.3	34
Tobacco		
Barley	4.5	21
Flue cured	3.4	50
Fruit, sugar, and vegetable crops		
Beets		
Sugar	67	50
Table	56	46
Cabbage	78	72
Irish potatoes	56	27
Onions	67	41
Oranges	52	31
Pineapple	40	16
Sugarcane	224	96

Reproduced from Terman GL (1978) *Atmospheric Sulfur – The Agronomic Aspects.* Technical Bulletin no. 23. Washington, DC: The Sulfur Institute.

1. It is required for the synthesis of S-containing amino acids, cystine, and cysteine, and methionine, which are major components of proteins;

2. It is required for the synthesis of important metabolites, including coenzyme A, biotin, thiamine, or vitamin B_1, glutathione, sulfolipids;

3. It is a vital component of ferredoxin, a type of nonheme Fe–S protein present in chloroplasts. This is involved in photosynthesis, nitrite reduction, sulfate reduction, the assimilation of N_2 by roots of nodule bacteria and free-living N_2-fixing soil bacteria, and in mitochondrial electron transport;

4. It is an essential component (as cysteinyl, a–SH group) of the active sites of many enzymes;

5. Although it is not a component of chlorophyll, it is required for its synthesis by plants;

6. It occurs in volatile compounds responsible for the characteristic taste and smell of plants in the mustard and onion families.

Sulfur in Soils

Soil as a living system is a dynamic ecosystem supporting the life of microorganisms, plants, and animals. As is the case with carbon and nitrogen, S is continuously being cycled between organic and inorganic forms, resulting in synthesis and degradation of a variety of S compounds. The reactions and the nature of the S compounds formed depend on the environmental conditions of the soil. These include aeration, water content, pH, presence and absence of metal ions, tillage systems applied, crop residues incorporated, liming, type of crops, and crop rotations used.

The relative proportion of inorganic and organic forms of S in soils varies with the soil type, depth of sampling, cropping system, season, and soil conditions (field-moist or air-dried). No direct chemical method is available for determination of total organic S in soils, and this is normally determined from the difference between total S and inorganic S determined separately. The inorganic S fraction is small, ranging from 2 to 6% of the total S in soils from the humid and semihumid regions. This is normally present as inorganic SO_4^{2-} in well-aerated soils. In calcareous soils, much of this fraction is occluded by precipitate of calcium carbonate, leading to a smaller fraction of the total S in organic combinations. Similarly, because of lack of percolating water, inorganic SO_4^{2-} accumulates in surface layers of arid soils in the form of gypsum, leading to a small fraction of organic S in such soils. Organic S concentration decreases with depth in the soil profile, and this decrease is associated with a decrease in organic C, except in soils where organic matter accumulates in subsurface or B horizons such as podzolic soils.

The chemical nature of soil organic matter in soils is important, because this is the reservoir that supplies SO_4^{2-} to plants. Even though complete chemical forms of S soil organic matter are not completely characterized, three broad groups of S compounds are recognized. These are:

1. Hydriodic acid-reducible S. This fraction consists of organic S that is reduced to H_2S upon boiling with hydriodic acid (HI) under N_2 gas. This form of S is bonded to a carbon atom through an oxygen atom (i.e., ester sulfate, C–O–S). Examples of such forms of S are phenolic sulfate, sulfated polysaccharides, choline sulfate, and sulfated lipids. On average, approximately 50% of the organic S in soils is in this form; it can range from 30 to 60%, and values as high as 95% have been reported for some Iowa subsurface

soils. It is mineralized to SO_4^{2-} through the enzyme arylsulfatase, as follows:

$$R\text{–}O\text{–}SO_3^- + H_2O \xrightarrow{\text{Arylsulfatase}} R\text{–}OH + HSO_4^- \quad [1]$$

2. Carbon-bonded S. This fraction consists of S atoms directly bonded with carbon atoms (i.e., C–S). This form of S is hydrolyzed by Raney Ni (50% each of Al and Ni powder) under alkaline conditions (NaOH). When treated with HCl, this mixture releases the S atoms in the form of H_2S. Sulfur-containing amino acids, cystine and methionine, are the main components of this fraction, which accounts for approximately 20–30% of the soil organic S. Other S-containing compounds such as sulfoxide, sulfones, sulfenic, sulfinic, and sulfonic acids, and S-containing heterocyclic compounds are also constituents of this fraction;

3. Inert or residual S. This fraction is the organic S that is not reduced by either of the reagents described above. This fraction represents approximately 30–40% of the total organic S in soils. Because this fraction is resistant to drastic chemical treatment, it is of little importance as a source of S for plants.

Sources of Mineralizable Sulfur in soils

It is believed that both HI-reducible and carbon-bonded S fractions of soil organic S are sources of SO_4^{2-} for plant uptake.

Organic Nitrogen and Sulfur Relationship in Soils

The presence of large concentrations of organic S in surface soils is indicated by the close relationship between organic N and S. In many of the soils studied in North America and other parts of the world, the mean ratio of organic N:S is approximately 8:1. But, this ratio may range from 5:1 to 13:1 (Table 2). The ratio appears to be the same for virgin soils and their cultivated counterparts (Table 3), suggesting that the organic forms of these elements mineralize in about the same ratio as they occur in soil organic matter. However, results of some studies indicate that organic S in soils is depleted faster than that of organic N.

As is true of N, when land is first cultivated, the S content of the soil declines rapidly and an equilibrium level is reached which is influenced by climate, cultural practices, and soil type. At the equilibrium level, soil organic matter essentially ceases to act as a source of S for plant growth. Before reaching this equilibrium, however, the rate of S mineralization is so slow that it cannot cope with the plant's need for this element. This results in the appearance of S-deficiency symptoms on the plants. Because S is a mobile element in plants, the S-deficiency symptoms appear on the older leaves. Recent studies indicate that lack of S to meet plant's requirements may not only reduce yields, but it can reduce quality (reduction of the amino acids cysteine and methionine in protein) in grain legumes by changing gene expression of storage proteins in developing seeds.

Table 2 Mean ratios of nitrogen to sulfur in surface horizons of soils

Location	*Description*[a]	*N:S ratio*
Virgin soils		
Alberta, Canada	Brown Chernozems (4)	7.1
	Black Brown Chernozems (4)	7.7
	Gray Wooded (7)	12.5
	Gleysols (6)	5.0
USA (several states)	Not specified (10)	8.7
Cultivated soils		
NSW, Australia	Pasture soils (5)	8.5
	Clover pasture, podzolic (44)	7.1
Eastern Australia	Acid soils (128)	8.3
	Alkaline soils (27)	6.6
Sao Paulo and Parana, Brazil	Agricultural, varied (6)	7.7
Canterbury, New Zealand	Grassland, unfertilized	7.7
North Scotland	Agricultural, noncalcareous (40)	7.1
USA		
Iowa	Agricultural, varied (37)	6.5
Iowa	Agricultural, varied (6)	7.7
Minnesota	Brown Chernozems (6)	6.4
	Black prairie soils (9)	6.1
	Podzols (24)	8.5
Mississippi	Podzols (4)	13.1
Oregon	Agricultural, varied (16)	9.9
Several states	Agricultural, varied (10)	8.0

[a]Numbers in parentheses are number of soils examined.
Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Sulfur Availability Indexes

It is generally accepted that S is taken up by plants in the SO_4^{2-} form. Studies involving nutrient culture experiments have shown that low-molecular-weight organic-S compounds are utilized by plant. But, the availability of soil organic S depends primarily on its mineralization, which, in turn, depends on climatic factors such as temperature and moisture, and on the chemical nature of organic S present in soils.

Assessment of the plant-available S in soils is complicated by the fact that, in addition to soils, several other sources contribute to the S needs of plants. These sources include S in rainfall and irrigation

Table 3 Organic nitrogen and sulfur ratios in virgin soils (V) and their cultivated (C) counterparts

Location		*N:S ratio*
Big Springs, TX	V	12.0
	C	11.1
Colby, KS	V	6.5
	C	6.8
Mays, KS	V	8.0
	C	6.8
Moccasin, MT	V	7.6
	C	6.0
Dalbart, TX	V	9.6
	C	11.7
Madan, ND	V	9.0
	C	9.6
North Platte, NE	V	9.9
	C	9.0
Lawton, OK	V	10.3
	C	10.4
Archer, WY	V	8.5
	C	8.0
Havre, MT	V	9.1
	C	8.2
Mean		8.9

Reproduced from Stewart BA and Whitefield CJ (1965) Effects of crop residue, soil temperature, and sulfur on the growth of winter wheat. *Soil Science Society of America Proceedings* 29: 752–755.

water, S in the atmosphere, and S in fertilizers and pesticides. Of the sources of S, the contribution of S in rainfall and direct absorption by soils and plants from the atmosphere are the most difficult to evaluate. In addition, both plants and soils absorb SO_2, and most likely other S gases, directly from the atmosphere. For example, studies in Wisconsin in the early 1970s, by using radioactive ^{35}S, have shown that under optimal yield 14% of the S in alfalfa (*Medicago sativa* L.) is derived from atmospheric sources.

There is no general agreement on the best methods for using an index of plant availability. The procedures used fall into one of the following two groups: (1) plant analyses and (2) soil analyses. The use of plant analysis for assessing S nutrition of plants is based on the notion that an essential element should be present in the plant as a concentration just sufficient for unrestricted plant growth.

Numerous methods and procedures have been proposed for evaluation of the plant-available S in soils. These methods include extraction with water, extraction with various salts or acid solutions, S mineralization during incubation, microbial growth, and plant growth and composition. The concentration of S removed by the various extractants normally falls in one of the following groups: (1) readily soluble sulfate, (2) readily soluble and some of the adsorbed sulfate, and (3) readily soluble and some of the adsorbed sulfate and some of the organic S.

Incubation procedures for assessing plant-available S paralleling those used in the estimation of available N have met with little success, because precise determination of the small amount of inorganic SO_4^{2-} released during incubation is difficult. In addition, the amount of S mineralized during incubation is affected by the presence or absence of plants.

Sulfur Requirement of Plants

The requirement of S or any other nutrient by crops is defined as "the minimum content of that nutrient associated with the maximum yield" or "the minimum rate of the uptake of the nutrient associated with maximum growth rate." The first definition refers to that the total amount of the element in the crops (normally expressed in kilograms per hectare) or the concentration of the element in the plant or plant part. The second definition is related to the minimum concentration of the element taken up from the soil or nutrient solution that is associated with the maximum growth. Information on the S content of crops is useful in estimating the S-fertilizer requirement.

Both the uptake and requirements for S differ greatly among plant species, among cultivars within species, and with the stage of development of the crop. Assessment of the S requirement of crops is more complicated than that of any other nutrient. That is because: (1) there are several S sources for plants, (2) each source has a different efficiency for its utilization by crops, (3) S has limited reuse within the crop, and (4) relatively large concentrations of S can be accumulated within the crop. The supply of S during the growing period can be changed by the atmospheric conditions and precipitation, impurities in fertilizers, addition of pesticides, and mineralization of organic S in soil organic matter. In spite of the difficulties associated with assessment of S status on crops, several approaches have successfully been used for this purpose. These include: (1) chemical soil tests, (2) a number of plant-tissue chemical analyses, and (3) crop-deficiency symptoms. The first of these approaches is the most commonly used.

Sulfur Metabolism in Plants

S metabolism in plants is very complex. The significant steps are as follows: (1) sulfate is 'activated,' that is, enzymatically converted to adenosine-5′-phosphosulfate (APS) and 3′-phosphoadenosine-5′-phosphatsulfate (PAPS). In plants, APS functions as substrate for SO_4^{2-} reduction and as a precursor of PAPS, while PAPS acts as the SO_4^{2-} donor in the formation sulfate esters (organic sulfates); (2) the

activated sulfate is reduced; (3) reduced S is incorporated into cysteine; (4) the cysteine-S is transferred into methionine and other essential compounds; (5) methionine is transferred into S-adenosylmethionine, which is a methyl ($-CH_3$) donor and a precursor of important non-S-containing compounds; and (6) cysteine and methionine are incorporated into proteins. Sulfate transport in plants is inhibited by structurally related compounds or ions such as SO_3^{2-}, $S_2O_3^{2-}$, $S_2O_5^{2-}$, SeO_4^{2-}, CrO_4^{2-}, MoO_4^{2-}, and WO_4^{2-}. Studies have shown that SeO_4^{2-} is a competitive inhibitor of SO_4^{2-} transport in plants.

Sulfur-Containing Materials Added to Soils

A number of materials added to soils contain S. These range from animal manures to sewage sludge, crop residues, and irrigation water. The S concentration in waste materials and irrigation water varies with the sources and quality of the material. Therefore, knowledge of the composition of such materials is needed before its application or use on soils.

Fertilizers

Many sources of S-containing fertilizer are available, most of which are old and established sources of S. Their behavior in soils is determined by the chemical nature of the S atom and the reactions involved when added to soils. The following are examples of dry fertilizer containing sulfur:

1. Ammonium sulfate. This compound is widely used as a source of N, but $(NH_4)_2SO_4$ contains 21% N in the ammoniacal form and approximately 24% of S in the SO_4^{2-} form;
2. Single superphosphate. This compound has been manufactured by the fertilizer industry since 1850. It contains approximately 8% of P. It is produced by reacting sulfuric acid (H_2SO_4) with phosphate rock. The reaction is:

$$3[Ca_3(PO_4)_2] \cdot CaF_2 + 7H_2SO_4 = 3Ca(H_2PO_4)_2 + 7CaSO_4 + 2HF \quad [2]$$

3. Ammonium phosphate sulfate. This is essentially a mixture of monoammonium phosphate and ammonium sulfate. The most common of these products contains 16% N, 8% P, and 13% S;
4. Gypsum. Large quantities of gypsum ($CaSO_4 \cdot 2H_2O$) are produced as a by-product of manufacturing concentrated phosphate fertilizer. It is a low-analysis fertilizer (13–14% S in the by-product forms and 18.6% S in the pure form);
5. Elemental S. When added to soils, elemental S is converted to SO_4^{2-}. The reactions involved are affected by particle size (the finer the particles, the faster the reaction rate), temperature, moisture and aeration, soil pH, soil type and properties, fertilizer interaction, characteristics of the elemental S, rate and placement, and time of application. The reactions involved are:

$$S^0 \longrightarrow S_2O_3^{2-} \longrightarrow S_2O_6^{2-} \longrightarrow SO_3^{2-} \longrightarrow SO_4^{2-} \quad [3]$$

6. Sulfur-bentonite. This product consists of 90% of elemental S and 10% of bentonite. When this material is added to soils, the bentonite imbibes moisture, causing the granules to disintegrate, releasing very fine S particles, which are rapidly converted to SO_4^{2-};
7. Elemental sulfur suspension. This material is made of finely ground elemental S (40–60% of S^0) and a small amount of attapulgite clay. It can be applied directly to soils or mixed with other fertilizer materials;
8. Phosphate-elemental sulfur materials. This material is a mixture of elemental S and mono- or diammonium phosphate. It is a source of S and P for plants;
9. Other dry S-fertilizer mixtures are available; these include urea-ammonium sulfate, ammonium nitrate–ammonium sulfate, and potassium sulfate mixtures;
10. Urea–S. This material contains 20% of S and 36% of N. It has excellent physical properties and easy to handle.

Fluid fertilizers containing sulfur include:

1. Ammonium thiosulfate. This compound is a clear solution containing 26% of S and 12% of N. It is the most commonly used S fertilizer. This material can be used for direct application to soils and in irrigation waters;
2. Ammonium-polysulfide. Polysulfides are used as fertilizers, as soil conditioners, and for treatment of irrigation waters to improve percolation into soils. Ammonium polysulfides are used for all these purposes. When added to soil, the polysulfide is changed to colloidal S and sulfide. Calcium polysulfide is marketed as soil conditioner and for treatment of irrigation water. The reaction involved is:

$$CaS_6 + H_2O + CO_2 = H_2S + CaCO_3 + 4S^0 \quad [4]$$

3. Ammonium bisulfite. This is a clear liquid, low-analysis fertilizer, containing 17% of S and 8.5% of N. It has a strong odor of SO_2 and is marketed to a very limited extent;

4. Urea-sulfuric acid mixtures. Two grades of this material are available on the market. One grade contains 10% of N and 18% of S, and the other contains 28% of N and 9% of S. Because this material is highly acidic (pH between 0.5 and 1.0), it is essential that all application equipment is made from stainless steel and extreme precautions must be taken in handling the material.

In addition to the S-containing materials described above, aluminum sulfate is added to soil for soil acidification. Acid soil is required for plants such as azaleas and camellias. When added to soils, the reaction is:

$$Al_2(SO_4)_3 + 6H_2O = 2Al(OH)_3 + 3H_2SO_4 \quad [5]$$

When using any of the above materials, the user should adhere to the recommendations provided by the manufacturer in terms of safety, rate of application, method and time of application, and storage conditions.

Sulfur in the Atmosphere

The atmosphere is an important compartment of the S cycle for agricultural crops and forest ecosystems. Sulfur gases and aerosols are found in the atmosphere of urban, rural, and remote areas. The sulfur species in the atmosphere range from oxidized forms such as sulfur dioxide (SO_2) and sulfate (SO_4^{2-}) to reduced forms such as hydrogen sulfide (H_2S), carbonyl sulfide (COS), dimethyl sulfide ($(CH_3)_2S$, and carbon disulfide (CS_2). These are derived from natural and anthropogenic sources. The lifetimes of the S compounds in the atmosphere range from a few hours to many years and can be found in various chemical forms from the troposphere to the stratosphere. The relationship between atmospheric S and the terrestrial ecosystem is very complex. To understand the interactions, the following information is needed: (1) the rate and chemical forms of S entering the atmosphere, (2) the chemical transformations occurring in the atmosphere, (3) the transport from the emission sources to the receptor (plant), (4) the amount and rate of transfer to soil and plants, and (5) the impact on crop production. The natural sources of S in the atmosphere include: (1) volcanic activity, (2) input from ocean spray, (3) bacterial decomposition of plant material in soils, and (4) animal manure and sewage sludge. Man-made sources include: (1) combustion of S-containing fossil duels (coal, oil, and gas), and (2) reduced S gases released from many industrial processes (e.g., pulp mills).

Atmospheric S gases are absorbed by plants, surface water, and surface soils or deposited in precipitation in the SO_4^{2-} form which, in turn, is absorbed by plants, retained in soils, or leached through groundwater.

Sulfur in Precipitation

Sulfur in the atmosphere occurs in gaseous, solid, and liquid form, with transformation among the different forms accomplished by chemical, biological, and photometric processes. As is the case with N, S is highly transitory among these forms in air and water. The reactions involved in these transformations determine the concentration of S in precipitation, which, in turn, affects its presence in surface and groundwater, with ultimate impact on soils and plants.

The data available on S content of rainfall indicate wide variation by geographic locations. Expressed in kilograms of sulfur per hectare, the annual addition of S in precipitation in North American ranges from 0.5 to 14. In many areas, most of this (approximately 60–80%) additional S is deposited during the crops' growing season. Laboratory and greenhouse experiments have shown that many soils of the USA do not contain sufficient plant-available S to meet the crops' requirements, yet no S-deficiency symptoms have been reported. This is because, in many areas, uptake of atmospheric SO_2 and of that in precipitation can compensate for the soil S deficiency. Indeed, it has been demonstrated that sunflowers can use SO_2 and H_2S as their only source of S without their normal growth being affected.

See also: **Acid Rain and Soil Acidification**; **Enzymes in Soils**; **Fertility**; **Leaching Processes**; **Minerals, Primary**; **Organic Matter:** Principles and Processes; **Organic Residues, Decomposition**; **Soil–Plant–Atmosphere Continuum**; **Sulfur in Soils:** Overview; Biological Transformations

Further Reading

Freney JR (1967) Sulfur-containing organics. In: McLaren AD and Peterson GH (eds) *Soil Biochemistry*, vol. 1, pp. 229–259. New York: Marcel Dekker.

Germinda JJ and Gupta VVSR (1992) Biochemistry of sulfur cycling in soil. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 1–53. New York: Marcel Dekker.

McLachlan KD (ed.) (1975) *Sulphur in Australasian Agriculture*. Sydney, Australia: Sydney University Press.

Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.

Stewart BA and Whitefield CJ (1965) Effects of crop residue, soil temperature, and sulfur on the growth of winter wheat. *Soil Science Society of America Proceedings* 29: 752–755.

Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Tabatabai MA (1985) Effect of acid rain on soils. *CRC Critical Reviews in Environmental Control* 15: 65–110.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Agronomy monograph no. 27. Madison, WI: American Society of Agronomy.

Tabatabai MA (1987) Physicochemical fate of sulfate in soils. *Journal of the Air Pollution Control Association* 37: 34–38.

Tabatabai MA (1994) Soil enzymes. In: Weaver RW, Angle JS, and Bottomley PS (eds) *Methods of Soil Analysis*, part 2, pp. 775–783. *Microbiological and Biochemical Properties*. Soil Science Society of America Book Series No. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA (1996) Sulfur. In: Sparks DL (ed.) *Methods of Soil Analysis*, part 3, pp. 921–960. *Chemical Methods*. Soil Science Society of America Book Series No. 5, Madison, WI: Soil Science Society of America.

Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.

Terman GL (1978) *Atmospheric Sulfur – The Agronomic Aspects*. Technical Bulletin no. 23. Washington, DC: The Sulphur Institute.

Tisdale SL, Nelson WL, and Beaton JD (1985) *Soil Fertility and Fertilizers*, 4th edn. New York: Macmillan.

SURFACE COMPLEXATION MODELING

S Goldberg, USDA–ARS, Riverside, CA, USA

Published by Elsevier Ltd.

Introduction

Adsorption is the process through which ions are removed from solution and accumulate at a solid surface. The ion accumulation takes place at the interface between the surface and the solution forming a two-dimensional structure. If adsorption continues and leads to a three-dimensional structure, the process is called precipitation. The general loss of ions from solution to a surface is called sorption.

Adsorption can occur either specifically or nonspecifically. Specific adsorption occurs when ions have a high affinity for the surface and it results in the formation of inner-sphere surface complexes. Inner-sphere surface complexes are complexes that contain no water molecules between the adsorbing ion and the surface functional group. Examples of surface functional groups are reactive hydroxyl groups on oxide or clay minerals and carboxyl or phenol groups on organic matter. Such surface functional groups are a source of solid surface charge since they undergo dissociation and/or protonation reactions as a result of changes in solution pH. Specific anion adsorption occurs via ligand exchange where the adsorbing ion replaces a reactive surface hydroxyl from the surface functional group. Nonspecific adsorption is dominated by electrostatic attraction and results in outer-sphere complex formation or in adsorption in the diffuse ion swarm. Adsorption in the diffuse ion swarm is the weakest type of adsorption since the ion does not attach to a specific surface functional group but remains free in the aqueous solution, neutralizing surface charge only by its proximity to the charged solid surface. Outer-sphere surface complexes are also formed through nonspecific adsorption and contain at least one water molecule between the adsorbing ion and the surface functional group.

A model is a simplified representation of reality that considers only those characteristics of the system that are pertinent to the problem at hand. A chemical model provides a description of a chemical system consistent with its chemical properties while simultaneously being as simple and as chemically correct as possible. The ideal chemical model is realistic, effective, comprehensive, and predictive. A realistic model conforms to accepted theories of chemical behavior, an effective model closely describes experimental observations, a comprehensive model applies to a wide range of experimental conditions without modification, and a predictive model can be applied to various different chemical conditions.

Description of Models

Surface complexation models are chemical models that give a molecular description of adsorption phenomena using an equilibrium approach. Analogous to complex formation in solution, surface complexation models define surface species, chemical reactions, equilibrium constants, mass balances, and charge balances and their molecular features can be given thermodynamic significance. One of the major

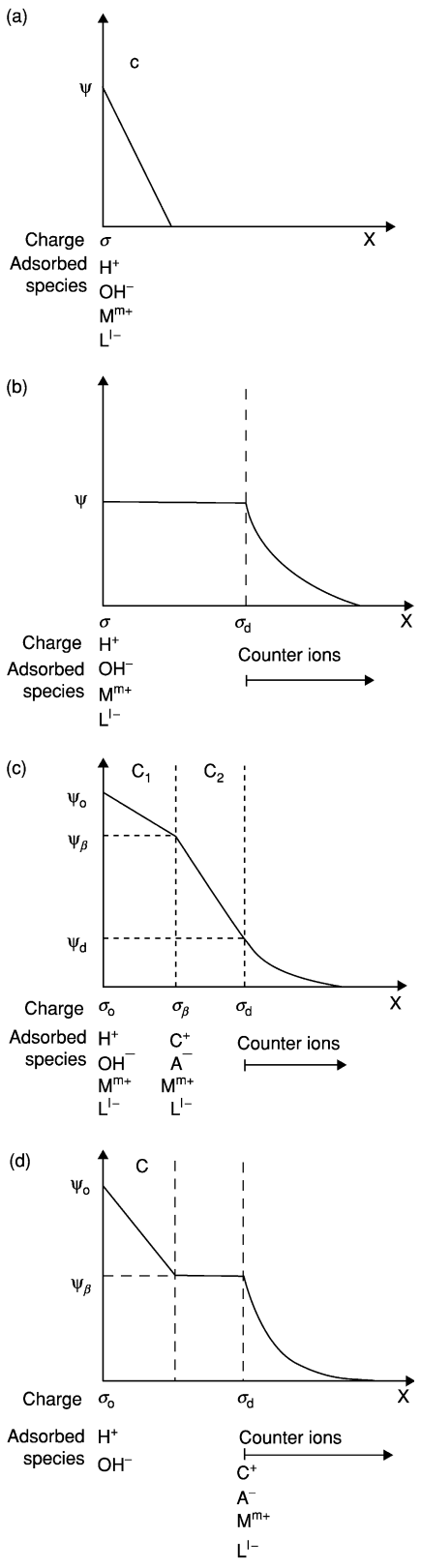


Figure 1 Placement of ions, potentials, charges, and capacitances for the (a) constant capacitance model, (after Westall (1986); (b) diffuse layer model, (after Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*. New York: John Wiley); (c) triple layer model, after Westall (1980); (d) one-pK model, (after Westall (1986)). (a) and (d) from Davis JA and Hayes K (eds) *Geochemical Processes at Mineral Surfaces*, ACS symposium Series 323: 54–78, Copyright (1986) American Chemical Society; (b) reprinted with permission of John Wiley from Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*; (c) reprinted with permission from Kavanaugh MC and Leckie JO (eds) *Particulates in Water Characterization, Fate, Effects and Removal*. ACS Advances in Chemistry Series 189: 33–44, Copyright (1980) American Chemical Society.

advancements of surface complexation models is consideration of the charge on both the adsorbing ion and the solid adsorbent surface. Surface complexation models constitute a family of models having many common characteristics and adjustable parameters. The models differ in their structural representation of the solid–solution interface, i.e., the location and surface configuration of the adsorbed ions.

Surface Configuration of the Solid–Solution Interface

Members of the surface complexation model family include the two-pK models: constant capacitance, diffuse-layer, triple-layer. Two-pK models are based on a reactive surface functional group, SOH, that undergoes both protonation and dissociation:

$$\mathrm{SOH} + \mathrm{H^+} \leftrightarrow \mathrm{SOH_2^+} \qquad [1]$$

$$\mathrm{SOH} \leftrightarrow \mathrm{SO^-} + \mathrm{H^+} \qquad [2]$$

hence the term two-pK model.

Comparable models can be written based on the one-pK concept. So far, the one-pK model has been developed based on the Stern model. In the one-pK model, surface functional groups carry either one or two protons, SOH and $\mathrm{SOH_2}$, respectively. Surface charging can be represented with one reaction:

$$\mathrm{SOH^{1/2-}} + \mathrm{H^+} \leftrightarrow \mathrm{SOH_2^{1/2+}} \qquad [3]$$

The location and surface configuration of the adsorbed ions for the various surface complexation models are presented in Figure 1. In the constant capacitance model and the diffuse layer model all surface complexes are inner-sphere and are located in a single surface plane (Figure 1a, b). The diffuse-layer model includes a diffuse layer that commences at the d-plane and extends into solution. In the triple-layer model ions forming inner-sphere surface complexes adsorb in the surface o-plane and ions forming outer-sphere surface complexes adsorb in a β-plane located between the o-plane and the d-plane (Figure 1c). In the representation of the one-pK model based on the Stern model indicated in Figure 1d, protons and hydroxyls form inner-sphere surface complexes

located in the o-plane; all other ions form outer-sphere surface complexes and are located in the d-plane.

Surface Complexation Reactions

In the two-pK models chemical reactions for inner-sphere surface complexation are eqns [1], [2], and:

$$\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{SOM^{(m-1)}} + \mathrm{H^+} \quad [4]$$

$$2\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{(SO)_2M^{(m-2)}} + 2\mathrm{H^+} \quad [5]$$

$$\mathrm{SOH} + \mathrm{L^{l-}} + \mathrm{H^+} \leftrightarrow \mathrm{SL^{(l-1)-}} + \mathrm{H_2O} \quad [6]$$

$$2\mathrm{SOH} + \mathrm{L^{l-}} + 2\mathrm{H^+} \leftrightarrow \mathrm{S_2L^{(l-2)-}} + 2\mathrm{H_2O} \quad [7]$$

$$\mathrm{SOH} + \mathrm{L^{l-}} + 2\mathrm{H^+} \leftrightarrow \mathrm{SHL^{(l-2)-}} + \mathrm{H_2O} \quad [8]$$

where M is a metal ion of charge m^+ and L is a ligand of charge l^-. Surface complexes can be monodentate or bidentate. In monodentate complexation one bond is formed between the adsorbing ion and the surface functional group. Bidentate complexes contain two bonds between the adsorbing ion and two surface functional groups. Equations [4] to [7] are used in the constant capacitance model. In the diffuse-layer model, reaction [8] is used in place of reaction [7] since bidentate complexes for adsorbed anions have not been considered.

The following chemical reactions for outer-sphere surface complexation are considered in the triple-layer model in addition to eqns [1], [2], and [4] to [7]:

$$\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{SO^-} - \mathrm{M^{m+}} + \mathrm{H^+} \quad [9]$$

$$\mathrm{SOH} + \mathrm{M^+} + \mathrm{H_2O} \leftrightarrow \mathrm{SO^-} - \mathrm{MOH^{(m-1)}} + 2\mathrm{H^+} \quad [10]$$

$$\mathrm{SOH} + \mathrm{H^+} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{L^{l-}} \quad [11]$$

$$\mathrm{SOH} + 2\mathrm{H^+} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{LH^{(l-1)-}} \quad [12]$$

$$\mathrm{SOH} + \mathrm{C^+} \leftrightarrow \mathrm{SO^-} - \mathrm{C^+} + \mathrm{H^+} \quad [13]$$

$$\mathrm{SOH} + \mathrm{H^+} + \mathrm{A^-} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{A^-} \quad [14]$$

where C^+ is the cation and A^- is the anion of the background electrolyte.

In the one-pK models, chemical reactions for surface complexation are eqn [3] and:

$$\mathrm{SOH^{1/2-}} + \mathrm{M^+} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{M^{m+}} \quad [15]$$

$$\mathrm{SOH^{1/2-}} + \mathrm{M^{m+}} + \mathrm{H_2O} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{MOH^{(m-1)}} + \mathrm{H^+} \quad [16]$$

$$\mathrm{SOH_2^{1/2+}} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^{1/2+}} - \mathrm{L^{l-}} \quad [17]$$

$$\mathrm{SOH^{1/2-}} + \mathrm{C^+} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{C^+} \quad [18]$$

$$\mathrm{SOH_2^{1/2+}} + \mathrm{A^-} \leftrightarrow \mathrm{SOH_2^{1/2+}} - \mathrm{A^-} \quad [19]$$

Equilibrium Constants for Surface Complexation

The equilibrium constants describing inner-sphere surface complexation in the two-pK models are:

$$K_+ = \frac{[\mathrm{SOH_2^+}]}{[\mathrm{SOH}][\mathrm{H^+}]}\exp[\mathrm{F}\Psi/RT] \quad [20]$$

$$K_- = \frac{[\mathrm{SO^-}][\mathrm{H^+}]}{[\mathrm{SOH}]}\exp[-\mathrm{F}\Psi/RT] \quad [21]$$

$$K_\mathrm{M}^1 = \frac{[\mathrm{SOM^{(m-1)}}][\mathrm{H^+}]}{[\mathrm{SOH}][\mathrm{M^{m+}}]}\exp[(\mathrm{m}-1)\mathrm{F}\Psi/RT] \quad [22]$$

$$K_\mathrm{M}^2 = \frac{[\mathrm{(SO)_2M^{(m-2)}}][\mathrm{H^+}]^2}{[\mathrm{SOH}]^2[\mathrm{M^{m+}}]}\exp[(\mathrm{m}-2)\mathrm{F}\Psi/RT] \quad [23]$$

$$K_\mathrm{L}^1 = \frac{[\mathrm{SL^{(l-1)-}}]}{[\mathrm{SOH}][\mathrm{L^{l-}}][\mathrm{H^+}]}\exp[-(\mathrm{l}-1)\mathrm{F}\Psi/RT] \quad [24]$$

$$K_\mathrm{L}^2 = \frac{[\mathrm{S_2L^{(l-2)-}}]}{[\mathrm{SOH}]^2[\mathrm{L^{l-}}][\mathrm{H^+}]^2}\exp[-(\mathrm{l}-2)\mathrm{F}\Psi/RT] \quad [25]$$

$$K_\mathrm{L}^3 = \frac{[\mathrm{SHL^{(l-2)-}}]}{[\mathrm{SOH}][\mathrm{L^{l-}}][\mathrm{H^+}]^2}\exp[-(\mathrm{l}-2)\mathrm{F}\Psi/RT] \quad [26]$$

where F is the Faraday constant, Ψ is the surface potential, R is the molar gas constant, T is the absolute temperature, and square brackets represent concentrations. The exponential terms are correction factors accounting for the effect of surface charge on surface complexation. Equations [20] to [25] are considered in the constant capacitance model where $\Psi = \Psi_o$. In the diffuse layer model, eqn [25] is replaced with eqn [26] and $\Psi = \Psi_d$.

Equilibrium constants for outer-sphere surface complexation in the triple layer model are:

$$K_{\mathrm{M}}^{3} = \frac{[\mathrm{SO^{-} - M^{m+}}][\mathrm{H^{+}}]}{[\mathrm{SOH}][\mathrm{M^{m+}}]} \exp[\mathrm{F}(m\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [27]$$

$$K_{\mathrm{M}}^{4} = \frac{[\mathrm{SO^{-} - MOH^{(m-1)}}][\mathrm{H^{+}}]^{2}}{[\mathrm{SOH}][\mathrm{M^{m+}}]} \exp[\mathrm{F}((m-1)\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [28]$$

$$K_{\mathrm{L}}^{4} = \frac{[\mathrm{SOH_2^{+} - L^{l-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}][\mathrm{L^{l-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - l\Psi_{\beta})/RT] \quad [29]$$

$$K_{\mathrm{L}}^{5} = \frac{[\mathrm{SOH_2^{+} - LH^{(l-1)-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}]^{2}[\mathrm{L^{l-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - (l-1)\Psi_{\beta})/RT] \quad [30]$$

$$K_{\mathrm{C}} = \frac{[\mathrm{SO^{-} - C^{+}}][\mathrm{H^{+}}]}{[\mathrm{SOH}][\mathrm{C^{+}}]} \exp[\mathrm{F}(\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [31]$$

$$K_{\mathrm{A}} = \frac{[\mathrm{SOH_2^{+} - A^{-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}][\mathrm{A^{-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - \Psi_{\beta})/RT] \quad [32]$$

In the one-pK model the equilibrium constants for surface complexation are:

$$K_{\mathrm{H}} = \frac{[\mathrm{SOH_2^{1/2+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{H^{+}}]} \exp[\mathrm{F}\Psi_{\mathrm{o}}/RT] \quad [33]$$

$$K_{\mathrm{M}}^{1} = \frac{[\mathrm{SOH^{1/2-} - M^{m+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{M^{m+}}]} \exp[\mathrm{mF}\Psi_{\mathrm{d}}/RT] \quad [34]$$

$$K_{\mathrm{M}}^{2} = \frac{[\mathrm{SOH^{1/2-} - MOH^{(m-1)}}][\mathrm{H^{+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{M^{m+}}]} \exp[(m-1)\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [35]$$

$$K_{\mathrm{L}} = \frac{[\mathrm{SOH_2^{1/2+} - L^{l-}}]}{[\mathrm{SOH_2^{1/2+}}][\mathrm{L^{l-}}]} \exp[-l\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [36]$$

$$K_{\mathrm{C}} = \frac{[\mathrm{SOH^{1/2-} - C^{+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{C^{+}}]} \exp[\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [37]$$

$$K_{\mathrm{A}} = \frac{[\mathrm{SOH_2^{1/2+} - A^{-}}]}{[\mathrm{SOH_2^{1/2+}}][\mathrm{A^{-}}]} \exp[-\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [38]$$

Mass and Charge Balances

The mass balance expression for the surface functional group, SOH, in the two-pK models is:

$$\begin{aligned} S_{\mathrm{T}} = {} & [\mathrm{SOH}] + [\mathrm{SOH_2^{+}}] + [\mathrm{SO^{-}}] + [\mathrm{SOM^{(m-1)}}] \\ & + 2[(\mathrm{SO})_2\mathrm{M^{(m-2)}}] + [\mathrm{SL^{(l-1)-}}] + 2[\mathrm{S_2L^{(l-2)-}}] \\ & + [\mathrm{SHL^{(l-2)-}}] + [\mathrm{SO^{-} - M^{m+}}] + [\mathrm{SO^{-}} \\ & - \mathrm{MOH^{(m-1)}}] + [\mathrm{SOH_2^{+} - L^{l-}}] + [\mathrm{SOH_2^{+}} \\ & - \mathrm{LH^{(l-1)-}}] + [\mathrm{SO^{-} - C^{+}}] + [\mathrm{SOH_2^{+} - A^{-}}] \end{aligned} \quad [39]$$

The mass balance for the surface functional groups, SOH and SOH_2, in the one-pK model is:

$$\begin{aligned} S_{\mathrm{T}} = {} & [\mathrm{SOH^{1/2-}}] + [\mathrm{SOH_2^{1/2+}}] + [\mathrm{SOH^{1/2-} - M^{m+}}] \\ & + [\mathrm{SOH^{1/2-} - MOH^{(m-1)}}] + [\mathrm{SOH_2^{1/2+} - L^{l-}}] \\ & + [\mathrm{SOH^{1/2-} - C^{+}}] + [\mathrm{SOH_2^{1/2+} - A^{-}}] \end{aligned} \quad [40]$$

The mass balance represents a summation of all surface species considered in the particular surface complexation model.

The charge balance expressions for the two-pK models are:

$$\begin{aligned} \sigma_{\mathrm{o}} = {} & [\mathrm{SOH_2^{+}}] + [\mathrm{SOH_2^{+} - L^{l-}}] + [\mathrm{SOH_2^{+} - LH^{(l-1)-}}] \\ & + (m-1)[\mathrm{SOM^{(m-1)}}] + (m-2)[(\mathrm{SO})_2\mathrm{M^{(m-2)}}] \\ & + [\mathrm{SOH_2^{+} - A^{-}}] - [\mathrm{SO^{-}}] - [\mathrm{SO^{-} - M^{m+}}] \\ & - [\mathrm{SO^{-} - MOH^{(m-1)}}] - (l-1)[\mathrm{SL^{(l-1)-}}] \\ & - (l-2)[\mathrm{S_2L^{(l-2)-}}] - (l-2)[\mathrm{SHL^{(l-2)-}}] \\ & - [\mathrm{SO^{-} - C^{+}}] \end{aligned} \quad [41]$$

$$\begin{aligned} \sigma_{\beta} = {} & m[\mathrm{SO^{-} - M^{m+}}] + (m-1)[\mathrm{SO^{-} - MOH^{(m-1)}}] \\ & + [\mathrm{SO^{-} - C^{+}}] - l[\mathrm{SOH_2^{+} - L^{l-}}] - (l-1) \\ & [\mathrm{SOH_2^{+} - LH^{(l-1)-}}] - [\mathrm{SOH_2^{+} - A^{-}}] \end{aligned} \quad [42]$$

$$\sigma_{\mathrm{o}} + \sigma_{\beta} + \sigma_{\mathrm{d}} = 0 \quad [43a]$$

$$\sigma_{\mathrm{o}} + \sigma_{\mathrm{d}} = 0 \quad [43b]$$

where σ is the surface charge. The charge balances, eqns [41] and [42], represent the summation of all charge contributions in a particular plane of adsorption. All of the models consider charge balance in the surface plane, eqn [41]. Charge balance in the β-plane, eqn [42], is restricted to the triple-layer model. Charge balance eqn [43a] is considered in the triple-layer model while eqn [43b] is used in the diffuse-layer model. The charge balance expressions for the one-pK model are eqn [43b] and:

$$\sigma_o = \frac{1}{2}([SOH_2^{1/2+}] - [SOH^{1/2-}]) \quad [44]$$

Charge-Potential Relationships

All surface complexation models contain relations between surface charges and surface potentials. In the constant capacitance model the charge-potential relationship is:

$$\sigma = C\Psi \quad [45]$$

where C is the capacitance. The charge potential relationship for the diffuse layer model is:

$$\sigma_d = -(8\epsilon_o DRTI)^{1/2} \sinh(F\Psi_d/2RT) \quad [46]$$

where ϵ_o is the permittivity of vacuum, D is the dielectric constant of water, and I is the solution ionic strength. In the triple-layer model the charge potential relationships are eqn [46] and:

$$\sigma_o = C_1(\Psi_o - \Psi_\beta) \quad [47]$$

$$\sigma_d = C_2(\Psi_d - \Psi_\beta) \quad [48]$$

The charge-potential relationships for the one-pK model are eqn [46] and:

$$\sigma_o = C(\Psi_o - \Psi_d) \quad [49]$$

Obtaining Values of Adjustable Parameters

Surface Site Density

The total number of reactive surface functional groups, S_T is an important adjustable parameter in the surface complexation models and is related to the surface site density:

$$S_T = \frac{Sa10^{18}}{N_A} N_S \quad [50]$$

where S is the surface area, a is the particle concentration, and N_A is Avogadro's number. Experimental methods for determining surface site density include: tritium exchange, potentiometric titration, fluoride adsorption, and maximum adsorption. Values of this parameter can also be calculated from crystal dimensions or optimized to fit experimental adsorption data. Various determinations of surface site density vary by an order of magnitude; the lowest values are obtained from crystallographic calculations while tritium exchange yields the highest values. Uncertainty in the value of the surface site density is a major limitation in the use of surface complexation models since the ability of the models to describe adsorption is sensitively dependent on this value. To standardize surface complexation modeling, a fixed value of 2.31 sites nm^{-2} has been used for many natural materials. Applications of the diffuse-layer model to metal adsorption have split the total number of reactive surface functional groups into a 'strong,' S_s, and a 'weak,' S_w, set of adsorption sites. This approach greatly increases the number of adjustable parameters since each set of sites, S_i, has its own protonation, dissociation, and metal surface complexation constants.

Capacitances

Some values of capacitance (C in the constant capacitance and one-pK model and C_1 in the triple layer model) can be obtained graphically from slopes of protonation-dissociation constants versus surface charge. Alternatively, both capacitances, C_1 and C_2, in the triple-layer model can be obtained using an electrokinetic extrapolation technique. Capacitance values obtained experimentally usually exhibit great variability; therefore, capacitances have generally been optimized to fit the titration data.

Surface Complexation Constants

Values of the protonation and dissociation constants in the constant capacitance model and the triple-layer model can be obtained from the same graphs used to obtain values of capacitance. These constants can also be obtained by optimizing titration data using a computer program. Values of the surface complexation constants for ion adsorption are obtained using computer optimization. An advantage of computer optimization, in addition to ease of use, is that it yields bias-free parameters with standard deviations and quality-of-fit criteria. Individual optimized equilibrium constant values can be weighted to obtain overall best estimates of the parameter:

$$\overline{\log K} = \sum \frac{(1/\sigma_{\log K})_i}{\sum(1/\sigma_{\log K})_i} [\log K]_i \quad [51]$$

For the diffuse layer model a set of best estimates of $\log K$ are available for a variety of adsorbing cations and anions. The advantage of this data set is that the surface complexation constants are all self-consistent; i.e., all ion surface complexation constants were optimized using the same values of protonation-dissociation constants and surface site density. This is an important point since parameter values in the surface complexation models are interdependent. Additionally, since each surface complexation model contains a different set of assumptions for the solid–solution interface, surface complexation

constants from one model must not be used in any other model.

Applications to Ion Adsorption on Natural Samples

All surface complexation models were originally developed to describe charging behavior and ion adsorption of ions on oxide minerals. Various curves are commonly used to describe adsorption behavior: adsorption isotherms, adsorption edges, and adsorption envelopes. Adsorption isotherms describe ion adsorption as a function of equilibrium ion concentration, usually at fixed solution pH. Adsorption edges and adsorption envelopes both describe ion adsorption as a function of solution pH at a fixed total ion concentration. Adsorption edge is the term generally applied to cation adsorption, while the term adsorption envelope is used to describe anion adsorption.

The most commonly studied oxide surfaces with surface complexation models have been the iron oxides goethite and ferrihydrite. Subsequently, the models were extended to include adsorption on clay minerals, organic materials, and soil samples. In extending the models to natural samples certain approximations and modifications are necessary. In the application to natural systems, such as clay minerals or soils, the assumption is made that ion adsorption occurs through interaction with the hydroxyl groups at the edges of the clay particles. The effect of permanent negatively charged sites at the clay basal planes on this adsorption process is ignored. This simplification may not be appropriate, especially for anions whose edge adsorption may be affected by this negative charge.

The surface complexation models contain the assumption that ion adsorption takes place on one or at most two sets of reactive surface sites. This is clearly an oversimplification since even simple oxide minerals contain several sets of reactive hydroxyl groups. However, this simplification is necessary to maintain the number of adjustable parameters at a reasonable level. Natural materials such as soils are complex, multisite mixtures having a variety of reactive surface functional groups. Thus surface complexation constants determined for soils represent average composite values for all these sets of reactive surface functional groups.

Constant Capacitance Model

The constant capacitance model has been used to describe adsorption on silicon, aluminum, iron, and titanium oxides, kaolinite, montmorillonite, and illite clay minerals, plant cell walls, and soils. Adsorbing ions that have been investigated include the cation and metal ions: calcium, cesium, lead, copper, cadmium, zinc, nickel, cobalt, aluminum, iron, manganese, silver, mercury, lanthanum, europium, ytterbium, and the anions: phosphate, sulfate, arsenate, arsenite, selenite, selenate, borate, molybdate, silicate, fluoride, phthalate, salicylate, benzoate, citrate.

Examples of the fit of the constant capacitance model to trace metal adsorption edges are provided in Figure 2 for iron, lead, copper, and cadmium adsorption on silica. As for many trace metal cations, the amount of adsorption increases rapidly from 0 to 100% over a narrow pH range. The model is well able to describe these changes in adsorption for the four different metal ions. Figure 3 indicates the ability

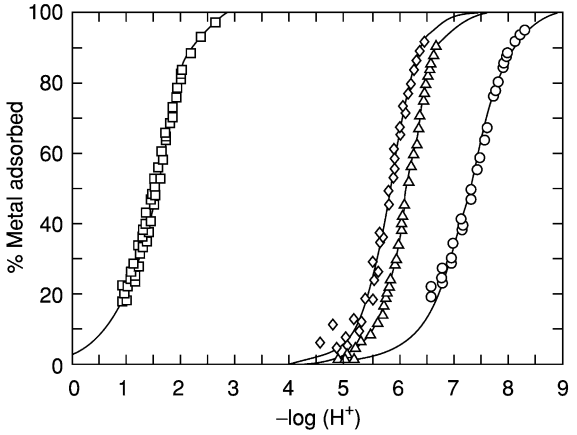


Figure 2 Fit of the constant capacitance model to metal adsorption on silica. Model results are represented by solid lines. □ Fe; ◇ Pb; △ Cu; ○ Cd. Reprinted with permission from Schindler PW, Fürst B, Dick R, and Wolf PU (1976) Ligand properties of surface silanol groups, I. Surface complex formation with Fe^{3+}, Cu^{2+}, Cd^{2+}, and Pb^{2+}. *Journal of Colloid Interface Science* 55: 469–475.

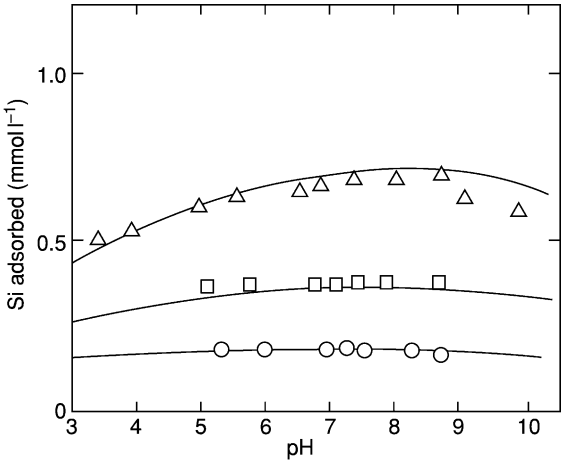


Figure 3 Fit of the constant capacitance model to silicate adsorption on goethite. Model results are represented by solid lines. ○ 2×10^{-4} M; □ 4×10^{-4} M; △ 8×10^{-4} M. Reprinted from Sigg LM and Stumm W (1981) The interaction of anions and weak acids with the hydrous goethite (α-FeOOH) surface. In: *Colloids and Surfaces*, vol. 2, pp. 101–117. Amsterdam: Elsevier Science.

of the constant capacitance model to fit adsorption envelopes for the anion, silicate, on to the iron oxide, goethite at various initial silicate concentrations. Silicate adsorption is nearly constant over most of the pH range. The model is able to describe the adsorption, including the pH dependence observed at the highest initial silicate concentration.

Applications of the constant capacitance model to soils have been restricted to anions: phosphate, selenite, borate, and molybdate. For the application of the model to selenite adsorption by soils, two sets of reactive surface functional groups were postulated; monodentate surface complexes were formed on one set of sites and bidentate surface complexes on the other. The model was initially applied to one Californian soil. As can be seen in Figure 4a, the fit of the model to the data is good. Subsequently, the model parameters obtained in fitting this soil were used to predict adsorption on additional Californian soils. Figure 4b shows that this prediction was qualitatively successful and indicates some predictive capability of the model for soils of somewhat similar chemical and physical characteristics.

An alternative approach has been developed for describing borate adsorption on soils. From the fitted surface complexation constants for a set of soils, a

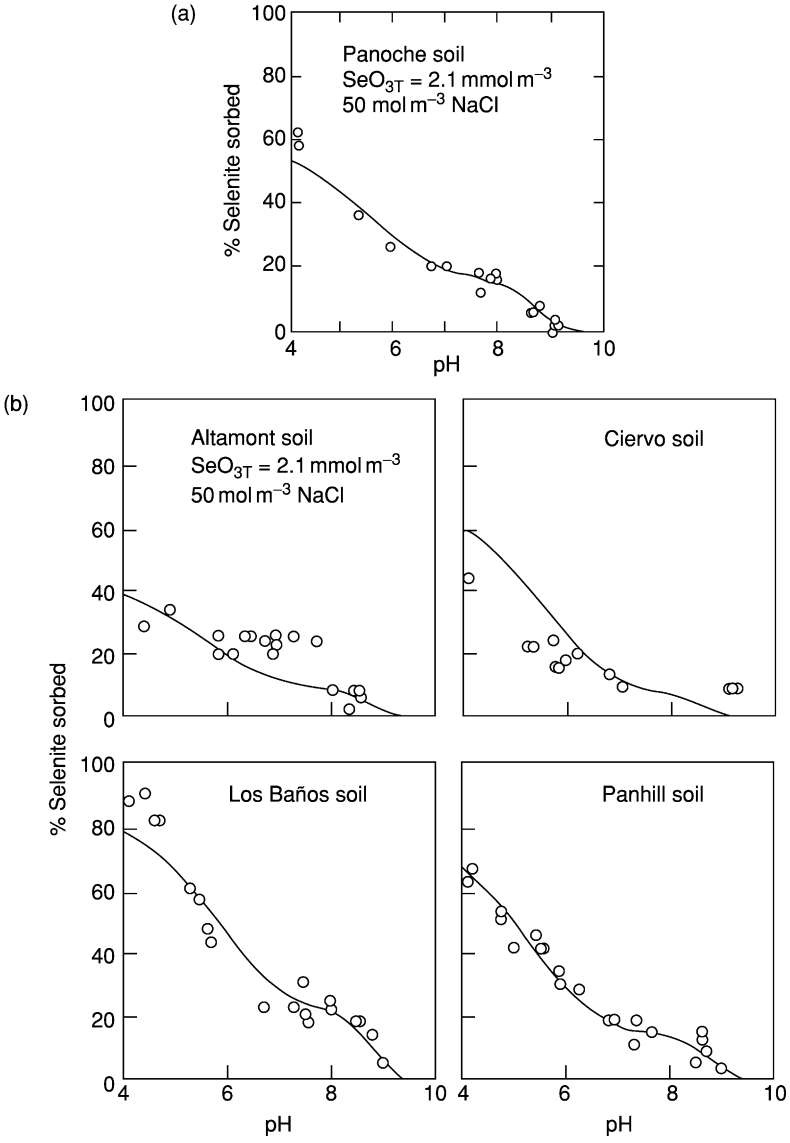


Figure 4 (a) Fit of the constant capacitance model to selenite adsorption on a California soil. Model fit is represented by a solid line. (b) Constant capacitance model predictions of selenite adsorption by California soils. Model predictions are represented by solid lines. Reprinted with permission from Sposito G, de Wit JCM, and Neal RH (1988) Selenite adsorption on alluvial soils: III. Chemical modeling. *Soil Science Society of America Journal* 52: 947–950.

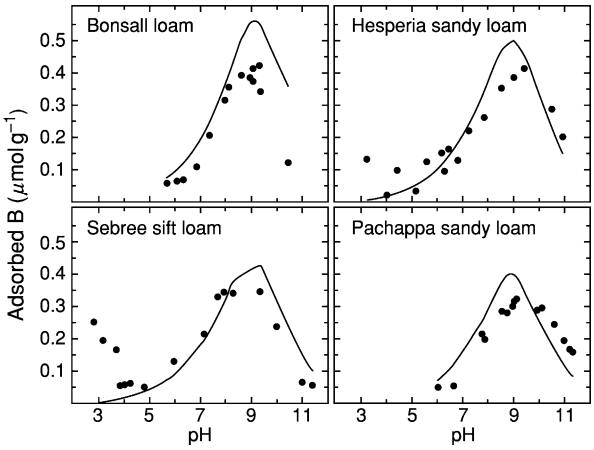


Figure 5 Constant capacitance model predictions of boron adsorption by soils of various soil chemical properties and diverse soil orders. Model predictions are represented by solid lines. Modified from Goldberg S, Lesch SM, and Suarez DL (2000) Predicting boron adsorption by soils using soil chemical parameters in the constant capacitance model. *Soil Science of Society of America Journal* 64: 1356–1363.

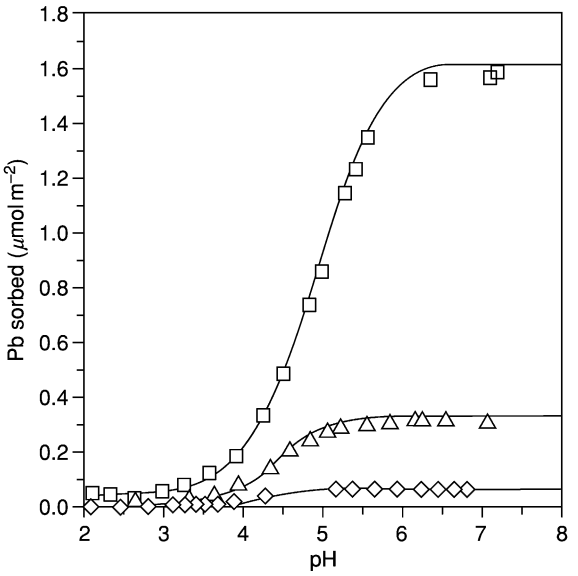


Figure 6 Fit of the diffuse layer model to lead adsorption on hematite. Model fits are represented by solid lines. □ 100 μmol l^{-1}; △ 20 μmol l^{-1}; ◇ 4 μmol l^{-1}. Reproduced with permission from Christl DI and Kretzschmar R (1999) Competitive sorption of copper and lead at the oxide–water interface: implications for surface site density. *Geochimica et Cosmochimica Acta* 63: 2929–2938.

general regression model was obtained that predicts the surface complexation constants for new soils from easily measured chemical parameters: surface area, organic carbon content, inorganic carbon content, and aluminum oxide content. These surface complexation constants were then used in the constant capacitance model to predict borate adsorption on the new soils. This approach constitutes a completely independent model evaluation that was able to predict borate adsorption on a diverse set of soils having a wide range of chemical properties, as indicated in Figure 5.

Diffuse-Layer Model

The diffuse-layer model has been used to describe adsorption on iron, aluminum, manganese, titanium, and silicon oxides, kaolinite, montmorillonite and biotite minerals, natural organic matter, bacterial cell walls, and sediments. Adsorbing ions that have been investigated include the cation and metal ions: calcium, strontium, barium, copper, nickel, zinc, cadmium, lead, cobalt, aluminum, chromium, silver, mercury, uranium and the anions: phosphate, sulfate, selenite, selenate, arsenate, arsenite, borate, chromate, fluoride, vanadate, thiosulfate, oxalate, phthalate, salicylate, benzoate, and fulvate.

The ability of the diffuse-layer model to describe metal adsorption edges is indicated in Figure 6 for lead adsorption on the iron oxide, hematite. The model was able to describe the data very well at three significantly different initial lead concentrations over a wide range of solution pH values. Figure 7 demonstrates the ability of the diffuse-layer model to fit

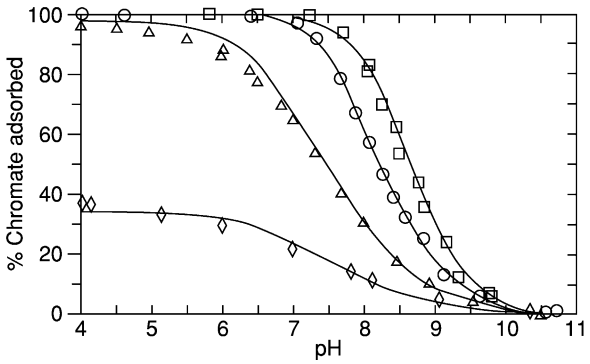


Figure 7 Fit of the diffuse layer model to chromate adsorption on goethite. Model fits are represented by solid lines. □ 0.01 mmol l^{-1}; ○ 0.05 mmol l^{-1}; △ 0.2 mmol l^{-1}; ◇ 0.8 mmol l^{-1}. Reprinted with permission from Mesuere K and Fish W (1992) Chromate and oxalate adsorption on goethite: I. Calibration of surface complexation models. *Environmental Science and Technology* 26: 2357–2364. Copyright American Chemical Society.

adsorption envelopes for the chromate anion on goethite. In order to fit the two intermediate chromate concentrations it was necessary to add a third chromate surface complex, [$SCrO_4^{3-}$]. With this addition, the model is well able to describe chromate adsorption at all initial chromate concentrations over a wide range of solution pH.

Applications of the diffuse-layer model to soil systems have not been carried out to date. Adsorption

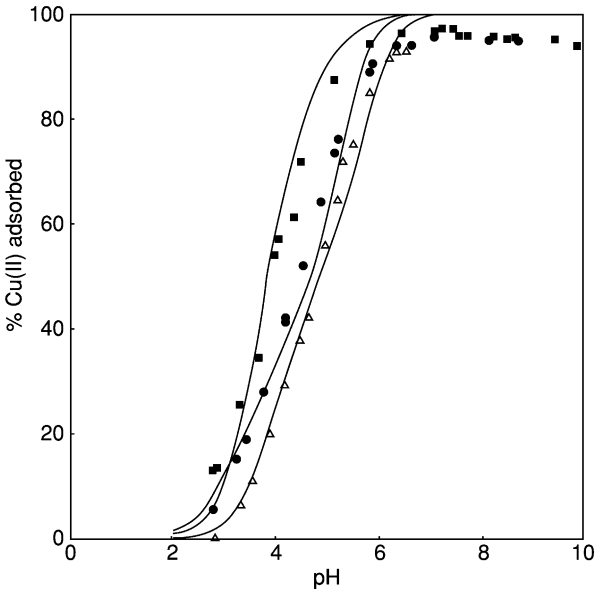


Figure 8 Fit of the diffuse layer model to copper adsorption on lignocellulose extracted from wheat bran. Model fits are represented by solid lines. △ $[Cu(II)]_{ini} = 2.10^{-4}$ mol l^{-1}, I = 0.1; ● $[Cu(II)]_{ini} = 2.10^{-4}$ mol l^{-1}, I = 0.01; ■ $[Cu(II)]_{ini} = 2.10^{-5}$ mol l^{-1}, I = 0.1. Reproduced with permission from Ravat C, Dumonceau J, and Monteil-Rivera F (2000) Acid/base and Cu(II) binding properties of natural organic matter extracted from wheat bran: modeling by the surface complexation model. *Water Research* 34: 1327–1329.

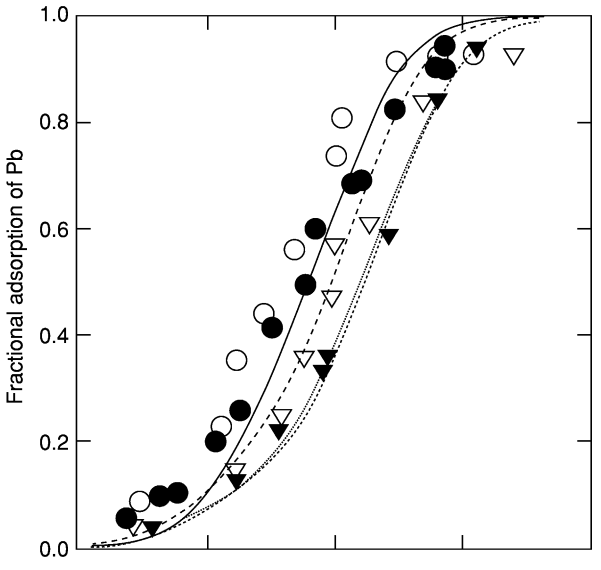


Figure 9 Fit of the triple-layer model to lead adsorption on goethite. Model fits are represented by solid lines. ○ 9.65 μmol l^{-1}; ● 24.1 μmol l^{-1}; 48.3 μmol l^{-1}; ▼ 72.4 μmol l^{-1}. Reproduced with permission from Kooner ZS, Cox CD, and Smoot JL (1995) Prediction of adsorption of divalent heavy metals at the goethite/water interface by surface complexation modeling. *Environmental Toxicology and Chemistry* 14: 2077–2083. Copyright SETAC, Pensacola, Florida, USA.

of various metal ions has been investigated on a lignocellulose organic substrate extracted from wheat bran. In these applications, two sets of reactive surface functional groups are considered representing carboxylic and phenolic sites. Figure 8 presents the ability of the diffuse-layer model to describe copper adsorption edges on this organic material as a function of initial copper concentration $[Cu_{ini}]$, solution pH, and ionic strength, I. In this application, the model is fitted to the data for $[Cu_{ini}] = 2 \times 10^{-4}$ M, $I = 0.1$. The model parameters resulting from this optimization were then used to predict the remaining model results depicted in Figure 8. The diffuse layer model is well able to describe copper adsorption on this natural material.

Triple-Layer Model

The triple-layer model has been used to describe adsorption on iron, aluminum, manganese, and silicon oxides, kaolinite and smectite clay minerals, and soils. Adsorbing ions that have been investigated include the cation and metal ions: sodium, potassium, calcium, magnesium, lead, zinc, cadmium, copper, cobalt, silver, mercury, uranium, plutonium, thorium, neptunium, and the anions: chloride, nitrate, perchlorate, sulfate, selenite, selenate, arsenate, arsenite, molybdate, borate, chromate, silicate, fluoride, carbonate, oxalate, phthalate, salicylate, lactate, acetate, formate, and humate.

The ability of the triple-layer model to describe metal ion adsorption edges is depicted in Figure 9 for lead adsorption on goethite. In this application an inner-sphere surface complex is assumed. The triple-layer model was well able to describe lead adsorption at various initial concentrations as a function of solution pH. Figure 10 shows the fit of the triple-layer model to anion adsorption envelopes for selenate adsorption on the iron oxide, ferrihydrite. With the assumption of two outer-sphere surface complexes, the model describes the adsorption data quantitatively for four vastly differing initial selenium concentrations as a function of solution pH.

The triple-layer model was able to fit calcium, magnesium, and sulfate adsorption on a Brazilian oxisol and sulfate adsorption on an acidic forest soil. Molybdate adsorption on two Californian soils as a function of solution pH and ionic strength could be described using either an inner-sphere or an outer-sphere adsorption mechanism. The quality of the fit was slightly better with the inner-sphere adsorption mechanism. In this application it was assumed that aluminol groups on the clay edges are the reactive surface functional groups in the soils and that

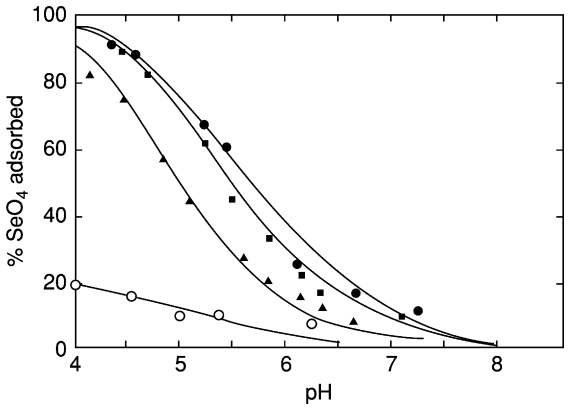


Figure 10 Fit of the triple-layer model to selenate adsorption on amorphous iron oxide. Model fits are represented by solid lines. Total Seo_4 added ● 2×10^{-7} mol l^{-1}; ■ 2×10^{-5} mol l^{-1}; ▲ 2×10^{-4} mol l^{-1}; ○ 2×10^{-3} mol l^{-1}. Reproduced with permission from Davis JA and Leckie JO (1980) Surface ionization and complexation at the oxide/water interface. III. Adsorption of anions. *Journal of Colloid Interface Science* 74: 32–43.

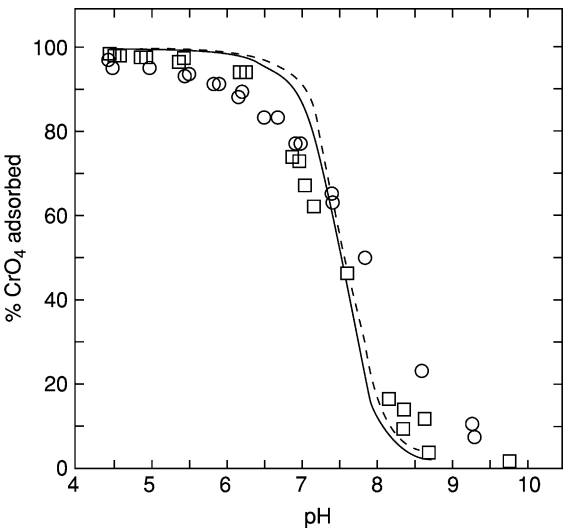


Figure 11 Fit of the triple-layer model to chromate adsorption on two soils. Model fits are represented by solid lines. □ Holston/Cloudland series soil; ○ Cecil/Pacolet series soil. Reprinted with permission from Zachara JM, Ainsworth CC, Cowan CE, and Resch CT (1989) Adsorption of chromate by subsurface soil horizons. *Soil Science of America Journal* 53: 418–428.

surface complexation constants determined for reactive surface hydroxyls of aluminum oxide can be used to describe the surface complexation reactions undergone by these aluminol groups.

In a similar approach, surface complexation constants previously determined for the reactions undergone by aluminum-substituted goethite were used to describe chromate adsorption by two soils. It was assumed that only the iron sites on the aluminum-substituted goethite are involved in chromate

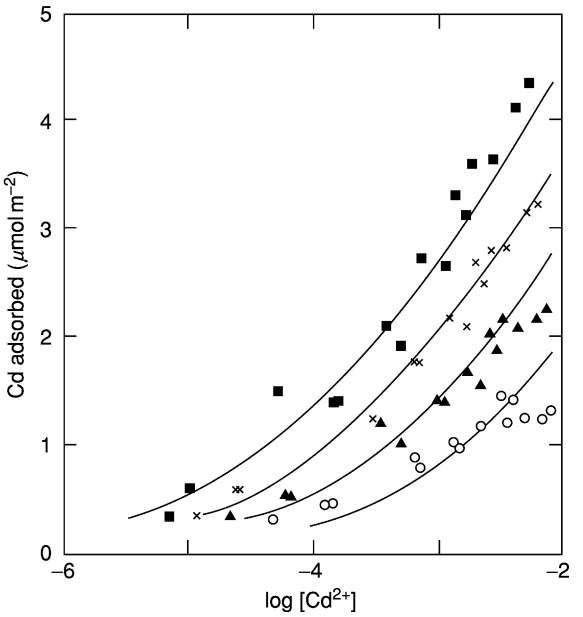


Figure 12 Fit of the one-pK model to cadmium adsorption on amorphous iron oxide. Model fits are represented by solid lines. pH ○ 7.0; ▲ 7.4; × 7.7; ■ 8.0. Reproduced with permission from van Riemsdijk WH, de Wit JCM, Koopal LK, and Bolt GH (1987) Metal adsorption on heterogeneous surfaces: adsorption models. *Journal of Colloid Interface Science* 116: 511–522.

adsorption, forming monodentate outer-sphere surface complexes. The ability of the model to describe chromate adsorption on the two soils as a function of solution pH is indicated in Figure 11. The fit of the model on these heterogeneous materials is qualitatively correct.

One-pK Model

The one-pK model has been used to describe adsorption on iron and aluminum oxides. The vast majority of studies to date have used goethite as the adsorbent material. Adsorbing ions that have been investigated include the cations: potassium, calcium, cadmium, copper, and anions: phosphate, arsenate, selenite, sulfate, chromate, lactate, oxalate, malonate, phthalate, citrate, and fulvate.

The ability of the one-pK model to fit cation adsorption isotherms is indicated in Figure 12 for cadmium adsorption on hematite. The model was well able to describe cadmium adsorption at various solution pH values and initial cadmium concentrations. In this application, consideration of background electrolyte surface complexes was neglected. Figure 13 indicates the ability of the one-pK model to fit anion adsorption isotherms for phosphate adsorption on goethite. The model was well able to describe phosphate adsorption over a very wide range of solution pH values

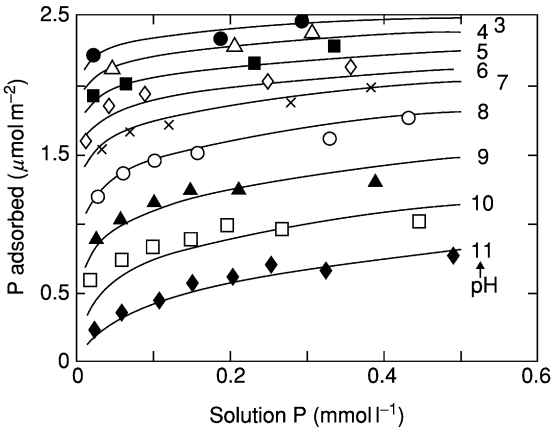


Figure 13 Fit of the one-pK model to phosphate adsorption on goethite. Model fits are represented by solid lines. Reproduced with permission from van Riemsdijk WH and van der Zee SEATM (1991) *Comparisons of Models for Adsorption, Solid Solution and Surface Precipitation*. Dordrecht, the Netherlands: Kluwer (based on experimental data of Bowden *et al.* (1980)) In: Bolt GH *et al.* (eds) *Interactions at the Soil Colloid–Soil Solution Interface*, pp. 241–256. Kluwer.

and initial phosphate concentrations. In this application, a surface complexation constant for potassium adsorption was also optimized.

List of Technical Nomenclature

ϵ_o	Permittivity of vacuum [$C^2\,J^{-1}\,m^{-1}$]
σ	Surface charge density [$C\,m^{-2}$]
σ_β	Surface charge density in the β-plane [$C\,m^{-2}$]
σ_d	Surface charge density in the d-plane [$C\,m^{-2}$]
σ_o	Surface charge density in the o-plane [$C\,m^{-2}$]
$(\sigma_{logK})_I$	Standard deviation for logK of the Ith data set
Ψ	Surface potential [V]
Ψ_β	Surface potential in the β-plane [V]
Ψ_d	Surface potential in the d-plane [V]
Ψ_o	Surface potential in the o-plane [V]
A^-	Anion of the background electrolyte
a	Particle concentration [$g\,l^{-1}$]
C	Capacitance [$F\,m^{-2}$]
C_1	Capacitance between the o- and the β-plane [$F\,m^{-2}$]
C_2	Capacitance between the β- and the d-plane [$F\,m^{-2}$]
C^+	Cation of the background electrolyte
D	Dielectric constant of water
F	Faraday constant [$C\,mol^{-1}$]
I	Ionic strength
K_I	Equilibrium constant
L	Ligand
l^{l-}	Charge on the ligand
M	Metal ion
m^{m+}	Charge on the metal
N_A	Avogadro's number
N_S	Maximum surface site density [sites nm^{-2}]
R	Molar gas constant [$J\,mol^{-1}\,K^{-1}$]
S	Surface area [$m^2\,g^{-1}$]
SOH	Reactive surface hydroxyls bound to a metal ion S
S_s	High-affinity sites
S_T	Total number of reactive surface functional groups
S_w	Low-affinity sites
T	Absolute temperature [K]

See also: **Clay Minerals**

Further Reading

Bolt GH and van Riemsdijk WH (1982) Ion adsorption on inorganic variable charge constituents. In: Bolt GH (ed.) *Soil Chemistry, Part B. Physicochemical Methods*, pp. 459–503. Amsterdam: Elsevier.

Davis JA and Kent DB (1990) Surface complexation modeling in aqueous geochemistry. *Reviews in Mineralogy* 23: 177–260.

Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*. New York: John Wiley.

Goldberg S (1992) Use of surface complexation models in soil chemical systems. *Advances in Agronomy* 47: 233–329.

Goldberg S (1993) Constant capacitance model: chemical surface complexation model for describing adsorption of toxic trace elements on soil minerals. *American Chemical Society Symposium Series* 518: 278–307.

Goldberg S (1995) Adsorption models incorporated into chemical equilibrium models. In: Loeppert R, Schwab AP, and Goldberg S (eds) *Chemical Equilibrium and Reaction Models*. Soil Science Society of America Special Publication 39, pp. 37–60.

Goldberg S (1998) Ion adsorption at the soil particle-solution interface: modeling and mechanisms. In:

Huang PM, Senesi N, and Buffle J (eds) *IUPAC Series on Analytical and Physical Chemistry of Environmental Systems*, vol. 4. *Structure and Surface Reactions of Soil Particles*, pp. 377–412. Chichester: John Wiley.

James RO and Parks GA (1982) Characterization of aqueous colloids by their electrical double-layer and intrinsic surface chemical properties. *Surface and Colloid Science* 12: 1119–216.

Sposito G (1983) Foundations of surface complexation models of the oxide–aqueous solution interface. *Journal of Colloid and Interface Science* 91: 329–340.

Sposito G (1984) *The Surface Chemistry of Soils*. Oxford: Oxford University Press.

SUSTAINABLE SOIL AND LAND MANAGEMENT

J L Berc, USDA Natural Resources Conservation Service, Washington, DC, USA

Introduction

The concept of sustainability has evolved to meet changing needs, changing realities, technological development, and new understanding of ecosystem functions. At any time in history it is likely that people have hoped to manage their land to sustain its productivity throughout their lifetime and for their offspring. In a new land with abundant resources, it is difficult to imagine depleting resources. Early national policy in the USA encouraged people to settle in the west, to claim the territory, and relieve population pressures in the east. If land was worn out, people could move to the western frontier. By the end of the nineteenth century, however, most of the land was settled, the frontier was 'closed,' and people could no longer count on more available land to replace worn-out land. Many families lacked sufficient knowledge, land, and water to sustain their livelihood on the land. Land was often pushed to produce crops that the soil and climate could not support. By the 1930s the nation was faced with much ruined and abandoned land in the east and a dust bowl and more abandoned farms in the west. From this ecological and human disaster came the public recognition of the need to help farmers and ranchers better balance long-term conservation needs and short-term economic needs so that agricultural production could be sustainable.

The Birth of Soil Conservation in US Agriculture

The vision of the early soil conservationists in the USA still inspires us today:

> Nature treats the earth kindly. Man treats her harshly. He overplows the cropland, overgrazes the pastureland, and overcuts the timberland. He destroys millions of acres completely. He pours fertility year after year into the cities, which in turn pour what they do not use down the sewers into the rivers and the ocean. The flood problem insofar as it is man-made is chiefly the result of overplowing, overgrazing, and overcutting of timber. This terribly destructive process is excusable in a young civilization. It is not excusable in the United States in the year 1938.
>
> *Henry A. Wallace, Secretary of Agriculture*
>
> Among the causes of this misuse of the soil and consequent waste of human resources, lack of knowledge on the part of individuals undoubtedly plays a large part. But even if all farmers thoroughly understood the consequences to themselves of the type of land use they are practicing and had perfect knowledge of soil-conservation techniques, a large number would still be unable to put the knowledge fully into practice. Social and economic limitations would prevent them from doing certain things they knew ought to be done.
>
> History is largely a record of human struggle to wrest the land from nature, because man relies for sustenance on the products of the soil. Yet too frequently man's conquest of the land has been disastrous: over extensive areas, his culture of the earth has resulted in extreme impoverishment or complete destruction of the soil resource on which he is dependent.... Conservation of the soil, in a national sense, requires the adoption of sound land-use principles and practices by agriculture as a whole. The attainment of this objective involves the widespread use of physical measures of land defense and the adjustment of certain economic and social forces tending to encourage exploitation of the soil.
>
> *Hugh H. Bennett, Chief, Soil Conservation Service, and W.C. Lowdermilk, 1938*
>
> Thou shalt inherit the Holy Earth as a faithful steward, conserving its resources and productivity from generation to generation. Thou shalt safeguard thy fields from soil erosion, thy living waters from drying up, thy forests from desolation, and protect thy hills from overgrazing by thy herds, that thy descendants may have abundance forever. If any shall fail in this stewardship of the land thy fruitful fields shall become sterile stony ground and wasting

> gullies, and thy descendants shall decrease and live in poverty or perish from off the face of the earth.
>
> *Walter Clay Lowdermilk, 1939*

W.C. Lowdermilk wrote this 'eleventh commandment' after reporting on the 'graveyard of empires' throughout the 'Old World,' where city after city was ruined not by climate change, but by the failure of their populations to conserve the soil and water resources from which they grew. His report, *Conquest of the Land Through 7000 Years*, was undertaken to help the USA in its struggle with the Dust Bowl and the gullied South. He warned that, if Americans failed to conserve their soil and water resources, the future of the nation would be at risk. He described erosion control and water management practices used at the time to increase water infiltration and reduce storm flow damage; and he called for improvement and adaptation of practices to address increasingly intensive land use.

The concept of sustainability in the 1930s focused on maintaining farm productivity to supply adequate food and fiber for the growing nation. Agricultural production was dependent on the inherent soil and water qualities of the land. Great progress was made during this era, with the introduction of contour cultivation, cover crops, and use of manure and crop rotations to restore soil nutrients and organic matter. The need for zoning, education, financial relief, and improved practices was recognized at the time to address holistically the roots of problems causing soil degradation and loss.

Emerging Agricultural Technology Shifts National Conservation Priorities

The synthetic fertilizer industry was in its early stages of development at the time of the beginning of soil conservation policy in the USA. Over the next generation, commercial nutrients began replacing crop rotations to maintain productivity. Water management at the watershed level was introduced, and soil drainage, irrigation, and land-leveling became major conservation efforts. By the late 1950s, more than 60 million hectares of farmland had been drained. Along with mechanization and new improved crop breeds, these technologies contributed to the doubling of crop and animal production between 1950 and the early 1990s. In this new technology frontier, the link between productivity and land quality was muted, and the impetus to conserve inherent soil productivity declined.

Erosion control was still encouraged to maintain the storage capacity of dams being constructed in watersheds throughout the country to prevent flooding and maintain adequate water supplies. The concept of 'tolerable soil loss' was developed in the 1950s from the perspective of the cost to the farmers of nutrients that would be lost from the eroded soil. The *1957 Yearbook of Agriculture: Soil* recommended that in the Corn Belt, depending on soil depth, mean annual soil losses should be less than 4.5–11 t ha^{-1} to avoid damaging the land or causing excessive silting on lower fields or in streams and reservoirs.

Nevertheless, erosion remained a major concern. The 1957 yearbook noted that "in the Great Plains alone, about 14 million acres not suited for permanent cultivation were cultivated in 1955. Much of this land offers low returns and is subject to severe erosion even in average years." The acreage damaged from erosion was equal to that damaged in the dust storms of the 1930s. Despite the long-term economic benefits to the farmer of conservation practices and because increasing numbers of farms were on rented land, long-term conservation investment was often not made.

The Conservation Reserve Program of the Soil Bank was established in 1956 to set aside cropland into soil-conserving crops such as grass and legumes for up to 10 years. By 1960 approximately 11.6 million hectares were enrolled. Objections from rural communities about disruptions to their economies caused by the enrollment of whole farms into this program caused it to be ended. The Agricultural Conservation Program, under which farmers received cost-share assistance for conservation, was made a permanent program in 1957. Conservation practices that were promoted included rotation cropping, terraces, shifts from row crops to pasture, improved grazing management, crop-residue management, and grassed waterways, as well as irrigation and drainage.

The importance of land-use planning and the use of the soil survey to develop land suitable for its intended purpose were introduced in the early 1960s. Concerns related to salt buildup in irrigated crops also emerged at that time.

Recognition of New Environmental Issues

The consolidation and specialization of farms, separation of crop from animal production, and increasing use of agrichemicals to manage production continued into the 1970s. No-till agriculture was introduced in conservation systems and helped to achieve conservation goals without relying on crop rotations or contour cultivation. Agricultural production systems today have further evolved, for example, incorporating genetically modified crops that tolerate herbicides used in reduced tillage. At the same time,

environmental awareness has grown. Soil and water quality are of concern not only for on-farm productivity, but also for human health on the farm, and in both rural and urban communities. The health of the ecosystem itself has been raised as a concern. However, public investment in conservation has not kept pace with these emerging conservation concerns.

The *1970 Yearbook of Agriculture, Contours of Change* noted that total conservation investment had not increased since 1962. A conservation-needs inventory was conducted, and it was found that almost two-thirds of cropland, pastureland, and rangeland needed conservation treatment; water pollution was severe.

In the *1983 Yearbook of Agriculture, Using Our Natural Resources*, erosion was cited as a national problem, despite 50 years of government programs to combat it. Concerns about maintaining farmland in agricultural production were discussed along with economic and environmental issues regarding irrigation and agrichemical use. Fish and wildlife habitat degradation resulting from reduced water levels from irrigation expansion, wetland conversion to cropland, planting fence row to fence row, and removing windbreaks was also raised. The advances of stubble-mulch and no-till management systems (developed in the 1950s) were discussed. Further, the Land Evaluation and Site Assessment system to aid land-use decision-making was introduced to support the 1981 National Farmland Protection Act, designed to slow urbanization of agricultural land.

The *1987 Yearbook of Agriculture, Our American Land* noted that, while technical assistance for conservation had remained stable since 1969, financial assistance declined by 77% between 1969 and 1985. New 'carrot and stick' provisions of the 1985 Farm Bill protected highly erodible land and wetlands through conservation compliance, and a new Conservation Reserve Program (CRP) removed approximately 15 million hectares of highly erodible cropland from production for 10 years.

Defining Sustainable Development

In 1987 The World Commission on Environment and Development (The Brundtland Commission) defined sustainable development as actions taken "... to meet the needs of the present without compromising the ability of future generations to meet their own needs." Agenda 21, The United Nations Programme of Action, developed at the 1992 Rio Earth Summit, declared:

> the destruction and degradation of agricultural and environmental resources is a major issue. Techniques for increasing production and conserving soil and water resources are already available but are not widely or systematically applied. A systematic approach is needed for identifying land uses and production systems that are sustainable in each land and climate zone, including the economic, social and institutional mechanisms necessary for their implementation.

The Agenda set the objective:

> To prepare and implement comprehensive policies and programmes leading to the reclamation of degraded lands and the conservation of areas at risk, as well as improve the general planning, management and utilization of land resources and preserve soil fertility for sustainable agricultural development.

In 1994, the US President's Council on Sustainable Development, Sustainable Agriculture Task Force, set forth the goal of management of agricultural activities to protect soil, air, and water quality, and biodiversity to increase agriculture's long-term productivity and profitability and enhance human health and well-being. The task force recommended that assistance to farmers be linked to voluntary implementation of farm and ranch plans for natural resource conservation in agricultural production. They also recommended implementation of land-use policies to preserve the prime land base for US agriculture.

The 1990 Farm Bill defined sustainable agriculture as:

> an integrated system of plant and animal production practices having a site-specific application that will, over the long term, satisfy human food and fiber needs; enhance environmental quality and the natural resources base upon which the agricultural economy depends; make the most efficient use of nonrenewable resources and on-farm/ranch resources; and integrate, where appropriate, natural biological cycles and controls; sustain the economic viability of farm/ranch operations; and enhance the quality of life for farmers/ranchers and society as a whole.

Policy Aspects

During the early 1990s, the US Department of Agriculture (USDA) also developed policy on sustainable development:

> The USDA is committed to working toward the economic, environmental, and social sustainability of diverse food, fiber, agriculture, forest, and range systems. USDA will balance goals of improved production and profitability, stewardship of the natural resource base and ecological systems, and enhancement of the vitality of rural communities. USDA will integrate these goals into its policies and programs, particularly through collaboration, partnerships and outreach.

In the 1991 yearbook, *Agriculture and the Environment*, Secretary of Agriculture Edward Madigan noted that national expenditure for pollution control increased fourfold from 1972 to 1990. Erosion had been greatly reduced by the CRP, and adjustments in the commodity programs in the 1990 Farm Bill were designed to encourage farmers to plant alternative crops and use crop rotations to reduce soil erosion and the use of agrichemicals. Advances in conservation tillage were reported, and associated environmental concerns such as increased application of agrichemicals were addressed.

A holistic, ecosystem-based concept of sustainable soil and land management was developed in technical and policy arenas throughout the decade. The Soil and Water Conservation Society published a collection of essays, *Soil Management for Sustainability*, in 1993. In it, the editors discussed applying agroecologic principles to farming to contribute positively to the quality of air, water, and soil resources and to meet the economic and social needs of the food and fiber producer. They offered the concept of farming according to the specific productivity and vulnerability of the soil in its immediate and larger landscape position.

In 1993, the National Research Council's Committee on Long-Range Soil and Water Conservation published *Soil and Water Quality, An Agenda for Agriculture*. Threats to soil resources were identified and criteria to guide soil management were recommended. The fate and transport of agricultural chemicals were analyzed to identify changes in farming systems required to improve water quality. Policy and program options to improve long-term conservation of soil and water quality were recommended. The committee identified these objectives for soil and water resource management: conserve and enhance soil quality as a fundamental first step to environmental improvement; increase nutrient, pesticide, and irrigation use efficiencies in farming systems; increase the resistance of farming systems to erosion and runoff; and make greater use of field and landscape buffer zones to capture otherwise uncontrolled off-site flow of agrichemicals.

The USDA released *Food and Agricultural Policy: Taking Stock for the New Century* in 2001. In the section on conservation and the environment, the report pointed out that, although soil erosion had declined by 40% over the previous 15 years, farmland was still losing 1.72 billion metric tons of soil every year. It stated that the array of conservation issues had grown with changes in the structure of agriculture, in farm practices, and in public concerns. Soil was recognized as an emerging challenge to manage as a national strategic asset. More than one-third (52.25 million hectares) of US cropland was reported to be in need of improved soil quality. It recognized that the biological, chemical, and physical processes that drive agricultural productivity cannot be bypassed with inputs. Soil degradation, such as compaction, crusting, salination, and loss of organic matter, needed to be addressed to maintain multiple soil benefits, including reduced runoff and erosion, increased carbon sequestration, and improved productivity and sustainability. The need to increase support for conservation on land in agricultural production was recognized, and innovative, market-based environmental benefits and/or stewardship policy tools were suggested.

The 2002 Farm Bill increased conservation on land in agricultural production and introduced a new program to pay farmers directly for the environmental benefits they produce with their conservation systems. The Farm Bill also directs a large portion of conservation resources to address environmental issues caused by livestock operations.

Sustainable and Organic Agricultural Systems

Sustainable agricultural systems strive to restore and enhance inherent soil qualities and productivity through improved conservation and management practices, while reducing use of commercial inputs. The USDA has supported a Sustainable Agriculture Research and Education Program since 1988. A number of colleges and universities throughout the USA have developed sustainable agriculture programs during this time frame.

Interest in organic agriculture has also grown dramatically. The production of organic food and fiber is largely dependent on building fertile, productive soils without the use of synthetic inputs. The public interest in agricultural production and products free of fertilizers and pesticides, expressed through the marketplace, both for human health and ecosystem health, has caused organic and sustainable production systems to flourish. US retail sales of organic agricultural products have increased by 20–25% each year since 1990 to US $7.8 billion in 2000, with global sales topping US $17.5 billion. US land in organic production grew from approximately one-quarter of a million hectares in 1992 to close to one million hectares in 2001. In 2002 the USDA National Organic Label Standard went into effect. This is expected to expand markets for organic products further.

Inventories and Assessment

Soil Loss

Soil losses recognized in the 1930s included both physical and chemical soil properties: loss of structure, loss of nutrients by crop removal and leaching, and loss of the soil itself through erosion by wind or water. Good conservation practices included primarily maintaining soil organic matter by crop rotations, cover crops, contour cultivation, and use of manure.

The 1937 appraisal of the 712.3 million hectares of agricultural land in the USA reported 114 million (16%) ruined or severely damaged by erosion, and another 314 million (44%) moderately damaged. Of the 168 million hectares of cropland then in production (1935 Census of Agriculture), 81 million were ruined, severely damaged, or had lost at least half of the topsoil, and erosion had begun on another 40.5 million hectares. Due to soil depletion from overgrazing of the 295 million hectares of rangeland in 1935 in the western USA, 239 million hectares (80%) were eroding more or less seriously, further reducing productive capacity.

Wind erosion was reported each year from 1954 to 1957 as damaging more than 6.1 million hectares each year, and destroying another 0.5–2 million hectares of crops. The annual cost of water-borne sediment damage in the mid-1980s was estimated between US $4 and 5 billion. The 1982 USDA National Resources Inventory (NRI) reported that 57.2% of nonfederal rural land needed conservation for erosion, drainage, irrigation management, or vegetative cover.

Changes in USDA conservation programs initiated in the 1985 Farm Bill created new progress in erosion control through the late 1980s into the 1990s. Conservation goals, however, were tied to earlier concepts of rates of tolerable soil loss (T), which emphasized farm-level economics more than ecosystem-wide sustainability. Conservation programs were targeted to the most highly erodible land, and ignored land that was eroding at close to T-values. Despite progress in controlling 'excess erosion,' the 1992 NRI reported that 82 755 000 ha of cropland and 23 661 000 ha of pastureland needed conservation treatment. The 1997 NRI reported that of the total 152 684 000 ha of cropland, 26 204 000 ha were eroding at more than

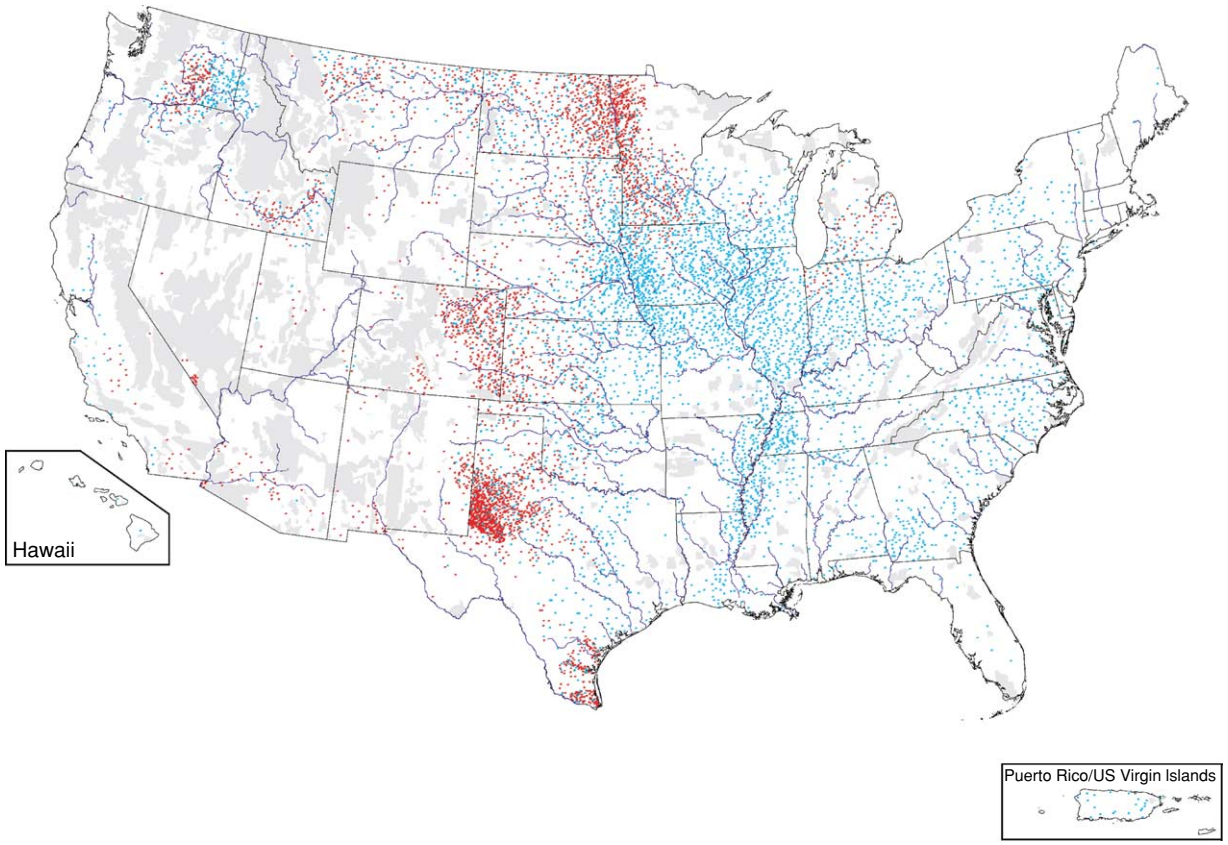


Figure 1 This dot map shows tons of erosion due to wind and water on cropland and Conservation Reserve Program land in the US. Each blue dot represents 970 million metric tons of mean annual erosion due to water. Each red dot represents 760 million metric tons of mean annual erosion due to wind. The combined erosion for the USA is 1.72 billion metric tons per year.

T owing to water erosion, and 19 372 000 ha of cropland were eroding at more than *T* owing to wind erosion (Figure 1).

Water Quality

Water pollution by nutrients or other agrichemicals from agricultural soils was not nationally recognized as a concern until the 1970s. From 1974 to 1981, river and stream water quality monitoring data showed that nitrate concentration increases were widespread nationally. From 1980 to 1989, the data showed that concentrations tended to decrease as often as they increased. These fluctuations are attributable to several farm economy crises during the period that severely curtailed application of purchased inputs. However, the cropland application of commercial nitrogen fertilizers continues to contribute to excess nitrogen in ground and surface waters. One example of the impact is the hypoxia, or 'dead' zone in the Gulf of Mexico, considered symptomatic of excessive nitrogen runoff (Figure 2).

From 1982 to 1992, commercial phosphorus consumption dropped 22% nationally, and river and stream monitoring data from 1982 to 1989 showed widespread declines in total phosphorus concentrations. However, increasing concentration of animal agriculture production systems resulted in increased localized problems associated with animal waste. As livestock production shifted to fewer, larger operations, more manure nutrients were spread on smaller areas. The number of counties without enough land to spread safely the manure generated within their boundaries increased significantly.

The 2002 Report, *State of the Nation's Ecosystems, Measuring the Lands, Water, and Living Resources of the United States*, evaluated chemical, physical, and biological conditions on US farmland. Approximately 20% of groundwater wells and 10% of stream sites on farmlands sampled by the US Geological Survey National Water Quality Assessment Program had nitrate concentrations that exceeded the federal drinking water standard (10 ppm or $10\,\mu g\,ml^{-1}$). About three-quarters of sampled farmland stream sites had phosphorus levels above the maximum concentration (0.1 ppm or $0.1\,\mu g\,ml^{-1}$) recommended to prevent algal growth in streams. More than 80% of the farmland streams had at least one pesticide above aquatic life guidelines and about 4% had one or more compounds that exceeded human health standards or guidelines. Approximately 60% of groundwater wells sampled in farmland areas had at least one detectable pesticide. It was noted that drinking water or aquatic standards or guidelines do not exist for approximately half of the 76 pesticides analyzed. In addition, mixtures of pesticides and intermittent high-concentration events are

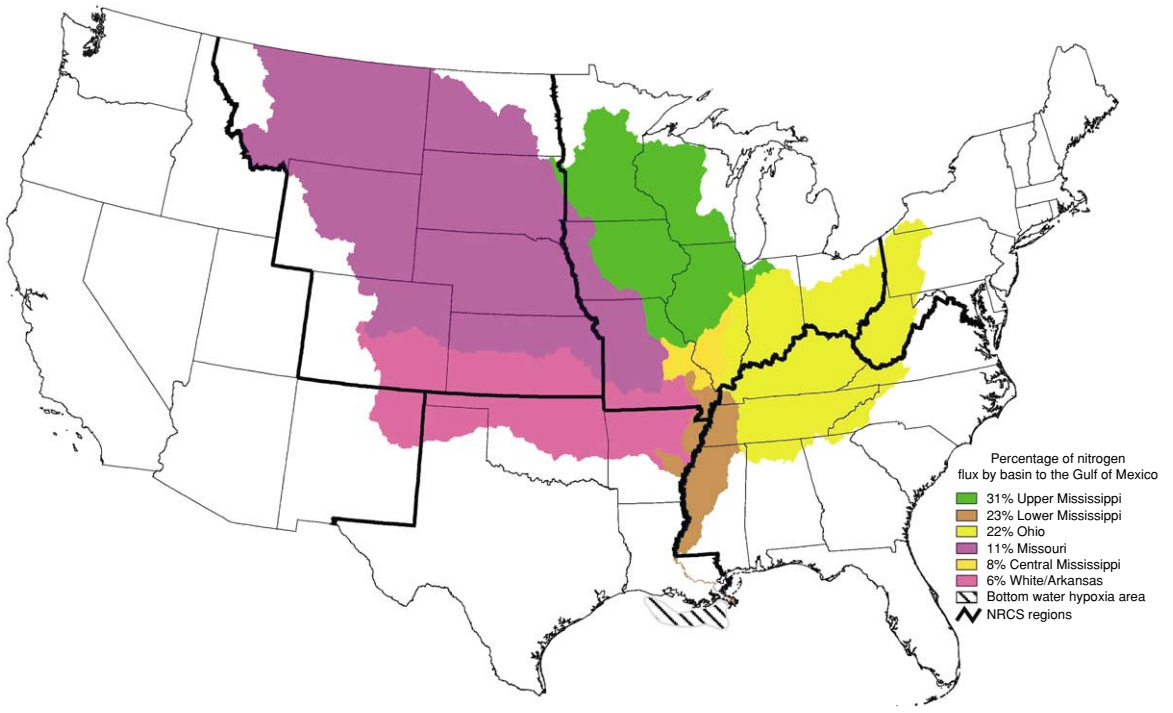


Figure 2 Areas where nitrogen leaves the soil and joins the water system, creating the 'bottom water hypoxia area' in the Gulf of Mexico. NRCS, Natural Resources Conservation Service.

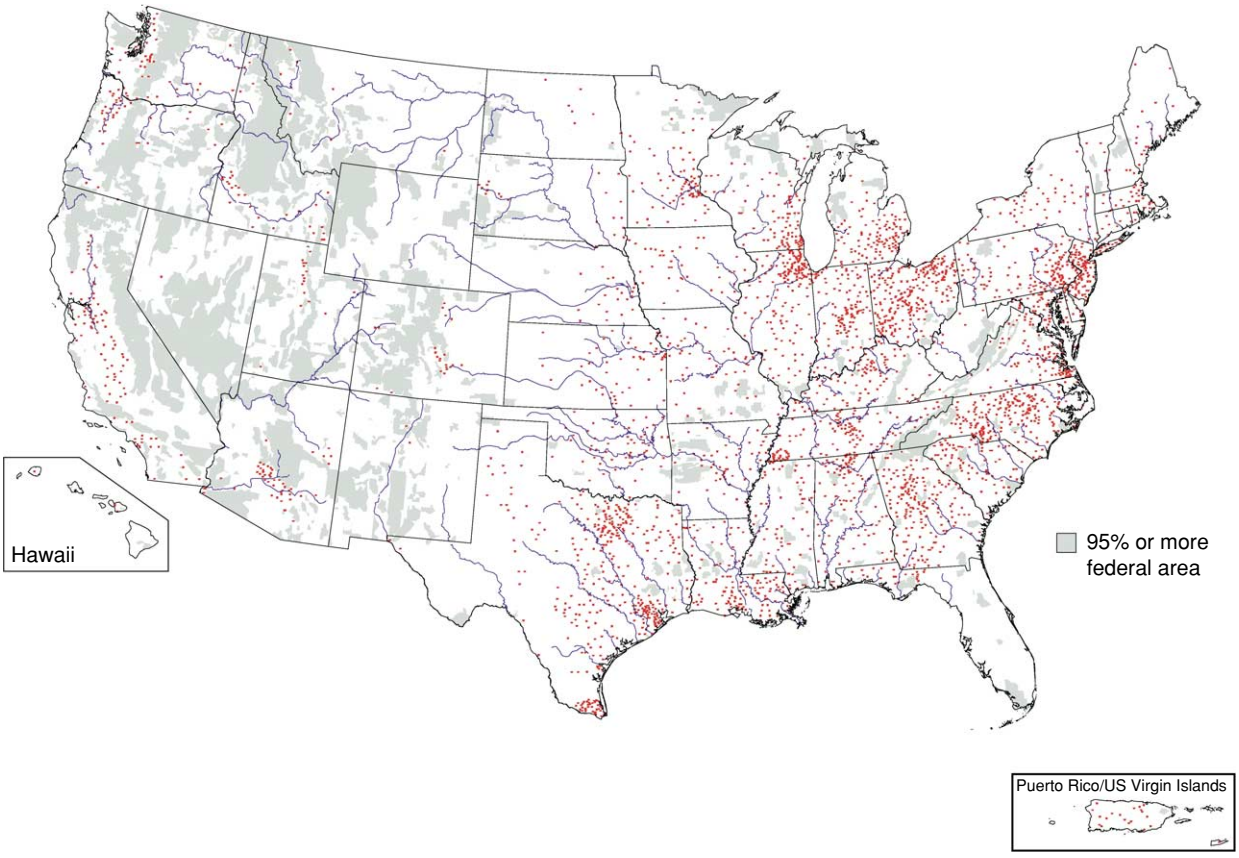


Figure 3 Of the 170 500 000 ha of cropland in 1982, 2 975 500 ha of prime farmland were developed for nonagricultural use between 1982 and 1997. Each red dot represents 810 hectares of newly developed land.

not covered or well understood. The report further identified the need for additional data and indicators to characterize soil organic matter, soil biology, and soil salinity.

Loss of Farmland

The loss of prime farmland to nonagricultural development emerged in the 1980s as a key concern to sustainable soil and land management. In 15 years, 8% of the 93 500 000 ha of prime farmland farmed in 1982 were lost, primarily due to nonagricultural development. This loss caused marginal land with less-favorable characteristics to be brought into production to meet agricultural demand. Production on marginal land requires more inputs and can create more negative offsite environmental impacts. A Federal Farmland Protection Program and state and county programs to preserve agricultural land were implemented and preserved about 0.5 million hectares of farmland by 2002, but the accelerating loss of agricultural land has not been significantly slowed (Figure 3).

Contributing to concerns about the supply of agricultural land is the recognition of the ecologic value of wetlands to sustain water quality, fish, and wildlife. Once considered a conservation practice, the drainage of wetlands is no longer supported or condoned. There is an effort to restore wetlands on agricultural lands, to achieve no net loss of wetlands in the USA. This puts further pressure on remaining croplands to provide the agricultural production demanded by increasing population throughout the world.

At the same time, the rising demand for and profitability of locally produced food, demonstrated by the growth of farmers' markets, 'pick-your-own' farms, and Community Supported Agriculture (contracts between farmers and consumers to supply farm production throughout the growing season for an annual fee), have helped to preserve some farmland, primarily located near urban consumers. These small, direct-market farms often use sustainable or organic production systems.

See also: **Civilization, Role of Soils**; **Degradation**; **Desertification**; **Land-Use Classification**; **Pesticides**; **Salination Processes**; **Salinity:** Management

Further Reading

Heinz Center (2002) *The State of the Nation's Ecosystems. Measuring the Lands, Water, and Living Resources of the United States*. Cambridge, UK: Cambridge University Press.

Kellogg RL, Lander CH, Moffitt DC, and Gollehon PEN (2000) *Manure Nutrients Relative to the Capacity of Cropland and Pastureland to Assimilate Nutrients: Spatial and Temporal Trends for the United States*. Washington, DC: US Department of Agriculture.

Lal R and Pierce FJ (eds) (1991) *Soil Management for Sustainability*. Ankeny, IA: Soil and Water Conservation Society.

Lowdermilk WC (1953) *Conquest of the Land Through 7000 Years*. Agriculture Bulletin No. 99, US Department of Agriculture, Natural Resources Conservation Service. Washington, DC: US Government Printing Office.

Sanders JH, Southgate DD, Lee JG (1995) *The Economics of Soil Degradation: Technological Change and Policy Alternatives*. SMSS Technical Monograph No. 22. US Department of Agriculture, Natural Resources Conservation Service, World Soil Resources. Washington DC: US Government Printing Office.

Sorensen AA, Greene RP, and Russ K (1997) *Farming on the Edge*. Washington, DC: American Farmland Trust.

UN Department of Public Information (1994) *Agenda 21: Earth Summit. The United Nations Programme of Action from Rio, 1992*. New York: UN Publications.

US Department of Agriculture (2001) *Food and Agricultural Policy: Taking Stock for the New Century*. Washington, DC: US Government Printing Office.

US Department of Agriculture, Soils and Men (1938) *1938 Yearbook of Agriculture*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1957) *1957 Yearbook of Agriculture: Soil*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1983) *Using Our Natural Resources: 1983 Yearbook of Agriculture*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1998) *Ten Years of SARE: A Decade of Programs, Partnerships and Progress in Sustainable Agriculture Research and Education*. Washington, DC: US Government Printing Office.

US National Research Council Committee on Long-Range Soil and Conservation (1993) *Soil and Water Quality: An Agenda for Agriculture*. Washington, DC: National Academy Press.

US President's Council on Sustainable Development (1996) *Sustainable America: A New Consensus for Prosperity, Opportunity, and a Healthy Environment*. Sustainable Agriculture Task Force Report. Washington, DC: US Government Printing Office.

USDA–NRCS (1996) *America's Private Land: A Geography of Hope*. Washington, DC: US Government Printing Office.

USDA–NRCS (1997) *Water Quality and Agriculture*. Working Paper No. 16. Washington, DC: US Government Printing Office.

USDA–NRCS/Iowa State University Statistical Laboratory (2000) *Summary Report 1997 National Resources Inventory*. Washington, DC: US Government Printing Office.

World Commission on Environment and Development (The Brundtland Commission) (1987) *Our Common Future*. Oxford, UK: Oxford University Press.

SWELLING AND SHRINKING

D Smiles, CSIRO Land and Water, Canberra, Australia
P A C Raats, Wageningen University, Wageningen, The Netherlands

Introduction

After a century of study, hydrologic theory of non-swelling soils is applied routinely to irrigation and drainage, and to other problems of land and water management where quantitative insights and predictions relating to soil water and solute movement are required. By contrast, water movement in soils that change volume with water content is not well understood, and management of swelling soils remains problematic.

Difficulties arise because volume change complicates measurement of material balance, and both the liquid and the solid phases must be considered; volume change also introduces ambiguity, for example, in water profile measurements based on neutron probes when compared with, say, time-domain reflectometry, where the probes may 'float' with the solid phase. Then, swelling or shrinking accompanying soil water content change results in vertical displacement of the wet soil, which involves gravitational work and contributes an overburden component to the total potential of the soil water. In addition, the total flux of the water has a component advected with the moving solid as well as the water potential-driven, darcian flux relative to the solid. Finally, many swelling soils crack and the network of cracks provides

alternative, preferential pathways for rapid flow of water which prejudice application of theory simply based on darcian flow.

Swelling Soils

The Vertisol and the Histosol orders in Soil Taxonomy represent soils that change volume with change in water content. The Vertisols are the more significant, with great areas in India (approx. $7.9 \times 10^5\,km^2$), Australia (approx. $7 \times 10^5\,km^2$), and Sudan (approx. $5 \times 10^5\,km^2$), while China, Ethiopia, and the USA each have approximately $1.5 \times 10^5\,km^2$. They tend to be chemically fertile, but physically they are generally very intractable; very sticky when wet and strong when dry. Characteristically, they have high contents of smectite clay mineral. In Australia in the Murray Darling Basin alone, there are approximately $2.5 \times 10^5\,km^2$ of these soils. They represent approximately 10% of the basin and almost 50% of its irrigated soils, and they support irrigated and dryland crops which are currently worth more than A\$$1.5 \times 10^9$ annually.

The Histosols of concern include at least $10^6\,km^2$ of low-lying clay soils of marine or estuarine origin. These soils, which are mainly of the Hemist suborder, are characterized by permanently high water tables and the possible presence of sulfuric or sulfidic materials. They occupy low-lying positions of the landscape and many have been laid down over the past 10 000 years in the littoral zone of most continents and islands. Their natural water volume fraction may be higher than 0.8 and their bulk density may be as low as $0.2\,t\,m^{-3}$. They occupy areas of high and increasing agricultural and urban importance. Their development often involves consolidation and the sulfidic soils among them generate acid drainage.

The hydrology of these soils is not well understood, and neither is the reaction and 'chromatographic' chemistry associated with transfer of nutrients, or the residence and degradation of agricultural chemicals. The physicochemical consequences of oxidation of sulfides in the Histosols are also unclear.

In addition to soils, there are also particulate industrial and mining slurries that present local environmental problems. The physics of swelling soils has much to offer with respect to dewatering and consolidating these materials to rehabilitate the landscape; it also has much to offer in engineering soils to support built structures.

Volume change accompanying water movement is central to these activities and, in the case of the Histosols, oxidation of organic matter may also contribute significantly to irreversible volume change accompanying development. Present approaches to hydrology of swelling soils, however, are generally based in nonswelling soil theory; they rarely account for volume change and its consequences, and their use may result in significant errors in estimations of local water and solute flux, water balance, and aquifer recharge. The consequences of oxidation of part of the soil matrix, for example, on material balance estimates are rarely considered.

History

Schübler measured soil volume change in the early nineteenth century and Tempany provided a contemporary perspective early in the twentieth century. Tempany introduced the concept of the shrinkage curve and formally related linear and 'cubical' contraction of soil aggregates in relation to water loss. At the same time, anchored rods were used to measure profile swelling and water content change.

One-dimensional flow of water in soils that change volume with water content was first analyzed in a modern context by Terzaghi, who sought to describe soil consolidation following the imposition of loads associated with built structures. He perceived that the soil water initially carries an imposed load in a saturated soil, and that the soil gradually compresses as this load is transferred to the soil matrix when the water redistributes. Terzaghi thus associated consolidation with water flow and his approach anticipated that of Richards for rigid unsaturated soils, in being based on material continuity and the Darcy law. The outcome was a linear diffusion equation that Terzaghi solved using Fourier series. A decade later and independently, Nicholson and Childs applied a diffusion approach to water flow in clay soils in an agricultural context.

Terzaghi's presentation was obscure: papers still appear in civil engineering journals speculating on what he meant; and civil engineers did not appreciate his formulation and use of solid-based, space-like material coordinates. As a result, the theory was recast in physical space specifically identified with small strain behavior and Terzaghi's classic textbook discarded the material insights of the 1923 work. In 1960, however, Terzaghi's solid-based, space-like coordinates were explicitly reintroduced by McNabb to deal with soil consolidation. Raats and Klute and Smiles and Rosenthal independently developed this approach and the latter also experimentally tested it. In addition, Philip analyzed the one-dimensional problem using conventional spatial coordinates. He dealt well with mechanics of swelling, but his formulation was complicated and he subsequently abandoned it in favor of the simpler mathematics associated with the material approach.

Focus on water flow relative to the solid, described by the Darcy law, lay at the core of these analyses. It also emerged that, when the Darcy law was combined with material continuity using solid-based space-like coordinates, a flow equation analogous to the Richards equation resulted so, at least for one-dimensional problems, solutions to the Richards equation became transferable to flow in swelling systems.

Contemporaneously, chemical engineering schools in Japan and the USA working on filtration of very wet particulate suspensions and civil engineers working on soil consolidation identified similar approaches. Focus on flow of water relative to the solid, rather than solid movement relative to the water, also provides a unifying principle that extends the analysis to sedimentation in very wet particulate systems.

Elements of Theory

Water flow in unsaturated nonswelling soils is described by combining the Darcy law with an equation of continuity that accounts for the water in the system during steady and unsteady flow; the Richards equation results. These equations are formulated at the macroscopic scale and they apply to materials where the water potential and the hydraulic conductivity are well-defined functions of the water content.

Two distinct but complementary philosophies emerge to deal with soil profile volume change and the nature of flow in swelling soils. The first approach is mesoscopic and it focuses on behavior of individual soil-structural elements. It envisages water movement through macropores to or from these elements which then swell or shrink three-dimensionally to produce one-dimensional (vertical), macroscopic profile swelling or shrinkage. The approach is concerned with darcian or nondarcian flow in the fractures and three-dimensional volume change of aggregates, which ultimately results in vertical soil profile movement.

The second approach is macroscopic. It recognizes three-dimensional volume change of soil-structural units, but argues that the detail of this behavior is irrelevant and that we need concern ourselves only with the net one-dimensional, vertical volume change and flow that is observed macroscopically. It asserts that if the area of cross section is large enough and if there is no lateral net transfer of material from the control area, then behavior of structural entities beneath it may be described by a representative average vertical displacement. This approach is analogous to the Darcy scale approach to soil water movement in nonswelling soil, which also relates to volume averages and is agnostic about flow detail in individual pores and aggregates. It results in a one-dimensional analysis analogous to that of Richards.

Swelling and shrinking of individual structural elements need to be considered explicitly if the characteristic time of these mesoscopic processes approximates that of the macroscopic processes. In formulating a basic theory, it is presumed that, in general, they do not concur, and focus is on the macroscopic approach. That approach has been extensively tested with saturated materials and, to a lesser extent, with unsaturated soils. It exemplifies the principal issues of equilibrium and flow in a swelling material.

The practical challenges in applying the macroscopic approach are to:

1. Define a scale of discourse that permits meaningful measurements of macroscopic properties of the soil, such as water content and bulk density;
2. Meet Darcy-like requirements that the water potential and the hydraulic conductivity of the soil are well-defined functions of the water content.

To use the Richards equation, the potential and conductivity characteristics are again necessary but, in addition, a 'shrinkage curve' is needed to define the way the volume of soil per unit area of cross section, i.e., the elevation, changes with the soil water content.

Analysis is based on the material balance equations for the solid as well as the water, because both may be in motion relative to an observer, and on the Darcy law, which describes flow of water relative to the particles in response to a space gradient of total head. Both these elements are critical, but material balance must be retained even when the applicability of the Darcy equation is uncertain.

Material Balance in Swelling Systems

One-dimensional flow of water in a swelling system requires material balance equations for both the aqueous and solid phases:

$$\frac{\partial \theta_w}{\partial t} = -\frac{\partial F_w}{\partial z} \qquad [1]$$

$$\frac{\partial \theta_s}{\partial t} = -\frac{\partial F_s}{\partial z} \qquad [2]$$

where z is a distance coordinate, t is the time, F_w and F_s are the volume flux densities of the water and of the solid (cubic meters per square meter per second) relative to the observer, and θ_w and θ_s are the volume fractions of the water and solid. The volume fractions are defined per unit area of horizontal cross section of the soil and the reference volume includes the cracks.

Recognizing that the flux of water, seen by the observer, occurs relative to and with the moving soil solid:

$$F_w = u + \theta_w F_s/\theta_s = u + \vartheta F_s \quad [3]$$

where u is the volume flux of water relative to the particles, and the second term on the right describes transfer of water advected with the moving solid. The ratio of volume fractions, ϑ, is called the moisture ratio:

$$\vartheta = \theta_w/\theta_s = \theta_g \rho_s \quad [4]$$

with θ_g the water mass fraction (kilograms per kilogram) and ρ_s the specific gravity of the soil solid. In saturated soils, ϑ is equal to the void ratio, e.

Eqns [1], [2], and [3] may be manipulated to derive a material balance Eqn [5] for the water based on a solid-based, space-like material coordinate, $m(z, t)$, viz:

$$\frac{\partial \vartheta}{\partial t} = -\frac{\partial u}{\partial m} \quad [5]$$

For one-dimensional flow, the coordinate, $m(z, t)$, is formally defined by:

$$dm(z,t) = \frac{\partial m}{\partial z}dz + \frac{\partial m}{\partial t}dt = \theta_s dz - F_s dt \quad [6]$$

This definition of $m(z, t)$ satisfies the material balance Eqn [2] for the solid and reflects the distribution of the solid in space. Furthermore, $m(z, t)$ is independent of soil displacement that might accompany water content change and can be determined by integrating Eqn [6]. This integration is simplified if it is based on a surface, e.g., $z = 0$, where $F_s = 0$. In the absence of sedimentation or erosion, the soil surface provides such a datum. Then:

$$m = \int_0^z \theta_s dz - \int_0^t F_{s_{z=0}} dt = \int_0^z \theta_s dz = \int_0^z (\rho/\rho_s) dz \quad [7]$$

where m is the cumulative volume of solid, per unit area of cross section, measured from $z = 0$ or the cumulative oven-dried mass per unit area measured away from the soil surface and divided by ρ_s. Thus m is, for example, the cumulative amount of oven-dried solid per unit area measured from the top of a continuous field core sample taken for water content measurement. The final integral in Eqn [7] shows how m is related to the soil specific gravity profile, $\rho(z)$.

Flux Laws in Swelling Systems

The Darcy law describes flow of water relative to the solid in response to a space gradient of the total potential, Φ, of the water in the soil:

$$u = F_w - \vartheta F_s = -k(\theta_w)\frac{\partial \Phi}{\partial z} \quad [8]$$

in which the hydraulic conductivity, $k(\theta_w)$, is a function of the water content but also of the structure. Here it is identified as a function just of θ_w. In nonswelling soils the total potential, Φ, consists of a gravitational component and a capillary component, the latter arising as a result of the interaction of the water with the soil surfaces and their geometry. If Φ and its components are defined as work per unit weight of water, they have dimensions, L, and units meters of H_2O. The hydraulic conductivity, $k(\theta_w)$, then has dimensions LT^{-1} and units meters per second.

Defined this way, Φ is identical to the engineer's hydraulic head. Specifically:

$$\Phi = z + p_w \quad [9]$$

with the gravitational head, z, the elevation of the point of concern in the soil above a convenient datum, and p_w the pressure in the water relative to the pressure of the ambient gas phase measured with a manometer. Eqn [9] applies to both swelling and nonswelling materials. In unsaturated, nonswelling soils, the pressure head, p_w, is negative and reflects only the capillary potential.

Water Potential in a Swelling Soil

The components of Φ in swelling systems have been considered from a thermodynamic perspective and mechanistically. The latter approach argues that Φ represents the work involved in transferring unit weight of water from a reference flat surface of water at atmospheric pressure to a swelling material at some height, z, relative to that reference. Thus Φ is the sum of:

1. The gravitational potential, z, representing elevation of the element of water from the datum to the height z;
2. The unloaded matric or capillary potential, ψ, representing interaction between the water and the soil solid surfaces and their geometry;
3. The overburden potential, Ω, representing vertical displacement of the wet soil accompanying unit change in water content at z.

If the soil is saturated, unit change in water content produces unit change in elevation of the soil and, if the volume of the system is parameterized by the void ratio, e, this implies that $de/d\vartheta = 1$. If the soil is unsaturated, $de/d\vartheta$ tends to be less than 1, unit change in water content does not produce unit height change, and the work involved in elevating the wet profile is

less than it would be in the saturated case. The total potential, Φ, is then defined by the equation:

$$\begin{aligned}\Phi &= z + p_w = z + \psi + \Omega \\ &= z + \psi + \alpha\left(\int_z^T \gamma dz + P\right)\end{aligned} \quad [10]$$

The unloaded capillary potential ψ is the potential a tensiometer would measure at the existing moisture ratio in the absence of overburden. In Eqn [10], γ is the wet specific gravity of the soil, P is any static load on the soil surface, and α reflects the degree to which the soil is elevated by unit change in water content. In a saturated soil, $\alpha = 1$; when the soil is unsaturated $1 \geq \alpha$; and in a nonswelling soil, $\alpha = 0$. In civil engineering contexts, P may be very important, but we neglect it here without any loss of generality. Notice that, in Eqn [10], p_w includes the overburden component and may be positive.

The overburden, Ω, is related to the civil engineers' effective stress according to the equation

$$\Omega = \alpha\sigma = \sigma' + p_w \quad [11]$$

in which σ is the total normal stress and σ' is the effective or interparticle stress. Comparison of Eqns [10] and [11] shows that $\sigma' = -\psi$, and hence is related to ϑ through $\psi(\vartheta)$.

The nature of α has generated some controversy. At one time civil engineering practice equated it with the slope of the shrinkage curve. A more correct but more complicated definition followed some argument in which it was accepted that α must be a value of $de/d\vartheta$ averaged from zero load to that actually experienced at the point in question. The mathematical consequences of this definition are cumbersome, however, and it seems to be agreed that the civil engineering approximation remains the most practicable way forward.

It has been argued that, because the manometric pressure, p_w, is measurable, while ψ, an 'unloaded' capillary potential, is experimentally inaccessible, the water content should be defined as a function of p_w rather than ψ. The world remains indifferent to the issue. Figure 1 compares water content profiles and components of the total potential of the soil water for idealized nonswelling and swelling soils.

Potential Gradient

It suffices that the driving force in the Darcy equation is the space gradient of hydraulic head (or total potential), which is the sum of the manometric pressure (which may be negative relative to atmospheric pressure) plus a gravitational component relative to a convenient height datum. In this approximation, and if $\vartheta(\psi)$ is single-valued, $\partial\Phi/\partial z$ from Eqn [10] becomes:

$$\begin{aligned}\frac{\partial\Phi}{\partial z} &= \frac{\partial}{\partial z}\left(\psi + z + \alpha\int_z^T \gamma dz\right) \\ &= \theta_s\left(\frac{\partial\psi}{\partial m}\right) + (1 - \alpha\gamma) \\ &\quad + \theta_s\left(\frac{d\alpha}{d\vartheta}\right)\left(\int_m^0 \gamma/\theta_s dm\right)\left(\frac{\partial\vartheta}{\partial m}\right)\end{aligned} \quad [12]$$

In Eqn [12], the second equality on the right-hand-side is derived by differentiating the first equality by parts, and using the definition of m, i.e., $dm/dz = \theta_s$ to eliminate z.

Darcy Law and Hydraulic Conductivity

Substitution of the last of Eqns [12] in the Darcy law, in m-space, yields:

$$\begin{aligned}u = -k(\theta_w)\theta_s\Bigg(&\left(\frac{\partial\psi}{\partial m}\right) + \left(\frac{1 - \alpha\gamma}{\theta_s}\right) \\ &+ \left(\frac{d\alpha}{d\vartheta}\right)\left(\int_m^0 \gamma/\theta_s dm\right)\left(\frac{\partial\vartheta}{\partial m}\right)\Bigg)\end{aligned} \quad [13]$$

In Eqn [13], $k(\theta_w)\theta_s = k_m$ is identified as a 'material' hydraulic conductivity. Philip derived this equation 30 years ago, while Sposito derived the same equation from thermodynamic considerations.

The hydraulic conductivity, k_m, has conventionally been treated as a material characteristic parameterized by ϑ. However, it must be a function of both the structure, parameterized by e as well as the water content, ϑ, in a swelling material. No difficulty arises in saturated swelling systems where $e = \vartheta$, but, when $e > \vartheta$, the hydraulic conductivity must be represented by a surface, $k(e, \vartheta)$. This issue has escaped analysis hitherto and the magnitude of the effect needs to be established.

Flow Equation

Substitution of Eqn [13] in Eqn [5] yields an equation analogous to the Richards equation for a nonswelling soil. For saturated clays subject to large, constant imposed loads, gravity can be neglected and $\alpha = 1$. This latter condition characterizes many important practical chemical engineering situations. Numerical methods have been used to solve the full equation together with a nonlinear shrinkage curve to illustrate the solution for flow in unsaturated marine clays.

The principal physical differences, which distinguish equilibrium and flow in saturated swelling systems from those in nonswelling soils, arise from the overburden effect in these systems. Overburden

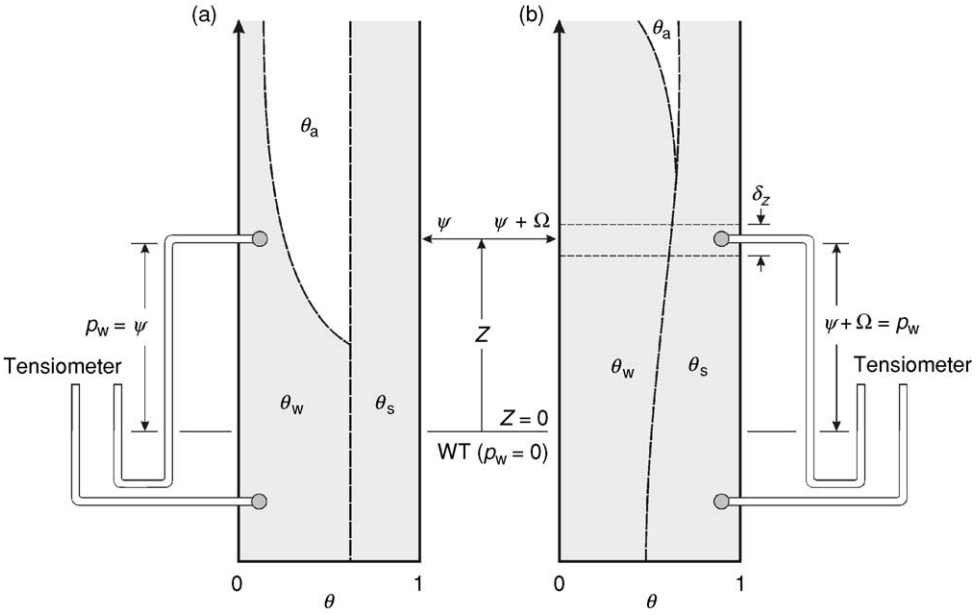


Figure 1 Comparison of the vertical, static equilibrium water content profiles in nonswelling soils (a) and swelling soils (b) in contact with a free water surface at atmospheric pressure. The *z*-coordinate is positive upwards, with its origin at the free water surface, and θ_w, θ_s, θ_a are the volume fractions of the water, solid, and air; they sum to unity. In the swelling soil, θ_s increases with depth toward some maximum density determined by the weight of the wet soil it supports and the water table height. A corresponding decrease is observed in θ_w. Desaturation in a swelling soil above the water table represents cracks and then desaturation of aggregates. The components of total potential, Φ, of the water in the two systems differ. Thus the arrows show the steps in transferring unit amount of water from the free water surface, WT, to the system at elevation *z* above it. Defined per unit weight of water, *z* is then the gravitational potential relative to the free water surface at WT; and p_w the pressure a manometer measures when inserted at height *z*. The manometric pressure is equal to ψ, the 'capillary' potential in the nonswelling soil, but is the sum of an unloaded 'capillary' or moisture potential and the overburden potential, Ω, which arises because of the vertical displacement of the wet profile, δz, that must accompany insertion of the water at height *z*. In the saturated region, unit displacement accompanies insertion of unit amount of water. In the unsaturated region of the swelling soil, the vertical displacement is less than unity. At static equilibrium, the manometer/tensiometer water level must have the same elevation as the water table. Reproduced from Smiles DE (2000) Hydrology of swelling soils: a review. *Australian Journal of Soil Research* 38: 501–521 (corrigendum 40: 1467).

reverses (and diminishes) the effect of gravity in systems where $\rho_s > 1$:

1. Infiltration into these systems resembles capillary rise in nonswelling soils;
2. Steady upward flow of water proceeds more readily in swelling than nonswelling soils;
3. Static equilibrium water-content profiles in the saturated region are drier at the bottom than at the top, in contrast to profiles in nonswelling soils;
4. The period of time for which a 'gravity-free' analysis applies to transient flow may be greatly extended;
5. The shape of the $\psi(\vartheta)$ characteristic together with a relatively weakly varying $k_m(\vartheta)$ often results in a material moisture diffusivity that decreases with increasing water content, so desorption in muds and slurries resembles absorption in nonswelling soil.

Overview

Water flow accompanied by volume change remains one of the most challenging areas of porous material physics. Some areas of application have been well tested, at least in one-dimensional flow, and theory has been usefully applied to a number of unit processes in chemical engineering and to consolidation in civil engineering. Examples of unit processes include filtration under constant pressure and constant rate conditions and effects of filter membranes that significantly resist flow. They also include centrifugation and centrifugal filtration of wet particulate slurries. Some of these analyses are based on well-known solutions of the Richards equation, but some required and tested solutions that had not previously been available in soil science. Constant-rate filtration and filtration through an impeding membrane exemplify the latter class of problem. Most of this work has involved saturated systems.

In soil science, however, one-dimensional theory remains incomplete and its test is uncertain. Curiously, it represents a bridging problem between well-founded and tested analysis of flow in unsaturated nonswelling systems and that in saturated swelling ones. In fact, Eqns [1]–[13] apply equally well to swelling and nonswelling materials, although application of the concepts and terminology of swelling soil theory to nonswelling soils is unusual. Thus, both 'limiting' systems are described well by the Richards equation and both demand that $\psi(\vartheta)$ and $k_{\mathrm{m}}(\vartheta)$ be well defined, although the swelling system requires a shrinkage curve also. The consequences of volume change in unsaturated soils, however, remain the challenge. It is useful to identify some important issues of concern.

Scale of Discourse

Because of the size of the structural units in swelling soils, the scale of discourse that permits application of an approach analogous to the approach of Richards for nonswelling soils, where properties represent measurable spatial averages, is much larger than is generally the case in nonswelling soils. As a result, the scale of measurement of water content and water potential requires careful consideration and it may be that many current methods fail to provide appropriate volume averages at the scales involved. The existence of cracks need not complicate measurement, however, provided the cracks are included in the representative volumes and areas of cross section. Water content defined per unit amount of solid, consistent with a solid-based, space-like coordinate, is a preferred basis for water balance in these soils. This convention ensures that volume change does not prejudice estimates of water content change.

Material Balance and Coordinates

In addition to requirements that material properties be consistent with the scale of the space-like coordinate if integral quantities such as the profile water content are to be interpreted without ambiguity, differences over time will be incorrect if the change in volume of a profile is ignored. Volume change accompanying water content change is irrelevant and material balance is exact, however, if the water content is expressed per unit amount of soil solid (mass or volume) and 'space' is defined as the cumulative amount to soil solid per unit area, measured, for example, from the soil surface. This approach extends to organic soils where the organic matter is being oxidized and to acid sulfate soils where drainage consolidates the profile as well as removing organic matter. The 'irreducible' mineral content of the soil provides a secure reference frame in these circumstances. It also provides an appropriate basis for intrinsic expression of the concentrations of water, organic matter, or sulfide.

Observations of consistent systematic behavior of heavy clay soils over almost 1000 days confirm that α of the soil profile is relatively constant at approximately one-third and is thus consistent with three-dimensional normal volume change of the soil aggregates. The effect is not greatly depth-dependent and permits quite accurate estimate of the soil water balance from surface displacement measures.

Water Flow

Water flow in swelling systems generally occurs in response to a space gradient of the total head. At the same time, preferential flow in macropores of cracked soils may dominate flow in the early stages of heavy rain or flood irrigation when the soil is very dry. Both phenomena and their relative importance have been considered in the mesoscopic analysis of flow, although a general approach has yet to be formulated. In addition, laboratory measurement must recognize and mimic the constraints experienced by swelling field soils, and procedures common in civil engineering may be obligatory in laboratory studies of flow and volume change in these soils. Early theory tended to identify these constraints carefully. Some recent experiments have been less careful and experimental artifacts inconsistent with theory have prejudiced analyses.

Multidimensional Flow

Application and test of theory have been confined largely to one-dimensional flow and volume change, although a long-available, three-dimensional analysis has also been used in some approximate analyses. Cracked systems have also been dealt with by using a single parameter to characterize the transversely isotropic deformation of the solid phase and combining this with one-dimensional, vertical flow of the aqueous phase. Experiments that model the constraints on, and test theory of, multidimensional flow remain elusive, however.

Measurement

Problems of measurement and its interpretation are endemic. For example, there has been concern about tensiometry and its interpretation in swelling systems with anisotropic constraints which have not been resolved. There is great need for laboratory and field measurement, with specific focus on noninvasive measurement of the time course of the spatial distributions of solid, liquid, and gaseous phases in the soil.

In the laboratory, civil engineers in particular systematically use equipment such as triaxial consolidation cells that are designed to characterize volume change with well-defined constraints in engineering soils. Their terminology and theory may not always be familiar to soil scientists but the technology offers much to soil physics. At the same time, many soil physical principles are unfamiliar to engineers, but each subdiscipline can learn from the other to the ultimate benefit of both.

Water Retention and Hydraulic Conductivity Functions in Unsaturated Swelling Soils

Both the water retention and hydraulic conductivity functions in swelling systems still present significant challenges for resolution. In relation to water retention, four parameters, e, ϑ, p_w, and P, of which three are independent, characterize an unsaturated, swelling porous material. The relations $e(P, \vartheta)$ and $p_w(P, \vartheta)$ have been preferred, mainly because of the wealth of experimental data that exists in relation to $e(P, \vartheta)$; this approach results in the notion of an unloaded moisture potential, as shown in Eqn [10]. The nature of the hydraulic conductivity function remains to be resolved. For nonswelling soils, hydraulic conductivity seems to be a nonhysteretic function of the volumetric water content, θ_w, with the fixed volume fraction of the solid, θ_s, as a parameter. Analogy suggests that the hydraulic conductivity, k, for an unsaturated swelling system must be similarly defined in a $k-e-\vartheta$ diagram. Few, if any, data are available to illustrate the nature of this surface. At the same time, there is an extensive data set for saturated montmorillonite in equilibrium with a range of solution salt concentrations. These data confirm that $k_m(\vartheta)$ is not unique in pure clays but reflects colloid structures associated with the cation suite and concentrations with which the clay is in equilibrium. At the same time, these data reveal strong unifying benefits across systems of different solution salt concentrations, when k is related to ψ. These observations await systematic study and explanation.

Summary

Water flow in swelling soils is described within a macroscopic, one-dimensional theory because field volume change, in the large, is vertical. The theory accounts for material balance of both the solid and liquid phases of the system, both of which tend to be in motion. It also recognizes that the total flux of water has a darcian component relative to the solid as well as a component advected with the moving solid. The theory identifies an overburden component of the water potential in addition to those due to gravity and capillarity. Material properties required by the approach require unambiguous measurement at the scale of application. Several issues of theory and measurement remain unresolved and skills currently available in civil and chemical engineering would benefit the science greatly.

List of Technical Nomenclature

α	Load factor
γ	Specific gravity of wet soil
θ_s	Volume fraction of solid
θ_w	Volume fraction of liquid
ϑ	Moisture ratio
ρ	Soil specific gravity
ρ_s	Soil solid specific gravity
σ	Total stress (L)
σ'	Effective stress (L)
Φ	Total water potential (L)
ψ	Capillary potential (L)
Ω	Overburden potential (L)
e	Void ratio
F_s	Volume flux density of solid (LT^{-1})
F_w	Volume flux density of water (LT^{-1})
k	Hydraulic conductivity (LT^{-1})
k_m	Material hydraulic conductivity (LT^{-1})
m	Solid-based space-like coordinate (L)
p_w	Manometric pressure (L)
t	Time (T)
u	Volume flux of water relative to particles (LT^{-1})
z	Vertical distance and gravitational potential (L)

See also: **Clay Minerals**; **Stress–Strain and Soil Strength**; **Structure**

Further Reading

Ahmad N and Mermut M (eds) (1996) *Vertisols and Technologies for their Management. Developments in Soil Science*. Amsterdam, the Netherlands: Elsevier.

Bouma J and Raats PAC (eds) (1984) *Proceedings of ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*. International Institute for Land Reclamation

and Improvement Publication No. 37. Wageningen, The Netherlands: International Institute for Land Reclamation and Improvement.

McGarity JW, Hoult EH, and So HB (eds) (1984) *The Properties and Utilisation of Cracking Clay Soils. Reviews in Rural Science 5*. Armidale, Australia: University of New England Press.

McNabb A (1960) A mathematical treatment of one-dimensional consolidation. *Quarterly of Applied Mathematics* XVIII: 337–347.

Philip JR (1968) Kinetics of sorption and volume change in clay-colloid pastes. *Australian Journal of Soil Research* 6: 249–267.

Philip JR (1969) Hydrostatics and hydrodynamics in swelling soils. *Water Resources Research* 5: 1070–1077.

Philip JR (1970) Hydrostatics and hydrodynamics in swelling soils: reply to Youngs and Towner. *Water Resources Research* 6: 1248–1251.

Philip JR (1995) Phenomenological approach to flow and volume change in soils and other media. *Applied Mechanics Review* 48: 650–658.

Raats PAC (1998) Kinematics of subsidence of soils with a non-conservative solid phase. *Proceedings of a Symposium on New Concepts and Theories in Soil Physics and Their Importance for Studying Changes Induced by Human Activity*. The 16th World Congress of Soil Science, Aug. 20–26 1998, Montpellier, France. CD-ROM.

Raats PAC (2002) Flow of water in rigid and non-rigid, saturated and unsaturated soils. In: Capriz G, Ghionna VN, and Giovine P (eds) *Modeling and Mechanics of Granular and Porous Materials*, pp. 181–211. *Modeling and Simulation in Science, Technology and Engineering*. Boston, MA: Birkhäuser.

Raats PAC and Klute A (1968) Transport in soils. *Soil Science Society of America Proceedings* 32: 161–166 and 452–456.

Smiles DE (1986) Principles of constant-pressure filtration. In: Cheremisinoff NP (ed.) *Encyclopedia of Fluid Mechanics*, pp. 791–824. Houston, TX: Gulf Publishing Company.

Smiles DE (2000) Hydrology of swelling soils: a review. *Australian Journal of Soil Research* 38: 501–521 (corrigendum: 40: 1467).

Sposito G (1975) Steady vertical flows in swelling soils. *Water Resources Research* 11: 461–464.

Tempany HA (1917) The shrinkage of soils. *Journal of Agricultural Science* 8: 312–330.

TEMPERATE REGION SOILS

E A Nater, University of Minnesota, St. Paul, MN, USA

Introduction

A multiplicity of definitions for temperate regions exist, with a central concept that these regions have continental climates with significant seasonal differences in temperature. The US Department of Agriculture–Natural Resources Conservation Service (USDA–NRCS) defines temperate regions, and hence temperate region soils, as having mesic, thermic, or hyperthermic temperature regimes; iso regimes (those experiencing less than a 6°C annual difference in soil temperature at 50 cm depth) are normally excluded. Although aridic and xeric moisture regimes are often considered to occur in temperate regions, mainly subhumid (ustic) and humid (udic and perudic) temperate region soils are discussed here.

Temperate regions, and hence temperate region soils, fall mainly in the mid- to lower-latitude regions of the globe. The length and temperature of the growing season are sufficient for growth of annual grain and oilseed crops. In humid and subhumid temperate regions, precipitation is sufficient for growth of crops without irrigation, though there may be a dry season toward the end of the growing season in the drier parts of this region, thus necessitating early-maturing crops such as winter wheat. The majority of lands cultivated by intensive, mechanized agricultural practices occur in temperate regions.

Temperate regions are bounded by boreal regions, where soil temperatures are lower and the growing season is shorter, and by tropical regions, where temperatures are warmer and frosts rarely, if ever, occur. Humid and subhumid temperate regions are bounded by arid regions, where moisture is insufficient to support the growth of most crops without irrigation, and Mediterranean regions, where the majority of precipitation occurs during the cool season and the growing season is generally very dry. Overall, temperate regions constitute almost 40% of the Earth's ice-free terrestrial surface (Table 1).

State Factors

Climate

Climate is the initial bounding parameter for any discussion of temperate regions. Temperate regions encompass gradients in both moisture and temperature, but also in atmospheric moisture deficit, which has profound effects on evapotranspiration and consequently on soil moisture status and soil properties. In addition, the relative timing between annual temperature cycles, wet and dry cycles, and atmospheric moisture deficit are also important determinants for the plant communities that thrive in these landscapes. Climate strongly influences the development of soil properties, both directly through its effect on soil moisture status and indirectly through its effect on organisms.

Other factors being equal, soils in wetter parts of the region tend to have deeper sola and experience stronger leaching, lower base saturation, losses of

Table 1 Areal distribution of soil orders worldwide and in temperate regions

	Area		*Proportion*	
	Earth's surface (10^6 km^2)	*Temperate regions (10^6 km^2)*	*Earth's surface (%)*	*Temperate regions (%)*
Gelisols	11.26	–	8.61	–
Histosols	1.53	0.121	1.17	0.23
Spodosols	3.35	0.592	2.56	1.14
Andisols	0.91	0.202	0.7	0.39
Oxisols	9.81	0.197	7.5	0.38
Vertisols	3.16	1.642	2.42	3.17
Aridisols	15.73	11.560	12.02	22.29
Ultisols	11.05	3.405	8.45	6.57
Mollisols	9.01	4.491	6.89	8.66
Alfisols	12.62	4.868	9.65	9.39
Inceptisols	12.83	5.383	9.81	10.38
Entisols	21.14	14.355	16.16	27.68
Shifting sands	5.32	4.68	4.07	9.02
Rock	12.08	0.363	10.0	0.70
Totals		51.865	39.65	100.00

soluble salts, gypsum, and carbonates, and higher rates of mineral weathering. Soils in drier parts of the region, particularly those with high seasonal atmospheric moisture deficits, tend to have shallower sola, less leaching, and consequently higher base saturation, and are more likely to have accumulations of secondary carbonates and gypsum. Salts more soluble than gypsum tend to accumulate only in hydrologic discharge areas in these regions, though their accumulation is more common in aridic moisture regimes. Drier soils also tend to have lower rates of mineral weathering.

Likewise, soils in warmer parts of the region tend to be more weathered, have lower base saturation, and typically have less organic matter accumulation in the A horizon due to increased rates of organic matter decomposition. Increases in soil temperature enhance the rate of many chemical and biochemical pedogenic processes, particularly soil mineral weathering. Typically, a 10°C increase in temperature will produce a two- or three-fold increase in the rate of many biochemical reactions.

Soils in cooler parts of the region tend to have less mineral weathering, higher base saturation, and higher organic matter contents in the A horizon. These trends are reinforced by the occurrence of large areas of younger parent materials in the cooler parts of the region due to direct or indirect effects of continental glaciation, and also by the fact that the soil surface may be frozen for a significant portion of the year in colder climates, thus reducing the duration of action of many biologically and hydrologically driven pedogenic processes.

The balance between precipitation (or, more correctly, infiltration) and evapotranspiration controls the balance between leaching and chemical precipitation, and thus accumulation of carbonates, gypsum, and more soluble salts. Where precipitation exceeds evapotranspiration, leaching of relatively soluble materials occurs as water in excess of field capacity can move out of the system in response to gravity. Over time, leaching will remove all soluble salts, carbonates, and gypsum from soils, even those formed on carbonate bedrock.

Where precipitation is less than evapotranspiration, soil moisture status is less likely to exceed field capacity, and less water is available to be transported out of the solum. Consequently, carbonates, gypsum, and soluble salts will be moved downward in the profile with the wetting front, but may not be leached entirely out of the solum. Over time, horizons of carbonate, gypsum, or even soluble salts may form under horizons that have lost most, if not all, of the carbonates, gypsum, or soluble salts by leaching. Because of their high solubility, and hence mobility, salts more soluble than gypsum are easily removed from soils.

This simple analysis should be taken only as a rough guide; episodic precipitation events, snow melt, noncoincidence of precipitation and high evapotranspirative demand, and even bypass flow may all produce significant leaching even when annual total precipitation is much less than evapotranspiration.

Parent Materials

Because of their large areal extent, temperate region soils encompass a broad array of parent materials, ranging from residuum to glacially derived sediments to loess, alluvium, and colluvium. Although this broad array exists, a few generalities can be observed.

In the colder parts of temperate regions, particularly in the northern hemisphere, a significant proportion of parent materials were derived either directly or indirectly from processes associated with late Quaternary glaciation. Many of these soil landscapes formed on glacial drift, glaciofluvial, or glaciolacustrine sediments. Extraglacial areas near glacial margins were strongly affected by periglacial processes and may have experienced extensive wind erosion or fluvial reworking. Still other landscapes, often far removed from the glacial margins, were blanketed by thin-to-thick loess deposition derived mainly from riverine sediment sources associated with the wastage of continental ice sheets. Landscapes formed on recent glacial sediments have deranged (disconnected) drainage networks with numerous closed depressions and poor internal drainage.

In warmer parts of the region, soil landscapes not affected by recent glaciation or loess deposition form mainly on residuum, alluvium, or colluvium. Properties of these parent materials are closely related to local bedrock sources and thus may vary widely from one region to another. Soil landscapes in warmer temperate regions may show considerably more development and weathering due to the more stable nature of the surfaces, the longer duration of their exposure to pedogenic processes that alter them, and warmer soil temperatures. These landscapes generally have evolved highly dissected, well-connected, dendritic surficial drainage networks with few closed depressions.

The majority of andic parent materials are associated with continental margins, and particularly the Pacific rim. Restriction of this discussion to humid and subhumid temperate regions eliminates xeric, Mediterranean climates strongly affected by oceanic thermal masses, and thus effectively excludes nearly all of the soils dominated by andic parent material. The remaining parent materials, then, were derived from crystalline rocks.

Organisms

The majority of landscapes in humid and subhumid temperate regions are dominated by one of two ecosystem types: prairie or steppe, and deciduous or mixed deciduous–coniferous forest.

Low-relief landscapes, particularly in the drier parts of the region, are dominated by prairies or steppes, which range from shortgrass prairies, with a significant annual grass component in drier temperate regions, to tallgrass prairies, dominated by perennial bunch grasses in wetter temperate regions. Prairies can withstand and more readily regenerate from frequent fires than most forests can. Fires started in low-relief landscapes are often windswept and can cover very large areas, whereas higher-relief landscapes tend to slow or stop the spread of fires.

Deciduous or mixed deciduous–coniferous forests are also common in humid and subhumid temperate regions, particularly in areas of higher relief. They transition to coniferous or broadleaf evergreen forests in the warmer parts of the region, and to coniferous forests near the boreal region border.

Other ecosystem types such as spruce bog or fen, and savanna, are interspersed within temperate regions, but the prairie and forest general types predominate. Bounding this region are chaparral or desert scrub ecosystems in more arid regions; savanna, chaparral, or annual grasslands in Mediterranean regions; coniferous forests in boreal regions; and broadleaf evergreen forests or savannas in the tropics. These ecosystem types strongly affect soil properties.

The formation of thick, dark, organic-rich A horizons (mollic epipedons) is common under grassland- or savanna-type vegetation. Grasses have a deep, fibrous root distribution and a rapid root turnover rate (approximately one-third of grass roots die and are replaced annually) which, in combination, produce significant inputs of organic matter at depth, requiring no translocation to move the organic matter deep into the solum. In addition, the efficient base-cycling present in grassland ecosystems leads to high base saturation, a requirement for mollic horizon classification. Soils with mollic epipedons typically do not have identifiable E horizons owing to the lack of strong leaching in subsurface horizons and to the depth of the A horizon which masks them if present. Notable exceptions occur in Albolls, where distinct albic E horizons are present, often bisecting the mollic epipedon into a surficial and a subsurface portion.

Deciduous and mixed deciduous–coniferous forests typically have roots that are distributed much closer to the soil surface and have a much lower root-turnover rate. The majority of organic inputs in forests are from leaf litterfall, which occurs at the soil surface. Incorporation of leaf litter into the soil requires the activity of worms or other invertebrates. Where leaf materials are not incorporated, the soil typically develops a thin-to-thick layer of partially decomposed organic materials at the soil surface, termed the forest floor or the litter layer. This O horizon can be several centimeters thick and usually rests on an underlying E horizon formed by eluviation of clays and chelation, and subsequent leaching of iron and other coatings from the surface of silt and clay grains in the matrix. It may also rest on a thin A horizon. If earthworms or other invertebrates incorporate the forest floor into the underlying mineral soil horizons, then a relatively thin, light-colored A horizon (ochric epipedon) is typically observed. Introduction of nonnative earthworm species into forested landscapes in the northern USA has all but eliminated the thick O horizons once commonly observed in many forests.

Likewise, B horizon development is also strongly related to the organism factor. B horizon development is often minimal in prairie soils. Many have no B horizon at all (A/C morphology) or else have cambic B horizons that display a moderate degree of development. Older, more well-developed prairie soils may have argillic or natric horizons, but they are not nearly as common under prairie as they are under forests. Several factors tend to inhibit clay dispersal and movement in prairie soils, including the presence of organic coatings in the A horizon that tend to bind clays together; higher exchangeable Ca contents and higher pH generally present in grasslands, which promote flocculation; and the lack of stemflow common to deciduous forests, which tends to concentrate a high proportion of the precipitation falling on a tree directly under its bole, thus producing a localized effect similar to greatly increased precipitation.

Soils formed under deciduous or mixed deciduous–coniferous forest typically display prominent B horizon development. Argillic horizons are commonly observed under forest, with kandic horizons present in the more highly weathered soils or in those that formed on acidic parent materials; spodic horizons occur in coarse-textured soils dominated by coniferous or other vegetation that contributes large amounts of soluble fulvic acids to the soil surface.

Soils under coniferous forests may have spodic B horizons, resulting from podzolization processes. Coniferous leaf litter is more acidic than that of most deciduous forests and releases large quantities of fulvic acids, which are soluble in acidic environments. These organic compounds can chelate Fe, Mn, and Al from metal (hydr)oxide coatings on the soil

grains. In relatively coarse-textured soils, the fulvic acid–metal chelates are transported with the soil solution as long as they remain soluble, producing a distinct E horizon where metal (hydr)oxides have been removed. The chelated metals are deposited and accumulate lower in the profile due to changes in pH or other factors, producing distinctive red or black horizons of accumulation of Fe, Mn, and Al chelates.

Topography

Soil properties and soil-landscape relations are strongly affected by topography, in large part because topography exerts significant control on soil moisture relations and on erosional and depositional processes. There is a clear interaction between rates of infiltration and precipitation that strongly influences these relationships. Where the infiltration rate is lower than the rate of precipitation, runoff will occur. Runoff will be minimal on relatively flat summit positions due to their low slope, thus maximizing time for infiltration to occur. Runoff increases on shoulder and backslope positions, where slopes are higher, thus decreasing infiltration. These landscape positions are also the most susceptible to erosion by water due to the higher runoff and steeper slopes. Water and materials transported from the upper landscape positions accumulate on footslope and toeslope landscape positions. These are often the wettest landscape positions and are sites for deposition of eroded sediments. In more arid regions, they may also be sites for accumulation of soluble salts leached from higher landscape positions.

Soils in summit positions often show relatively more development than soils in other landscape positions due to increased infiltration and limited erosion in the summit position. These soils typically have deeper sola and are more leached than other soils in other landscape positions.

Infiltration into soils on shoulder and backslope landscape positions is proportionally lower than on summits due to increased runoff. Soils in these landscape positions are generally less developed than others in the same landscape, owing in part to their lower effective moisture status and to erosion. Erosion in these landscape positions is high because of the steeper slopes and the volume of runoff that occurs on them. These soils usually have limited B horizon development, thinner A horizons, and shallower sola than soils at the summit.

Soils in footslope and toeslope positions accumulate water and materials that were transported from above. In drier climates, these soils may show the greatest degree of development because they receive the most water. In wetter climates, footslope and especially toeslope landscape positions may be seasonally or permanently saturated. They may also be sites for accumulation of eroded sediments or soluble salts. Often the sola are very thick owing to addition of eroded materials.

If the rate of infiltration exceeds that of precipitation, runoff will not occur. If soil moisture exceeds field capacity, it will move in response to gravity. Depending on the presence or absence of subsurface strata or horizons that restrict the flow of water, water will either move downward to groundwater or may move laterally through the soil or underlying sediments by throughflow or lateral flow. Far more detailed hydrologic analyses are available now using digital terrain modeling and flownet analyses.

Landscape drainage patterns can significantly affect soil landscapes and the catenary relationships between soils. They have less effect on soils in upper landscape positions, more in lower landscape positions.

Relatively young landscapes typically have poorly dissected, deranged drainage patterns, where surficial connectivity is limited, drainage is mostly internal, and numerous closed depressions exist that collect water during the wet season and form seasonal-to-permanent depressional wetlands with hydric soils and wetland vegetation. This is particularly true in glacial landscapes, where the hummocky terrain commonly associated with glacial drift produces numerous closed depressions. The hydrology of landscapes with deranged drainage patterns is often quite complex. Soil properties are strongly affected not only by landscape position, but also by position with respect to recharge or discharge hydrology. Soils in footslope and toeslope positions are generally saturated for significant periods during the year and experience alternating reducing and oxidizing environments. The soils commonly have high organic matter contents and display a variety of redoximorphic features. Wetland rims may be water discharge sites and can sometimes contain significant salt, gypsum, or carbonate accumulations.

As landscapes evolve and mature, their surficial drainage patterns become more highly dissected, intersecting and draining closed depressions. They develop dendritic drainage patterns that have higher-order streams, increasingly more surficial connectivity, and fewer closed depressions. Lakes, ponds, and depressional wetlands are less common. The landscapes have more surficial and less internal drainage.

Significant differences between deranged and dendritic landscapes occur in the footslope and toeslope positions. In deranged landscapes, toeslopes are typically seasonal or permanent wetlands, with high concentrations of poorly decomposed organic materials, at least seasonal anaerobic conditions, and accumulation of erosional materials. Footslope

morphologies may be quite different depending on the hydrologic flow associated with them.

Recharge hydrology (soil water moves vertically to the groundwater) in wetlands is associated with leaching of soluble materials and removal of colloids from horizons or soils. Argillic horizon formation is a common result of recharge hydrology. Discharge hydrology (water moves from the groundwater table to or near the soil surface) is associated with chemical precipitation and is best observed in semiarid regions where wetland rims may display high salt concentrations resulting from capillary rise from saline groundwaters. Seasonal flow reversals are also common.

In dendritic landscapes, footslope and toeslope landscape positions often terminate in or near seasonal-to-perennial streams or alluvial floodplains that seasonally accumulate sediments. These landscape positions may also experience seasonal or permanent saturation, but it is less common than in deranged drainage systems.

Soils subjected to seasonal-to-permanent flooding or saturation can develop strongly anaerobic conditions because the rate of oxygen diffusion in the aqueous phase in soils is much lower than it is in the gas phase. If oxygen demand in the soil, generated mainly by root and microbial respiration, exceeds the rate of oxygen diffusion into the soil, reducing conditions can develop. Reduction requires both saturation of the soil and sufficient microbial activity to deplete oxygen from the soil so that alternate electron acceptors such as Fe^{3+} become favored. Saturation for insufficient duration or when microbial respiration is very slow (low soil temperatures or lack of suitable microbial substrate) will not produce sufficient oxygen demand in soils to consume free oxygen and thus will not produce reducing conditions or redoximorphic features.

Reducing conditions can lead to the dissolution of Fe and Mn (hydr)oxides, transport of reduced Fe^{2+} and Mn^{2+} with soil solution, and reprecipitation of Fe and Mn (hydr)oxides in other locations in the soil where pH and redox conditions favor precipition. Both Fe^{3+} and Mn^{4+} (hydr)oxides are very insoluble; the reduced forms (Fe^{2+}, Mn^{2+}), however, are very soluble in the aqueous phase and thus can be transported with the soil water and either removed entirely from a location in the soil, thus producing redox depletions, or moved to a different zone or region in the soil, where they are reoxidized to form redox concentrations. Because even small amounts of Fe and Mn coatings can mask the color of the underlying soil matrix, their removal provides a clear visual indicator of redoximorphic processes.

Saturated soils typically have higher organic matter contents than soils in more aerobic environments due to lower rates of organic-matter decomposition under anaerobic conditions. Histic epipedons may develop over mineral soil surfaces under these conditions. If the duration of saturation extends throughout most of the year, Histosols (organic soils) may develop.

Age

Soil age refers to the time since exposure of a surface and, as such, is a measure of the length of time a soil or soil landscape has been affected by pedogenic processes. Soil surfaces in whole regions may be exposed by processes such as continental glaciation or loess deposition, or over much smaller areas as a result of smaller-scale processes such as hillslope erosion or deposition.

Initially, soil properties strongly reflect the properties of the parent materials on which they form. Over time, they are altered by pedogenic processes and, given sufficient time, may little resemble their initial condition. Formation of horizons takes periods ranging from a few hundred years (formation of A horizons) to millennia (argillic, spodic, calcic horizons), to hundreds of thousands or even millions of years (oxic horizons).

Mineral fractions in young soils are derived from the parent materials and are largely unaltered. As soils age, mineral weathering occurs and secondary minerals begin to form. Over time the mineral fraction may be significantly altered from its initial state and be more closely related to weathering products controlled by thermodynamic stability fields. Over sufficiently long periods of time, soils with relatively different parent materials may end up with relatively similar mineralogies and compositions. Kaolin group minerals, which form by intense weathering of primary silicate minerals, often dominate the clay mineral fraction in highly weathered soils. In high-intensity, highly leached, base-poor environments, less thermodynamically stable clays such as illites and smectites are removed by chemical weathering and disappear entirely from older landscapes unless protected by microenvironments that are more suitable for their stability.

Effects of Humans

Humans also affect soils and their properties, particularly through agricultural and construction activities, but also through more subtle means. Temperate region soils have undergone extensive change due to the high density of intensive, mechanized agriculture. Agricultural tillage has caused accelerated erosion on many landscapes by leaving soil surfaces poorly protected against raindrop impact. Tillage and agricultural drainage of wetlands or seasonally wet soils have increased aeration of the solum and enhanced the rate of organic matter oxidation, releasing vast

quantities of CO_2 to the atmosphere. Many agricultural soils in temperate regions have lost more than 50% of their precultivation organic matter contents, with resultant losses in tilth and native fertility. Soil acidification is another widespread effect, resulting both from atmospheric deposition of nitric and sulfuric acid related to fossil fuel burning, but also from the application of nitrogen fertilizers.

Classification

Table 1 provides a generalized view of the areal extent of soils by US taxonomic order. The definition of 'temperate region' used in Table 1 is broadly defined to include both aridic and xeric moisture environments.

The dominant soil orders in temperate regions, as broadly defined, are Entisols and Aridisols, which comprise 27.7% and 22.3% of the total area, respectively. Other common orders include Inceptisols (10.4%), Alfisols (9.4%), Mollisols (8.7%), and Ultisols (6.6%). Nonsoil surfaces (shifting sands and rock) make up almost 10% of the total area of temperate regions. The remaining orders in toto constitute approximately 6% of the region.

Similar data portraying the areal extent of soil orders only in humid and subhumid temperate regions are not readily available. Because of the uncertainties in distribution of many of these orders between humid and subhumid and aridic and xeric moisture regimes, it is difficult to provide accurate relative distributions of these orders. However, some trends in soil distribution in these regions can be estimated from knowledge of soil distributions in aridic and xeric moisture regimes. Excluding soils from aridic moisture regimes would eliminate all Aridisols and nearly all of the shifting sands, along with large areas of Entisols. All of these soils and/or terrestrial surfaces are common in aridic moisture regimes. Mollisols, Vertisols, and Oxisols are also present in aridic moisture regimes, but constitute a smaller proportion of the total areas. Alfisols and Inceptisols are, by definition, excluded. It is more difficult to make generalizations about the areal extent of soil orders in xeric moisture regimes as compared to humid and subhumid temperate regimes, as all of the orders common to one region are also present in the other, though in varying proportions. The only clear distinction is that the majority of Andisols are located along coastal margins, particularly around the Pacific Rim; excluding xeric moisture regimes from the discussion would effectively exclude nearly all Andisols also.

For the most part, humid and subhumid temperate regions are dominated by Inceptisols, Entisols, Alfisols, Mollisols, and Ultisols. Other orders are present but generally constitute a smaller relative portion of humid and subhumid temperate regions.

Some general trends for the distribution of soil orders in humid and subhumid temperate regions follow. Alfisols and Ultisols are mainly associated with forested landscapes and Mollisols with prairies. Ultisols occur in warmer, wetter, and more stable landscapes of the region where soils have weathered more and base saturations are consequently lower. Alfisols are more common under forests in cooler, drier climates and younger landscapes where weathering, leaching, and removal of bases is not as extensive. Spodosols occur near the boundaries with boreal and tropical regions, where they are much more common.

Entisols and Inceptisols are interspersed throughout the region on erosional or depositional landscape positions, particularly shoulder and backslopes, where soil development is limited or new parent materials have been exposed or deposited. They are also common in wetter portions of the landscape where soil development processes are slow. Histosols occur in closed depressions and other very poorly drained landscape positions, particularly in glaciated regions, where closed depressions abound. Vertisols develop in clayey soils with high shrink–swell clays and seasonally wet and dry periods. They are common in alluvial and lacustrine settings, or form on decomposing shales, and are more common in the drier portions of the region.

List of Technical Nomenclature

Deranged drainage	Drainage patterns characterized by a lack of surficial connectivity and the presence of numerous closed depressions. Typically associated with relatively young landscapes
(Hydr)oxides	A term used to describe, in aggregate, oxides, hydroxides, and oxyhydroxides of metals, particularly iron and manganese, in soils. Used particularly when one is referring to materials which may fall in any of those three classes

All taxonomic terms and soil diagnostic horizon terms are from USDA–NRCS *Soil Taxonomy*.

See also: **Cold-Region Soils**; **Mediterranean Soils**; **Tropical Soils:** Arid and Semiarid; Humid Tropical

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. London, UK: Oxford University Press.
Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 4th edn. Ames, IA: Iowa State University Press.
Daniels RB and Hammer D (1992) *Soil Geomorphology*. New York: John Wiley.
Jenny H (1980) *The Soil Resource*. New York: Springer-Verlag.
Ruhe RV (1975) *Geomorphology*. Boston, MA: Houghton Mifflin.
Soil Science Society of America (1997) *Glossary of Soil Science Terms*. Madison, WI: Soil Science Society of America.
Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. A basic system of soil classification for making and interpreting soil surveys. USDA Soil Conservation Service, Agricultural Handbook No. 436. Washington, DC: US Government Printing Office.
Wright HE Jr. (ed.) (1983) *Late-Quaternary Environments of the United States*, vol. 2. The Holocene. Minneapolis, MN: University of Minnesota Press.

Temperature Regime *See* **Thermal Properties and Processes**

TENSIOMETRY

T K Tokunaga, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Introduction

The practical application of equilibrium and non-equilibrium concepts to the hydrostatics and hydrodynamics of soil water in laboratory and field environments depends on the ability to measure relevant potentials. In unsaturated soils, the matric and gravitational potentials are usually regarded as the hydraulically important energy components. A tensiometer is a device used to measure the matric potential, ψ_m, the component of the soil water chemical potential that is directly dependent on water content. When the local soil gas-phase pressure differs from that of the reference pressure (usually taken as the local atmospheric pressure), the tensiometer reading also includes this difference as an additional air pressure potential. Thus, the tensiometer most generally equilibrates to the sum of the matric and air pressure (pneumatic) potentials, also referred to as the tensiometer potential. In addition, tensiometers can be sensitive to shrinking and swelling of soils. For simplicity, this discussion will be restricted to conditions where the soil gas phase is in equilibrium with the atmosphere, and soil volume changes are negligible, such that the tensiometer equilibrates simply with the soil matric potential.

The basic components of a tensiometer consist of a porous tip or cup, an internal cavity, and a device to measure pressure within the cavity (Figure 1). The water-saturated porous boundary (the tensiometer tip or cup) is placed in contact with soil. This permits exchange of soil water between the region of interest and the tensiometer cavity until equilibrium is reached. Thus, the measurement of matric (or capillary) potential is generally achieved through determining the pressure of a fluid phase within the tensiometer, after equilibrium with the surrounding soil has been established. If the water table rises up to or above the tip, the tensiometer provides a measurement of the local pressure potential rather than the matric potential. Because most tensiometers rely on measuring the pressure of liquid water inside the tensiometer body, the lowest attainable equilibrium reading is imposed by the vapor pressure of water. Although this suggests that the lower operational limit ranges from -100 to -95 kPa for temperatures ranging from 5 to 35°C, a practical lower limit of approximately -85 kPa is commonly encountered because of exsolution of gases (air). The lower operating limit should be taken into account in selecting the porous tensiometer tip. Hydrophilic ceramics are most commonly used, but water-wettable porous glass, stainless steel, plastic, cloth, and even paper are used for tensiometer tips in various applications. The air-entry value of the porous tip is the gauge pressure at which the air phase displaces water-filled pores. Thus, when the matric potential decreases

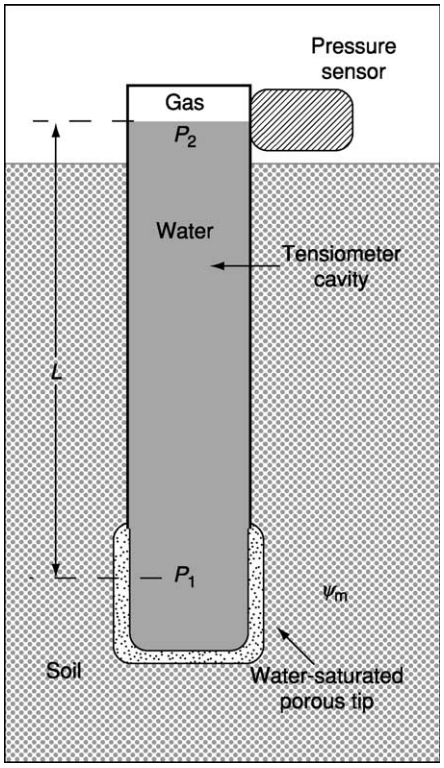


Figure 1 Basic components of a tensiometer. When significant, the hydrostatic pressure difference acting over the vertical distance L between the measurement point and the tensiometer tip must be accounted for. P_1, gauge pressure at elevation of interest.

(becomes more negative) such that its magnitude is larger than the air-entry value, air leaks in through the tip, causing the tensiometer to fail. Ceramics with air-entry values of approximately −100 kPa are typically used for tensiometer tips in field applications. More permeable ceramics with lower-magnitude air-entry suctions (e.g., −50 to −5 kPa) are used in some laboratory applications.

The original tensiometers, developed in the early twentieth century, consisted of a water-filled porous interface connected to a manometer. Since introduction of these early instruments, a wide variety of tensiometer designs have been developed. In field applications, tensiometers are used for irrigation scheduling, characterization of flow in soil profiles and deeper unsaturated sediments, and monitoring of contaminated soils. In the laboratory, tensiometers are used in soil physics, agricultural, and vadose-zone hydrologic research.

Equilibrium

By measuring the pressure of a fluid phase within a tensiometer, a measurement of the soil matric potential is obtained. The reliability of this measurement depends on establishing local equilibrium between the tensiometer and the soil, and on the condition that the matric potential is within the instrument's operational range, criteria that are discussed later. Equilibrium between the soil water and tensiometer can be viewed mechanically as a balance between capillary and adsorptive forces retaining water in the soil, and the hydrostatic pressure (suction) of water within the tensiometer. More generally, net exchange of soil water between the tensiometer (system) and soil (reservoir) occurs until differences in soil water chemical potential vanish. At this state, the gauge pressure in the tensiometer at the elevation of interest (denoted P_1 in Figure 1) is equal to ψ_m. When the pressure is recorded at an elevation different from that of the tensiometer tip, a hydrostatic correction is applied. In the example shown in Figure 1, $\psi_m = P_2 + \rho_w g L$, where ρ_w is the density of water, and g is the acceleration due to gravity. The time necessary to attain such equilibrium is generally determined by hydraulic properties of the soil, the soil–tensiometer contact, and response characteristics of the tensiometer.

Response Time

The time required to attain soil–tensiometer equilibrium depends on two basic factors: the rate of soil water transfer, and the change in tensiometer pressure per unit mass of soil water exchanged. These two factors can be parameterized in terms of the tensiometer–soil conductance and tensiometer gauge sensitivity. The overall conductance is determined by the (saturation-dependent) soil hydraulic conductivity, hydraulic contact between the soil and tensiometer tip, and the conductance of the tensiometer tip. Here, only the strictly tensiometer-dependent factors of tip conductance and gauge sensitivity will be addressed. In order to obtain a tensiometer reading rapidly that accurately reflects the local soil matric potential, it is desirable to maximize both the tip conductance and the gauge sensitivity. As shown below, the product of these two parameters is equal to the inverse of the instrument's response time.

The tensiometer conductance, K, is a measure of the volumetric rate that water can flow at through a tensiometer tip per unit hydraulic potential difference. When the potential difference (outside minus inside of the tensiometer) is expressed in terms of hydraulic head, K has dimensions of (L^2T^{-1}). The preferred maximization of K in order to shorten tensiometer response times can be achieved through increasing the permeability and bulk area of the porous tip, and decreasing the thickness of the tip. Each of these strategies is subject to constraints. Increasing the permeability of the tip material is achieved by

increasing its characteristic pore-size (since permeability scales to the square of pore diameter), and narrowing its pore-size distribution. However, the extent to which the characteristic pore-size can be increased is severely limited by the fact that the air-entry pressure is inversely proportional to pore-size. Thus more permeable tensiometer tips will not permit measurements of lower (more negative) matric potentials. The bulk surface area of the ceramic can be increased in order to provide more area through which soil water is exchanged. This option is limited when spatially resolved measurements of potential distributions are needed, and when constraints are associated with installation (labor and/or disturbance of soil profiles). The final option available for increasing the tensiometer-tip conductance involves decreasing the thickness of the porous ceramic or membrane. Compromised mechanical integrity imposes the primary limit on this option. In addition, thinner membranes may require more structural support to minimize flexing in response to pressure changes. This latter effect compromises the tensiometer's sensitivity.

The change in a tensiometer's pressure reading per unit volume (V_w) exchange of soil water defines its gauge sensitivity, S. When the pressure reading is expressed in head ($h(L)$) units, S has dimensions of (L^{-2}). Different types of pressure indicator have vastly different gauge sensitivities. Assuming the tensiometer cavity is rigid and completely water-filled, a pressure transducer can provide $S > 10^9\,m^{-2}$. In contrast, a simple water manometer (practical only for low-magnitude ψ_m measurements) has $S = dh/dV_w = A^{-1}$, where A is the cross-sectional area of the manometer tube. A 2-mm inner-diameter water manometer thus has a sensitivity of $3 \times 10^5\,m^{-2}$. Improving the sensitivity of manometer-based tensiometry through decreasing A is limited due to also increasing the uncertainty of the contact angle-dependent capillary rise (water) or capillary depression (mercury). Although the aforementioned values span a very wide range, most mechanical (Bourdon tube vacuum gauge) and electronic pressure transducers used in tensiometry have S in the range of 10^7–$10^9\,m^{-2}$. It should be noted that, because gases are highly compressible, the presence of air within tensiometers decreases S. The S associated with a gas cavity of volume V is equal to $P(\rho_w g V)^{-1}$, where P is the absolute pressure of the gas. Thus, a 1-cm^3 volume of air inside a tensiometer will allow S to be no greater than $10^7\,m^{-2}$, even when a much more sensitive pressure transducer is used to obtain that pressure reading.

The tensiometer's response time, τ, indicates how rapidly the instrument equilibrates with its environment when the rate of soil water exchange is controlled by the tip conductance (rather than the soil or soil–tensiometer contact). Under such conditions of tensiometer-limited response, the time parameter is related to K and S through $\tau = (KS)^{-1}$. The rate at which the tensiometer reading R relaxes towards its equilibrium value R_{eq} is then proportional the instantaneously magnitude of disequilibrium:

$$\frac{dR}{dt} = -\frac{(R - R_{eq})}{\tau} \quad [1]$$

In most tensiometers, τ is practically constant over the operational range, such that R_{eq} is approached according to:

$$\frac{R - R_{eq}}{R_0 - R_{eq}} = e^{-t/\tau} \quad [2]$$

where R_0 is the initial tensiometer reading. Thus τ is equal to the time require to diminish the difference between an arbitrary initial reading and the equilibrium reading to e^{-1} (approximately 37%) of the original discrepancy. By $t = 5\tau$, the tensiometer reading is within 1% of R_{eq} relative to the original discrepancy. Although tensiometers relying on pressure transducers or vacuum gauges have τ-values in the range of fractions of seconds to seconds, much slower responses are obtained in most applications due to low, unsaturated hydraulic conductivity of the soil and poor tensiometer–soil hydraulic contact.

Range of Applications

Given the wide variety of environments in which it is desirable to have information on matric potentials, it is worth addressing limitations in ranges of tensiometer applications as well as approaches that have been developed to diminish or circumvent limitations. Here, lower ranges in matric potential, tensiometry in deep unsaturated zones, and temperature-related problems are briefly discussed.

Although water-filled tensiometers have a lower practical limit of about -85 kPa, much lower matric potentials can be measured with osmotic tensiometers. In an osmotic tensiometer, a polymer solution within the tensiometer body equilibrates with soil water through a semipermeable membrane (ideally impervious to the polymer). The polymer solution equilibrates at a positive gauge pressure (in order to compensate for water potential lowering resulting from interactions with the polymer), thereby avoiding negative gauge pressure limitations of conventional tensiometers. Problems arising from high sensitivity to thermal fluctuations and slow leakage of the polymer have severely limited use of osmotic tensiometers, but recent advances may lead to wider interest in this alternative.

Interest in flow through deep unsaturated soils and sediments has grown over the past two decades, and this has necessitated the development of deep tensiometry methods for field applications. Use of a conventional field tensiometer with a water-filled column to provide the hydraulic connection between the buried tensiometer tip and the pressure gauge at the soil surface (Figure 1) permits a decreasing range of matric potential measurement with deeper installations because of the elevation-dependent hydrostatic pressure decrease. Such a conventional tensiometer used for depths greater than about 8 m will permit little to no useful measurements of matric potentials, because the upper section of the column exists at low absolute pressure and rapidly becomes filled with water vapor. Approaches used to circumvent this problem eliminate use of the water-filled column and rely instead on either burying pressure transducers at or near the depth of the tensiometer tip or use of an air-filled column. The latter option is not as suitable for monitoring transient flow conditions, since the air-filled cavity imparts low gauge sensitivity, hence a low response time.

Temperature variations in the field have long been known to make interpretation of tensiometer readings difficult. The phenomenon is complex and often sensitive to instrument design, since thermal effects on volume changes of tensiometer components and fluids can be substantial. This can be especially problematic when instrument sensitivity is high and the overall tensiometer–soil conductance is low. Keeping all instrument components below the soil surface (where maximum temperature variations occur) can diminish temperature-induced changes in tensiometer readings. When tensiometers are used in field locations exposed to freezing temperatures, protection from damage by ice formation is necessary. This can be accomplished to an extent by adding methanol or antifreeze to the water in the tensiometer, or by burying the instrument below the depth of soil freezing.

List of Technical Nomenclature

ρ_w	Density of water ($kg\,m^{-3}$)
τ	Response time (s)
ψ_m	Matric potential (Pa)
g	Acceleration due to gravity ($m\,s^{-2}$)
h	Head (m)
K	Conductance ($m^2\,s^{-1}$)
L	Length (vertical) (m)
P	Pressure (Pa) or (kPa)
R	Reading (tensiometer) (kPa)
S	Sensitivity (gauge) (m^{-2})
t	Time (s)
V	Volume (cm^3) or (m^3)

See also: **Hydrodynamics in Soils**; **Thermodynamics of Soil Water**; **Vadose Zone:** Hydrologic Processes

Further Reading

Cassel DK and Klute A (1986) Water potential: tensiometry. In: *Methods of Soil Analysis*, 2nd edn, pp. 563–596. Monograph No. 9. Madison, WI: American Society of Agronomy.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

Or D (2001) Who invented the tensiometery? *Soil Science Society of America* 65: 1–3.

Peck AJ and Rabbidge RM (1969) Design and performance of an osmotic tensiometer for measuring capillary potential. *Soil Science Society of America Proceedings* 33: 196–202.

Richards LA (1928) The usefulness of capillary potential to soil moisture and plant investigators. *Journal of Agricultural Research* 37: 719–742.

Richards LA (1949) Methods of measuring soil moisture tension. *Soil Science* 68: 95–112.

Sposito G (1981) *The Thermodynamics of Soil Solutions*. Oxford, UK: Oxford University Press.

Stannard DI (1990) Tensiometers – theory, construction, and use. In: Nielsen DM and Johnson AI (eds) *Ground Water and Vadose Zone Monitoring*, pp. 34–51. ASTM STP 1053. Philadelphia, PA: American Society for Testing and Materials.

Towner GD (1980) Theory of time response of tensiometers. *Journal of Soil Science* 31: 607–621.

Termites *See* **Fauna**

Terra Rossa *See* **Mediterranean Soils**

TERRACES AND TERRACING

G R Foster, Bryan, TX, USA

Terraces are runoff management structures used to control soil erosion by water. All terraces are topographic modifications of hillslopes. Terraces reduce rill and interrill erosion by shortening the overland flow path (slope) length or reducing effective landslope steepness. Some terraces reduce net erosion by depositing a portion of eroded sediment on the hillslope. Terraces capture runoff and convey it off the hillslope without causing excessive erosion. Captured runoff increases soil moisture and thus crop production. Terraces are designed, installed, and maintained to fit local climate, soil, topography, and crop management in accordance with available resources, preferred land use, and other socioeconomic factors.

Benefits and Limitations

Terrace systems are highly effective at preventing excessive rill erosion, eliminating ephemeral gully erosion, reducing sediment yield, conserving soil moisture, protecting landscape quality, and increasing land value.

Terraces are topographic modifications that require soil displacement to construct them. They work best on deep soils, such as loess soils. Terraces require a significant investment to build and maintain. Farming with terraces may be inconvenient, and they may limit the choice of farming practices. Terrace systems that do not fit local conditions can be worse than no terraces at all.

Runoff and Erosion on Hillslopes without Terraces

Runoff begins as overland flow spread uniformly around the hillslope. Overland flow is terminated by draws (concentrated flow areas) on the landscape where flow becomes channelized. These draws are formed during landscape evolution. Figure 1 illustrates these flow paths.

Soil erosion by water on the overland flow areas is known as rill and interrill erosion. Rill erosion is the detachment of soil particles by surface flow, and it occurs in a series of rills, which are small, parallel channels a few tens of millimeters wide. Location of rills is determined by soil surface microtopography. Interrill erosion occurs on interrill areas, which is the space between rills. Interrill erosion is detachment of soil particles by raindrop impact. Detachment is the separation of soil particles (sediment) from the soil mass. Deposition is the accumulation of sediment on the soil surface. Sediment load is the amount of the sediment that is transported downslope. Most downslope sediment transport is by surface runoff. Detachment increases sediment load in a downslope direction, while deposition decreases sediment load.

Erosion in concentrated flow areas is known as concentrated flow erosion. This erosion is also known as ephemeral gully erosion if tillage fills the channels and runoff reforms the channels during each cropping cycle. Rills are obliterated by tillage. Rills are typically reformed in different locations depending on the soil-surface microtopography left by tillage. Ephemeral gullies always occur in the same location because of the dominance of landscape macrotopography.

Types of Terraces

The major types of terraces are gradient, parallel-impoundment, bench, and ridge. Most terrace systems involve runoff interceptors that collect the overland flow and redirect it around the hillslope. Most terrace systems include water-conveyance structures that receive the intercepted flow and convey it to low points on the hillslope.

Gradient Terraces

Construction Gradient terraces use an embankment and associated channel to intercept overland flow and direct it around the hillslope. Flow from the terraces is discharged into a protected channel that conveys the flow downslope, as illustrated in Figure 2. These terraces are typically used on hillslopes less than 10% in steepness. Gradient terraces are constructed tens of meters apart and typically use low embankments and shallow channels so that the entire area can be farmed with tractors and wide farm

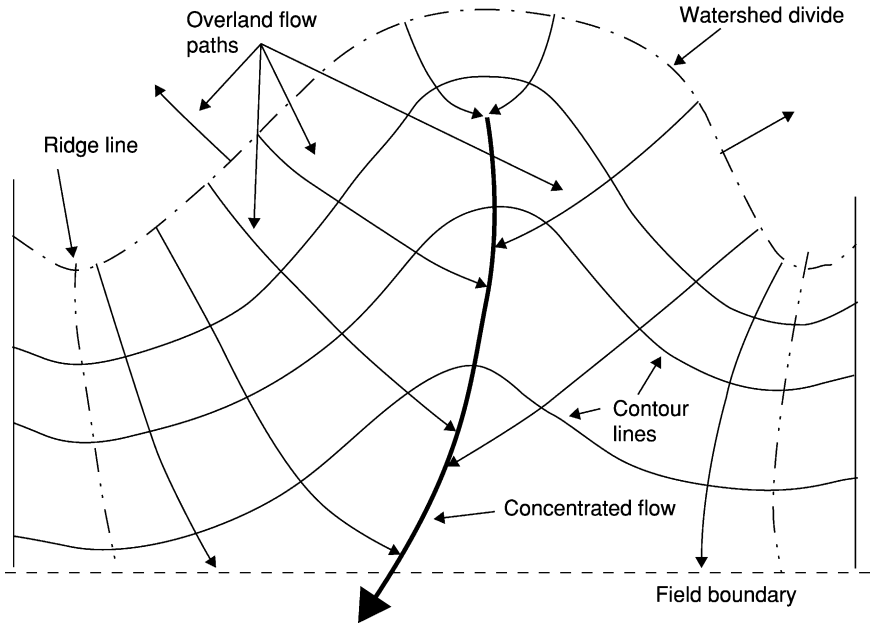


Figure 1 Overland flow and concentrated flow paths on a typical hillslope without terraces.

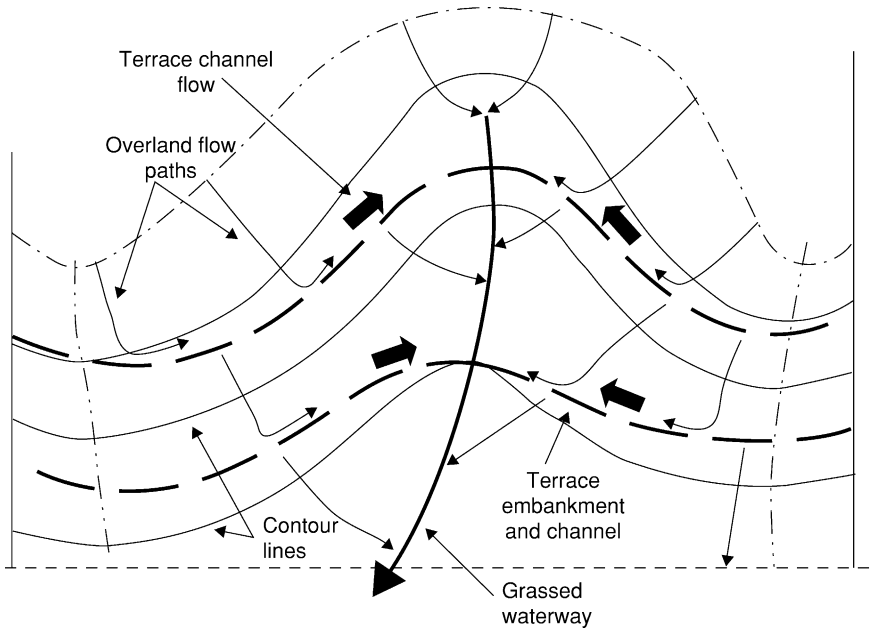


Figure 2 Overland flow and concentrated flow paths on a typical hillslope with gradient terraces on a uniform grade.

equipment. These terraces are referred to as broad-base terraces. Gradient terraces can be constructed with road graders, earthmovers, special terracing machines, and plows used in farming. Cuts and fills are balanced to minimize construction costs and the soil depth that is disturbed. Deep cuts expose subsoil and reduce crop production.

Gradient terraces shorten the 'slope length' of the overland flow path. Rill and interrill erosion vary approximately with the square root of slope length. Adding two gradient terraces to a 100 m slope length creates three slope lengths of 33 m minus the distance from the center of the terrace embankment to the upper edge of the terrace channel as illustrated in Figures 2 and 3. These terraces reduce rill and interrill erosion by ~40%.

Spacing Terraces are spaced to avoid excessive rill erosion. Equations for terrace spacing have been developed based on field experience. A rule of thumb is

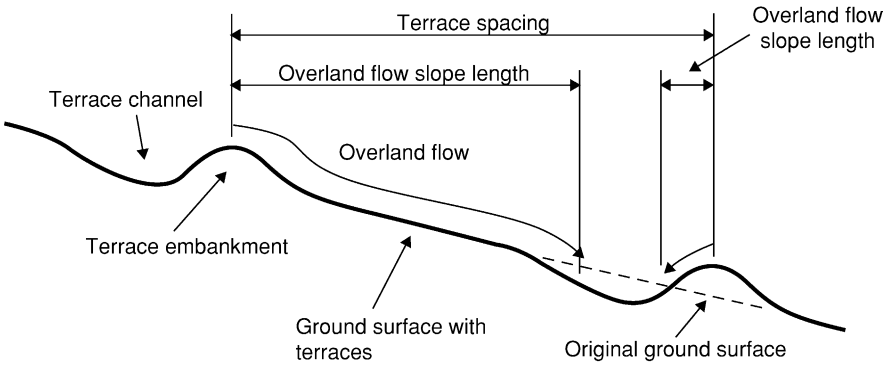


Figure 3 Terrace embankment and channel cross section for a broad base terrace.

that rill erosion is excessive when average annual rate for rill and interrill erosion exceeds 15 t ha^{-1} year^{-1}. Terrace spacing is decreased where rainfall amount and intensity are high and where the soil has a low infiltration rate, both of which produce high runoff rates. Terrace spacing is also decreased on steep hillslopes. High runoff rates and steep slopes increase runoff erosivity and the potential for rill erosion. Terrace spacing is decreased for soils that are susceptible to rill erosion.

Vegetative cover including plant stems, roots, and residue, and soil surface roughness reduce runoff erosivity. Terrace spacing is based on the most vulnerable period for rill erosion during a cropping cycle, which is often the seedbed preparation period, especially for clean-tilled row crops such as maize or soybean.

Terrace spacing can also be based on the average annual soil erosion rate estimated using erosion prediction technology such as the Revised Universal Soil Loss Equation (RUSLE). The terraces are spaced so that the estimated erosion rate on the inter-terrace interval is less than the soil loss tolerance value assigned to the site-specific soil. Soil loss tolerance values range from about 4 to 11 t ha^{-1} year^{-1}, based on the degree that erosion is judged to harm a particular soil.

Contouring, a practice where the direction of tillage and crop rows is on the contour around the hillslope, reduces rill and interrill erosion by about 50% on mild slopes. Contouring fails when runoff erosivity becomes too great. The failure location is known as the critical slope length, which is computed with RUSLE. Terrace spacing should not exceed the critical slope length.

Deposition and sediment delivery Gradient terraces also control erosion by deposition in the terrace channel. Deposition occurs when the sediment load reaching the channel from the inter-terrace area is greater than transport capacity of the flow in the channel. Transport capacity in the channel increases in direct proportion to increases in channel grade and channel flow rate. If transport capacity in the channel is greater than the incoming sediment load, no deposition occurs.

Sediment delivery ratio (SDR) is the ratio of the sediment load leaving the terrace channel to the incoming sediment load from the inter-terrace area. The sediment delivery ratio is about 0.2 and 0.4 for terraces on a 0.2% and 0.5% grade, respectively, in clean-tilled row crops. If terrace grade is greater than about 1.0%, no deposition occurs (as a rule of thumb).

Deposition also depends on sediment characteristics. Less very fine sediment is deposited than coarse sediment. Sediment eroded from most cropped soils is a mixture of primary particles and aggregates (conglomerates of primary particles). Silt-textured soils (e.g., 5% clay, 87% silt, 8% sand) are poorly aggregated and produce a high proportion of fine sediment particles. An SDR value for a silt soil might be 0.34. A silt loam soil (e.g., 15% clay, 65% silt, 20% sand) produces sediment having a balanced mixture of primary particles and aggregates. The SDR for that soil might be 0.26. A high sand soil (e.g., 4% clay, 6% silt, 90% sand) is eroded mainly as large, easily deposited primary particles. The SDR might be 0.12. In contrast, a high clay soil (e.g., 60% clay, 20% silt, 20% sand) erodes as sediment having a significant amount of very fine primary clay particles. The sediment also has a large fraction of large aggregates that are readily deposited. The easily deposited aggregates make up for the large fraction of primary clay in the sediment that is not readily deposited. The SDR for the clay soil is 0.24, about the same as for the silt loam soil. Clay is a bonding agent such that a large fraction of the sediment eroded from high clay soils is large aggregates.

Periodic maintenance is required to clear terrace channels of deposited sediment to maintain flow capacity. Otherwise, flow overtops the terrace and causes serious gully erosion that requires major

repairs. Deposited sediment removed from the terrace channel is spread over the field near the terrace to replace soil loss by erosion and soil displaced by the terrace construction. The ridges and channels are maintained so that farming over them can continue. Placing the deposited sediment on the terrace ridges makes them too high.

Only partial credit is given to deposition caused by terraces as soil saved relative to protecting the entire hillslope, because the deposition is on a relatively small area. Also, the deposition is remote from the location on the hillslope that produced the sediment. Erosion can degrade source areas even though much deposition occurs downslope. The credit given to deposition as soil saved decreases as terrace spacing increases because the fraction of the total area benefited by deposition decreases. No credit is given for deposition for terraces spacing greater than 100 m. A maximum credit of one half is given for spacing closer than 30 m. Very little credit is given for deposition that occurs in terraces located near the bottom of the hillslope.

Outlet channels Gradient terrace systems include outlet channels that collect flow discharged at the outlet of the terrace channels and convey the flow downslope. These channels must be well protected to prevent gully erosion. Flow in them is erosive because of a steep channel grade. Typically, these channels are grassed waterways. Rock can be used to line outlet channels where the flow is especially erosive. A major benefit of terrace systems is the elimination of ephemeral gully erosion, but this important benefit is not realized if stable outlet channels are not provided.

The capacity and stability of terrace and outlet channels are based on a design storm with a 24 h duration that occurs with a return period of 10 years, referred to as a 10 year-24 h storm. Runoff is computed based on the condition during the cropping cycle that produces the most runoff. Channel flow capacity is based on the condition that provides the greatest hydraulic resistance, e.g., unmowed grass in outlet channels and heavy residue cover in terrace channels. Channel stability is based on the condition when the channel is most vulnerable to concentrated flow erosion, e.g., freshly mowed grass in outlet channels and tilled seedbed in terrace channels.

Moisture conservation Gradient terraces are used in low precipitation regions to conserve soil moisture and increase crop production. A level terrace grade is used and the bottom of the terrace channel is widened to increase water capture and expand the benefited area (*See* **Water Harvesting**).

Parallel-Impoundment, Underground-Outlet Terraces

A terrace system suited for farming with wide equipment is one where the terraces are nearly parallel. Uniform-grade terraces are difficult to farm because the terraces are not parallel. Wide equipment does not work well when crossing terraces and tends to destroy them. Parallel terraces, illustrated in Figure 4,

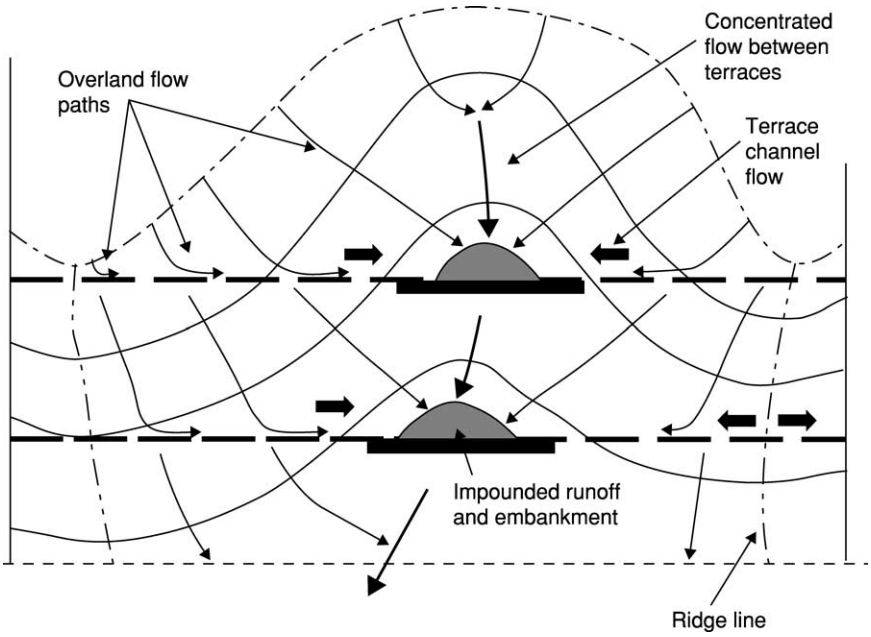


Figure 4 Overland flow and concentrated flow paths on a typical hillslope with parallel-impoundment, underground-outlet terraces.

require that grade varies along the terraces. Parallel-impoundment terraces are highly effective at controlling sediment yield and reducing ephemeral gully erosion, and are convenient to farm with large equipment. These terraces are less effective than gradient terraces at controlling rill and interrill erosion. Parallel-impoundment terrace systems include overland-flow interceptors, impoundments, and underground water-conveyance systems.

Overland-flow interceptors The overland-flow interceptors for parallel terraces are essentially the same as the embankments and channels used with gradient terraces. The slope lengths for the overland flow paths with a gradient terrace system are generally uniform along the terrace. However, the slope length of overland flow varies along parallel terraces. Slope lengths vary from essentially zero to the slope length of the natural landscape if the terraces are spaced far apart.

Channel grade varies along parallel terraces. The grade is zero where the channels cross a ridge and is steepest between the ridge and the impoundment in the draw. Deposition occurs uniformly along a uniform grade terrace channel, but deposition only occurs in parallel terrace channels where channel grade is sufficiently flat. Maximum channel grade is limited to ensure that no concentrated flow erosion occurs in the terrace channels. Terrace embankments and channels are relocated uphill to reduce terrace channel grade. The terrace spacing and hydraulic design procedures used for gradient terraces are also used to design parallel terraces.

Impoundments Impoundments located across the draws on the hillslope receive the flow from the overland-flow interceptors. These impoundments are sized to retain the runoff from a 10 year-24 h storm assumed to occur when runoff potential is greatest. The impoundment size is increased to store the sediment that will accumulate over the design life of the impoundment, usually 10 years. Maintenance can be performed periodically to remove and spread the deposited sediment over the hillslope to extend the life of the impoundments.

Runoff is retained in the impoundment to provide time for sediment to be deposited by settling to the bottom. Twenty-four hours is usually sufficiently long for most of the sediment to be deposited. The retention time is kept short to minimize inundation of crops and to speed drying so that farming operation can resume with minimal inconvenience.

Impoundment terraces very efficiently trap sediment. For example, an impoundment terrace will trap ~94% of the sediment eroded from a silt loam soil on the inter-terrace area for an SDR of 0.06. The SDR values for sediment eroded from sand, silt, and clay soils are 0.01, 0.07, and 0.14, respectively. The order of sediment delivery ratios among the soil textures differs from the SDR order for gradient terraces. This difference results from the sediment being continuously added along the gradient terrace channel, while the sediment is added to the impoundment at a point. Also, deposition mechanics in still water differ from deposition mechanics in channel flow.

The embankments for these impoundments are often so high that they cannot be crossed with farm equipment, which removes a portion of the field from crop production. The embankments are constructed with a steep backslope, illustrated in Figure 5, to minimize this loss. Dense grass is used on the backslope to prevent erosion. A bulldozer is used to construct the embankments.

Water-conveyance system Water is drained from the impoundments using a perforated riser pipe that projects above the floor of the impoundment. Water enters the riser and falls vertically into an underground pipe. Discharge rate is controlled using an orifice in the bottom of the riser to provide the desired retention time. Water flows from the riser into an underground pipe that runs up the draw and collects flow from all the terraces that cross that particular draw. A perforated underground pipe is sometimes used to provide subsurface drainage in the draw where seepage can collect.

Bench Terraces

Bench terraces reduce steep hillslopes to ones that are effectively flat. Bench terraces consist of a steep non-erodible section and a flat or nearly flat bench that is cultivated, as illustrated in Figure 6. The most impressive example of bench terraces is in China, where steep, highly erodible land is farmed. Reducing the effective landslope from 30% to 1% decreases erosion by ~99%.

Flat bench terraces are used in low rainfall areas to capture rainfall to increase soil moisture and crop production. Reduced runoff also reduces soil erosion (*See* **Water Harvesting**).

Flat benches with vertical backslopes Among bench terraces, flat benches with vertical backslopes provide the greatest land area for cultivation. However, constructing vertical walls for bench backslopes requires considerable resources. Vertical walls are easiest to construct if the wall is not high. However, a low wall results in a narrow flat bench area, which limits the cropping-management practices and size of equipment that can be used on the benches. A well-designed

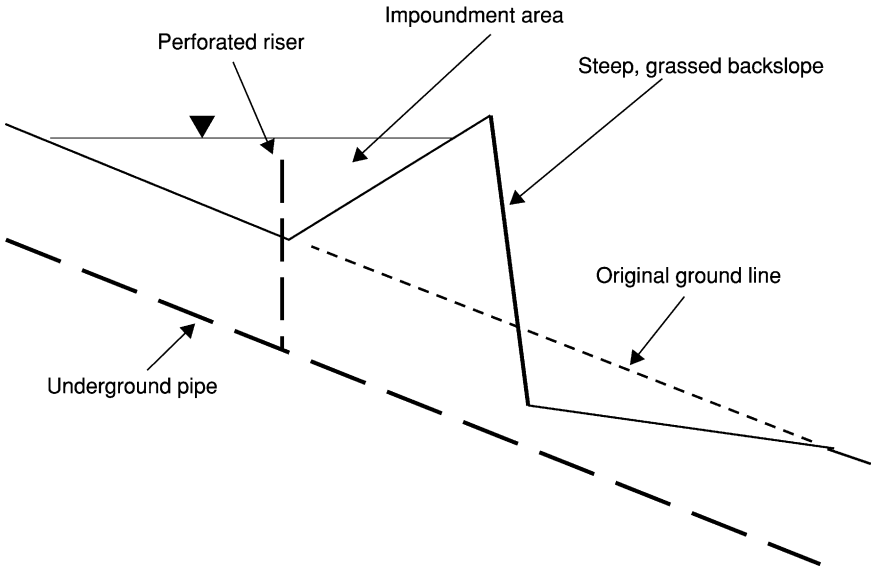


Figure 5 Impoundment embankment cross-section for an underground-outlet terrace.

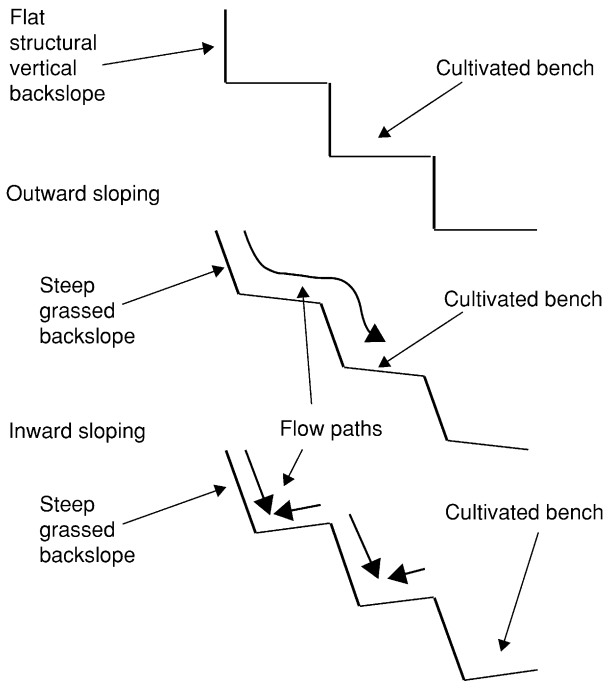


Figure 6 Types of bench terraces.

water-conveyance system is included with the terrace system to collect excess water and convey it downslope in stable channels.

Outward-sloping benches An alternative bench terrace system has a steep, grassed backslope (1:1 slope) and a relatively flat bench (less than 2%) wide enough (e.g., 3 m) to accommodate mechanically powered equipment. The bench can be outward or inward sloping. If the bench is outward sloping, as illustrated in Figure 6, overland flow originates at the top of the hillslope and cascades down the slope. Erosion is low because the grass essentially prevents erosion on the steep backslope, and erosion is low on the nearly flat bench.

While bench terrace systems theoretically perform well with water cascading from one terrace to the next, runoff must be carefully controlled to avoid concentrated flow and gully erosion. Maintaining bench terraces with the front edge perfectly on a level grade so that runoff flows uniformly over the bench edge is difficult. Also, the grass cover on the steep backslope must be well maintained to prevent areas where gully erosion can begin. Runoff should be collected at regular intervals downslope and brought to nonerodible channels that convey the runoff downslope.

Inward-sloping benches Inward-sloping benches, illustrated in Figure 6, eliminate the cascading flow down the slope. Each bench and its backslope are individual watersheds. The outer edge of the bench is left higher than the back edge so that runoff flows back toward the hillslope. The only runoff in each bench watershed is from the rainfall that falls directly on that watershed. The benches can be put on a slight grade around the hillslope so that runoff flows around the hillslope at the base of the backslopes to nonerodible collection channels that convey the runoff downslope.

Inward-sloping bench terraces have less rill and interrill erosion than do outward-sloping bench terraces. However, the difference is not great because at

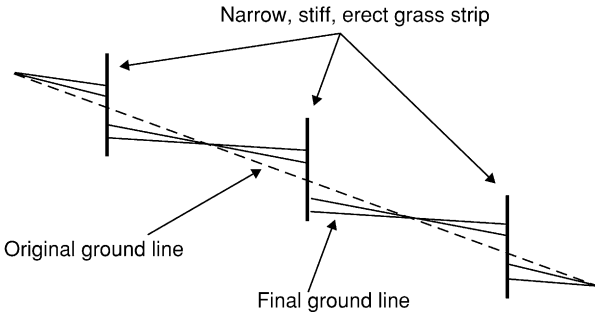


Figure 7 Narrow, stiff grass strips enable erosion and deposition to create bench terraces.

1% slope steepness and less, most of the erosion is caused by interrill erosion, which is not affected by slope length. For example, reducing slope length from 50 m to 3 m results in only a 20% reduction in erosion for a 1% steep slope.

Naturally formed benches Soil must be mechanically moved to construct the terraces described above. Permanent, narrow grass strips can be used to let nature build bench terraces. Narrow, approximately 0.5 m wide, strips of dense erect, stiff grass are planted very carefully on the contour. These dense strips slow the overland flow to cause deposition in backwater ponded on the upper side of the grass strip. Over time, the hillslope slowly evolves as illustrated in Figure 7. Reducing steepness from 10% to 2% for a 10 m hillslope segment by rotating the segment about its midpoint requires 0.35 m of erosion at the segment's upper end and 0.35 m of deposition at the segment's lower end. Approximately 100 years is required for this adjustment assuming that all sediment eroded on the segment's upper portion is deposited on the segment's lower portion, an average annual erosion rate of 50 t ha^{-1} $year^{-1}$, and a specific soil bulk density of 1.3. The process is often slow. However, an immediate benefit is that the grass spreads the runoff so that rill erosion is reduced for a few meters downslope from the strip.

Ridge Ridges and beds on the contour around the hillslope are actually a set of small terraces where the slope length is half the spacing of the ridges. The slope length of the overland flow is the distance from the top of the ridge to the middle of the furrow between the ridges. Almost all of the erosion on the ridges is interrill erosion because of short slope lengths.

The furrows between the ridges act like terrace channels. A flat furrow grade, less than 2% for clean-tilled row crops, deposits much of the sediment eroded on the ridges. Deposition in the furrows is given full credit, as soil saved protects the entire hillslope because the deposition is very near to the source of the sediment.

To function as terraces, ridges must be sufficiently high to avoid overtopping and be carefully constructed and maintained to ensure that they perform well. Erosion from failed ridges can be greater than if ridges had not been used. Overtopping in a local area is like a dam break. Failures cascade downslope, resulting in serious ephemeral gully erosion. Maintaining a sufficient height along the ridge and avoiding adverse furrow grades that pond water prevent overtopping in localized areas.

Ridges are most effective when arranged perfectly on the contouring. Maximum moisture is retained for crop utilization. When ridges overtop, the runoff flows over the ridges uniformly along the ridges if the ridges are perfectly level. The flow depth will be very shallow and much less erosive than if the runoff concentrates at low places and breaks over in a concentrated flow. Ridges perfectly on the contour reduce ephemeral gully erosion. When ridges are perfectly on the contour, the head of ephemeral gullies is further down the hillslope than if ridges and furrows are on a grade. Ridges and furrows on a slight grade concentrate overland flow far up the hillslope, which extends ephemeral gully erosion further up the hillslope than with ridges perfectly on the contour, as illustrated in Figures 8 and 9.

Straight-ridge alignment facilitates farming with large equipment. However, the furrow grade along straightened ridges is non-uniform, just as with parallel terraces. The grade must be sufficiently flat so that rill erosion does not occur in the furrows on areas adjacent to concentrated-flow areas. Runoff collects in these natural waterways and flows downslope. These concentrated-flow areas must be protected with a grass or another stable lining to prevent serious ephemeral gully erosion. If these concentrated-flow areas are not protected, ephemeral gully erosion can be greater with ridging slightly off the contour than with ridges directly up and downhill. However, wide grass waterways take land out of production, reduce farming convenience, and require significant maintenance.

Summary

Terraces are important structures used to safely convey runoff, control rill erosion, and prevent ephemeral gully erosion, especially on steep hillslopes with long slope lengths. Terraces are also used to conserve soil moisture and thus to increase crop production in low rainfall areas. Terraces intercept overland flow and redirect it around the hillslope where protected channels convey the runoff downslope.

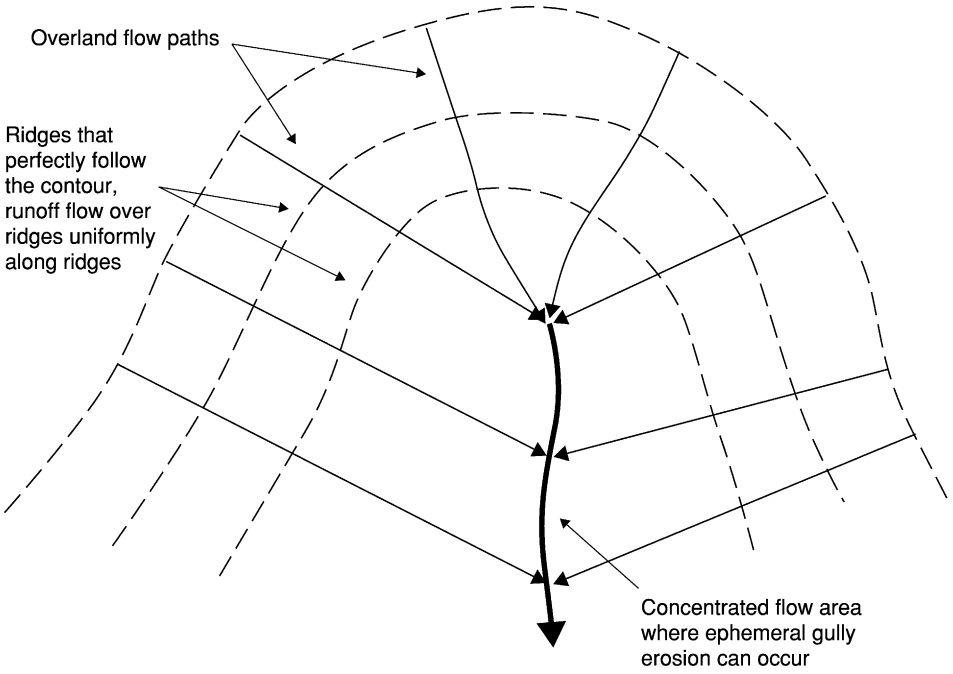


Figure 8 Overland flow paths and location of the head of ephemeral gullies when ridges are perfectly on the contour.

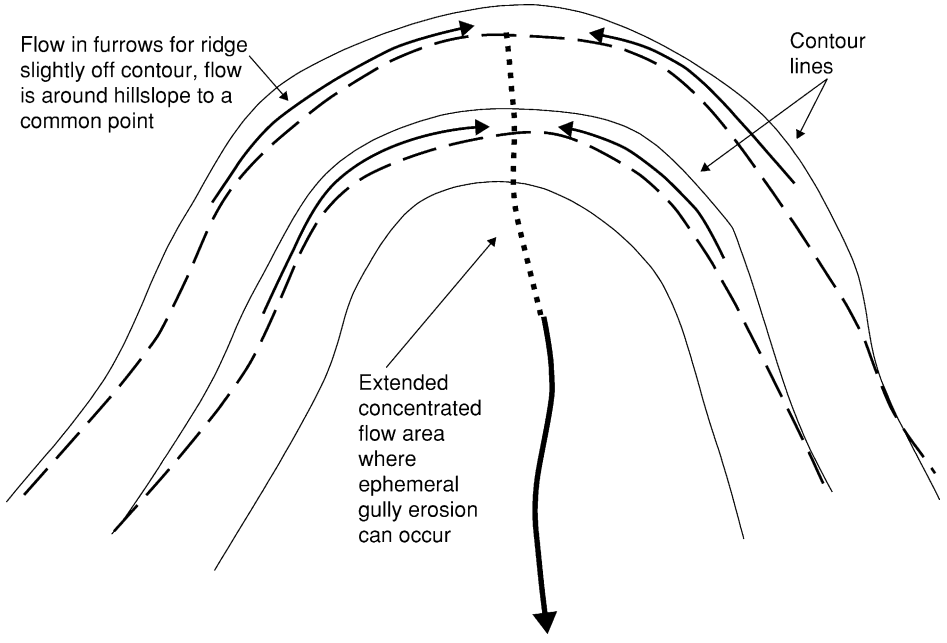


Figure 9 Overland flow paths and location of the head of ephemeral gullies when ridges are on a slight grade.

Terraces are used in a wide range of agroecosystems ranging from small plots farmed with human and animal power to large fields farmed with large machines. Terraces are best suited to deep soils where soil can be moved to create benches and ridges without seriously affecting the soil productivity. Terraces require significant resources of capital or labor to construct and must be maintained. Terraces on steep hillslopes in China graphically illustrate how terraces can be used to make maximum use of the landscape for crop production.

See also: **Aggregation:** Physical Aspects; **Conservation Tillage**; **Crop-Residue Management**;

Drainage, Surface and Subsurface; **Erosion:** Water-Induced; **Overland Flow**; **Precipitation, Watershed Analysis**; **Water Harvesting**; **Watershed Management**

Further Reading

Beasley RP, Gregory JM, and McCarty TR (1984) *Erosion and Sediment Control*, 2nd edn. Ames, Iowa: Iowa State University Press.

Foster GR and Highfill R (1983) Effect of terraces on soil loss: USLE P factor values for terraces. *Journal of Soil and Water Conservation* 38: 48–51.

Haan CT, Barfield BJ, and Hayes JC (eds) (1994) Hydrologic frequency analysis and rainfall-runoff estimation in stormwater computations. In: *Design Hydrology and Sedimentation for Small Catchments*, pp. 5–103. San Diego, California: Academic Press.

Hurni HA (1981) Nomograph for design of labour-intensive soil conservation in rain-fed cultivations. In: Morgan RPC (ed.) *Soil Conservation: Problems and Prospects*, pp. 185–210. New York: John Wiley.

Morgan RPC (1986) *Soil Erosion and Conservation*. New York: John Wiley.

Toy TE, Foster GR, and Renard KG (2002) *Soil Erosion: Processes, Prediction, Measurement, and Control*. New York: John Wiley.

Troeh FR, Hobbs JA, and Donahue RL (eds) (1991) Conservation structures. In: *Soil and Water Conservation*, 2nd edn, pp. 247–275. Engle Wood Cliffs, New Jersey: Prentice-Hall.

US Department of Agriculture-Natural Resources Conservation Service (2003) Conservation practice standards. In: *Field Office Technical Guide*. Available on Internet at www.usda.gov/technical/efotg/, Washington, DC.

Zhengsan F, Piehua Z, Qiande L, Baihe L, Letian R, and Hanxiong Z (1981) Terraces in the Loess Plateau of China. In: Morgan RPC (ed.) *Soil Conservation: Problems and Prospects*, pp. 481–513. New York: John Wiley.

Zingg AW (1940) Degree and length of land slope as it affects soil loss in runoff. *Agricultural Engineering* 2: 59–64.

TESTING OF SOILS

A P Mallarino, Iowa State University, Ames, IA, USA

Introduction

Soil testing is an important diagnostic tool for determining the nutrient needs of plants and for environmental assessments. Some soils are inherently deficient in plant nutrients. Other soils had sufficient levels of nutrients in the past, but removal with crop harvest has depleted the reserves. Thus, soil testing is widely accepted and used in most advanced crop-production areas of the world to determine fertilization needs for crops. Soil testing can also be used to identify application rates of waste materials containing nutrients or other elements that could harm the environment. Waste materials such as animal manures and industry by-products may provide various plant nutrients. However, high application rates to soils designed to dispose of the material at a low cost may result in nutrient loads that are harmful to plant, animal, or human health. Nutrient management regulations are being developed to address land application of waste materials. Soil testing is required in many regulations and management guidelines to assess environmentally harmful levels of certain elements and to determine limits to application rates.

Soils are tested routinely for the primary nutrients phosphorus (P), potassium (K), and nitrogen (N). In some regions, soils are also routinely tested for other primary nutrients such as calcium (Ca), magnesium (Mg), and sulfur (S), and for other nutrients required in very small amounts by crops such as boron (B), copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), and zinc (Zn). Soils receiving waste materials are also tested for elements such as arsenic (As), cadmium (Cd), nickel (Ni), lead (Pb), mercury (Hg), and selenium (Se) among others.

Two of the primary plant nutrients, N and P, may have harmful effects on the environment when applied to soils in excessive amounts. Excessive N and P applications to agricultural fields and ineffective nutrient, soil, and water conservation practices are increasing nutrient pollution in many regions of the world. However, the basic concepts of soil testing also apply to other elements. Phosphorus is included with the group of elements with relatively low mobility in soils (Ca, Cd, Cu, Fe, Mg, Mn, Mo, Ni, Pb, and Zn among others). N, especially in the nitrate form, is included with the group of elements with greater mobility (which include B, Cl, S, and others). Important concepts include the meaning of a soil-test value, soil-testing quality, use of soil testing to determine economically optimum nutrient application rates, and use of soil testing for environmental assessments.

Soil-Test Extractants and Methodologies

Phosphorus Soil Tests

Soil tests for P and other elements have been developed based on an understanding of the chemical forms in which the elements exist in the soil and empirical work. In most soils, P is primarily associated with aluminum (Al), Fe, and Ca. P precipitation and/or adsorption to Fe and Al oxides generally predominates in acid (pH < 7.0) to neutral soils, whereas reactions with Ca predominate in soils with significant amounts of calcium carbonate. The degree and strength to which P is bound in soils are largely determined by the amount and types of Fe and Ca compounds present and by other soil properties such as pH, organic matter, clay mineralogy, and the amount of P currently present in the soil. The influence of Fe increases and that of Ca decreases as rainfall, temperature, and weathering increase. These factors are also important in determining the plant availability and solubility of other immobile elements.

Phosphorus soil-test methods for agronomic use employ dilute strong acids, dilute weak acids, complexing ions, and/or buffered alkaline solutions. Most tests have been developed to reflect the soil properties related to P sorption that predominate in a region. For example, the Bray P-1 test was developed for use in the acid-to-neutral soils of the north-central region of the USA; the Olsen (or sodium bicarbonate) test was developed primarily for use on calcareous soils of the same region; and the Mehlich-1 test was developed for the low cation-exchange capacity and highly weathered soils common to southeastern regions. The Mehlich-3 P extractant was developed more recently. It has been adopted because it is suitable for a wider range of soil properties than other P tests and also for other nutrients (such as K, Ca, Mg, Cu, Fe, and Zn).

Increasing concerns about nonpoint-source P pollution from agriculture has prompted questions about the suitability of agronomic soil P tests for environmental purposes. Several soil tests have been proposed as environmental soil P tests because they may provide better estimates of dissolved P delivery to water resources. Two of these tests are being evaluated extensively. One test based on FeO-impregnated paper uses a sink approach to extract soil-bound P that is most likely available to aquatic organisms. The other test is based on weaker desorption reactions and uses either deionized water or a diluted calcium-chloride solution to extract soil P. These tests are environmentally sound for laboratory use because few hazardous chemicals are needed.

Although the results of the different soil P tests are often highly correlated for soils with similar properties, the actual quantities extracted can differ greatly. For example, the Mehlich-3 test extracts 1.5 to 2 times the amount of P as the Mehlich-1 and Olsen tests. The FeO-impregnated paper environmental P test usually measures 4–10 times the amount of P measured by the water extractant. These observations highlight the importance of understanding the meaning of a soil-test value.

Nitrogen Soil Tests

The cycling of N in soils is different from other nutrients and requires a different soil-testing approach. More than 97% of N in soils is present in organic forms and is converted to inorganic forms available to plants mostly through microbial activity. The rate of N release from organic to inorganic forms is difficult to predict because it depends on temperature, moisture, aeration, organic matter type, soil pH, and many other factors. The predominant inorganic end-result of mineralization is nitrate-N. This N form is highly mobile in soils and is subject to loss by leaching and to gaseous losses through denitrification. Because of these factors, N fertilizer recommendations for crops are frequently made on the basis of crop yield potential coupled with an N-credit system and less frequently on the basis of soil testing. The N-credit system is based on empirical information of the effects of previous crops, several soil properties, and climate on crop response to N fertilization.

Soil testing for nitrate-N has been used for a long time in low-rainfall regions and is becoming more common in humid areas of the USA, mainly for maize production. In low-rainfall areas, where nitrate leaching potential is minimal, the amount of nitrate-N in the soil profile before planting crops can be credited directly against the N needs of the crop. In many humid regions, measurements of soil profile nitrate or ammonium before planting have been unreliable to predict N needs of maize. However, nitrate-N testing early in the growing season has been shown to provide an index of the soil N-supplying capability for maize. Soil samples are collected deeper than for other nutrients (usually from a 0- to 30-cm depth) after planting the crop and before plants reach a height of 15–30 cm.

The nitrate-N soil-test interpretation is usually complemented by consideration of the previous crop, rainfall during the 2- to 4-week period before collecting soil samples, and soil properties that influence nitrate leaching. As work on this type of testing for nitrate in humid regions continues, it is becoming a useful diagnostic tool to decrease N fertilizer rates and reduce potential N pollution of water resources. As with all soil tests, local or

regional field calibrations are being conducted to determine critical concentration ranges and fertilizer recommendations.

Calibration of Soil Tests for Crop Production

Soil tests seldom extract the total amount of nutrients or elements in a soil sample. Soil tests have been developed to measure a fraction of the total soil nutrient concentration that correlates with plant growth. Interpreting a soil-test value requires an understanding of the impacts on test results of the extractant used, method of soil sampling, sample handling, and the intended use for the result. The amount of nutrient measured by various soil tests can vary widely depending on the extractant used. Important extractant properties include the concentration of the chemical compounds and the reaction time with the soil. The method used to measure the nutrient after extraction may be important for some nutrients. For example, colorimetric methods usually measure only orthophosphate P, while other methods may also measure other P forms. Also, the same soil test may measure widely different amounts of nutrients in soils with contrastingly different chemical and (or) mineralogical properties. Extractants used and interpretations of results often vary depending on soil properties and the purpose of soil testing (for example, measurement of total, soluble, or plant-available concentrations).

A soil-test method useful for predicting crop response to fertilization should produce values that are well correlated with plant nutrient uptake and growth. Greenhouse and field research is conducted to determine which soil-test extractant is best suited for a given combination of soil, crop, and growing conditions. Accurate interpretations of soil-test results and appropriate fertilizer recommendations require that the relationship between the amount of a nutrient measured by a given soil test and the crop response to the added nutrient must be known. The process of determining the probability of crop response at a given soil-test value is known as soil-test calibration and must be determined by field experimentation. The calibration procedure usually involves growing a crop or various crops in soils representative of the region on which the test will be used and the application of various fertilization rates. The soils should have a broad range of soil-test values that include deficient, optimum, and above-optimum values.

The crop yield response can be expressed as an absolute value or as a value relative to the yield achieved without nutrient addition. As an example, Figure 1 shows the relationships between the relative or absolute yield increase of maize and the amount of P extracted by a soil test. Although the specific shape of the relationship between soil-test values and crop growth or yield can differ, the general response is fairly consistent. At low soil-test values, crop yield is limited by nutrient deficiency. As soil-test values increase, yield increases at a reduced rate until a maximum yield value. At higher levels, there is no longer a relationship between the soil-test values and yield. The maximum yield value can remain approximately constant for a wide range of soil-test values or can decrease at excessively high nutrient levels.

The soil-test value at which crop growth or yield reaches a maximum is called the soil-test critical concentration. This is the soil-test value that best separates soils where a positive crop yield response to an added nutrient is likely from those soils where a response is not likely. Various mathematical models can calculate different critical concentration values and some models are asymptotic to a maximum. There is no reasonable agronomic consideration that justifies a single critical concentration. Moreover, economic

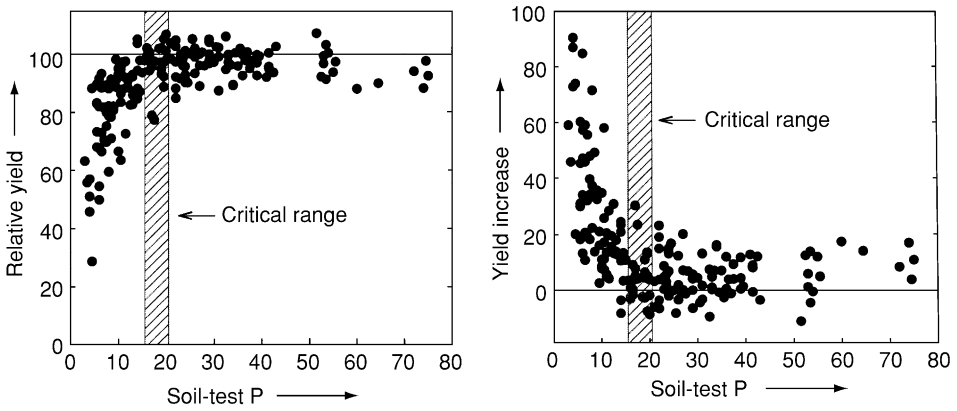


Figure 1 Example of the relationships between relative or absolute crop yield response and soil-test P values used for soil-test calibration.

considerations suggest that maximum economic yield perhaps should be the basis for determining critical concentrations instead of a physical maximum. Thus, critical concentration ranges are usually determined by using a combination of models and a variety of both agronomic and economic considerations. The soil-test response curve is often used to divide soil-test levels into several categories such as very low, low, optimum, high, or excessive.

A thorough understanding of crop response to nutrients and factors such as sampling date (season of the year), sampling method (mainly sampling depth), method of nutrient application (for example, band or broadcast), and application timing is needed to interpret soil-test results correctly. Also, other factors can influence crop growth (climatic factors, plant population, levels of other nutrients in the soil, and crop cultivars, among others). These factors may influence crop growth without affecting the amount of nutrient needed to produce a certain yield level or they may interact with soil nutrient levels.

Climatic factors are especially important for nutrients that are highly mobile in the soil, such as nitrate-N, B, Cl, and S. High nutrient mobility results from limited retention of nutrient chemical forms by soil organic or mineral constituents. High nutrient mobility results in more difficult soil-test calibrations and introduces more uncertainty when interpreting soil-test results. Levels of these nutrients in soil change markedly during the year as a result of changes in factors such as soil temperature and rainfall. Soil samples for relatively immobile nutrients (such as P, K, Ca, Mg, Cu, Zn) are often collected from the top 15–20 cm of soil, from where most of the nutrient uptake takes place. Other nutrient measurements, such as subsoil testing, are sometimes useful for the more immobile nutrients such as P or K but are more useful for mobile nutrients.

Calibration of Soil Tests for Environmental Purposes

The general concepts for soil-test calibration described for agronomic soil tests also apply to soil tests used for environmental purposes. With agronomic tests, the most important consideration is related to the meaning of a soil-test value for crop growth or yield. With environmental tests, the most important consideration relates to the meaning of a soil-test value for predicting environmental impacts. For several reasons, calibration and interpretation of soil tests used to predict potential environmental impacts are more complex than for agronomic uses. The most important reason is that a certain concentration of an element in the soil may have different impacts on the environment depending on the amounts absorbed by plants, amount retained by the soil, and potential delivery to environmentally critical areas. For example, soils with high clay content have a greater sorption capacity for P (and for other immobile elements) than coarse-textured soils, and a lesser amount of P is delivered as the distance from the source to channeled water flow increases. Establishment of nutrient loads that can result in unacceptable water-quality degradation will depend on factors such as the distance to a sensitive water body, the effectiveness of transport pathways for water and either total or dissolved nutrient fractions, and socioeconomic factors involving land or water use. A single, critical soil-test concentration range is unlikely to be found for nutrient impacts on the environment.

Not all areas of a field or landscape contribute equally to nutrient loss. For loss to occur, both a source of the nutrient and a mechanism for transporting it to water bodies or groundwater are needed. A key concept in effective management of nutrient pollution is to focus on the critical areas in which these two factors maximize the potential for loss. The P Index is an example of an assessment tool that considers soil-test values with other site-specific factors (such as hydrology, nutrient management, water-body proximity, and sensitivity to nutrient inputs) to determine the potential environmental risk of P levels in a field or parts of fields. The P Index includes source factors (soil-test values, as well as fertilizer and manure application rates, methods, and timing) and transport factors (such as soil erosion, surface runoff, subsurface drainage, and distance to streams or water bodies). Indexes are also being developed for N.

Agronomic soil-test methods designed to assess plant-available nutrients are sometimes used for environmental purposes. An effective environmental soil test should measure the nutrient forms most important to eutrophication or other negative environmental impacts. Eutrophication is a process whereby water bodies such as lakes, estuaries, or slow-moving streams receive excess nutrients that stimulate excessive aquatic plant or algae growth. This enhanced growth reduces dissolved oxygen in the water when dead plant or algae material decomposes, which causes ecosystem imbalances and may promote growth of toxic bacteria and unpleasant odor. Current research indicates that there are usually similar relationships between either agronomic or environmental soil P tests and dissolved P in surface runoff or subsurface drainage, except immediately after P additions. However, extending the concept of agronomic critical soil-test ranges to declare that nutrient concentrations above those needed for

optimum crop production will result in significant environmental impacts is not appropriate, because this ignores important aspects of nonpoint-source pollution. If soil tests are to be interpreted to predict the probability of nutrient pollution, calibrations relating soil-test results to specific measurements of environmental response (such as the concentration or load of P in surface runoff, for example) are necessary.

Quality of Soil Testing

Laboratory Quality Control

The quality of a laboratory soil-test result can be assessed on the basis of bias and precision. Bias refers to the deviation of the analytical result from the true value and measures accuracy of the result. Precision refers to the reproducibility of a given test value. A soil-testing laboratory could have good precision for a specific procedure but have bias in its results. A primary means of evaluating soil-test quality procedures as well as laboratory performance is through comparisons of tests performed on well-mixed subsamples of the same soil at different laboratories. In North America, a voluntary proficiency-testing program involving state and private laboratories was launched in the early 2000s to improve the quality of soil testing for agronomic purposes (the North American Proficiency Testing Program, NAPT). Results from this program indicate that, for example, the average variability of P tests is approximately 10–15% and that 65–70% of participating laboratories produced results within this level of variability. Variability is lower within any given laboratory. Data from the NAPT program indicate that precision levels for the P tests discussed are approximately 5–10% within each laboratory.

Soil Sampling for Soil Testing

For a soil-testing program to be effective, besides proper soil-test calibration and laboratory quality control, soil samples should be collected in a cost-effective manner and should accurately represent the nutrient levels in the area of interest. Sampling is a critical component of the soil-testing process because it usually represents the largest single source of error in soil testing. Many factors varying both spatially and temporally influence nutrient concentrations in soils. Sampling protocols must account for the great diversity in magnitude, structure, and spatial scale of nutrient variability present in agricultural fields and other ecosystems.

Soil nutrient variability within a crop field may be due to soil formation or management factors. Factors such as landscape position and soil parent material can cause great changes in soil texture, organic matter, drainage, and other properties. These properties affect nutrient levels directly through their influence on the amount of plant-available nutrient or indirectly though crop yield potential and, thus, the nutrient removal by crops. Variability caused by long-term history of management and land-use practices overlays that associated with soil-formation factors. The cumulative effects of nonuniform manure or fertilizer application are sources of potentially high soil-test variability. Proximity to livestock confinement areas, feed storage areas, and field boundaries are additional examples of historical factors that cause great variability in many fields. Small-scale variability usually predominates in fields with long histories of cropping and fertilizer or manure applications, especially when nutrients are applied using band methods. The challenge in these situations is to determine effective methods to delineate sampling areas within a field and the number of cores or borings needed for each composite sample to account for small-scale variability appropriately.

A variety of systematic and zone sampling approaches have been developed in different regions. The development of accurate and reliable global positioning technology, affordable geographic information systems, and variable-rate application equipment has led to widespread use of site-specific soil sampling approaches in North America. These sampling approaches are typically used to generate a soil fertility map to serve as an input to computer-controlled equipment for applying varying rates of one or more materials. One such approach is zone sampling, by which field subregions with more homogeneous properties than the field as a whole are delineated. Landscape position, soil color, soil mapping unit, and crop growth differences are examples of factors often used to help define management zones. Another site-specific approach involves systematic grid sampling, where soil-test patterns in a field are determined by means of a dense, systematic sample collection. A grid size of approximately 1 ha is most common today in the USA. Small-scale variability of P and K is so extreme in some fields that accurate within-field soil fertility mapping is impossible. Much uncertainty still exists regarding how best to perform site-specific soil sampling and generate accurate soil fertility maps.

Soil-test values may also change markedly with depth. This results from a combination of soil-forming and management factors and from the differential mobility of nutrients in soils. The elements with lower mobility tend to accumulate near the application point (often the most shallow soil layers), because they are more strongly retained by soil constituents than the more mobile nutrients. The tillage

system and the application method greatly influence vertical nutrient stratification. For example, the significant stratification of both P and K in no-till soils is well known, but deep nutrient application methods can reduce their concentration near the soil surface. Thus, the proper depth for soil sampling and its consistency are important considerations. Soil-test variability can also be the result of sampling-depth variability. Soil samples should be collected at the same depth used in the calibration research that serves as the quantitative basis for the soil-test interpretation and nutrient recommendations. Typically, this is a depth of 15–20 cm for the relatively immobile nutrients (such as P, K, Ca, Mg, Cu, Fe, Mn, Zn) and 30–60 cm for the more mobile nutrients (such as nitrate-N). However, the optimum soil-sampling depth varies with the objective of the soil testing. For example, soil sampling for predicting element loss with surface runoff may need to be shallower than for predicting uptake by plants or loss with subsurface drainage.

Use of Soil Tests to Determine Nutrient Loading Rates

Soil testing is becoming an established practice to determine whether environmentally unacceptable concentrations of plant nutrients or other elements are present in soils as a consequence of fertilization and waste applications. The term 'loading rate' implies that when excessive amounts of an element are applied to soils, it could be transported to environmentally critical areas and that the result may be harmful. Soils can be overloaded with a variety of nutrients, nonessential heavy metals, organic chemicals, and pathogenic organisms; and overloaded soils can affect the environment in various ways. For farmers and nutrient-management planners to make confident decisions about nutrient loading, soil tests must be able to predict loading rates and environmentally critical levels over a range of crops, soils, and environmental conditions.

In the USA and many European countries, regulatory initiatives are being established to mitigate water-quality problems associated with N and P water pollution. Other environmental efforts to which soil testing contributes include preventing nutrient toxicity to plants and protecting the food chain from accumulations of nutrients or elements harmful to animal and human health. These environmental problems usually occur at soil-test levels in excess of those required for optimal plant growth or as a result of clearly undesirable events (such as nutrient transport in floods or attached to eroded soil). Regarding phytotoxicity and feed or forage quality, soil tests have the potential to predict negative responses of plants to soils overloaded with a nutrient or nonessential elements.

Measurements and protocols for calibrating soil tests for environmental purposes are complex and are not well established. Soils can contribute to nonpoint-source pollution of water resources (mainly through erosion and surface runoff) even when soil-test levels are optimum for crop production. Nonincorporated surface applications of nutrient sources may result in high nutrient losses not correlated with soil-test levels. In this instance, nutrient transport may occur before any significant interaction between the nutrient source and the soil. A soil test alone does not permit accurate characterization of the risk of nutrient loss from soil to water. Estimates of the amount of nutrient directly lost from surface-applied materials are also needed. What happens between the field's edge and the surface water body, including the possibility for long-term retention of nutrients in buffer zones and any channel processes resulting in retention or release of nutrients, is a very important component of predicting loading rates. Complex diagnostic tools such as N and P risk indexes should be developed and tested for their efficacy to reduce nutrient pollution.

Summary

Soil testing is, and should be, an important component of more comprehensive environmental diagnostic tools. Researchers are attempting to define measurable parameters useful to improve estimates of the relationship between soil-test concentrations and loading rates to water resources. An example of such work is the ongoing international effort to develop better soil-testing procedures (including both methods of analysis and their interpretations) for N and P to identify watershed or field areas likely to contribute significant amounts of N and P to surface water and groundwater through erosion, surface runoff, or leaching. Growing evidence from field research demonstrates a strong relationship between soil-test values and concentrations of elements in surface runoff and subsurface drainage. However, local or regional calibrations are needed, because the relationships vary greatly across soil-test methods and soil properties. Moreover, the effective impact of a certain soil-test value on environmentally sensitive areas is strongly affected by the site hydrology and various transport mechanisms. Thus, if loading rate decisions regarding fertilizers or waste materials are made solely on the basis of soil-test concentrations, erroneous decisions will probably be made regarding potential element losses from fields. However, soil testing plays an important role

in optimizing crop production and protecting the environment.

See also: **Fertility**; **Fertilizers and Fertilization**; **Quality of Soil**

Further Reading

Carter MR (ed.) (1993) *Soil Sampling and Methods for Soil Analysis*. Boca Raton, FL: CRC Press.

Kamprath EG, Beegle DB, Fixen PE *et al.* (2000) *Relevance of Soil Testing to Agriculture and the Environment*. CAST Issue Paper 15. Ames, IA: Council for Agricultural Science and Technology.

Power JF and Dick W (eds) (2000) *Beneficial Uses of Agricultural, Municipal, and Industrial By-Products*. Madison, WI: Soil Science Society of America.

Steele K (ed.) (1995) *Animal Waste and the Land–Water Interface*. Boca Raton, FL: Lewis Publishers.

Tiessen H (ed.) (1995) *Phosphorus in the Global Environment: Transfers, Cycles, and Management*. Scientific Committee on Problems of the Environment (SCOPE), International Council of Scientific Unions (ICSU), United Nations Environment Programme (UNEP). Chichester, UK: John Wiley.

Westerman RL (ed.) (1990) *Soil Testing and Plant Analysis*, 3rd edn. Madison, WI: Soil Science Society of America.

TEXTURE

G W Gee, Pacific Northwest National Laboratory, Richland, WA, USA

Introduction

Soil consists of an assemblage of particles that differ widely in size and shape. Some particles are coarse enough to be seen easily with the naked eye, while others are so tiny they can only be seen with powerful electron microscopes. The term 'soil texture' is an expression of the dominant particle sizes found in soil and has both qualitative and quantitative connotations.

Qualitatively, 'soil texture' refers to the feel of soil material. A soil when rubbed by hand can have a feeling that ranges from coarse and gritty to fine and smooth. Experienced soil classifiers can press soil into their hands and tell its texture, coarse or fine, or gradations in between. The expressions 'light soil' and 'heavy soil' are frequently used to characterize the general physical behavior of different soils. A coarse-grained, sandy soil tends to be loose, well-aerated, and easy to cultivate, so it is called a 'light soil.' A fine-textured soil tends to absorb water easily and becomes plastic and sticky when wet, and tight, compact, and cohesive when dry, so it is called a 'heavy soil.' These terms are misleading, since coarse-textured soils are generally denser, having lower porosity, than fine-textured soils, and thus are heavier rather than lighter per unit mass. Table 1 shows the range of densities and porosities that can be found in soils of various textures.

Quantitatively, soil texture refers to the relative proportions of various sizes of soil particles, sometimes called 'size separates.'

Soil Particle Size: Measurements and Classification

The size of soil particles can vary over 6 orders of magnitude, ranging from large stones and rocks (greater than 0.25 m in size) to tiny, submicron clays (less than 0.0001 mm). The shape of soil particles also varies widely. Particle shape is often an indication of the mineral assemblage and the degree of soil weathering that has taken place. For beach sand, composed of quartz, particles are uniformly rounded and almost spherical. In contrast, most clay materials are far from spherical, many being platelets, looking more like playing cards (e.g., kaolinite) or tubes (e.g., halloysite) than marbles. In spite of the variation in shape, soil particles are sized according to their 'effective' particle diameter. Because of the huge size range, no single method can be used to measure particle diameters. Measurements are typically made with tape or ruler for large particles (boulders or cobbles) of more than a few centimeters in diameter, with sieves for particles as small as approximately

Table 1 Range of values found in bulk densities and associated properties for soils of various textures

Description	*Bulk density*[a] *($g\,cm^{-3}$)*	*Porosity*[b] *($cm\,cm^{-3}$)*	*Void ratio*[c] *($cm\,cm^{-3}$)*
Clay	1.20	0.54	1.17
Loam	1.40	0.49	0.96
Sand	1.60	0.39	0.64
Sandy loam[d]	1.90	0.28	0.39
Sandstone	2.12	0.20	0.25

[a]Bulk density, mass of solids per total volume.
[b]Porosity, total volume of pores (liquid and gas phase) per total volume.
[c]Void ratio, total volume of pores (liquid and gas phase) per solid volume.
[d]Compacted subsoil.

0.05 mm and with a variety of sedimentation methods (e.g., pipet, hydrometer, laser light-scattering) for particles less than 0.05 mm.

Size Measurements

Size measurements depend on separating the soil into primary particles. Methods for soil analysis consist of disaggregating by washing, vibrating, or otherwise dispersing the soil into single units. Typically the larger particles are measured directly by tape, ruler, or calipers. Particles from 100 mm to 0.05 mm are sieved. Materials less than 2 mm are generally washed through a nest of screens for a given period of time after treating with some kind of soil dispersant such as sodium hexametaphosphate, which swamps the clay surfaces with sodium, causing dispersion of the clay materials and accelerating disaggregation. For tropical soils or soils with high organic, iron, or aluminum contents, special treatments are deployed to remove these binding agents and release the primary particles. Soils of volcanic ash origin are notorious for aggregation, and pretreatment by chemical or mechanical dispersion can significantly affect the amount of clay produced. In one test, researchers have demonstrated that the clay (less than 0.002 mm) content of a volcanic ash soil varies from 1% to 56% depending on the amount of pretreatment imposed on the sample. Some tropical soils consist largely of clay-size particles yet behave much like sands, in terms of their water-storage and related physical properties, because of the high degree of natural aggregation.

Fine Soil Measurements

Fine soil measurements rely on sedimentation of a soil suspension. The sample is stirred in water and dispersing agent, then the suspension is allowed to settle. Sampling is done at preselected times at a fixed depth in the suspension using either a pipet or hydrometer. Pipet techniques take a physical sample after a given settlement period, while hydrometer methods rely on estimates of effective densities of the settling particles. Knowing settlement depths allows for a straightforward calculation of the effective particle diameter. The Stokes law for viscous drag on a settling spherical object is combined with buoyant and gravitational forces to obtain the settlement rate. Combining forces and solving for the particle velocity yields:

$$v = d^2 g\,(\rho_p - \rho_w)/(18\eta) \qquad [1]$$

where d is equivalent particle diameter, g is gravity acceleration, ρ_p is particle density, ρ_w, is solution density, and η is solution viscosity.

Figure 1 shows the pipet method and Figure 2 shows a typical hydrometer used for soil-solution density measurements. This equipment is relatively inexpensive and readily available and, for routine analysis in soils, sedimentation, and engineering laboratories, the pipet and hydrometer are the standard tools for fine-fraction analysis. In recent years, the need for diagnostic tools in soil genesis and mineralogy have led to more sophisticated techniques being developed and tested. Laser light-scattering techniques are now being used in soil science more routinely to obtain size distributions of soil materials. Figure 3 shows the primary components of a laser light-scattering device.

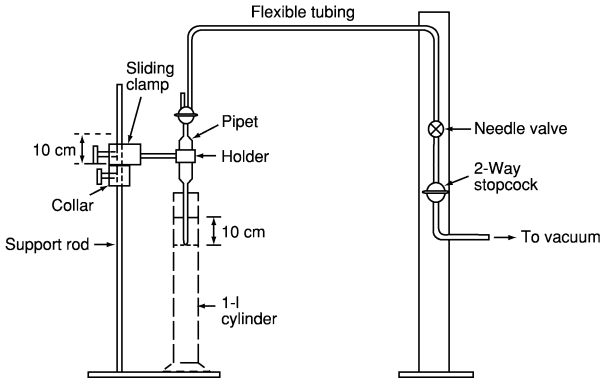


Figure 1 Schematic of pipet apparatus and stand for sedimentation analysis of soil-particle size.

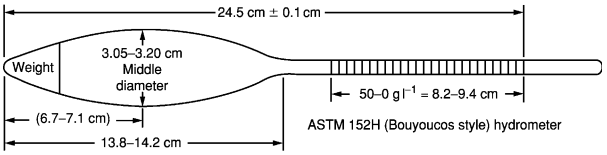


Figure 2 Schematic of ASTM 152H-type hydrometer used for soil particle-size analysis.

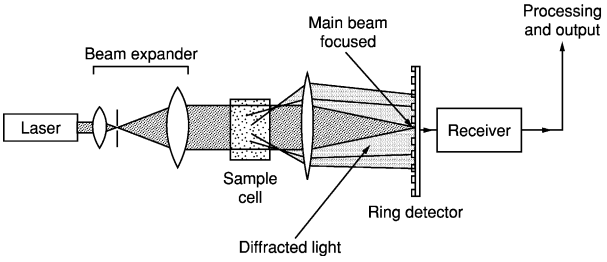


Figure 3 Laser diffraction particle-size analyzer showing the primary components: light sources, sample, focusing lens, detector, and processing system. (Reproduced with permission from Syvitski JPM (ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.)

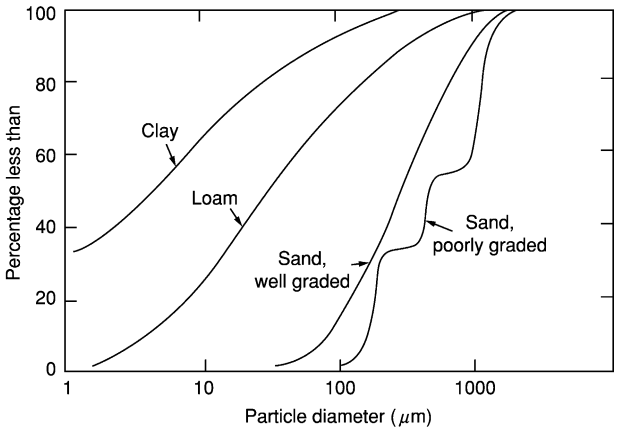


Figure 4 Particle-size distribution curves for several soil. (Adapted from Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.)

Figure 4 shows particles-size distribution for a range of soils. All but the poorly sorted sand exhibit log-normal characteristics and might be considered fractal in nature. Much has been made about the fractal nature of particle-size distributions of soils and the possibility of relating the fractal dimension of the particle size to an equivalent pore-size distribution and then subsequently obtaining estimates of hydraulic properties from these parameters. In spite of this temptation, recent analysis suggests that many soils are actually multifractal, that is, they do not have the same fractal dimension over the entire particle-size range, so caution must be exercised in extrapolating fractal parameters beyond the range over which they might apply. Nevertheless, particle-size distribution curves are very useful and can help diagnose the origins of a soil as well as estimate numerous physical and chemical properties.

Figure 5 shows how various methods (sieving, sedimentation, and laser methods) are combined to obtain the size distribution of a soil sample. It appears that, with proper calibration and similar pretreatments (i.e., chemical and physical pretreatment), different methods can yield comparable results.

Size and Textural Classification Schemes

There are more than a dozen different schemes for classifying size separates. Figure 6 shows four of the more common schemes, including the size limits used in each scheme and the mesh numbers for the screens used in sieving the particles. Note that the size separates have arbitrary size limits, and for the most part the differences between schemes are relatively small, with division between the International Society of

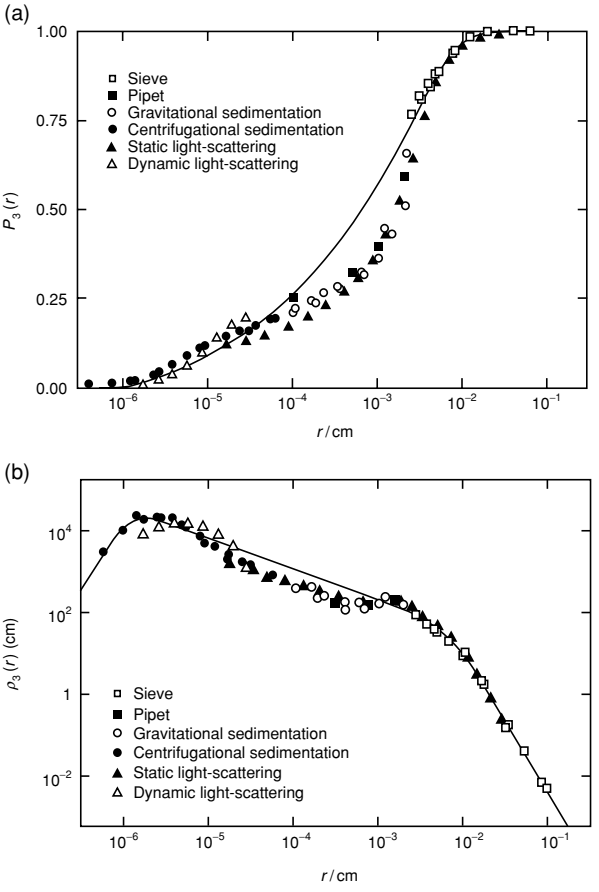


Figure 5 Cumulative mass distribution $P_3(r)$ (a) and number distribution density $\rho_3(r)$ for Buchberg soil as a function of particle radius for different measurement methods (solid lines represent empirical fit of three power-law functions). (Reproduced with permission from Wu Q, Borkovec M, and Sticher H (1993) On particle size distributions in soils. *Soil Science Society of America Journal* 57: 883–890.)

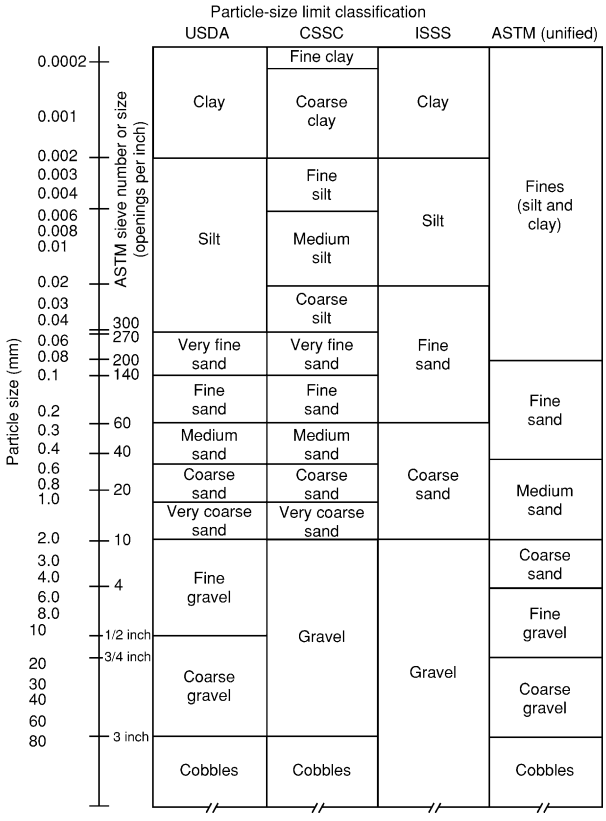


Figure 6 Particle-size limits according to several current classification schemes. USDA, US Department of Agriculture; CSSC, Canada Soil Survey Committee; ISSS, International Society of Soil Science; ASTM, American Society of Testing and Materials. (Adapted from Soil Survey Staff (1975); and McKeague JA (ed.) (1978) *Manual on Soil Sampling and Methods of Analysis*. Ottawa, Canada: Canadian Society of Soil Science.)

Soil Science (ISSS) and the US Department of Agriculture (USDA) classification being primarily the separation between the sand and clay fractions, with ISSS defining the uppermost silt limit at 0.02 mm and the USDA upper silt limit defined at 0.05 mm. Conversions from one system to another have been made and incorporated into user-friendly computer programs that translate one system readily into the another.

Textural Classification

Since soils rarely consist of one size range, textural classification is based on different combinations of size separates. The classification schemes that have been developed are as diverse as the applications for which the textural data are used. Two primary textural classifications are discussed, the universal classification scheme, used in engineering and adopted by the American Society of Testing and Materials (ASTM), and the USDA scheme that is widely used in agronomic science.

Universal classification (engineering) scheme Table 2 shows the engineering classification scheme used by the ASTM. In order to use this textural scheme, particle-size distribution is needed and certain engineering properties are required. From the size-distribution curve, the coefficient of curvature, C_c, can be obtained as:

$$C_c = (D_{30})^2/(D_{10} \times D_{60}) \quad [2]$$

where D_{60}, D_{30}, and D_{10} are the particle sizes corresponding to 60, 30, and 10% finer on the cumulative particle size distribution curve (Figure 7) and C_u, the coefficient of uniformity, is obtained as:

$$C_u = D_{60}/D_{10} \quad [3]$$

These properties are helpful in defining the types of gravel or sand that are present. C_u and C_c are shown in Figure 7 for a coarse soil. In this case, $C_u = 200$ and $C_c = 5.6$. If less than 5% of the sample passes the 200-mesh (0.075 mm) sieve, this soil is classed as a poorly graded gravel, GP. The other engineering properties are measures of the amount and type of fine materials present in the soil and are obtained from the liquid limit (LL) and plasticity index (PI). These properties (the so-called Atterberg limits) are obtained from standard methods and generally plotted on a PI versus LL diagram (Table 2). High LLs imply high clay content, and high PI implies the presence of swelling-type clays (e.g., smectites). Swelling clays such as smectites are typically more plastic than nonswelling clays (e.g., kaolinites). Kaolinitic clays are not plastic and grade out as materials that fall below the 'A' line on the PI versus LL curve (Table 2). In total, 16 different textural classes are used for engineering purposes, ranging from well-graded gravel (WG) to silts and clay (CH) and organic soils (PT).

USDA classification For agronomic and other soil-science applications, the textural designations are quite straightforward. It is assumed that the organic fraction is minimal and that only three size separates, sand, silt, and clay fractions, are being considered. In the USDA system, the fractions of the three size separates, sand (2–0.05 mm), silt (0.05–0.002 mm), and clay (0.002 mm) are plotted and 12 distinct textures are assigned according to the mix of separates (Figure 8). For the textural diagram, the assumption is:

$$m_d + m_t + m_y = 1 \quad [4]$$

where m_d is size fraction of sand, m_t is size fraction of silt, and m_y is size fraction of clay. Only two fractions are required to obtain texture from the textural triangle.

Table 2 Unified soil classification system, including plasticity chart

Criteria for assigning group symbols and names using laboratory tests[a]				*Soil classification*	
				Group symbol	*Group name*[b]
Coarse-grained soils more than 50% retained on No. 200 sieve	Gravels more than 50% of coarse fraction retained on No. 4 sieve	Clean gravels less than 5% fines[c]	$C_u > 4$ and $1 < C_c < 3$[e]	GW	Well-graded gravel[f]
			$C_u < 4$ and/or $1 > C_c > 3$[e]	GP	Poorly graded gravel[f]
		Gravels with fines more than 12% fines[c]	Fines classify as ML or MH	GM	Silty gravel[f,g,h]
			Fines classify as CL or CH	GC	Clayey gravel[f,g,h]
	Sands 50% or more of coarse fraction passes, No. 4 sieve	Clean sands less than 5%	$C_u > 6$ and $1 < C_c < 3$[e]	SW	Well-graded sand[i]
			$C_u < 6$ and/or $1 > C_c > 3$[e]	SP	Poorly graded sand[i]
		Sands with fines more than 12% fines[d]	Fines classify as ML or MH	SM	Silty sand[g,h,i]
			Fines classify as CL or CH	SC	Clayey sand[g,h,i]
Fine-grained soils 50% or more passes, No. 200 sieve	Silts and clays liquid limit less than 50	Inorganic	PI > 7 and plots on or above 'A' line[j]	CL	Lean clay[k,l,m]
			PI < 4 or plots below 'A' line[j]	ML	Silt[k,l,m]
		Organic	Liquid limit – oven-dried < 0.75 Liquid limit – not dried	OL	Organic clay[k,l,m,n] Organic silt[k,l,m,o]
	Silts and clays liquid limit 50 or more	Inorganic	PI plots on or above 'A' line	CH	Fat clay[k,l,m]
			PI plots below 'A' line	MH	Elastic silt[k,l,m]
		Organic	Liquid limit – oven-dried < 0.75 Liquid limit – not dried	OH	Organic clay[k,l,m,p] Organic silt[k,l,m,q]
Highly organic soils		Primarily organic matter, dark in color, and organic odor		PT	Peat

[a]Based on the material passing the 75-mm sieve.
[b]If field sample contained cobbles or boulders, or both, add 'with cobbles or boulders, or both' to group name.
[c]Gravels with 5–12% fines require dual symbols: GW-GM well-graded gravel with silt; GW-GC well-graded gravel with clay; GP-GM poorly graded gravel with silt; GP-GC poorly graded gravel with clay.
[d]Sands with 5–12% fines require dual symbols: SW-SM well-graded sand with silt; SW-SC well-graded sand with clay; SP-SM poorly graded sand with silt; SP-SC poorly graded sand with clay.
[e]$C_u = D_{60}/D_{10}$ and $C_c = (D_{30})^2/(D_{10} \times D_{60})$.
[f]If soil contains 15% or more sand, add 'with sand' to group name.
[g]If fines classify as CL-ML, use dual symbol GC-GM, or SC-SM.
[h]If fines are organic, add 'with organic fines' to group name.
[i]If soil contains ≥15% gravel, add 'with gravel' to group name.
[j]If Atterberg limits plot in hatched area, soil is a CL-ML, silty clay.
[k]If soil contains 15–29% plus No. 200, add 'with sand' or 'with gravel,' whichever is predominant.
[l]If soil contains ≥30% plus No. 200, predominantly sand, add 'sandy' to group name.
[m]If soil contains ≥30% plus No. 200, predominantly gravel, add 'gravelly' to group name.
[n]PI ≥4 and plots on or above 'A' line.
[o]PI <4 or plots below 'A' line.
[p]PI plots on or above 'A' line.
[q]PI plots below 'A' line.
(Reproduced with permission from ASTM (2000) *Standard Test for Classification of Soils for Engineering Purposes*. D 2487–98. 2000 Annual Book of ASTM Standards, 04.08: 238–247. Philadelphia, PA: ASTM.)

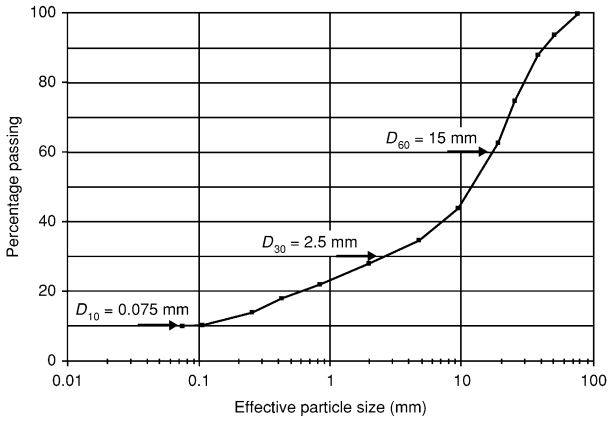


Figure 7 Particle-size distribution curve for a coarse soil, using sieve analysis. Also shown are the D_{60}, D_{30}, and D_{10} values required to compute the coefficient of curvature, C_c, and coefficient of uniformity, C_u. D_{60}, D_{30}, and D_{10} are the particle sizes corresponding to 60, 30, and 10% finer (percent passing).

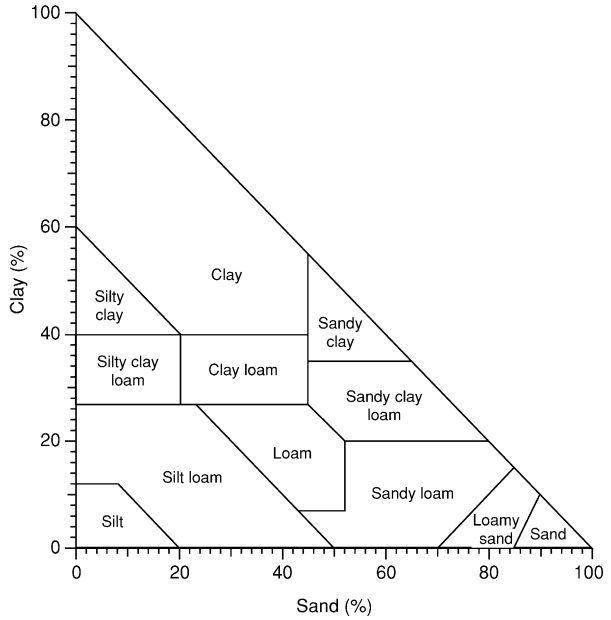


Figure 8 Textural triangle for USDA textural classification scheme.

Use of Texture Data for Estimating Hydraulic Properties

Soil scientists use soil-texture information to make qualitative judgments about a number of physical properties, but, until recently, there have been limited attempts to use texture for quantitative assessments of soil properties. Because the division of particles into size classes is entirely arbitrary, it is not surprising that texture data do not generally relate directly to fundamental physical properties. For example, it is not easy to determine how much silt it takes to equal a unit of clay, or if silt and sand fractions have any effect on a particular soil property of interest. In addition, there are underlying issues of soil mineralogy, organic versus inorganic fraction, particle-shape factors, and aggregation that complicate matters, so that the use of particle size data often requires some degree of empiricism in relating to such things as water-retention characteristics or effective surface area.

A more fundamental approach for describing soil texture is to use a statistical description for the particle-size distribution. A basic property of soil is the mean effective particle diameter. Particle-size distributions tend to be log-normal or fractal, so the appropriate means and standard deviations are log means and log standard deviations. These can be computed as:

$$d_g = \exp\left(\sum_i m_i \ln d_i\right) \quad [5]$$

$$\sigma_g = \exp\left(\sqrt{\sum_i m_i (\ln d_i)^2 - (\ln d_g)^2}\right) \quad [6]$$

where d_g is geometric mean diameter (in micrometers), m_i is mass fraction of component i (in grams), d_i is mean diameter of component i (in micrometers), and σ_g is geometric standard deviation (in micrometers).

When sand, silt, and clay fractions are known, the geometric mean diameter and the geometric standard deviation can be estimated in terms of the silt and clay fractions by using the following expressions:

$$d_g = \exp(5.756 - 3.454\, m_t - 7.712\, m_y) \quad [7]$$

$$\sigma_g = \exp[33.14 - 27.84\, m_t - 29.13\, m_y - (\ln d_g)^2]^{1/2} \quad [8]$$

where m_t is mass fraction of silt and m_y is mass fraction of clay.

Table 3 shows typical silt and clay fractions, geometric mean particle diameter, and geometric standard deviation for the 12 textural classes of the USDA classification scheme as identified in Figure 7. The standard textural triangle (Figure 7) can be transformed into a diagram based on the geometric mean particle size and the geometric mean standard deviation. From such a construct, textures can be interpreted in terms of more quantitative relationships. Coarser-soil soils can be easily incorporated into such a scheme. The standard textural triangle ignores the soil fraction greater than 2 mm. Frequently there is a need to incorporate this fraction into the size distribution, particularly when hydrologic properties such as water retention and unsaturated hydraulic conductivity are of interest.

Table 3 Typical silt and clay fractions and geometric mean particle diameter and geometric standard deviation for the 12 US Department of Agriculture textural classes

Texture	*Silt fraction*	*Clay fraction*	d_g (μm)	σ_g (μm)
Sand	0.05	0.03	212	4.4
Loamy sand	0.12	0.07	122	8.7
Sandy loam	0.25	0.10	62	12.2
Loam	0.40	0.18	20	16.3
Silt loam	0.65	0.15	11	9.6
Silt	0.87	0.07	10	4.1
Sandy clay loam	0.13	0.27	25	28.4
Clay loam	0.34	0.34	7	23.1
Silty clay loam	0.58	0.34	3	10.9
Sandy clay	0.07	0.40	11	39.7
Silty clay	0.45	0.45	2	13.8
Clay	0.20	0.60	1.5	22.8

Summary

Much effort has been expended in recent years to develop methods that relate particle-size distribution to pore size and subsequently to a host of other soil properties such as water retention and unsaturated hydraulic conductivity. While this is appealing, particularly when particle-size distribution data are available and other data are absent, this approach must be viewed with caution, because the particle-size distribution data obtained in most routine applications are at best empirical in nature. It needs to be pointed out that while particle-size distributions are important and offer what is often considered the most fundamental properties of a soil, the actual size distribution obtained from a laboratory analysis is almost always dependent upon the method used to obtain the particle-size distribution. As an example, researchers have found that some soils of volcanic ash origin have clay-sized particles that show great resistance to dispersion when subjected to gentle shaking but break down under high-energy source treatment. The chemical pretreatment and amount of mechanical work done on the soil are dictated by arbitrary decisions, so there is no absolute size distribution for a given soil. Because we are dealing with natural systems, we must rely on empirical methods and arbitrary limits and thus are constrained by the methods deployed to obtain size distributions and, resulting textural descriptions for soils.

Further Reading

Allen T (1981) *Particle Size Measurement*. London, UK: Chapman and Hall.

American Society for Testing and Materials (2000) *Standard Test for Classification of Soils for Engineering Purposes*, pp. 238–247. D 2487-98. 2000 Annual Book of ASTM Standards, vol. 04.08. Philadelphia, PA: ASTM.

American Society for Testing and Materials (2000) *Standard Test Method for Liquid Limit, Plastic Limit, and Plasticity Index of Soils*, pp. 546–558. D 4318-98. 2000 Annual Book of ASTM Standards, vol. 04.08. Philadelphia, PA: ASTM.

American Society for Testing and Materials (2000) *Standard Test Method for Particle-Size Analysis of Soils*, pp. 10–17. D 422-63 (1998). 2000 Annual Book of ASTM Standards 04.08. Philadelphia, PA: ASTM.

Christopher TBS and Mokhtaruddin AM (1996) A computer program to determine soil textural class in 1-2-3 for Windows and Excel. *Communications in Soil Science Plant Analysis* 27: 2315–2319.

Folk RL (1980) *Petrology of Sedimentary Rocks*, 2nd edn. Austin, TX: Hemphill.

Gee GW and Or D (2002) Particle size analysis. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, pp. 255–293. *Physical Methods*. SSSA Book Series No. 5. Madison, WI: Soil Science Society of America.

Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.

Kubota T (1972) Aggregate-formation of allophanic soils: effects of drying on the dispersion of the soils. *Soil Science and Plant Nutrition* 18: 79–87.

Loveland PJ and Whalley WR (2001) Particle size analysis. In: Smith KA and Mullins CE (eds) *Soil and Environmental Analysis: Physical Methods*, 2nd edn. New York: Marcel Dekker.

McKeague JA (ed.) (1978) *Manual on Soil Sampling and Methods of Analysis*. Ottawa, Canada: Canadian Society of Soil Science.

Muggler CC, Pape T, and Buurman P (1996) Laser grain-size determination in soil genetic studies. 2. Clay content, clay formation, and aggregation in some Brazilian oxisols. *Soil Science* 162: 219–228.

Shiozawa S and Campbell GS (1991) On the calculation of mean particle diameter and standard deviation from sand, silt, and clay fractions. *Soil Science* 152: 427–431.

Syvitski JPM (ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.

US Department of Agriculture (1982) *Procedures for Collecting Soil Samples and Methods of Analysis for Soil Survey*. Soil Survey Investigations Report No. 1. Washington, DC: Soil Conservation Service.

Warrick AW (2002) *Soil Physics Companion*. Boca Raton, FL: CRC Press.

Yong RN and Warkentin BP (1996) *Introduction to Soil Behavior*. New York: Macmillan.

THERMAL PROPERTIES AND PROCESSES

D Hillel, Columbia University, New York, NY, USA

Introduction

The soil temperature, as it varies in time and space, is a factor of primary importance in determining the rates and directions of soil physical processes and of energy and mass exchange with the atmosphere. Temperature governs evaporation and aeration, as well as the types and rates of chemical reactions that take place in the soil. Finally, soil temperature strongly influences biological processes such as seed germination, seedling emergence and growth, root development, and microbial activity.

Soil temperature varies in response to changes in the radiant, thermal, and latent energy exchange processes that take place primarily through the soil surface. The effects of these phenomena are propagated into the soil profile via a complex series of transport processes, the rates of which are affected by time-variable and space-variable soil properties.

Modes of Energy Transfer

In general, there are three principal modes of energy transfer: radiation, convection, and conduction. By radiation, we refer to the emission of energy in the form of electromagnetic waves from all bodies above 0 K. According to the Stefan–Boltzmann law, the total energy emitted by a body, J_t, integrated over all wavelengths, is proportional to the fourth power of the absolute temperature T of the body's surface. This law is usually formulated:

$$J_t = \epsilon\sigma T^4 \qquad [1]$$

where σ is a constant and ϵ is the emissivity coefficient, which equals unity for a perfect emitter (generally called a 'black body'). The absolute temperature also determines the wavelength distribution of the emitted energy. The Wien law states that the wavelength of maximal radiation intensity λ_m is inversely proportional to the absolute temperature:

$$\lambda_m = 2900/T \qquad [2]$$

where λ_m is in micrometers. The actual intensity distribution as a function of wavelength and temperature is given by Planck's law:

$$E_\lambda = C_1/\lambda^5[\exp(C_2/\lambda T) - 1] \qquad [3]$$

where E_λ is energy flux emitted in a given wavelength range and C_1 and C_2 are constants.

Since the temperature of the soil surface is generally of the order of 300 K (though it can range from below 273 K, the freezing point, to 330 K or higher), the radiation emitted by the soil surface has its peak intensity at a wavelength of approximately 10 μm and its wavelength distribution over the range of 3–50 μm. This is in the realm of infrared, or heat, radiation.

A very different spectrum is emitted by the sun, which acts as a black body at an effective surface temperature of approximately 6000 K. The sun's radiation includes the visible light range of 0.3–0.7 μm, as well as some infrared radiation of greater wavelength (up to approx. 3 μm) and some ultraviolet radiation ($\lambda < 0.3$ μm). Since there is very little overlap between the two spectra, it is customary to distinguish between them by calling the incoming solar spectrum 'short-wave' radiation, and the spectrum emitted by the Earth 'long-wave' radiation.

The second mode of energy transfer, called 'convection,' involves the movement of a heat-carrying mass, as in the case of ocean currents or atmospheric winds. An example more pertinent to soil physics would be the infiltration of warm waste water (from, e.g., a power plant) into an initially cold soil.

Conduction, the third mode of energy transfer, is the propagation of heat within a body by internal molecular motion. Since temperature is an expression of the kinetic energy of a body's molecules, the existence of a temperature difference within a body will normally cause the transfer of kinetic energy by the numerous collisions of rapidly moving molecules from the warmer region of the body to their neighbors in the colder region. The process of heat conduction is thus analogous to diffusion and, in the same way that diffusion tends in time to equilibrate a mixture's composition throughout, heat conduction tends to equilibrate a body's internal distribution of molecular kinetic energy – that is, its temperature.

In addition to the three modes of energy transfer described, there is a composite phenomenon which one may recognize as a fourth mode, namely latent heat transfer. A prime example is the process of distillation, which includes the heat-absorbing stage of evaporation, followed by the convective or diffusive movement of the vapor, and ending with the heat-releasing stage of condensation. A similar catenary process can also occur in transition back and forth

from ice to liquid water in soils subject to freezing and thawing.

Energy Balance for a Bare Soil

The radiation balance of a bare soil surface can be written:

$$J_n = (J_s + J_a)(1 - \alpha) + J_{li} - J_{lo} \quad [4]$$

Here J_n is the net radiation, that is, the sum of all incoming-minus-outgoing radiant energy fluxes; J_s the incoming flux of short-wave radiation directly from the sun and J_a the short-wave diffuse radiation from the atmosphere (sky); J_{li} the incoming long-wave radiation flux from the sky and J_{lo} the outgoing long-wave radiation emitted by the soil; and, finally, α is the albedo, or reflectivity coefficient, which is the fraction of incoming short-wave radiation reflected by the soil surface rather than absorbed by it. In the present context, all terms that do not pertain to the soil, namely J_s, J_a, and J_{li}, are disregarded.

The albedo α is an important characteristic of soil surfaces, and it can vary widely in the range of 0.1–0.4, depending upon the soil's basic color (whether dark or light), the surface's roughness, and the inclination of the incident radiation relative to the surface. In the short term, the albedo also depends on the changing wetness of the exposed soil. The drier the soil, the smoother its surface, and the brighter its color, the higher its albedo. To a certain extent, the albedo can be modified by various surface treatments such as tillage and mulching.

Apart from the reflected short-wave radiation, the soil also emits long-wave radiation. In accordance with eqn [1], the emitted flux J_{lo} depends primarily on soil surface temperature but is also affected by the soil's emissivity. This parameter, in turn, depends on soil wetness, but its range of variation is generally small, i.e., between 0.9 and 1.0.

The net radiation received by the soil surface is transformed into heat, which warms the soil and air and vaporizes water. We can write the surface energy balance as follows:

$$J_n = S + A + LE \quad [5]$$

where S is the soil heat flux (the rate at which heat is transferred from the surface downward into the soil profile), A is the 'sensible' heat flux transmitted from the surface to the air above, and LE is the evaporative heat flux, a product of the evaporative rate E and the latent heat per unit quantity of water evaporated, L.

The total surface energy balance (combining eqns [4] and [5]) is therefore:

$$(J_s + J_a)(1 - \alpha) + J_{li} - J_{lo} - S - A - LE = 0 \quad [6]$$

Conventionally, all components of the energy balance are taken as positive if directed toward the surface, and negative otherwise.

Conduction of Heat in Soil

The conduction of heat in solid bodies was analyzed as long ago as 1822 by Fourier, whose name is associated with the linear transport equations that have been used ever since to describe heat conduction. The first law of heat conduction, known as the Fourier law, states that the flux of heat in a homogeneous body is in the direction of, and proportional to, the temperature gradient:

$$q_h = -\kappa \nabla T \quad [7]$$

Here q_h is the thermal flux (i.e., the amount of heat conducted across a unit cross-sectional area in unit time), κ is thermal conductivity, and ∇T the spatial gradient of temperature T. In one-dimensional form, this law is written:

$$q_h = -\kappa_x \partial T/\partial x \quad \text{or} \quad q_h = -\kappa_z \partial T/\partial_z \quad [8]$$

Here $\partial T/\partial x$ is the temperature gradient in any arbitrary direction designated x, and $\partial T/\partial z$ is, specifically, the gradient in the vertical direction representing soil depth ($z = 0$ being the soil surface). The subscripts attached to the thermal conductivity term are meant to account for the possibility that this parameter may have different values in different directions (i.e., that it may be nonisotropic). The negative sign in these equations is due to the fact that heat flows from a higher to a lower temperature (i.e., in the direction of, and in proportion to, a negative temperature gradient).

Equation [7] is sufficient to describe heat conduction under steady-state conditions, that is, where the temperature at each point in the conducting medium is invariant and the flux is constant in time and space. To account for nonsteady (transient) conditions, we need a second law analogous to Fick's second law of diffusion. To obtain the second law of heat conduction, the principle of energy conservation in the form of the continuity equation is invoked, which states that, in the absence of any sources or sinks of heat, the time rate of change in heat content of a volume element of the conducting medium must equal the change of flux with distance:

$$\rho c_m \partial T/\partial t = -\nabla \cdot q_h \quad [9]$$

where ρ is mass density and c_m specific heat capacity per unit mass (called simply 'specific heat' and

defined as the change in heat content of a unit mass of the body per unit change in temperature). The product ρc_m (often designated C) is the specific heat capacity per unit volume, and $\partial T/\partial t$ is the time rate of temperature change. Note that ρ represents the total mass per unit volume, including the mass of water in the case of a moist soil. The symbol ∇ ('del') is the shorthand representation of the three-dimensional gradient. An equivalent form of eqn [9] is:

$$\rho c_m \partial T/\partial t = -(\partial q_x/\partial x + \partial q_y/\partial y + \partial q_z/\partial z)$$

where x, y, z are the orthogonal direction coordinates.

Combining eqn [9] with [7] gives the second law of heat conduction:

$$\rho c_m \partial T/\partial t = -\nabla \cdot (\kappa \nabla T) \quad [10]$$

which, in one-dimensional form, is:

$$\rho c_m \partial T/\partial t = \partial/\partial x(\kappa \partial T/\partial x) \quad [11]$$

Sometimes there is need to account for the possible occurrence of heat sources or sinks in the realm where heat flow takes place. Heat sources include such phenomena as organic matter decomposition, wetting of initially dry soil material, and condensation of water vapor. Heat sinks are generally associated with evaporation. Combining all these sources and sinks into a single term S, we can rewrite the last equation:

$$\rho c_m \partial T/\partial t = \partial/\partial x(\kappa \partial T/\partial x) \pm S(x,t) \quad [12]$$

in which the source–sink term is shown as a function of both space and time.

Volumetric Heat Capacity of Soils

The volumetric heat capacity C of a soil is defined as the change in heat content of a unit bulk volume of soil per unit change in temperature. Its units are calories per cubic centimeter per degree (Kelvin) or joules per cubic meter per degree. As such, C depends on the composition of the solid phase (mineral and organic constituents) of the soil, on bulk density, and on soil wetness (Table 1).

The value of C can be estimated by summing the heat capacities of the various constituents, weighted according to their volume fractions:

$$C = \sum f_{si} C_{si} + f_w C_w + f_a C_a \quad [13]$$

Here, f denotes the volume fraction of each phase: solid (subscripted 's'), water ('w'), and air ('a'). The solid phase includes a number of components subscripted 'i,' such as various minerals and organic matter, and the symbol Σ indicates the summation of the products of their respective volume fractions and heat capacities. The C value for water, air, and each component of the solid phase is the product of the particular density and the specific heat per unit mass (i.e., $C_w = \rho_w c_{mw}$; $C_a = \rho_a c_{ma}$; $C_{si} = \rho_{si} c_{msi}$).

Table 1 Densities and volumetric heat capacities of soil constituents (at 10°C) and of ice (at 0°C)

	Density ρ		*Heat capacity C*	
Constituent	*g cm^{-3}*	*kg m^{-3}*	*cal cm^{-3} K*	*J m^{-3} K*
Quartz	2.66	2.66×10^3	0.48	2.0×10^6
Other minerals (mean)	2.65	2.65×10^3	0.48	2.0×10^6
Organic matter	1.3	1.3×10^3	0.6	2.5×10^6
Water (liquid)	1.0	1.0×10^3	1.0	4.2×10^6
Ice	0.92	0.92×10^3	0.45	1.9×10^6
Air	0.001 25	1.25	0.003	1.25×10^3

Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Most of the minerals composing soils have nearly the same values of density (approx. 2.65 g cm^{-3} or 2.65×10^3 kg m^{-3}) and of heat capacity (0.48 cal cm^{-3} K or 2.0×10^6 J m^{-3} K). Since it is difficult to separate the different kinds of organic matter present in soils, it is tempting to group them all into a single constituent (with mean density of approximately 1.3 g cm^{-3} or 1.3×10^3 kg m^{-3}, and a mean heat capacity of approximately 0.6 cal cm^{-3} K or 2.5×10^6 J m^{-3} K).

The density of water is less than half that of mineral matter (approx. 1 g cm^{-3} or 1.0×10^3 kg m^{-3}); its specific heat is more than twice as large (1 cal cm^{-3} K, or 4.2×10^6 J m^{-3} K). Finally, since the density of air is only approximately 1/1000 that of water, its contribution to the specific heat of the composite soil can be neglected generally.

Thermal Conductivity of Soils

Thermal conductivity, designated κ, is defined as the quantity of heat transferred through a unit area of the conducting body in unit time under a unit temperature gradient. As shown in Table 2, the thermal conductivities of specific soil constituents differ very markedly (see also Table 3). Hence the space-averaged (macroscopic) thermal conductivity of a soil depends upon its mineral composition and organic matter content, as well as on the volume fractions of water and air.

Since the thermal conductivity of air is very much smaller than that of water or solid matter, a high air content (or low water content) corresponds to a low

thermal conductivity. Moreover, since the proportions of water and air vary continuously, κ is also time-variable. Soil composition is seldom uniform in depth, hence κ is generally a function of depth as well as of time. Unlike heat capacity, thermal conductivity is sensitive not merely to the volume composition of a soil but also to the sizes, shapes, and spatial arrangements of the soil particles.

The relationship between the overall thermal conductivity of a soil and the specific conductivities and volume fractions of the soil's constituents is very intricate, as it involves the internal geometry or structure of the soil matrix and the mode of transmission of heat from particle to particle and from phase to phase.

The dependence of thermal conductivity and diffusivity on soil wetness is illustrated in **Figure 1**. The influence of latent heat transfer by the diffusion of water vapor in the air-filled pores is proportional to the temperature gradient in these pores. It can be taken into account by adding to the thermal conductivity of air an apparent conductivity due to evaporation, transport, and condensation of water vapor (the so-called vapor-enhancement factor). This value is strongly temperature-dependent and rises rapidly with increasing temperature.

Because soil water potential depends on temperature, the development of a temperature gradient generally induces the movement of water as well as of heat. Hence techniques for measuring heat transfer through a soil sample based on steady-state heat flow between two planes maintained at a constant temperature differential involve the risk of changing the sample's internal moisture distribution and therefore its thermal properties. During the process of measurement, the soil near the warmer plane becomes drier, while that near the cooler plane becomes wetter. Early attempts to measure thermal conductivity failed to recognize this pitfall as they purported to maintain constant soil moisture conditions during prolonged, steady-state heat flow. Hence their results can only be considered approximations at best. While steady-state methods may be sufficiently accurate for measuring thermal conductivity of dry soils, short-term, transient heat-flow techniques are preferable, in principle, for moist soils.

Table 2 Thermal conductivities of soil constituents (at 10°C) and of ice (at 0°C)

Constituent	*mcal cm^{-1} sK*	*Wm^{-1} K*
Quartz	21	8.8
Other minerals (average)	7	2.9
Organic matter	0.6	0.25
Water (liquid)	1.37	0.57
Ice	5.2	2.2
Air	0.06	0.025

Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Simultaneous Transport of Heat and Moisture

The flows of water and of thermal energy under nonisothermal conditions in the soil are interactive phenomena: the one entails the other. Temperature gradients affect the moisture potential field and induce both liquid and vapor movement. Reciprocally, moisture gradients move water, which carries heat. The simultaneous occurrence of temperature gradients and of moisture potential gradients in the soil therefore brings about the combined transport of heat and moisture.

Two separate approaches to the combined transfer of heat and moisture have been attempted: (1) a

Table 3 Average thermal properties of soils and snow[a]

Soil type	*Porosity*	*Volumetric wetness*	*Thermal conductivity (10^{-3} cal cm^{-1} s^{-1}°C)*	*Volumetric heat capacity (cal cm^{-1} s^{-1}°C)*	*Damping depth (diurnal) (cm)*
Sand	0.4	0.0	0.7	0.3	8.0
	0.4	0.2	4.2	0.5	15.2
	0.4	0.4	5.2	0.7	14.3
Clay	0.4	0.0	0.6	0.3	7.4
	0.4	0.2	2.8	0.5	12.4
	0.4	0.4	3.8	0.7	12.2
Peat	0.8	0.0	0.14	0.35	3.3
	0.8	0.4	0.7	0.75	5.1
	0.8	0.8	1.2	1.15	5.4
Snow	0.95	0.05	0.15	0.05	9.1
	0.8	0.2	0.32	0.2	6.6
	0.5	0.5	1.7	0.5	9.7

[a]After van Wijk and de Vries (1963).
Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

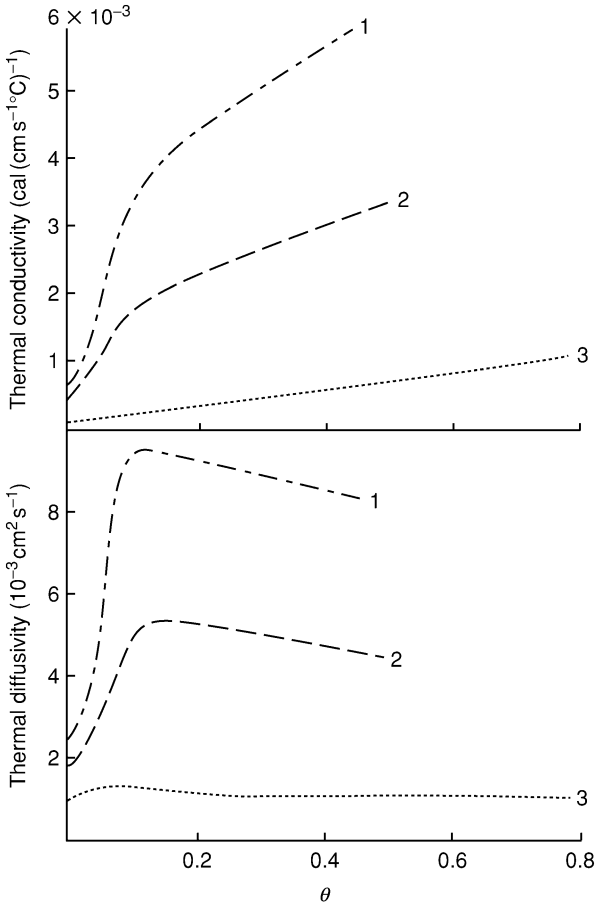


Figure 1 Thermal conductivity and thermal diffusivity as functions of volume wetness (volume fraction of water) for: (1) sand (bulk density 1460 kg m^{-3}, volume fraction of solids 0.55); (2) loam (bulk density 1330 kg m^{-3}, volume fraction of solids 0.5); and (3) peat (volume fraction of solids 0.2). (After de Vries, 1975.) (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

mechanistic approach, based on a physical model of the soil system, and (2) a thermodynamic approach, based on the phenomenology of irreversible processes in terms of coupled forces and fluxes. Though starting from different points of view, the two approaches have been shown to be related and, properly formulated, can be cast into an equivalent mold.

The mechanistic approach is based on the concept of viscous flow of liquid water under the influence of gravity and of capillary and adsorptive forces, and on the concept of vapor movement by diffusion. Local 'microscopic-scale' thermodynamic equilibrium between liquid and vapor is assumed to exist at all times and at each point within the soil. The general differential equation describing moisture movement in a porous system under combined temperature and moisture gradients for unidimensional vertical flow is, accordingly:

$$\partial\theta/\partial t = \nabla \cdot (D_T \nabla T) + \nabla \cdot (D_w \nabla\theta) - \partial K/\partial z \quad [14]$$

where θ is volumetric wetness, t time, T absolute temperature, D_T the water diffusivity under a temperature gradient (the sum of the liquid and vapor diffusivities), D_w the water diffusivity under a moisture gradient, K the hydraulic conductivity, and z the vertical space coordinate. The last term on the right-hand side is due to the gravity gradient and becomes positive if z is taken to be increasing downwards.

The heat transfer equation is, similarly:

$$C_v \partial T/\partial t = \nabla \cdot (\kappa \nabla T) - L\nabla \cdot (D_{w,vap} \nabla\theta) \quad [15]$$

Here C_v is volumetric heat capacity, κ apparent thermal conductivity of the soil, L latent heat of vaporization of water, and $D_{w,vap}$ diffusivity for heat conveyed by water movement (mostly vapor). Equations [14] and [15] are of the diffusion type, involving θ- and T-dependent diffusivities as well as gradients of both θ and T.

Taken together, eqns [14] and [15] describe the coupled transport of moisture and heat in soils. The assumption of local thermodynamic equilibrium links the vapor pressure p_v to the matric potential ψ by the following relation: $p_v = p_{vs}h = p_{vs}\exp(Mg\psi/RT)$, where p_{vs} is the saturated vapor pressure at the particular temperature T, h relative humidity, M molar mass, g acceleration due to gravity, and R the universal gas constant.

The approach based on the thermodynamics of irreversible processes formulates a pair of phenomenological equations in which the fluxes of moisture q_w and heat q_h are expressed as linear functions of the moisture potential (e.g., pressure) gradient dp/dz and the temperature gradient dT/dz:

$$\begin{aligned} q_w &= -L_{ww}(1/T)dp/dz - L_{wh}(1/T^2)(dT/dz) \\ q_h &= -L_{hw}(1/T)dp/dz - L_{hh}(1/T^2)(dT/dz) \end{aligned} \quad [16]$$

The four phenomenological coefficients occurring in these equations (L_{ww}, L_{wh}, L_{hw}, L_{hh}, relating water flow to the water potential gradient, water flow to the thermal potential gradient, heat flow to the water potential gradient, and heat flow to the thermal potential gradient, respectively) are unknown functions of p (or θ) and T. According to the Onsager theorem, the cross-coupling coefficients L_{wh} and L_{hw} are equal when the fluxes and forces are properly formulated. Thus, the number of coefficients that must be measured is reduced.

An apparent advantage of the irreversible thermodynamics approach is that it makes no *a priori* assumptions regarding the mechanisms of the transport phenomena formulated. Hence it would seem to be less restrictive than a physical theory whose validity

is constrained at the outset by its mechanistic assumptions. The disadvantage of this approach, however, is precisely its failure to address itself to, and provide insight into, the nature and internal workings of the phenomena considered.

Thermal Regime of Soil Profiles

In nature, soil temperature varies continuously in response to the ever-changing meteorological regime acting on the soil–atmosphere interface. That regime is governed by a regular periodic succession of days and nights, and of summers and winters. Yet the regular diurnal and annual cycles are perturbed by such irregular episodic phenomena as cloudiness, cold waves, heat waves, rainstorms or snowstorms, and periods of drought. Add to these external influences the soil's own changing properties (i.e., temporal changes in reflectivity, heat capacity, and thermal conductivity as the soil alternately wets and dries, and the variation of all these properties with depth), as well as the influences of geographic location, vegetative cover, and, finally human management, and one can expect the thermal regime of soil profiles to be complex indeed.

The simplest mathematical representation of nature's fluctuating thermal regime is to assume that at all depths in the soil the temperature oscillates as a pure harmonic (sinusoidal) function of time around an average value. Assume that, although soil temperature varies differently at different depths in the soil, the average temperature is the same for all depths. A starting time ($t = 0$) is chosen such that the surface is at the average temperature. The temperature at the surface can then be expressed as a function of time (Figure 2):

$$T(0,t) = T_{ave} + A_0 \sin \omega t \qquad [17]$$

where $T(0,t)$ is the temperature at $z = 0$ (the soil surface) as a function of time t, T_{ave} is the average temperature of the surface (as well as of the profile), and A_0 is the amplitude of the surface temperature fluctuation (the range from maximum, or from minimum, to the average temperature). Finally, ω is the radial frequency, which is 2π times the actual frequency. In the case of diurnal variation, the period is 86 400 s (24 h), so $\omega = 2\pi/86\,400 = 7.27 \times 10^{-5}\,s^{-1}$. Note that the argument of the sine function is expressed in radians rather than in degrees.

The last equation is the boundary condition for $z = 0$. For the sake of convenience, let us assume that at infinite depth ($z = \infty$) the temperature is constant and equal to T_{ave}. Under these circumstances, the temperature at any depth z is also a sine function of time, as shown in Figure 3:

$$T(z,t) = T_{ave} + A_z \sin[\omega t + \phi(z)] \qquad [18]$$

in which A_z is the amplitude at depth z. Both A_z and $\phi(z)$ are functions of z but not of t. They can be determined by substituting the solution of eqn [18]

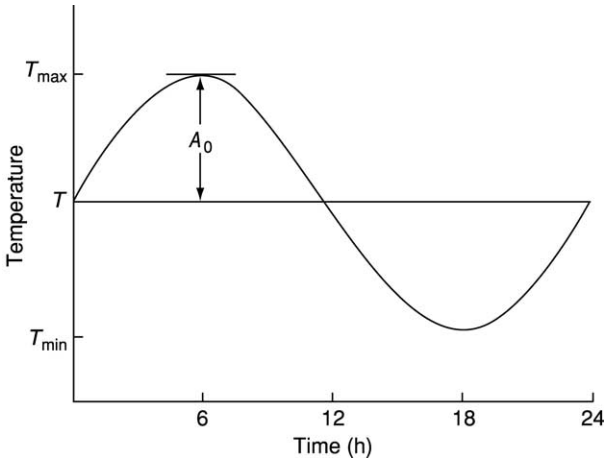


Figure 2 Idealized daily fluctuation of surface soil temperature, according to the equation: $T = T_{ave} + A_0 \sin(\omega t/p)$, where T is temperature, T_{ave} average temperature, A_0 amplitude, t time, and p period of the oscillation (in this case, p refers to the diurnal 24 h). (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

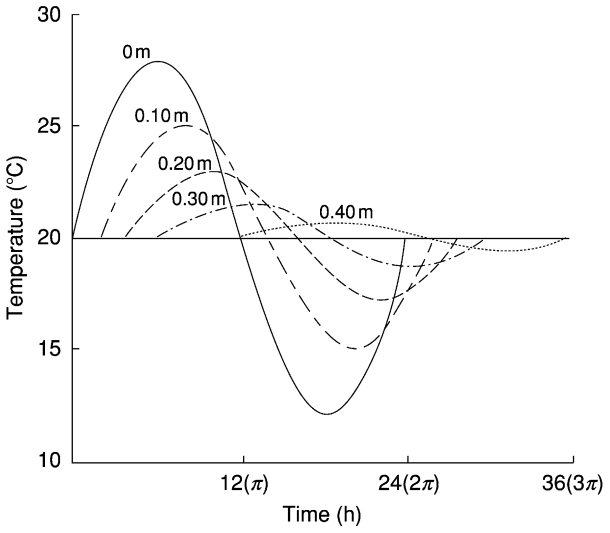


Figure 3 Idealized variation of soil temperature with time for various depths. Note that at each succeeding depth the peak temperature is damped and shifted progressively in time. Thus, the peak at a depth of 0.4 m lags about 12 h behind the temperature peak at the surface and is only about 1/16 of the latter. In this hypothetical case, a uniform soil was assumed, with a thermal conductivity of 1.68 $Jm^{-1}s^{-1}deg^{-1}$ (or 4×10^{-3} $cal\,cm^{-1}s^{-1}deg^{-1}$) and a volumetric heat capacity of 2.1×10^6 $Jm^{-3}deg^{-1}$ (0.5 $cal\,cm^{-3}deg^{-1}$). (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

in the differential equation $\partial T/\partial t = \kappa(\partial^2 T/\partial z^2)$. This leads to the solution:

$$T(z,t) = T_{ave} + A_0[\sin(\omega t - z/d)]/e^{z/d} \qquad [19]$$

The constant d is a characteristic depth, called the 'damping depth,' at which the temperature amplitude decreases to the fraction $1/e$ ($1/2.718 = 0.37$) of the amplitude at the soil surface A_0. The damping depth is related to the thermal properties of the soil and the frequency of the temperature fluctuation as follows:

$$d = (2\kappa/C\omega)^{1/2} = (2D_h/\omega)^{1/2} \qquad [20]$$

At any depth the amplitude of the temperature fluctuation A_z is smaller than A_0 by a factor $e^{z/d}$, and there is a phase shift (i.e., a time delay of the temperature peak) equal to $-z/d$. The decrease in amplitude and increase in phase lag with depth are typical phenomena in the propagation of a periodic temperature wave in the soil (Figure 4).

The physical reason for the damping and retarding of the temperature waves with depth is that a certain amount of heat is absorbed or released along the path of heat propagation when the temperature of the conducting soil increases or decreases, respectively. The damping depth is related inversely to the frequency, as can be seen from eqn [20]. Hence it depends directly on the period of the temperature fluctuation considered. The damping depth is $(365)^{1/2} = 19$ times larger for the annual variation than for the diurnal variation in the same soil.

The annual variation of soil temperature down to considerable depth causes deviations from the simplistic assumption that the daily average temperature is the same for all depths in the profile. The combined effect of the annual and diurnal variation of soil temperature can be expressed by:

$$T(z,t) = T_{ave,y} + A_y[\sin(\omega_y t + \phi_y - z/d_y)]/e^{z/d_y} + A_d[\sin(\omega_d t + \phi_d - z/d_d)]/e^{z/d_d} \qquad [21]$$

where the subscripted indices y and d refer to the yearly and daily temperature waves, respectively. Thus $T_{ave,y}$ is the annual mean temperature. The daily cycles are now seen to be short-term perturbations superimposed upon the annual cycle. Vagaries of weather (e.g., spells of cloudiness or rain) can cause considerable deviations from simple harmonic fluctuations, particularly for the daily cycles. Longer-term climatic irregularities can also affect the annual cycle. The soil temperature profile as it varies seasonally is shown in Figure 4.

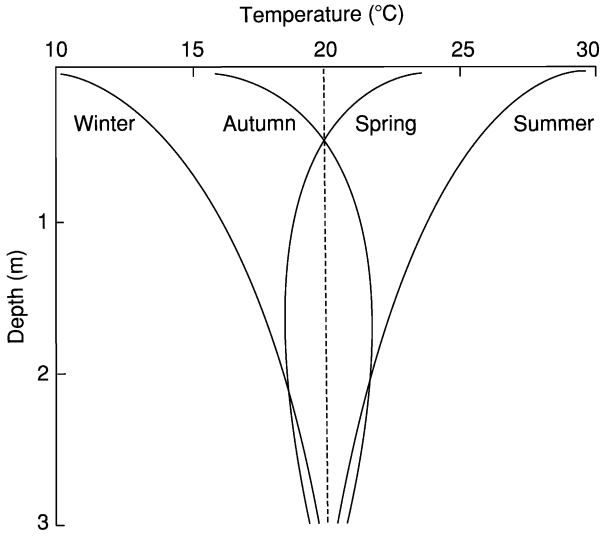


Figure 4 Soil temperature profile as it varies from season to season in a frost-free region. (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

An alternative approach is possible, with fewer constraining assumptions. It is based on numerical rather than analytical methods for solving the differential equations of heat conduction. Mathematical simulation models relying on digital computers now allow soil thermal properties to vary in time and space (e.g., in response to periodic changes in soil wetness), so as to account for alternating surface saturation and desiccation and for profile-layering. They also allow various climatic inputs to follow more realistic and irregular patterns. The surface amplitude of temperature need no longer be taken to be an independent variable, but one that depends on the surface energy balance and thus is affected by both soil properties and above-soil conditions.

Other innovations of practical importance involve the development of techniques for monitoring the soil thermal regime more accurately and precisely than was possible previously. One such technique is the infrared radiation thermometer for scanning or remote sensing of surface temperature for both fallow and vegetated soils without disturbance of the measured surface. Knowledge of the surface temperature and its variation in time is important in assessing energy exchange between the soil and the overlying atmosphere, as well as in determining boundary conditions for within-soil heat transfer.

See also: **Energy Balance**; **Evaporation of Water from Bare Soil**; **Radiation Balance**

Further Reading

Carslaw HS and Jaeger JC (1959) *Conduction of Heat in Solids*. Oxford, UK: Oxford University Press.

de Vries DA (1975) Heat transfer in soils. In: de Vries DA and Afgan NH (eds) *Heat and Mass Transfer in the Biosphere*, pp. 5–28. Washington, DC: Scripta Book Co.

Groenevelt PH and Bolt GH (1969) Non-equilibrium thermodynamics of the soil–water system. *Journal of Hydrology* 7: 358–388.

Hillel D (1977) *Computer Simulation of Soil–Water Dynamics*. Ottawa, Canada: International Development Research Centre.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Jackson RD and Taylor SA (1986) Heat transfer. In: *Methods of Soil Analysis*. Monograph No. 9, pp. 349–360. Madison, WI: American Society of Agronomy.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Cremingen, Germany: Catena-Verlag.

Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysical Union* 38: 222–228.

van Bavel CHM and Hillel D (1976) Calculating potential and actual evaporation from a bare soil surface by simulation of concurrent flow of water and heat. *Agricultural Meteorology* 17: 453–476.

THERMODYNAMICS OF SOIL WATER

P H Groenevelt, University of Guelph, Guelph, ON, Canada

Introduction

The word 'thermodynamics' is derived from $\vartheta\epsilon\rho\mu$os (heat) and $\delta\upsilon\nu\alpha\mu\iota$s (force, power). It presents the science of all forms of energy and mass, including entropic ('waste') heat contained in the mass at ambient temperature. The origin of the word suggests that this branch of science deals with both the statics (equilibrium) and the dynamics (nonequilibrium) of energy and mass. Statics (classical thermodynamics) deals with the state of the system in which no heat or mass transfer occurs (this state is extremely rare in natural soils). Dynamics (nonequilibrium thermodynamics) deals with transport processes of mass and heat.

'Soil water' is often used interchangeably with the term 'soil moisture.' Here we distinguish between the two terms: 'soil water' indicating the chemical component H_2O in the soil and 'soil moisture' meaning the soil solution. Thermodynamics distinguishes the different chemical components in the system. It deals with the interaction between and transfer of these components and heat.

Classical (Equilibrium) Thermodynamics of Soil Water

The birth of thermodynamics is rather confusing. The founders tried to deal with such things as the effectiveness of steam engines (S. Carnot). R. Meyer, who was the first (in 1842) to publish the equivalence of work and heat, based his conclusion on his observations of the degree of 'redness' of human blood. No wonder that thermodynamics had a shaky start. The first mathematical formulation of the 'First Law of Thermodynamics' (the law of conservation of energy) came from Helmholtz and Joule in 1847. These scientists were dealing with systems that were 'on the move.' Yet the system had to be at equilibrium and the 'changes' had to be infinitesimally small and 'reversible' (meaning that the changes can be reversed without any loss of useful energy at the ambient temperature, or entropy production).

One would expect the First Law of Thermodynamics to equate integral values of the different forms of energy. Yet the equations always come in differential form first. At first only 'closed' systems were considered, implying that no mass could move into or out of the system. It was not until J. Willard Gibbs published his two great treatises, *The Equilibrium of Heterogeneous Substances* (in the *Transactions of the Connecticut Academy*, 1876 and 1878), that the 'system' was opened up and many of the mysteries of the previous 50 years were clarified. The *Transactions of the Connecticut Academy* were not widely read in Europe, and it was not until C.N. Lewis published his famous book in 1923, 20 years after Gibbs' death, that Gibbs' works became widely known. In the year of his retirement, Gibbs remarked that, during the 30 years of his teaching at Yale University, he estimated that only half a dozen of his students had benefited from his lectures.

We start with the integrated Gibbs equation:

$$E = U + \Psi = TS - PV + \sum_{i} \mu_i m_i + \sum_{i} \psi m_i \qquad [1]$$

where E is the energy of the system, U is the 'internal' energy, and Ψ is the energy derived from external force fields.

U consists of three 'blobs' of different forms of energy: the first one (TS) being the entropic energy (heat in the system at the ambient temperature T); the second one ($-PV$) being the energy the system has lost in order to create a volume, V, for itself to exist at the ambient pressure P; the third one being the sum of the chemical energy of each component (i) in the system, with m_i being the mass of component i and μ_i being the chemical potential of component i.

Ψ is derived from external force fields such as gravity or an electrical or magnetic field (in the gravity field $\Psi = gh$, g being the gravitational acceleration, and h the height above the reference level). If other forms of energy (such as kinetic energy or viscous stress energy) play a role in the transformations of energy, they should be added to eqn [1]. Although kinetic and viscous stress energy play a dominant role in energy transformations in the soil solution on the Navier–Stokes scale, on the Darcy scale they are almost always negligible and will therefore be ignored here.

The 'free' energy, G, was defined by Gibbs as:

$$G \equiv U - TS + PV = \sum_i \mu_i m_i \qquad [2]$$

Thus, μ_i is the specific (per unit mass) Gibbs free energy of constituent i.

The first expressions of the First Law were all in differential form, even after Gibbs 'opened up' the system:

$$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V + \sum_i \mu_i \mathrm{d}m_i \qquad [3]$$

This is the differential form of the Gibbs equation. It is not trivial, because if one carries out the complete differentiation of eqn [1], one finds that, after subtraction of eqn [3] there is a 'leftover':

$$S\mathrm{d}T - V\mathrm{d}P + \sum_i m_i \mathrm{d}\mu_i = 0 \qquad [4]$$

This leftover is known as the Gibbs–Duhem equation.

From eqn [4] it follows that:

$$\mathrm{d}\mu_w = Sm_w^{-1}\mathrm{d}T - Vm_w^{-1}\mathrm{d}P + \sum_{(k\neq w)} m_k m_w^{-1}\mathrm{d}\mu_k \qquad [5]$$

This is the central equation in classical thermodynamics of soil water. It dictates how the chemical potential (specific Gibbs free energy) of water in the soil changes with the temperature, the pressure, and the composition of the soil solution.

The guiding principle is now that, when the system is in equilibrium, $\mathrm{d}\mu_w = 0$. The specific Gibbs free energy of water is the same everywhere, or, as far as water is concerned, there is no transport and no entropy production. In short, the water is at rest. The inverse statement, "when $\mathrm{d}\mu_w = 0$, the water is at rest," is not necessarily true and is not a fundamental principle of thermodynamics. It is possible that the last two terms of the right-hand side of eqn [5] (temporarily) balance each other, making $\mathrm{d}\mu_w = 0$, while both water and the solute are on the move and entropy is being produced. This is possible when 'leaky' semipermeable membranes are present. Heavy clay soils are excellent examples of leaky semipermeable membranes.

Therefore, the concept of the 'total' potential of the soil water (the sum of all the component potentials) is not very useful. When the whole system is at rest (complete equilibrium), then the total potential of the water is equal everywhere in the system.

But, when the total potential of the water is equal everywhere, this does not necessarily mean that the water is at rest. The gradient of the total potential is not the appropriate driving force in the flux equation of water. This will become clearer in the section on nonequilibrium thermodynamics, below.

The apprentice in thermodynamics often asks: "What is the practical use of all this?" Thermodynamics is still not a coherent, comprehensive science. Developments are still happening haphazardly. Some scientists, such as Clifford Truesdell, have complained bitterly about the inconsistencies and the lack of mathematical rigor, using words such as "thermodynamics was approached through detours in the fog of word-play." Here, only two of the practically useful results will be discussed.

The Tensiometer (and the Pressure Membrane Apparatus)

Early scientists such as L.A. Richards, G.H. Bolt, and M.J. Frissel, working on the thermodynamics of soil moisture, immediately realized that water in the soil is present in the soil solution and that the soil solution is never pure water. The soil solution contains dissolved salts and ions, as well as counterions, that is, a surplus of cations, necessary to counterbalance the surplus of electrons present in the lattices of clay particles. The electrical double layer acts as an osmometer and the counterions cause a volume element of the solution, close to a clay surface, to carry a volume charge. This gives the soil solution an additional opportunity to possess energy (in the presence of an electrical field). In addition, there are microforce fields acting on the solution that is in close proximity of solid surfaces (London–Van der Waals forces). All this implies that eqn [5] should contain a few more terms if the thermodynamic system were chosen on the Navier–Stokes scale.

The first requirement in soil physics is to formulate and assess the hydrostatics of soil water, i.e., establish

the characteristic relationships when the water is at rest. There are no 'membranes' in nature or in the laboratory that are semipermeable to heat. Thus, when a system is brought to equilibrium, the temperature will be the same everywhere. The energy level in the external force fields, such as gravity, can easily be established separately. In order to measure the energy level of soil water derived from internal force fields (menisci, electrical double layers, London–Van der Waals forces), it is now advantageous to bring the soil solution in equilibrium with either a compartment containing a solution that reflects the composition of the soil solution or one that contains pure water. There are membranes that are ('perfectly') semipermeable to all dissolved materials in the soil solution. In the latter case (using a perfect osmometer and measuring the hydrostatic pressure in the osmometer with respect to atmospheric pressure), the energy level of the soil water is measured at the ambient temperature and at the particular position in the external force field that results from all internal force fields plus all dissolved materials (the total osmotic pressure of the soil solution). In the former case, using a perfectly 'leaky' membrane such as the sintered ceramic cup (tensiometer cup) or a Visking membrane in a pressure plate apparatus, the solution inside the compartment (tensiometer cup) will reflect the composition of the 'free salts' in the soil solution and will show the 'equilibrium' solution. By measuring the hydrostatic pressure in a tensiometer with respect to atmospheric pressure, the energy level of all internal force fields is measured (including electrical double layers), excluding the free salt. This energy level is called the 'matric potential.' The counterions cannot equilibrate in the tensiometer cup. They act as internal miniosmometers and can be considered part of the matrix (the solid phase of the soil). Because most soils are quite 'leaky' as far as the exclusion of dissolved materials is concerned, the water in the soil moves largely as a solution. Therefore the gradient of the matric potential is the dominant driving force on the soil water (in addition to the force of the gravitational field). The imperfect semipermeability of a clay soil, causing osmotic pressure differences to drive the water, will be dealt with in the section Nonequilibrium Thermodynamics of Soil Water, below. The action of osmotic pressure gradients will then be formulated as 'capillary' osmosis and the imperfect semipermeability will be evaluated by the 'reflection coefficient.'

The Maxwell Relations and Their Use in the Hydrostatics of Swelling Soils

One of the many complaints about the lack of mathematical rigor lies in its origin: a closed pot on a bench in a chemistry laboratory. Temperature and pressure are allowed to change in steps, ΔT and ΔP, although only in small steps, δT and δP. But entropy production is not allowed and therefore the steps have to be infinitesimally small, dT and dP. They begin to look like a differential, but differential with respect to what? Time or space? Everything becomes cleaner when the denominator of the differential is specified. This first comes about, while still in the realm of classical thermodynamics, when a cross-differentiation is performed on the differential form of the Gibbs equation. This leads to the Maxwell relations, several of which can be constructed. Many more can be found after performing Legendre transforms on the Gibbs free energy. Most of the Maxwell relations are quite useless, but occasionally one finds a gem. One example is a Maxwell relation for shrinking soils. For many decades soil scientists have searched for an expression for the differential of the matric potential (tensiometer pressure), p, with respect to the load (or overburden pressure), P, at constant ratio of mass of water and mass of solids. Here, both the dependent and the independent variable are intensive variables. The appropriate Maxwell relation shows that the above differential is equal to the differential relationship between the void ratio (volume of voids per volume of solids), e, and the moisture ratio (volume of water per volume of solids) at constant load pressure, P. Here, both the dependent and the independent variable are extensive variables.

Thus:

$$\mathrm{d}p(\mathrm{d}P)^{-1} \text{ at constant } m_{\mathrm{w}}(m_{\mathrm{solids}})^{-1} = \mathrm{d}e(\mathrm{d}\vartheta)^{-1} \quad [6]$$

at constant P. This means that the behavior of the tensiometer reading can be predicted upon loading from the slopes of the shrinkage lines for different values of P. Because the shrinkage lines for different values of P are nearly parallel, one only needs to measure the unloaded shrinkage line in order to predict how the tensiometer reading changes when moving deeper into the soil profile or upon loading by machinery.

Nonequilibrium Thermodynamics of Soil Water

The thermodynamic 'system' is now chosen to be a unit volume element in the soil solution. All the extensive variables are expressed per unit volume. Subsequently, each of the terms of the differential form of the Gibbs equation (eqn [3]) are written as differentials with respect to time.

The mathematical sloppiness of classical thermodynamics disappears once the terms of the differential

form of the Gibbs equations are transformed into proper time differentials and the realm of nonequilibrium thermodynamics is entered.

Next, for each of the terms separately, an appropriate conservation (continuity) equation is constructed. The general form of such a conservation equation shows that the change of the content of the entity of concern, with time, is equal to the negative of the divergence of the flux of that entity (inflow minus outflow), plus or minus one or more source or sink terms. Thus, for the entropic energy term:

$$\mathrm{d}\left(SV^{-1}\right)\mathrm{d}t^{-1} = -\mathrm{div}\ j_s + \sigma \qquad [7]$$

where SV^{-1} is the entropy per unit volume of solution, j_s is the flux of entropy, and σ is the entropy production term. This latter term plays a central role in nonequilibrium thermodynamics. If all the processes and forms of energy are accounted for, nothing is counted twice, and all algebra is carried out correctly, then the entropy production term is nonnegative. As the absolute temperature is always positive, the energy dissipation term, $T\sigma$, must also be nonnegative. When σ and thus $T\sigma$ are zero, the system is at equilibrium and nothing moves. If anything moves, then entropy is produced, and σ and $T\sigma$ are positive. Energy is being dissipated; that is, energy is transformed from useful energy to waste energy (i.e., heat at the ambient temperature). This is the Second Law of Thermodynamics.

As the first term on the right-hand side of Eqn [3] is already multiplied by T, one finds, after replacing the time differentials of the different terms in Eqn [3] by the right-hand side of their appropriate conservation equations, the product $T\sigma$ as the source term of entropy conservation equation. After all the appropriate substitutions are made, the term $T\sigma$ is singled out (written explicitly) and the equation is then called 'the (energy) dissipation function.'

On the right-hand side of the dissipation function, one usually finds the sum of a number of products of fluxes and forces. One may reshuffle these forces and the fluxes, under the strict principles that the rules of algebra are carried out correctly and that nothing is forgotten or counted twice.

The resulting complete expression for the energy dissipation is then the sum of products of fluxes and conjugated forces:

$$T\sigma = -j_q(\mathrm{grad}\ T) - j^V(\mathrm{grad}\ H) - j^D(\mathrm{grad}\ \pi) - j^E(\mathrm{grad}\ E) \qquad [8]$$

where j_q is the caloric (Fourier) heat flux, j^V is the (Darcy) flux of the soil solution, j^D is the (Fick) diffusion flux, and j^E is the electric current (usually indicated by I). grad π is the gradient of the osmotic pressure, and grad E is the gradient of the electrostatic potential.

The primary forms of energy dissipation are the dissipation of heat, pressure, mixing, and electrical energy. If simple linear, homogeneous transport equations are constructed relating a flux to its conjugated flux (the one that is occurring in the same product), the transport equations of Fourier (1822), Darcy (1856), Fick (1855), and Ohm (1827) are obtained. The disciplines of physics and chemistry have long recognized that 'coupled' transport can occur. Examples of such coupled transport processes are osmosis, electro-osmosis, thermo-osmosis, the Peltier effect, etc. A coupled transport process takes place when a flux arises due to a nonconjugated force (that is, a force that occurs in a product other than the one in which the flux of concern occurs).

Nonequilibrium thermodynamics now postulates that each flux occurring in the energy dissipation function is a linear, homogeneous (no intercept) function of all forces occurring in the same equation for the total energy dissipation. Thus:

$$j_q = -L_{TT}(\mathrm{grad}\ T) - L_{TV}(\mathrm{grad}\ H) - L_{TD}(\mathrm{grad}\ \pi) - L_{TE}(\mathrm{grad}\ E) \qquad [9]$$

$$j^V = -L_{VT}(\mathrm{grad}\ T) - L_{VV}(\mathrm{grad}\ H) - L_{VD}(\mathrm{grad}\ \pi) - L_{VE}(\mathrm{grad}\ E) \qquad [10]$$

$$j^D = -L_{DT}(\mathrm{grad}\ T) - L_{DV}(\mathrm{grad}\ H) - L_{DD}(\mathrm{grad}\ \pi) - L_{DE}(\mathrm{grad}\ E) \qquad [11]$$

$$j^E = -L_{ET}(\mathrm{grad}\ T) - L_{EV}(\mathrm{grad}\ H) - L_{ED}(\mathrm{grad}\ \pi) - L_{EE}(\mathrm{grad}\ E) \qquad [12]$$

The 'phenomenological' coefficients in the transport equations [9–12] are indicated by the letter L with two subscripts, say K and M. The first one (K) indicates which entity is being transported. The second one (M) indicates which driving force causes the transport. The terms for which the coefficient has identical subscripts represent the 'straight,' well-known transport processes, the laws of Fourier (1822), indicated by L_{TT}, Darcy (1856), indicated by L_{VV}, Fick (1855), indicated by L_{DD}, and Ohm (1827), indicated by L_{EE}. All other coefficients are known as 'coupling' coefficients, representing 'coupling' processes. They come in pairs, a pair being indicated by L_{KM} and L_{MK}. The two coefficients of a pair are called 'twin' coefficients. They represent all possible coupling phenomena.

If the energy dissipation equation is complete and all algebra has been carried out correctly, then the two twin coefficients are equal:

$$L_{KM} = L_{MK} \quad [13]$$

Thus the matrix of coefficients in eqns [9–12] is symmetrical. In the above matrix there are six of these pairs of equal twins. These equalities are known as the Onsager reciprocal relations (ORRs).

The fundamental value of the procedure outlined above is that all possible forms of energy dissipation are accounted for, even if they have never been observed.

For soil physics these forms are extremely important. After the merciless denigration of the ORRs by Truesdell, soil physicists should rise up and accept the validity and the great practical usefulness of ORRs. The framework of the matrix of coefficients in eqns [9–12] is comparable, in nature but not quite in stature, to the periodic system as proposed by Mendeleyev. For his systematic framework, Onsager received the Nobel Prize in Chemistry (1968), but Mendeleyev never received this accolade. The first such prize in chemistry was awarded to Jacobus van't Hoff, even though Mendeleyev was still alive. The refusal by the Nobel Committee to award the prize to Mendeleyev continued until he died in 1907. The framework makes it possible, in case the measurement of a certain coefficient is difficult or expensive, to measure its twin, which often appears to be easier or cheaper to measure.

When building models for transport processes in clay soils, e.g., based on the electrical double-layer theory, these ORRs can be used to verify the correctness of the model: if the cross-coefficients are not equal, researchers can be assured that somewhere they have made a mistake.

The actual occurrence of a coupled transport process always relies on some kind of a selection mechanism, such as a mechanism that can select between molecules of different chemical nature, e.g., water and salts, or a mechanism that can select between 'hot' and 'cold' molecules, e.g., a liquid–gas interface. It should be noted that convection (advection) is never a selection mechanism.

The coupled transport processes are discussed here in pairs of twin phenomena and indicated by their (equal) coefficients:

- L_{TV} (thermofiltration) and L_{VT} (thermo-osmosis). Both these phenomena are very common in soils. The transport of heat due to a water potential gradient in the absence of a temperature gradient and the transport of water due to a temperature gradient in the absence of a water potential gradient are occurring constantly in the soil. In unsaturated soil they are due to evaporation and condensation (the selection mechanism here is the heat of vaporization–condensation). In saturated soils the magnitude of these phenomena is very small (the selection mechanism is now the heat of wetting). In frozen soils they are caused by freezing and melting (here the selection mechanism is the heat of freezing–melting and solidification–sublimation);
- L_{TD} and L_{DT} (thermodiffusive processes). These phenomena are known as, respectively, the Dufour effect and the Soret effect. The magnitude of these effects in soils is very small, but their occurrence is quite probable;
- L_{TE} and L_{ET} (thermoelectric processes). These phenomena are known as, respectively, the Peltier effect and the Seebeck effect. The magnitude of these phenomena in soils is small, but their occurrence is quite likely;
- L_{VD} (osmosis, also called capillary osmosis) and L_{DV} (reverse osmosis or salt sieving). These phenomena are very common in soils. The magnitude is directly related to the clay content of the soil. (The selection mechanism lies with the electrical double layer.) Clay particles expel negative ions, and therefore they expel dissociated salts (negative adsorption). The longstanding conflict as to whether the osmotic pressure (potential) should be added to the hydraulic potential of water to produce the 'total' potential of water, the gradient of which then serves as the driving force on the water, is here resolved. As the value of L_{VD} is almost always smaller then the value of L_{VV} (except for a perfectly semipermeable membrane), the concept of the total potential is useless. The ratio L_{VD} to L_{VV} is known as the 'reflection coefficient';
- L_{VE} (electro-osmosis) and L_{EV} (streaming current). These electrokinetic phenomena also find their cause in the existence of electrical double layers. The most commonly observed result of the effects is the streaming potential in clay soils;
- L_{DE} (electrophoresis) and L_{ED} (diffusion current). These phenomena, together with osmosis and reverse osmosis, electro-osmosis, and streaming current are extensively discussed in the literature. The magnitude of all six coefficients on the basis of electrical double-layer theory has been calculated.

Modern Development

The branch of science called 'nonequilibrium thermodynamics' is now split into two subbranches. The first one is now called 'linear nonequilibrium thermodynamics.' Indeed all transport equations presented

here are linear. This subbranch of science deals with processes that are 'not far from equilibrium.' The second subbranch is now called 'nonlinear nonequilibrium thermodynamics.' Its great proponent is Ilya Prigogine, who received the Nobel Prize in Chemistry for his work in 1977. This subbranch of science deals with 'violent' processes ('far away from equilibrium'), such as Liesegang rings and atom bombs. The natural transport processes in the soil are never violent, and, as long as the Darcy law (one of the primary energy-dissipating processes) holds, all coupling processes fall in the park of 'not far from equilibrium.' However, if there are cases where the Darcy law breaks down and turns into the nonlinear Forchheimer equation (that is, when the water in the soil becomes turbulent), soil physicists may have to turn their attention to this latest branch of thermodynamics.

See also: **Darcy's Law**; **Heat and Moisture Transport**; **Hydrodynamics in Soils**

Further Reading

Bolt GH and Frissel MJ (1960) Thermodynamics of soil moisture. *Netherlands Journal of Agricultural Science* 8: 57–78.

Gibbs JW (1961) *The Scientific Papers of J. Willard Gibbs,* vol. I, *Thermodynamics*. Mineola, NY: Dover Publications.

Glansdorff P and Prigogine I (1971) *Thermodynamic Theory of Structure, Stability and Fluctuations*. Chichester, UK: John Wiley.

Groenevelt PH (1971) Onsager's reciprocal relations. *Search* 2: 264–267.

Groenevelt PH (1979) Transport processes. In: Fairbridge RW and Finkl CW (eds) *The Encyclopedia of Soil Science*, part 1, pp. 578–581. Stroudsberg, PA: Daudon, Hutchison & Ross.

Groenevelt PH and Bolt GH (1969) Non-equilibrium thermodynamics of the soil-water system. *Journal of Hydrology* 7: 358–388.

Groenevelt PH and Bolt GH (1972) Water retention in soil. *Soil Science* 113: 238–245.

Groenevelt PH and Elrick DE (1976) Coupling phenomena in saturated homo-ionic montmorillonite. II. Theoretical. *Soil Science Society of America Journal* 40: 820–823.

Groenevelt PH and Grant CD (2001) Re-evaluation of the structural properties of some British swelling soils. *European Journal of Soil Science* 52: 469–477.

Groenevelt PH and Grant CD (2002) Curvature of shrinkage lines in relation to the consistency and structure of a Norwegian clay soil. *Geoderma* 106: 235–245.

Groenevelt PH and Kay BD (1974) On the interaction of water and heat transport in frozen and unfrozen soils. II. The liquid phase. *Soil Science Society of America Proceedings* 38: 400–404.

Groenevelt PH, Grant CD, and Semetsa S (2001) A new procedure to determine soil water availability. *Australian Journal of Soil Research* 39: 577–598.

Katchalsky A and Curran PF (1965) *Non-Equilibrium Thermodynamics in Biophysics*. Cambridge, MA: Harvard University Press.

Kay BD and Groenevelt PH (1974) On the interaction between water and heat transport in frozen and unfrozen soils. I. Basic theory: the vapor phase. *Soil Science Society of America Proceedings* 38: 395–400.

Pippard AB (1964) *The Elements of Classical Thermodynamics*. Cambridge, UK: Cambridge University Press.

Truesdell C (1969) *Rational Thermodynamics*. New York: McGraw-Hill.

Tillage *See* **Conservation Tillage**; **Cultivation and Tillage**; **Zone Tillage**

TILTH

D L Karlen, USDA Agricultural Research Service, Ames, IA, USA

Introduction

Tilth is defined as the physical condition of soil as related to its ease of tillage, fitness as a seedbed, and its promotion of seedling emergence and root penetration. It is also an important indicator of soil quality, because it relates to soil structure in terms such as 'mellowness' or 'friability.' Soils with poor tilth are described as appearing lifeless, resembling brick or concrete, or being cloddy. If soils with poor tilth have a high percentage of sand, they will often disaggregate or separate into their primary sand-, silt-, and clay-size particles with little or no tendency to bind together. Soils with poor tilth can also be massive and difficult to till, work into a suitable seedbed or seed. Soils with medium tilth are characterized as being

somewhat cloddy, with a tendency to 'ball up' and be 'rough pulling' (i.e., having a high draft or requiring a greater energy input) when worked into a seedbed. Soils with good tilth are described as being crumb-like, easy to slice (like cutting butter), and spongy. Defining the descriptive terminology for soil tilth is a challenging task for those specializing in soil management. Tilth is difficult, if not impossible, to quantify. A common description of this dilemma is that tilth is what all farmers can measure through the soles of their boots, but no soil scientist can define.

History of Tilth

The concept of soil tilth predates modern agriculture, but a clear definition has never evolved because of differing perceptions. For example, an old childhood story explaining the concept of soil tilth describes how, in his dying moments, an elderly man called his sons to his side to tell them that if they would dig diligently in the garden, they would find a hidden treasure. The sons did so, but alas, they found no treasure of silver or gold. The subsequent harvest from this much-worked garden, however, was so large that the father's meaning gradually dawned upon his sons and they realized the wisdom of his words. A sixteenth-century book entitled *Boke of Husbandry*, written to provide instructions for growing peas and beans, tells the farmer how to determine whether the soil is ready for planting. It states that he should walk on the plowed ground and:

> if it synge or crye, or make any noyse under thy fete then it is to wete to sowe; and if it make no noyse, and wyll beare thy horses, thanne sowe in the name of God.

In the 1930s, several authors described how climatic conditions and various soil-, crop-, and animal-management practices affected soil tilth. By sieving soil samples and separating the material into different sizes, they showed how freezing, drying winds, rainfall, tillage, animal compaction, and other management factors affected tilth. Eventually, 'soil tilth' was used as a 'blanket' term, describing all the soil conditions that determine the degree of fitness of a soil as an environment for the growth and development of a crop plant. Soil structure was identified as a key factor in determining soil tilth, because structure is a soil property that can be altered rapidly through tillage operations and changes in environmental factors such as rainfall or temperature.

Soil tilth gradually came to be recognized as a dynamic condition. As virgin grassland or forest soils are cultivated, the organic matter concentrations, base (cation) saturation, and porosity decrease, while bulk density increases. These changes reduce the granulation or tendency of the soil to form stable aggregates and gradually degrade soil structure. Declining soil tilth is associated with increased runoff, erosion potential, need for tillage to prepare an adequate seedbed, and fertilizer to sustain reasonable and profitable crop yields.

Soil-management research through the mid twentieth century focused on identifying basic properties and processes that affect soil structure and tilth (i.e., flocculation, cementing by organic and inorganic colloids, wetting and drying cycles, freezing and thawing, organic matter cycling, biological activity, and tillage). One result of these efforts was that, in 1959, the 86th US Senate passed Senate Document No. 59, confirming the need:

> to conduct basic research on the relation of the physicochemical nature of soils, the role of organic matter, the activities of microbes, and the effects of mechanical manipulation upon the structural attributes of soils; [directed] toward predicting the effects of soil management practices [and] providing an optimum environment for root growth of crops in different kinds of soils.

However, 30 years elapsed before the US Department of Agriculture (USDA) Agricultural Research Service established the National Soil Tilth Laboratory in Ames, Iowa. Significant progress has been made in understanding many of the scientific aspects of soil structure, organic matter cycling, activities of microorganisms, and effects of various soil- and crop-management practices. However, with regard to adoption of improved soil-management practices, economic pressures and political forces beyond the control of most farmers often contribute to declining soil tilth.

Soil Organic Matter and Tilth

Organic matter is one of the main factors affecting soil tilth because of its role in aggregation (**Figure 1**). Studies with disturbed and undisturbed soils show that, as organic matter increases, compaction decreases and there is more space for air and water exchange. The constant addition of organic matter throughout the soil profile is one reason why virgin grassland soils generally have excellent aggregation and structure. Those soils, especially in humid regions, generally have a high organic matter content, a high percentage of water-stable aggregates, and a structure that accommodates rapid plant growth with unlimited root proliferation. These soil characteristics occur because in humid regions grasses replace most of their roots and top growth each year. Also, when the aboveground plant material dies, it falls on the soil surface, where it is decomposed or mixed into the upper part of the soil by earthworms and other soil organisms. The dead

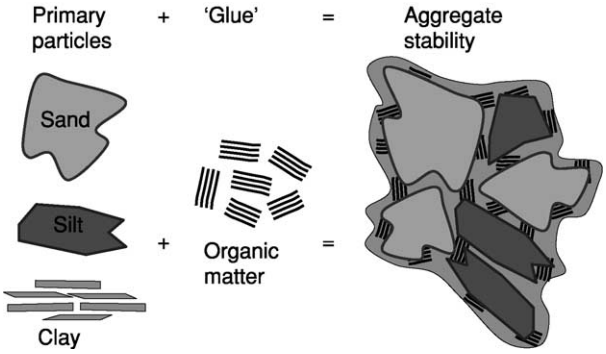


Figure 1 A conceptual illustration of soil aggregation, defined as the process whereby primary soil particles (sand, silt, and clay) are bound together, usually by natural forces and substances (glues) derived from root exudates, microbial activity, and other organic matter sources.

roots and plant residues provide a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates, which modify effects of soil texture with regard to water and air relationships, and root penetration. Increased water-stable aggregation, annual proliferation of plant roots, and the associated wetting and drying cycles gradually develop a soil structure that is stable and, with good management, generally resists degradation.

The regular addition of organic materials can improve tilth by creating larger and more stable aggregates that are not easily eroded by wind or water. Organic matter, both as residue on the soil surface and as a binding agent for aggregates near the surface, strongly influences water infiltration, retention, and the potential for particle detachment and erosion. Surface residues intercept rainfall or irrigation water, and dissipate the energy before the drops hit the surface, detach soil particles through the splash, or begin to form a surface crust by filling the voids and cracks between the aggregates with fine particles. Organic materials on the soil surface also slow the water as it flows across the field, thus increasing opportunities for the water to infiltrate. Another characteristic of soils with good tilth is the presence of large channels or pore space between the water-stable aggregates. These channels greatly enhance the ability of soil to conduct water from the surface into the subsoil, where, because of enhanced root proliferation (another characteristic of good tilth), plants can subsequently use the water to support growth and development.

Soil-Management Effects on Tilth

Tilth, as an indicator of the physical condition of a soil, is strongly influenced by any soil management practice that quantitatively affects soil structure, aggregation processes, or soil organic matter cycling. Incorporating perennial crops into extended rotations, reducing tillage intensity and frequency, including cover crops, and optimizing nutrient and water management are among the most common practices that can be used to create the desired condition.

Surface mulch from crop residues, cover crops, animal manure, municipal sludge, leaves, or other carbonaceous sources provides both physical protection from raindrop impact and a food source for earthworms and other soil microorganisms. The surface mulch also moderates soil temperature and moisture extremes at the soil surface. If soils are left bare and exposed, surface temperatures can be very high, causing the soil to dry out. When this occurs, earthworms and other soil insects move deeper into the soil, leaving a surface zone that contains very few active organisms. Bacteria and fungi that live in thin films of water die or become inactive, slowing the natural process of organic matter cycling and aggregation.

Tillage Effects on Soil Tilth

Although tillage can improve tilth over the short term by loosening surface soil, disrupting crusts, and creating a more favorable soil–water–air environment for plants, it is also a primary cause for the long-term deterioration of tilth. The loosening process that often has positive short-term plant-production benefits has negative long-term effects, because it increases the rate of organic matter decomposition through chemical and microbial mineralization. This occurs because previously protected organic matter is exposed by the tillage and subsequently used as a food source by microorganisms. The loss of organic matter decreases aggregate stability, increasing the potential for surface crusting, poor seedling emergence, runoff, erosion, and other indicators of poor soil tilth (e.g., decreased infiltration of rainfall or irrigation water, decreased soil water retention, and decreased nutrient cycling). Tillage itself or the wheel traffic associated with the operations can further reduce the quality of tilth by increasing compaction. This reduces total pore space and frequently increases water-filled pore space, leading to reduced aeration, slower nutrient cycling, and a decreased volume of soil for plant root proliferation. These processes obviously occur in fields used for crop production, but they have similar negative effects on tilth at construction sites, in forests during logging operations, on campgrounds, athletic fields, or even desert areas where off-road recreational vehicles are allowed.

Tillage has been described as being 'addictive' for soils because of the vicious cycle that can be

established. Soils are tilled to disrupt surface crusts, compacted zones, or other real or perceived limitations to crop production. Doing so may provide short-term improvements in tilth and nutrient cycling, but these periods of improvement become shorter and shorter as more of the organic matter is lost through oxidation. Ultimately the soils may be 'burned out' and must either be abandoned until natural processes (e.g., wetting and drying; plant root proliferation, exudation, death, and decay; weathering) can restore the tilth or management practices (e.g., crop rotation, tillage, water management) are changed. As a result of these interactions, monitoring of soil tilth may provide a useful indicator of the sustainability of an agricultural system.

In situations where nitrogen is the most limiting factor, increasing nitrogen fertilization rates and therefore crop productivity has been shown to increase soil organic matter, even with moldboard plowing. However, without risking increased nitrogen leaching and potential water-quality degradation, development and adoption of reduced- or no-tillage agricultural practices can generally improve soil tilth even more by slowing organic matter decomposition, maintaining aggregate stability, and preserving soil structure. The critical management practice for these systems generally focuses on crop-residue management. Maintaining crop residues on the soil surface and the lack of loosening through tillage reduces dispersion of the surface aggregates by rainfall or runoff. By leaving crop stubble on untilled soil, the potential effects of wind erosion are minimized. Reducing or eliminating tillage also diminishes tillage erosion and keeps soil from being moved downhill by tillage tools.

Studies from around the world illustrate how tillage can affect soil tilth. In one study where direct drilling was compared with several methods of cultivation for wheat production, five soil series from southern England were examined. At each site, the tillage treatments had been in place for at least 4 years. The results showed that aggregates from the surface were more stable after direct drilling than after plowing. Most of the effects were attributed to a positive interaction between aggregate stability and organic matter content (i.e., two critical indicators of soil tilth). Direct drilling did decrease both aggregate stability and organic matter content in soils where grass had been grown before the treatments were imposed, but the decrease was less than that observed with plowing. Benefits of direct drilling were assumed to be greater root density in surface layers and less organic matter mineralization than in plowed soils. A similar response was observed in the USA when moldboard plowing and no-till crop-production practices were evaluated on land being taken out of the Conservation Reserve Program.

Extended Crop Rotations

Soil tilth is affected in two ways by incorporating forage or hay crops into an extended rotation. First, soil organic matter decomposition decreases because the soil is not being disturbed or tilled each year; second, grass and legume sods develop extensive root systems, part of which will naturally die each year, adding new organic matter to the soil. The dead roots provide a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates and improved tilth.

To demonstrate the effect of perennial root systems on soil aggregation, researchers have compared aggregate size distributions for soils under continuous corn production with those where corn is grown in rotation with oat and meadow. The data show that, with continuous corn production, aggregate size is less than half that for corn grown in rotation. The studies also show that aggregation decreases slowly but consistently over a 7-year period of continuous corn production. In related studies, comparisons among crop sequences show that aggregation increases by switching to a corn–oat–meadow rotation after 11 years of continuous corn production. Also, it takes only 4 years of continuous corn production to decrease the aggregation established under either bluegrass or alfalfa to less than that found after 18 years of a corn–oat–meadow rotation. The highly significant correlation between crop yield and aggregate size confirms that soil tilth and productivity are both affected by crop rotation.

Cover Crops

Growing cover crops can improve soil tilth by decreasing erosion, increasing infiltration, and adding organic residues to the soil. These benefits are achieved because cover crops are often grown only during seasons when the soil is especially susceptible to erosion, although in orchards and vineyards the cover crops are frequently grown year-round. The leaves and stems of the cover crop intercept rainfall and dissipate its energy, while roots bind the soil and hold it in place. The amount of benefit from cover crops depends on the above- and belowground biomass and rooting structure of the plant being grown and the length of time before the soil is prepared for the next crop. For example, in the southeastern USA, crimson clover can be grown successfully as a cover crop prior to cotton. Initially planted in late autumn, the clover grows slowly throughout the winter and early spring, often reseeding itself for the next year

before the cotton crop is planted. In more northern areas such as the US Corn and Soybean Belt, where cold temperatures significantly shorten the growing season, warm-area crops such as crimson clover cannot survive. Oat, rye, and hairy vetch are some of the possible cover crops for the northern states of the USA, but the amount of protection against soil erosion and especially the amount of nitrogen supplied by the vetch are generally much less than in the southern USA. Also, in semiarid regions, where water is often the most limiting factor, growing cover crops may not be a viable management practice.

The quantity of organic residues returned to the soil by cover crops varies depending on plant species and the length of the growing season. Depending on temperature and moisture conditions, production can vary from less than 1000 kg ha^{-1} to more than 10 000 kg ha^{-1}. With low amounts of residue, there may be some detectable changes in active organic matter (e.g., particulate organic matter), but, because the material decomposes so rapidly, there generally will be no detectable change in total soil organic matter. However, a 5-year study in California, USA, has shown that growing a clover cover doubles total organic matter in the top 5 cm of soil (13–26 g kg^{-1}) and increases it by 2 g kg^{-1} at the 5–15-cm depth.

Additional benefits of cover crops include sequestering nutrients (especially nitrate N) so they cannot be leached below the crop root zone, supplying nutrients (usually N) to the following crop, suppressing weeds, breaking pest cycles, and providing habitat for beneficial insects and other fauna. However, before a farmer or land manager decides to grow a cover crop to maintain or improve soil tilth, it is important to determine the objectives. Is the main purpose to add available nitrogen to the soil or to provide large amounts of carbon? Is erosion control during the late autumn and early spring the primary goal? Does the soil have a compaction problem that needs to be alleviated? Will the combination of cover crop, climate, and water-holding capacity of the soil result in excess water depletion that adversely affects the primary crop? Once the primary purpose for growing a cover crop has been determined, decisions on plant species (i.e., legume or grass), the optimum date of planting, and when to eliminate the cover crop will be easier to make.

Evaluating Soil Tilth

Several factors can cause soil tilth to change, including tillage, compaction due to wheel traffic or grazing, reconsolidation due to rainfall or irrigation, and cultivation. Qualitative observations of surface crusts, cloddiness, poor seedling emergence, ponding, restricted plant rooting, soil organism diversity and population (especially earthworms), and increased runoff, erosion, or compaction have traditionally been used to evaluate soil tilth. Among the more quantitative measurements that have been used are aggregation (e.g., mean weight diameter, percentage of water-stable aggregates, or aggregate-size distributions), penetration resistance, infiltration rates, and crop performance.

Penetration resistance is measured most precisely with a penetrometer, although a stiff wire, spade, or tile probe can also be used. For precise measurements, the force required to push a rod with a cone-shaped tip through the soil to a known depth is recorded. To minimize the variation and provide the most quantitative data possible, measurements are taken 24–48 h after rainfall or irrigation so that the soil moisture content is relatively uniform throughout the measurement zone. Penetrometer data are also more reliable in nonstructured sandy or sandy loam soils. In general, if the resistance exceeds 2070 kPa (300 psi) the soil will be too hard for root penetration. Although it is not possible to be quantitative with a wire, tile probe, or spade, it is still possible to use these tools to observe soil structure and qualitatively describe soil tilth. If the soil has good tilth, it is fluffy and granular, and separates easily along the various structural planes of weakness. If the soil structure is massive or compacted, the tilth, by definition, is poor.

Measurements of infiltration rates with a rainfall simulator or other less-sophisticated tools (i.e., single- or double-ring infiltrometers) are also useful for evaluating soil tilth. These measurements are based on the time required for a known volume of water to enter the soil. If a soil has good tilth, water entry will be moderately fast for the first 5–10 cm of water. If water continues to enter at a very high rate, it could indicate that the soil has no ability to retain the water or that there are cracks, crevices, fractures, root channels, worm holes, or other macropores carrying the water away. If water enters the soil at an extremely slow rate or not at all, tilth is generally in a less-desirable condition, perhaps due to compaction or surface crusting. Under those conditions, rainfall and irrigation water usually runs off, increasing the potential for soil erosion and further degradation of soil tilth.

Tilth Indices

Efforts to develop a tilth index have been initiated to improve soil management and provide guidance for custom or prescribed tillage. The need for a tilth index was envisioned because even though an experienced person may be able to determine whether a soil has good tilth by sight and feel, there was no available method to measure or quantify it. Furthermore, if soil tilth can be quantified, it may be possible to measure

Table 1 Soil measurements and corresponding coefficients developed to compute a soil tilth index[a] for Mollisols and Alfisols in the Midwestern USA

Indicator	*Description*	*Tilth coefficient*
Bulk density	Mass per unit volume of dry soil	1.0, for $BD \leq 1.3\ g\ cm^{-3}$ $-1.5 + 3.87 \times BD - 1.5 \times BD^2$, for $1.3 \leq BD \leq 2.1\ g\ cm^{-3}$ 0.0, for $BD \geq 2.1\ g\ cm^{-3}$
Penetration resistance (cone index)	A measure of soil strength, indicating ease of root penetration, plant growth, and yield	1.0, for $CI \leq 1.0$ MPa $1.012 - 0.002 \times CI - 0.01 \times CI^2$, for $1.0 \leq CI \leq 10.0$ MPa 0.0, for $CI \geq 10.0$ MPa
Soil organic matter content	The organic fraction of soil excluding nondecayed plant and animal residue	1.0 for $OM \geq 50\ g\ kg^{-1}$ $0.59 + 0.122 \times OM - 0.008 \times OM^2$, for $10\ g\ kg^{-1} \leq OM \leq 50\ g\ kg^{-1}$ 0.70, for $OM \leq 10\ g\ kg^{-1}$
Aggregate uniformity coefficient	Shape of the grain-size (aggregate) distribution curve (all of equal size gives an AUC 1.0)	1.0, for $AUC \geq 5$ $0.348 + 0.245 \times AUC - 0.023 \times AUC^2$, for $2 \leq AUC \leq 5$ 0.75, for $AUC \leq 2$
Plasticity index	The difference in water content between the liquid limit and the plastic limit of a soil (cohesiveness)	1.0, for $PI \leq 150\ g\ kg^{-1}$ $1.02 + 0.0009 \times PI - 0.00016 \times PI^2$, for $150\ g\ kg^{-1} \leq PI \leq 400\ g\ kg^{-1}$ 0.80, for $PI \geq 400\ g\ kg^{-1}$

BD, bulk density; CI, cone index; UC, uniformity coefficient; OM, soil organic matter; PI, plasticity index; AUC, aggregate uniformity coefficient.
[a]Tilth index $= CF_{(BD)} \times CF_{(CI)} \times CF_{(UC)} \times CF_{(OM)} \times CF_{(PI)}$.

seasonal variation and to understand the subtle effects of soil- and crop-management practices such as conservation tillage or extended rotations.

The initial effort to develop an index assumed that tilth of a mineral soil could be characterized by examining bulk density, strength, aggregate characteristics, organic matter content, and consistency of a soil simultaneously. Bulk density, cone index (penetration resistance or strength), and an aggregate uniformity coefficient were chosen to reflect soil properties that respond to short-term manageable; organic matter content was chosen to reflect long-term management effects. The plasticity index was used as a measure of consistency, which reflects the relative ease with which a soil can be deformed or ruptured. This index relates to soil water properties and is strongly influenced by inherent soil characteristics (predominantly clay type and amount). These five measurements or indicators were also chosen because they can be measured easily in the field or with a routine soil test (e.g., for organic matter; see Table 1).

The search for a better tilth index continued with the addition of soil-water content. Concurrently, and with an even greater emphasis on the biological processes that affect tilth, the concept of soil quality has emerged as a strategy to quantify the effects of various soil-management practices on physical, chemical, and biological properties and processes occurring within the soil (*See* **Quality of Soil**). In the USA, the concept evolved during the 1990s, with two distinct areas of emphasis, education and assessment, both based on principles of soil science. Information sheets about the various indicators (or measurements) that can be used to evaluate soil quality, a *Soil Biology Primer* and *Guidelines for Soil Quality Assessment in Conservation Planning*, were among the educational materials developed for those with minimal knowledge of soil resources. For qualitative evaluation, scorecards were developed to provide farmers with a self-assessment of their current soil and crop management practices. The scoring is relatively simple (e.g., poor, fair, good) and generally based on observations of tilth, earthworms, runoff or ponding of water, plant vigor, ease of tillage, crop growth and development, and yield. Soil-quality test kits have been developed to provide semiquantitative indicator data and guidelines have been written to help interpret the data. A more comprehensive framework that uses scoring curves to interpret a full suite of potential indicators has been developed to provide a more comprehensive evaluation of soil tilth or soil quality. The most important aspect of these efforts is that quantitative methods for characterizing soil tilth do appear to be feasible and with further refinement may be useful for guiding soil and crop management decisions toward more sustainable land-use practices.

Summary

The concept of soil tilth is often easier for a farmer to recognize than for a scientist to describe. Traditionally, tilth has been described in many ways,

usually focusing on the physical condition of a soil and how it responds to tillage, functions as a seedbed, or affects seedling emergence and plant root growth and development. Soil tilth is strongly influenced by soil organic matter and soil structure. Soil management, especially tillage practices, cover crops, and crop rotations, often affects soil tilth because of its effect on soil organic matter. Soil scientists, ecologists, and agricultural engineers are continuing to evaluate different biological, chemical, and physical measurements or indicators that can be used individually or combined into index values to describe tilth quantitatively. That information may then help to improve soil-management practices through technologies such as no-tillage agriculture, the use of cover crops, and more diversified and extended crop rotations.

See also: **Compaction**; **Conservation Tillage**; **Infiltration**; **Organic Matter:** Principles and Processes; **Quality of Soil**; **Structure**; **Texture**

Further Reading

Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn. Boca Raton, FL: CRC Press.

Karlen DL, Erbach DC, Kaspar TC *et al.* (1990) Soil tilth: a review of past perceptions and future needs. *Soil Science Society of America Journal* 54: 153–161.

Karlen DL, Ditzler CA, and Andrews SS (2003) Soil quality: why and how? *Geoderma* 114: 145–156.

Magdoff F and van Es H (2000) *Building Soils for Better Crops*. Burlington, VT: Sustainable Agriculture Publications.

Schjønning P, Elmholt S, and Christensen BT (eds) (2004) *Managing Soil Quality – Challenges in Modern Agriculture*. Wallingford, UK: CABI.

Schwab GO, Fangmeier DD, Elliot WJ, and Frevert RK (1993) *Soil and Water Conservation Engineering*, 4th edn. New York: John Wiley.

Singh KK, Colvin TS, Erbach DC, and Mughal AQ (1992) Tilth index: an approach to quantifying soil tilth. *Transactions of the American Society of Agricultural Engineers* 35(6): 1777–1785.

TIME-DOMAIN REFLECTOMETRY

G C Topp, Agriculture and Agri-Food Canada, Ottawa, ON, Canada
T P A Ferré, University of Arizona, Tucson, AZ, USA

Introduction

The vital role of water in maintaining the life of the landscape requires the development of techniques to monitor and sustain water supply and its quality. Time-domain reflectometry (TDR) is a recently developed technology for use in soil and landscape processes, which has become widely used in highly diversified applications. Despite the importance of water and ionic solutes in the mass and energy balances of the soil profile, it has been only in the last 20 years that rapid, *in situ*, nondestructive measurement of soil water content and ionic solute concentration has become possible in the form of TDR. This is an electromagnetic (EM) technique using radar principles at radio frequency where estimates of water content and electrical conductivity of soil can be made separately from the same wave. The effect of the soil and water on the propagation velocity of the EM wave is analyzed to provide a reliable measure of the water content. The bulk electrical conductivity (EC) is determined from an analysis of the rate of decrease in amplitude during the propagation of the wave in soil. The principles of EM wave propagation are also used for the design of soil probes and to indicate the precision of measurements. The strong interaction between water content and electrical conductivity usually hinders the independent and separate determination of these properties in soil. This difficulty is overcome with TDR where both measurements are from the same sampled region and with the same EM wave.

Water molecules have unique electrical properties that determine the electrical properties of soil. Because of the strong dependence of soil electrical properties on the amount of water in a soil and on the quality of that water, electrical and electromagnetic measurements can be very useful for characterizing the soil water content. One method with which these properties can be measured is TDR. TDR can be thought of as one-dimensional radar using radio waves that propagate along a wave-guide. This wave-guide is constructed from parallel metal rods that are inserted in the soil. As radio waves travel through any material, the characteristics of the wave are continuously changed by the medium through which the wave is traveling. These changes in wave properties are measured by TDR and a wave equation analysis is used to estimate two electrical parameters

of the soil: dielectric permittivity (dielectric constant) and electrical conductivity. These properties are then used to estimate the water content and the electrical conductivity of the soil, respectively.

As the name implies, time-domain reflectometry relies on the measurement of the travel time of reflected waves. In this way, it is similar to other radar applications. In TDR, the wave velocity is measured by recording the time difference between the entry of a wave into the soil and the return of the reflected wave from the end of the TDR probe. Currently available TDR instruments operate in the range of radio frequencies (10 to 1000 MHz). In this frequency range, the dependence of the wave velocity on the water content is not strongly influenced by the electrical conductivity of the pore water or the electrical properties of the soil solids. This is supported by direct measurements made since the mid-1970s that show that, while other soil factors such as density, texture, temperature, and soluble salts affect the TDR wave velocity, their effects are less pronounced than are the effects of water content. That is, the soil water content is the most important soil factor affecting the travel time measured by TDR. This property of TDR makes the method very effective and efficient for the measurement of soil water content.

Soil salinity is an issue of great importance in regard to irrigation and agricultural production in arid and semiarid regions. It has long been realized that the electrical conductivity of pore water can be related to pore-water salinity. However, the bulk soil electrical conductivity is not solely dependent upon the concentration and composition of salts in solution. This property is also highly dependent on the soil texture, the temperature, and the water content. TDR can be used to infer the bulk electrical conductivity of a soil. The energy loss of the transmitted signal is related to the bulk electrical conductivity of the soil. This energy loss is determined from the relative amplitudes of the signal source and that of the wave reflected from the end of the probe. TDR has the advantage of simultaneously measuring both the water content and the bulk electrical conductivity in approximately the same volume of porous medium. Under some conditions, this information can be used to infer the bulk electrical conductivity, and thereby the salinity, of the pore water.

In the 20-plus years since TDR was first applied to soil measurements, the method has become a *de facto* standard for soil water content measurement and monitoring. The popularity of the method for environmental monitoring and research arises from a combination of its accuracy in a wide range of soils and its relative ease of use compared with many other available techniques. TDR provides real-time, *in-situ* soil water content measurements. Multiple soil probes can be incorporated into a switching network connected to a single instrument and data logger allowing for remote automated monitoring. For most soils, the accuracy of measurements of volumetric water content change is within $\pm 0.02\,m^3\,m^{-3}$ without the need for soil-specific calibration, and similar absolute water contents can be achieved with calibration. There is considerable flexibility in the design and placement of TDR probes, allowing users to modify water content measurement networks to conform to the requirements of any specific study. Finally, because TDR measures the volumetric water content, the data are directly applicable to hydrologic water balance analyses with no need for the measurement of supporting soil parameters such as bulk density.

Measurement of Soil Water Content Using TDR

Measuring the volumetric water content of a soil is difficult. As a result, methods have been developed that infer the volumetric water content based on the water potential, bulk electrical conductivity, or dielectric permittivity. The water potential–volumetric water content relationship is nonlinear and hysteretic. The bulk electrical conductivity depends on many factors in addition to the water content. In contrast, the bulk dielectric permittivity of a soil is highly correlated with the volumetric water content. For these reasons, recent advances in the development of indirect methods of measuring the volumetric water content have focused on methods that rely on the measurement of dielectric permittivity. TDR has long been used as a laboratory method to measure the dielectric permittivity of liquids. In an unrelated field, TDR has been widely applied to the location of faults in buried cables. However, it was not until it was demonstrated that a single correlation existed between the bulk dielectric permittivity and the volumetric water content for a wide range of porous materials that these two distinct uses came together to give rise to the TDR method for soil water content measurement. Over the past 30 years, TDR has been applied at many scales and under a broad range of conditions spanning agricultural and forestry environments and seasonal conditions from summer to winter. In addition, the method has been improved and modified.

Most TDR instruments launch a fast-rise voltage step (rise time <200 ps) along a transmission line that is connected to a probe that is inserted in the soil or medium of interest. The voltage pulse propagates as a planar EM wave, traveling in the soil and guided by the conductors that form the probe. The properties of

the soil that govern the propagation of the TDR pulse are described collectively by the propagation constant of the soil, from which velocity of propagation, v ($m\,s^{-1}$), and attenuation coefficient (α) are derived. The TDR instrument records and displays the voltage returning to the instrument as a function of elapsed time following the launch of the voltage pulse. This gives rise to a characteristic plot known as the waveform.

The waveform is a record of the voltage that arrives back at the instrument after being reflected from the end of the probe. Commonly, the returning voltage, V, is displayed as a normalized voltage, known as the reflection coefficient (ρ), as a function of time (t) (Figure 1, where $V/V_0 = 1 - \rho$), where V_0 is the voltage of the transmitted pulse. From the TDR waveform it is possible to determine both v and α. As discussed above, v is used to infer the volumetric water content while α is used to determine the electrical conductivity. A portion of the TDR step pulse travels a distance, L_{cable} (m), along the connecting cables to the probe and then reflects back to the instrument. A portion of the energy that enters the probe travels an additional distance, L (m), along the probe, and then travels the total distance back to the instrument. Taking the difference in these travel times to be t(s), the velocity of the wave can be determined from the known probe length, L, as:

$$v = \frac{2L}{t} \qquad [1]$$

From electromagnetic wave propagation theory, v is given in terms of relative dielectric permittivity and electrical conductivity of soil, which can be expressed as:

$$v = \frac{c}{\sqrt{\epsilon_{ra}}} \qquad [2]$$

where c ($= 3 \times 10^8\,m\,s^{-1}$) is the velocity of light and other EM waves in vacuum, and ϵ_{ra} is the apparent relative dielectric permittivity or apparent dielectric constant that controls the rate of propagation of the step pulse. Combining eqns [1] and [2] gives:

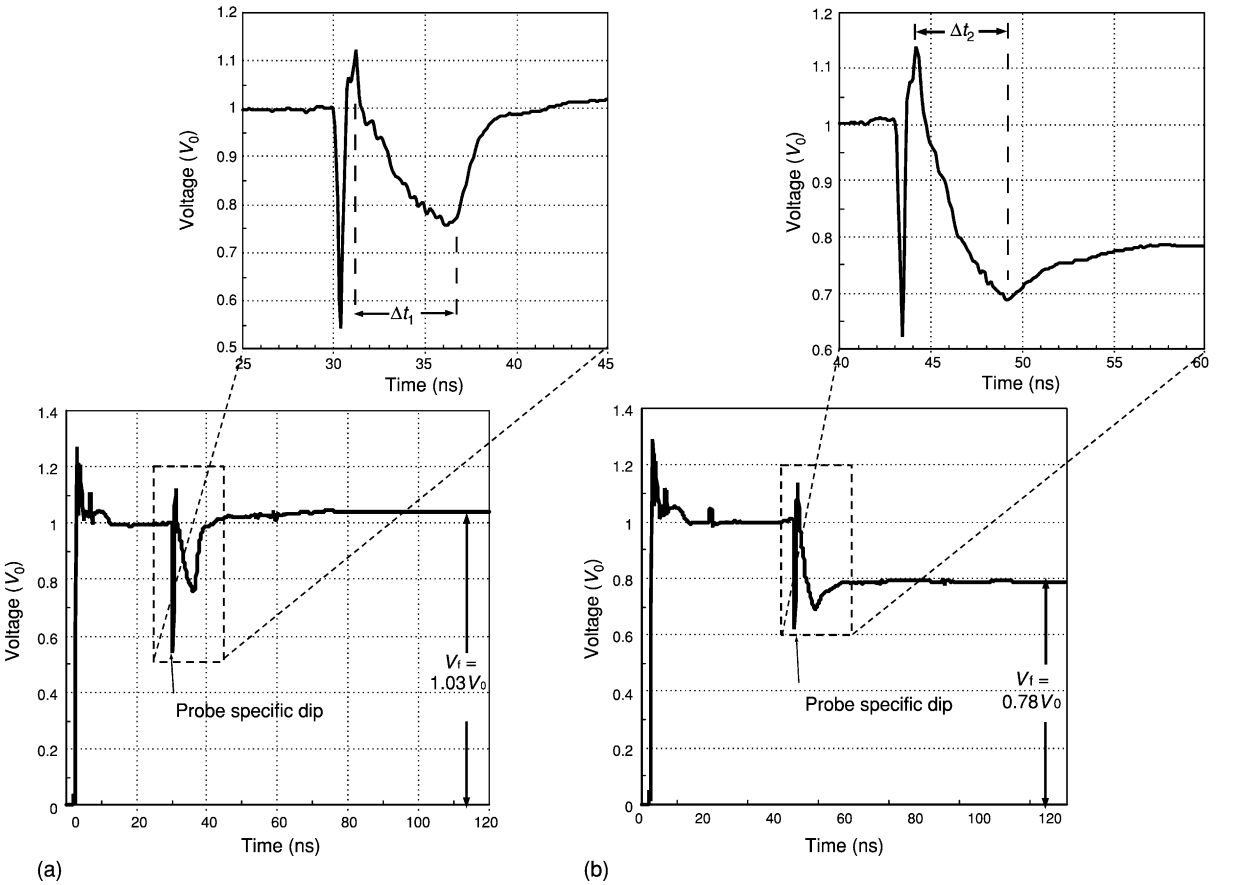


Figure 1 Two time-domain reflectometry curves from 20-cm probes in silty clay loam soil. The soils are at similar water content but the soil solution is more conductive in (b), giving a smaller return reflection and resulting lower V_f. In (a) $\theta_v = 0.304\,m^3\,m^{-3}$ and $\sigma_0 = 57\,mS\,m^{-1}$ and in (b) $\theta_v = 0.271\,m^3\,m^{-3}$ and $\sigma_0 = 95\,mS\,m^{-1}$.

$$\epsilon_{ra} = \left(\frac{ct}{2L}\right)^2 \quad [3]$$

More complete analyses of the EM wave equations show that ϵ_{ra} is primarily the result of the relative dielectric permittivity (real component) or dielectric constant of the soil, ϵ'_r. Also, ϵ_{ra} is affected to some extent by the electrical conductivity, σ_0 ($S\,m^{-1}$), the frequency of the signal, ω ($2\pi\,s^{-1}$), and the relative dielectric loss or imaginary component of relative dielectric permittivity, ϵ''_r. The effect of these factors on EM waves is expressed as:

$$\epsilon_{ra} = \frac{\epsilon'_r}{2}\left[1 + \sqrt{1 + \left(\frac{\epsilon''_r + \frac{\sigma_0}{\omega\epsilon_0}}{\epsilon'_r}\right)^2}\right] \quad [4]$$

where ϵ_0 is the dielectric permittivity of free space ($= 8.85 \times 10^{-12}\,F\,m^{-1}$). Most TDR applications in soil have assumed that $(\epsilon''_r + \sigma_0/\omega\epsilon_0)/\epsilon'_r << 1$, which leads to $\epsilon_{ra} \approx \epsilon'_r$. This assumption allows for the determination of the volumetric water content without prior knowledge of the bulk electrical conductivity.

The high relative dielectric permittivity of water imparts to wet soil a very strong dependence of ϵ_{ra} on the water content. The early laboratory work in TDR adopted an empirical approach, making no assumption about the 'state' of water in porous materials. Instead it related the TDR-measured ϵ_{ra} to the volumetric soil water content, θ_v ($m^3\,m^{-3}$), based on oven drying to 105°C. This early work gave a consistent empirical relationship between relative dielectric permittivity and volumetric water content for a wide range of soils and became a model for calibration of the TDR as a method for measurement of water content. Furthermore, it was found that the relationship is essentially independent of soil bulk density, texture, ambient temperature, and salt content, within the range of these parameters most often encountered. The equation now used widely for calibration is:

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2}\epsilon_{ra} - 5.5 \times 10^{-4}\epsilon_{ra}^2 + 4.3 \times 10^{-6}\epsilon_{ra}^3 \quad [5]$$

This relationship has extremely broad applicability, providing good descriptions for the water content of soils of widely varying properties, such as iron-rich volcanic soils, the unfrozen water content in frozen soils, soils with up to 50% gravel, oil shale waste, and crushed concrete. The widely obtained agreement between observations and eqn [5] has not been true for observations made in low-density soils, mineral soils with high organic or clay contents and artificial soils, such as glass beads. The widespread use of TDR has resulted in a number of applications in which eqn [5] cannot be used, and some effort has been devoted to finding alternative relationships between ϵ_{ra} and θ_v. However no relationship has emerged that shows more general applicability without the requirement for other supporting measurements (e.g., clay content, temperature).

Alternatives to empirically derived calibrations are often based on dielectric mixing models. The application and use of mixing models require varying levels of prior knowledge about the soil, such as porosity, density, and the relative dielectric permittivity of soil components. The application of mixing models for water content determinations has shown that 'square root' mixing models are most common. Such models are equivalent to a linear relationship between $\sqrt{\epsilon_{ra}}$ and the TDR travel time, t in eqn [3], which leads to a convenient linear calibration for water content. The linear relationship between θ_v and $\sqrt{\epsilon_{ra}}$ is supported by calibration data from numerous sources and is very similar to the polynomial expression in eqn [5] (Figure 2). Fitting a linear relationship where possible presents a significant improvement over fitting a

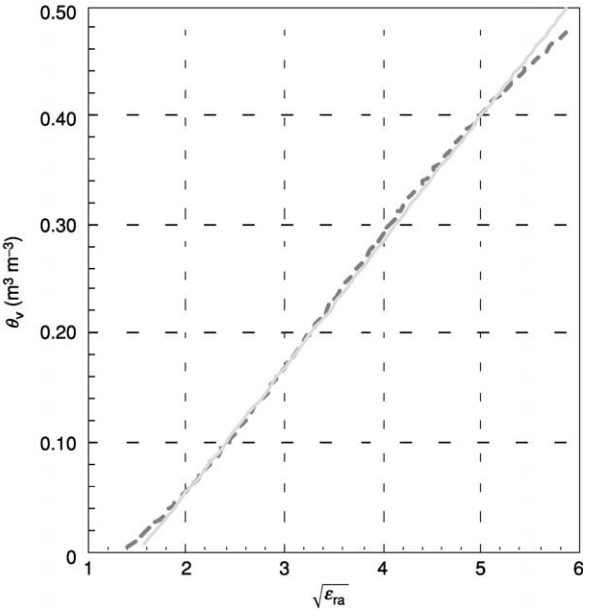


Figure 2 The similarity between eqn [5] (dashed line) and eqn [6] (solid line) over the water content range (0.05 to 0.45 $m^3\,m^{-3}$) applicable for most soils. Note the *x*-axis is $\sqrt{\epsilon_{ra}}$ and shows the linearity of eqn [6].

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2}\epsilon_{ra} - 5.5 \times 10^{-4}\epsilon_{ra}^2 + 4.3 \times 10^{-6}\epsilon_{ra}^3 \quad [5]$$

$$\theta_v = 0.115\sqrt{\epsilon_{ra}} - 0.176 \quad [6]$$

Reproduced with permission from White I, Knight JH, Zegelin SJ, and Topp GC (1994) Comments on ''Consideration of the use of time-domain reflectrometry (TDR) for measuring soil water content'' by Whalley WR. *European Journal of Soil Science* 45: 503–508.

polynomial calibration curve because there are only two parameters to define, making the fitting procedure much simpler. A linear regression of $\sqrt{\epsilon_{ra}}$ on θ_v gives rise to (Figure 2)

$$\theta_v = 0.115\sqrt{\epsilon_{ra}} - 0.176 \quad [6]$$

The constant term includes the electrical properties of the dry solids, primarily the bulk density and the relative permittivity of the constituent material. The slope coefficient in the first term on the right hand side embodies the effects of dissolved solutes and clay surfaces on the electrical properties of the water phase as seen by TDR. This slope changes relatively slowly as salt and clay contents increase. Clay and salt content do introduce some curvature to the relationship between $\sqrt{\epsilon_{ra}}$ and θ_v, indicating the importance of calibration checks to assure an appropriate calibration for high-resolution measurements and for unique soil conditions. Precise calibration equations become more important where absolute measures of water content are required as contrasted with differences or changes in water where more general calibrations usually suffice.

Instrumentation and Wave-Guides

The original and still widely used TDR instrument for soil was the portable cable-tester. This instrument displays the TDR waveform on a screen, allowing analyses to be performed and recorded manually. Alternatively, the data can be recorded and analyzed digitally on a personal computer (PC). Currently, a number of companies offer TDR instrumentation aimed directly at soil and environment studies. Most commercial instruments now offer automated analysis of the TDR waveform as a part of the basic instrument with network switching for multiple probes as an option. Most of these instruments make use of general calibration relationships that closely approximate eqns [5] or [6]. In addition, they allow for the calibration relationships to be customized for the desired soil characteristics.

One of the most important components of a TDR system is the soil probe or transmission line. The basic elements are the conductive components, often parallel metallic rods, which act as wave-guides. Currently the most common soil probes are of the parallel rod type, consisting of two or three parallel rods. These may vary in length from 0.1 to 1.0 m and with probe separations from 0.01 to 0.1 m. Probes that use bare metallic rods measure total water content along the path of the TDR pulse regardless of the distribution of water along this path. However, the distribution of electromagnetic energy is not uniform in the directions perpendicular to the rods. Furthermore, the distribution of energy depends on the number of rods, their diameters, and their separations. A rod separation to rod diameter ratio less than 10 is a practical rule of thumb for probe design. Further, the rod diameter should be large in relation to the dominant pore size, e.g., 10 times the dominant particle or aggregate size. We have found that 6-mm rods spaced at 50 mm have worked well in a variety of studies in tilled and untilled agricultural soil.

Many other probe configurations have come into current use, including profiling probes with the conductors mounted on opposite sides of a dielectric central core. The parallel wires mounted on a central core have been diversified to include a helical-wrapped pair on a single shaft probe and serpentine pathways on a plastic surface giving a surface probe. In all cases the energy density of the propagating wave decreases rapidly with distance from the conductors. Spatial weighting of the measurements in the cross-section perpendicular to the direction of wave propagation is a very important consideration, especially for probes that have nonmetallic components near the metal rods. Numerical analyses have been applied to the wave-equation response from parallel rod probes embedded in a porous medium. One aspect of this work was the definition of sample areas in the transverse plane that are not limited to predefined shapes, allowing for a more realistic description of the lateral sensitivity of different probe configurations. In general, TDR probes should be installed in a manner that minimizes variability of the water content within their sampling volume, especially in the direction that is transverse to the rods.

An important application of TDR relates to the measurement of shallow soil water content profiles. Three general approaches to water content profiling are depicted in Figure 3 using vertical and horizontal probes. In general, vertical rods (Figure 3a) give the most accurate measurement of cumulative water from the surface to the ends of the rods, regardless of the water content distribution with depth. Horizontal rods (Figure 3b) give higher resolution of the water content profile, with less accurate measurements of the total water stored to a given depth. It is possible to place probes depicted in Figure 3a and 3c at an angle off the vertical. Angled rods at say 45° give an optimal compromise between horizontal and vertical configurations. Another consideration is the potential for causing cracks in the soil or gaps around the rods. These features can have a significant impact on TDR measurements and on water movement, and vertical emplacements are most susceptible to these influences. In addition, water content profiles that are determined by the differencing of adjacent measurements (e.g., Figure 3a) introduce the error from both measurements, which can lead to a cumulative error

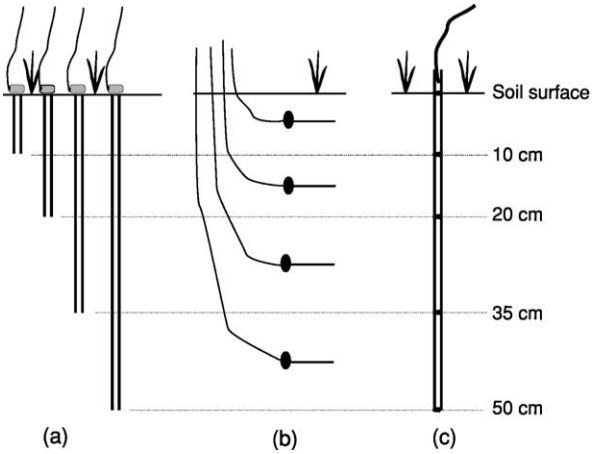


Figure 3 Three options for measurement of water content profiles. (a) Adjacent vertical; (b) horizontal; (c) segmented vertical. Probe lengths and intervals can be adjusted to the monitoring requirements. Installing probes in (a) and (c) off vertical results in a compromise between vertical and horizontal installation but optimizes the advantages from each. Reproduced with permission from Ferré PA and Topp GC (2002) Time domain reflectometry. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis Part 4 Physical Methods*. SSSA Book Series No. 5, Soil. Sci. Soc. Amer., Madison, WI.

as large as $\pm 0.03\,m^3\,m^{-3}$. On the other hand, horizontal rods require excavation for insertion at depth, which can also affect water movement. In addition, because the head of a horizontal probe is placed beneath the ground surface, it is important that it be watertight to avoid the influence of water entry. Practically, vertical rods are the simplest to install. However, their tendency to move vertically out of the soil during winter by the processes of frost-heave can be countered somewhat by the use of angled rods.

An alternative to the continuous-rod probes is the segmented probe depicted in Figure 3c. This probe configuration makes use of diode-shorting to achieve probe segmentation electronically and is available commercially. This probe is essentially a series of vertical probes placed in the same hole. As a result of its design, these probes provide both the water content profile and the total profile stored water from a single installation. Installation of this type of probe is of critical importance as any disturbance of the soil adjacent to the probe is in the most sensitive part of the measured region. Similarly, air gaps or preferential water flow along the probe will also affect measurements.

Impact of Salinity on Water Content Measurement

The explicit effect of the bulk electrical conductivity on TDR water content determinations has not been fully evaluated. Research is continuing with varied EM techniques to define conditions when the direct effect of bulk electrical conductivity (EC) results in unacceptable error in water content estimates. It appears, however, that the influence is very small for mineral soils having a clay content less than 50% and pore-water electrical conductivities less than $5\,dS\,m^{-1}$. For example, measurements in a soil of 55% clay gave an overestimate of water content of $0.014\,m^3\,m^{-3}$ due to the assumption that $\epsilon_{ra} = \epsilon'_r$. Typically, high EC conditions lead to difficulty in measuring the water content due to total extinction of the waveform through energy loss. This places a practical limit on the maximum length of TDR rods. However, to maintain acceptable water content accuracy, probe length should be at least 7.5 cm. If the EC is too high to use probes at least 7.5 cm in length, nonconductive rod coatings can be used to minimize the effects of signal loss. These coatings change the water content calibration and they should only be used when absolutely necessary.

Some of the limitations of TDR analysis and application may be addressed by a new approach to TDR waveform analysis. Standard analyses involve interpretation of the time of arrival of reflections. However, in addition to this time-based analysis, a TDR waveform can be analysed in the frequency domain using Fourier analysis. Conversion of TDR data into the frequency domain provides additional frequency-dependent information about the electrical properties of the soil and water. Frequency domain analyses of measurements made with 3-cm length probes may provide high-precision measurement of relative dielectric permittivity in saline soils where the time-domain analysis is not feasible. Further developments of this approach may extend the range of applicability of TDR.

Measurement of Bulk Electrical Conductivity Using TDR

As the TDR wave propagates through soil guided by the probe rods, its velocity is decreased by the presence of water. In addition, the magnitude of the wave is decreased by electrical losses associated with the bulk electrical conductivity of the soil, σ_0. The decrease in magnitude of the wave is expressed in the attenuation coefficient, α. It is possible to express σ_0 in terms of the values obtained from the TDR waveform (Figure 1):

$$\sigma_0 = \left(\frac{1}{120\pi L}\right)\left(\frac{Z_0}{Z_i}\right)\left(\frac{2V_0}{V_f} - 1\right) \qquad [7]$$

where Z_0 (ohms) is the characteristic impedance of the TDR probe (a factor of the probe geometry), Z_i is the output impedance of the TDR instrument, V_0 (volts) is the initial step voltage output and V_f is the

final voltage remaining after all multiple reflections from the probe, respectively. In practice, the quantity $(Z_0/120\pi LZ_i)$ is often treated as a calibration constant for the instrument, cable, and probe combination, which can be obtained from measurements in solutions of known EC. After the probe and TDR instrument have been characterized, the bulk EC is calculated from the value of V_f as the only independent variable at each measurement site or measurement time. Although the measurement is straightforward, it is important that V_f be determined well after all multiple reflections have returned and the TDR waveform has reached a final constant value.

The bulk EC measured by TDR has applications in research into solute transfer processes, particularly ionic solute breakthrough experiments. At a heterogeneous field site, using vertically installed TDR rods, solute mass flux measured with TDR was in good agreement with that obtained from solution samplers. The TDR-measured EC has been used to monitor nitrate levels in the field with a view to improving crop nutrient management and minimizing leaching to groundwater.

Impact of Water Content on Pore-Water Salinity Measurement

The EC of the soil solution is often desired for soil and environment studies, particularly in relation to soil salinity assessment or monitoring migration of polluting chemicals. The relationship between bulk EC and solution EC is strongly dependent on both the soil water content and the clay content of the soil. TDR offers the distinct advantage of measuring both water content and EC in the same sample and under identical conditions. Thus, TDR can be very useful for establishing both the water content and the pore-water EC simultaneously. TDR has been used in both direct and indirect approaches and to elucidate the relationship among water content, bulk EC, and solute EC. However, it is not always possible to determine the pore water EC if the water content varies along the TDR probes, as this will give a gradient of pore water EC along the probe.

Conclusion

In the 25 years since the early applications in soils, TDR has revolutionized the field measurement of soil water content. In 1991, a survey of soil physicists in four countries reported TDR was used one-third as often as gravimetric sampling and one-half as often as neutron moderation. It is most likely that TDR use had surpassed both these methods by the end of the 1990s, making it more ubiquitous than any other technique. The capability to log TDR data and the relatively nondestructive nature of the installations have allowed *in situ* monitoring, contributing to the much wider use of TDR. For water balance and water-use efficiency studies, TDR is particularly applicable because the probe designs and orientations can be chosen to meet a variety of conditions, such as agricultural row crops, drip irrigation patterns, and forested landscapes. Through the application of TDR in soil much has been learned about the soil's electromagnetic properties in the radio frequency range. This knowledge has assisted in the development of other electromagnetic techniques, such as ground-penetrating radar and remote sensing using either active or passive microwave radars. TDR is specifically valuable for 'ground-truthing' radar applications because both are interpreting measured EM properties of soil. As a consequence, TDR can be applied to defining the sampling regions of radar applications by appropriate choice of probe geometries for the TDR. Water content measurement by capacitance techniques has been accelerated recently, in part because appropriate frequencies and probe geometries were identified through TDR studies. Although capacitance devices are less expensive, they have not yet offered the ability to measure EC as can be done by TDR.

TDR is prominent among several emerging technologies being used to enable the long-neglected investigation of roots as the major agents of water and chemical transfer in soil. TDR water content measurements also improve the value and significance of other environmental parameters. For example, frequent measurements of temperature, water content by TDR, and oxygen concentrations in the rooting zone of maize can help to demonstrate when tillage and trafficking limit both O_2 supply and crop growth. Prior to TDR, measurements of O_2 supply for roots and other biota in the soil had not been possible. Time-domain transmissiometry (TDT), which is similar to TDR, has been combined with a soil penetrometer to measure both soil strength (compaction) and water content. The combination of sensors greatly enhances the value of the soil penetrometer as a diagnostic tool for soil compaction.

Regulations now play a leading role in the need to monitor soil contamination and harmful infiltration of contaminated solutions into groundwater. Cities in the USA such as Los Angeles, Tucson, Las Vegas, and Glendale are using TDR for such monitoring. Agencies from around the world have reported using TDR in environmental monitoring applications as diverse as nonaqueous liquid-phase concentrations (Sweden), subsurface pollutant detection (United Arab Emirates), changes in groundwater level and

crude oil thickness (USA), characterization of diesel-contaminated soil (Canada), and measuring water-table elevations within sidewalk systems (USA).

The application of TDR in soil is continuing to develop and diversify in ways that will enhance the value of TDR for soil and environmental monitoring. The analyses of spatial sensitivity within probes have facilitated designs for specific uses, often allowing combinations of TDR with other measurements such as with tensiometers and the cone penetrometer.

List of Technical Nomenclature

α	Attenuation coefficient
ϵ_0	Dielectric permittivity of free space ($F\,m^{-1}$)
ϵ_r	Relative dielectric permittivity (dielectric constant)
ϵ_r'	Real component of relative dielectric permittivity (dielectric constant)
ϵ_r''	Imaginary component of relative dielectric permittivity (relative dielectric loss)
ϵ_{ra}	Apparent relative dielectric permittivity (dielectric constant)
θ_v	Volumetric soil water content ($m^3\,m^{-3}$)
ρ	Reflection coefficient
σ_0	Electrical conductivity ($S\,m^{-1}$)
ω	Frequency of signal ($2\pi\,s^{-1}$)
c	Velocity of light ($m\,s^{-1}$)
L	Length (m)
t	Time(s)
V	Voltage (volt)
v	Propagation velocity of electromagnetic waves ($m\,s^{-1}$)
Z_0	Characteristic impedance of probe (ohms)
Z_i	Output impedance of instrument (ohms)

See also: **Neutron Scattering**; **Salination Processes**; **Water Content and Potential, Measurement**; **Water Potential**

Further Reading

Clothier BE and Green SR (1997) Roots: the big movers of water and chemical in soil. *Soil Science* 162: 534–543.

Dowding CH (ed.) (2001) *Second International Symposium and Workshop on Time Domain Reflectometry for Innovative Geotechnical Applications*. IL, USA: Infrastructure Technology Institute at Northwestern University Evanston.

Ferré PA and Topp GC (2000) Time-domain reflectometry sensor techniques for soil water content measurements and electrical conductivity measurements. In: Baltes H, Gopel W, and Hesse J (eds) *Sensors Update*, vol. 7, pp. 277–300. Weinheim: Wiley-VCH.

Ferré PA and Topp GC (2002) Time domain reflectometry. In: Dane J and Topp GC (eds) *Methods of Soil Analysis*, Part 4. 3rd edn, pp. 434–446, 534–545. Madison, WI: Soil Science Society of America.

O'Connor KM and Dowding CH (1999) *GeoMeasurements by Pulsing TDR Cables and Probes*. Boca Raton: CRC Press.

Topp GC and Reynolds WD (1998) Time domain reflectometry: a seminal technique for measuring mass and energy in soil. *Soil and Tillage Research* 47: 125–132.

Wraith JM (2002) Time domain reflectometry. In: Dane J and Topp GC (eds) *Methods of Soil Analysis*, Part 4. 3rd edn, pp. 1289–1297, 1311–1321. Madison, WI: Soil Science Society of America.

TROPICAL SOILS

Contents

Arid and Semiarid

H C Monger, New Mexico State University, Las Cruces, New Mexico, USA
J J Martinez-Rios, Universidad Juarez del Estado de Durango, Mexico
S A Khresat, Jordan University of Science and Technology, Irbid, Jordan

Introduction

Arid and semiarid soils occupy about one-third of the Earth's ice-free land surface. They are sources and sinks of atmospheric CO_2, sources and sinks (mainly sources) of global dust, and substrate that support high biodiversity of plants and animals. Arid and semiarid soils uniquely accumulate secondary minerals, such as calcite and gypsum, as the result of low rainfall and limited leaching. Because of sparse vegetative cover and high susceptibility to wind and water erosion, many arid and semiarid soils have low resistance and low resilience to disturbance. Hence land degradation (i.e., desertification) is common on most continents with arid and semiarid soils.

These dryland soils have been a factor of primary importance in human history. The oldest hominid fossils are found in east Africa in sediments with paleosols containing pedogenic carbonate indicating an arid or semiarid climate. The transition from hunting-and-gathering to agriculture took place in arid and semiarid Mesopotamia about 10 000 years ago. The production of surplus food on the floodplain soils of the Euphrates and Nile gave rise to early civilizations in Sumeria, Babylonia, and Egypt. Today several large urban centers are located on arid and semiarid soils around the world that have adequate groundwater or river water supplies. If irrigated, arid and semiarid soils are an important source of local and global food production. Still, most arid and semiarid land is sparsely populated, open, and often wilderness land.

Climatic Controls

The terms arid, deserts, semiarid, and steppes are used variously to describe dryland conditions. Arid (an adjective) describes a climatic condition of low rainfall – commonly taken to be less than 250 mm (10 in.) of mean annual precipitation (Figure 1a). Desert (a noun) is a region of the Earth's land surface within an arid climate. Likewise, semiarid is a climatic condition characterized by a mean annual precipitation between 250 and 500 mm (Figure 1a). A region of the Earth's surface within a semiarid climate is a steppe.

Closely linked to annual precipitation, and especially to soil moisture, is vegetation. The driest deserts are often barren of plant cover, but most deserts have scattered shrubs, cacti, forbs, and grasses. Though some steppe vegetation can occur in areas like the Badia of Jordan that receive as little as 100 mm of rainfall, generally steppe vegetation is characterized by higher amounts of rainfall in which short-grass prairie is bordered by desert vegetation on the arid side and tall-grass prairie, savanna, or woodlands on the subhumid side. Thus, by these definitions arid soils are synonymous with desert soils and semiarid soils are synonymous with steppe soils.

But to define deserts and steppes by precipitation alone is to ignore other important climatic variables, mainly temperature. One expression of the combined influences of both precipitation and temperature is the de Martonne aridity index based on the formula:

$$I_a = P_{mm}/(T^{\circ}C + 10) \qquad [1]$$

where the aridity index (I_a) is equal to the mean annual precipitation in millimeters (P_{mm}) divided by the mean annual temperature in degrees Celsius ($T^{\circ}C$) plus 10. According to this index, values below 5 characterize true deserts, values of approximately 10 demarcate dry steppes, values of about 20 represent prairies, and values above 30 typify forest. The boundary for the Chihuahuan desert of Mexico and the USA, for example, is based on an aridity index of 10 or lower.

The Köppen system also demarcates deserts and steppes as a function of both precipitation and temperature (Figure 1b). For example, some cold regions in the high latitudes of North America and Eurasia that get semiarid-amounts of precipitation are coniferous forest instead of steppes, and many cold

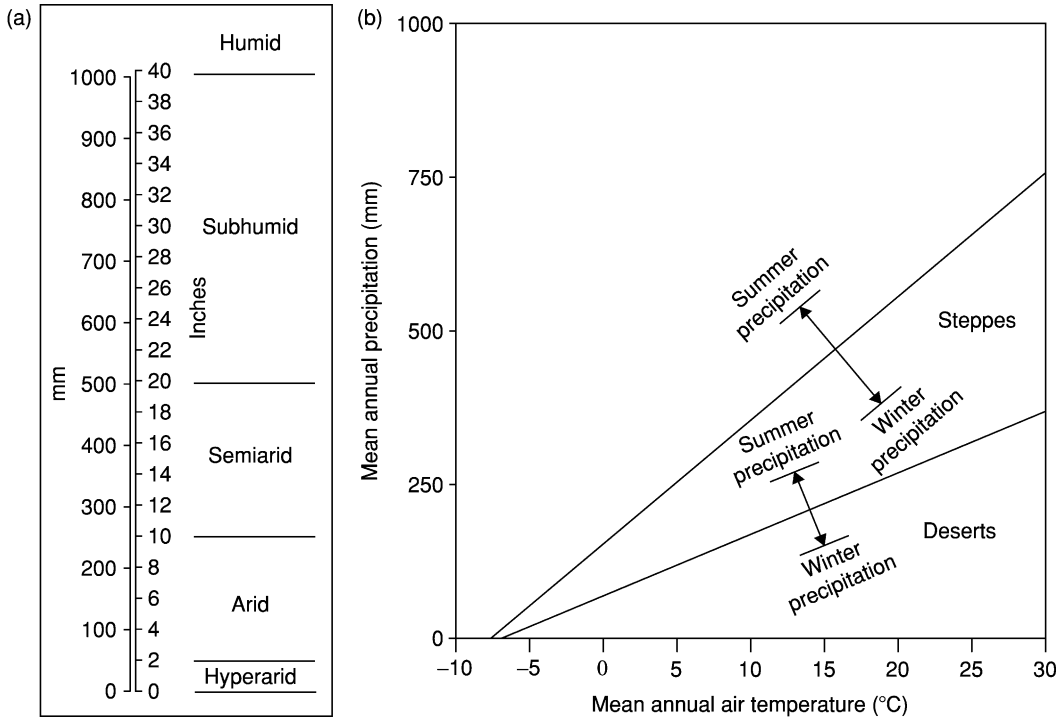


Figure 1 (a) Climate categories based on mean annual precipitation. (b) Desert and steppe boundaries as a function of both mean annual precipitation and temperature according to the Köppen system. Arrows show how boundaries shift according to whether precipitation falls mainly in the summer or winter.

areas that receive arid-amounts of precipitation are steppes instead of deserts.

Seasonality of precipitation is another climatic factor that affects desert and steppe boundaries. For a given mean annual temperature, the boundary of a steppe will extend into wetter climates if its precipitation falls mainly in the summer (Figure 1b). In other words, if precipitation falls in summer, the area of a steppe will be larger because summer evapotranspiration depletes soil moisture more thoroughly than winter evapotranspiration. Similarly, the size of a desert will be larger if its precipitation falls in the summer rather than in the winter.

Soil water potential is another way of defining boundaries of arid and semiarid soils. Soil moisture measured as soil water potential, which includes the influence of particle size and salts, is more important to vegetation than annual precipitation alone. The Soil Taxonomy system, for example, uses soil water potential to define moisture regimes as a criterion for classifying soils. The aridic moisture regime, for example, includes soils too dry to support nonirrigated crops, and is defined as soil that is moist (i.e., water held at tensions greater than −1500 kPa) for no more than 90 consecutive days when the soil temperature at a depth of 50 cm is above 8°C. Soils with the ustic and xeric moisture regimes are transitional between the aridic moisture regime and soils of humid climates that have the udic moisture regime. Semiarid soils occur within ustic and xeric moisture regimes, their drier subdivisions, and wetter subdivisions of the aridic moisture regime, namely, the aridic ustic, aridic xeric, xeric aridic, and ustic aridic regimes.

Processes of Soil Formation in Arid and Semiarid Soils

Water – its amount and depth-of-wetting – is the major driver that gives rise to differences between arid and humid soils. In humid soils >50% of the water entering the soil drains downward through the profile to groundwater. In arid soils, <10% flushes through the soil profile to the groundwater. A humid soil receiving 1300 mm of rain, for example, would have about 650 mm that percolated through its profile in a year. An arid soil receiving 200 mm would have <20 mm that percolated through its profile. Thus, nearly all water in arid soils and much water in semiarid soils enters and leaves via the soil surface. Notable exceptions are low-lying areas that receive runoff water (e.g., playas). In these topographic lows, soils are nonsaline because of deep leaching, if the water table is deep and the soils have high permeability. However, if the water table is shallow or soils have slow permeability, these topographic lows are zones of salt accumulation

As with humid soils, the processes of soil genesis (i.e., gains, transfers, transformations, and losses) operate in arid and semiarid soils, but the magnitude and direction of these processes are different. The shallow depth-of-wetting and incomplete leaching have a major impact on gains because authigenic minerals, such as calcite, silica, and gypsum, accumulate in the profile and give rise to the formation of calcic, petrocalcic, duripans, gypsic, and petrogypic horizons. Gains of dust are also important. Silicate clay dust, for example, contributes to the formation of argillic horizons, carbonate dust to the formation of calcic horizons, and gypsum dust to the formation of gypsic horizons. Gains of photosynthetic carbon in the form of soil organic matter are lower than in humid soils. But gains of photosynthetic carbon released as respired soil CO_2 that leads to HCO_3^- and $CaCO_3$ formation are higher than in humid soils.

Transfers of material down the profiles of arid and semiarid soils include illuvial clay, carbonate, and salts. Arid soils typically display a chromatographic pattern of an argillic horizon overlying a calcic horizon. If gypsum and soluble salts are also present, the profile can contain an argillic overlying a calcic overlying a gypsic overlying a salic horizon. Transfers of materials also occur up the profiles of some arid and semiarid soils. These include capillary rise of soluble material and particles moved upward by ants and termites. In some cases, desert pavements are formed by the upward movement of coarse fragments lifted by silts and fine sands that accumulate beneath them.

Important transformations in arid and semiarid soils include rock disintegration resulting from the crystallization of salts, thermal fluctuations, and chemical weathering. Other transformations involve the decomposition of organic matter and formation of clay minerals, such as palygorskites and sepiolites. The formation of kaolinite is common in parent materials containing feldspars. Losses are mainly in the form of erosional truncation of soil horizons.

Factors of Soil Formation in Arid and Semiarid Soils

The five soil-forming factors interact in assorted ways to produce arid and semiarid soils (Figure 2). Climate, especially water supply, is the defining factor. Water is not only the agent for mineralogical gains, transformations, and transfers, it is also essential for nutrient supply and direct use by vegetation. Vegetation, in turn, has a feedback link to soil by adding organic matter, translocating ions via bioaccumulation, and providing ground cover that protects the soil from erosion. Vegetation is also linked to animals by being their food supply, while animals are linked to vegetation by herbivory and seed dispersal (Figure 2). Animals affect soil by bioturbation – humans alone move about 40 GT of soil per year. Soil, in turn, affects animals by providing habitat, which is important for nematodes, gastropods, earthworms, crustaceans, mites, spiders, ants, termites, mice, moles, rabbits, gophers, birds, foxes, badgers, deer, bear, and even humans in some places of the arid world.

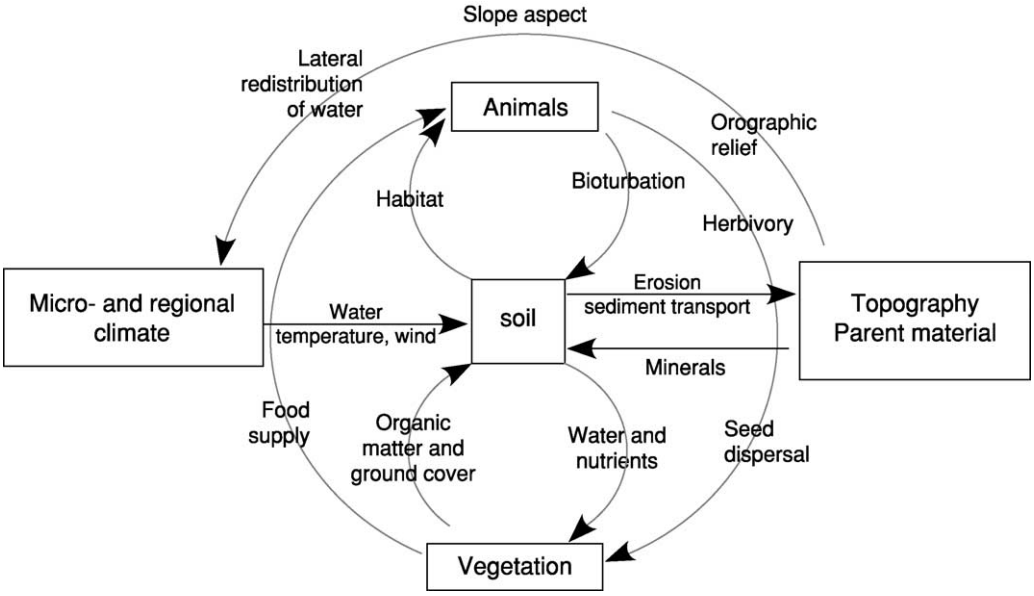


Figure 2 Illustration of the links among the soil-forming factors of climate, biota (vegetation and animals), topography, and parent material. The fifth factor, time, affects all of the illustrated links in proportion to the duration that the factors operate.

Parent material and topography are the main factors of the geologic setting that affect arid and semiarid soils. Parent material has a direct link to soil as a supplier of mineral detritus. Topography has its link to soil by its influence on regional climate at the orographic scale and by its influence on microclimate (resulting from slope aspect and laterally redistributed water) at the local scale. Over the long term, climate through its effects on soils alters the configuration of the landscape by erosion and sediment transport (Figure 2).

Properties of Arid Soils

Arid soils have surface horizons with several unique characteristics. Many arid soils, for example, are covered by desert pavement that overlies vesicular A and E horizons. Other arid soils are covered by salt efflorescence in areas where shallow groundwater has risen by capillarity and evaporated at the surface. Still other arid soils are covered by microbiotic crusts or by blankets of aeolian sand or silt. Nearly all arid soils have lower amounts of organic matter than their more humid counterparts. For classification purposes, the surface horizon (i.e., epipedon) that is ubiquitous for arid soils is the ochric epipedon. Other epipedons of arid soils with much smaller occurrences are the mollic, anthropic, and in very rare cases of grass sod over shallow basalt, the histic.

Subsurface horizons of arid soils are uniquely different from subsurface horizons of humid soils in some instances, yet similar in other instances. Subsurface horizons in arid soils that are uniquely different include horizons dominated by calcium carbonate, secondary silica, gypsum, and soluble salts, while horizons common to both arid and humid soils include those with weak structural and color development and accumulations of illuvial clay and sesquioxides. Diagnostic horizons unique to arid soils include the calcic, petrocalcic, duripan, gypsic, petrogypsic, natric, and salic horizons (Figure 3). Diagnostic horizons found in both arid and humid soils include the cambic, argillic, and in rare cases, the oxic.

In addition to diagnostic horizons, other soil properties, such as vertic, andic, lithic, climatic, anthropogenic, depth to groundwater, and particle-size characteristics are used to make taxonomic subdivisions in arid soils. At the highest taxonomic levels, arid soils include Leptosols, Gypsisols, Durisols, Calcisols, and Solonchaks in the World Reference Base (WRB) system; Desert gray brown, Desert takyr-like, Desert sandy, and Meadow desert soils in the Russian system; and Halosols and Aridisols in the Chinese system. In the Soil Taxonomy system, arid soils are classified as, in order of abundance, Aridisols, Entisols, Vertisols, Oxisols, and Andisols (Table 1). Moving dune fields also occupy large areas of deserts, especially in North Africa.

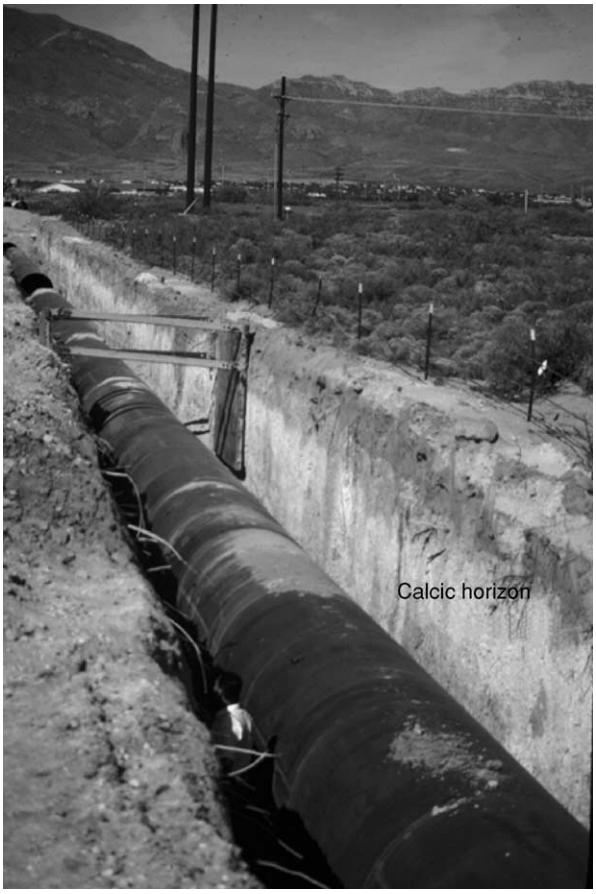


Figure 3 Calcic horizon in an arid soil in the Chihuahuan desert of North America. This site has an annual rainfall of 250 mm and an annual temperature of 16°C.

Properties of Semiarid Soils

Semiarid soils, owing to more rainfall and the homogenizing effects of greater vegetative cover, have surface horizons with more organic matter than arid soils and fewer unique features like desert pavements, vesicular horizons, and efflorescence. Although the ochric epipedon is still a common horizon of semiarid soils, the mollic epipedon is widespread and dominates the steppe areas that border tall-grass prairies.

Being transitional between arid and humid soils, semiarid soils have a large variety of subsurface horizons. As with arid soils, semiarid soils might have calcic, petrocalcic, duripan, gypsic, natric, and salic horizons. Yet, as with humid soils, semiarid soils might have albic, argillic, and kandic horizons as well as fragipans and plinthite.

Table 1 Global extent of arid and semiarid soils (km^2) based on the Soil Taxonomy system[a]

Soil order	*Suborder*	*Africa*	*Asia*	*Australia / Oceania*	*Europe*	*South America*	*Central America*	*North America*	*Global*
Arid soils									
Aridisols									15 798 100
	Cryids	653	417 587	–	103	165 909	–	499 351	1 083 601
	Salids	95 249	595 400	919	130	59 756	–	17 698	769 152
	Durids	–	–	–	–	–	–	–	–
	Gypsids	347 458	326 137	–	3454	–	–	–	677 050
	Argids	450 884	1 635 130	1 710 271	1218	520 175	1111	1 093 291	5 412 080
	Calcids	1 818 639	2 204 270	613 547	1013	98 680	–	202 235	4 938 384
	Cambids	842 681	1 125 259	346 253	5573	422 429	2259	173 378	2 917 832
Vertisols	Torrerts	196 570	65 210	602 630	–	6233	1754	42 435	914 832
Oxisols	Torroxes	9346	–	4247	–	16 550	–	–	30 143
Andisols	Torrands	834	–	–	–	95	–	150	1078
Entisols[b]	In aridic (torric) moisture regimes								12 600 308
Total arid soils									29 344 460
Semiarid soils[c]									
Mollisols									
	Ustolls	4371	1 587 526	21 689	303 635	409 796	–	1 744 438	4 071 455
	Xerolls	72 125	403 161	55 502	236 656	769	–	165 132	933 346
Vertisols									
	Usterts	722 630	600 544	14 863	29 001	38 107	18 247	114 387	1 763 776
	Xererts	11 692	47 348	20 258	17 105	1050	–	969	98 423
Oxisols	Ustoxes	1 718 701	18 457	39 249	–	1 338 805	1295	360	3 116 866
Andisols									
	Ustands	11 235	12 308	–	1975	11 414	11 004	7750	55 686
	Xerands	–	–	–	8937	2098	–	17 840	28 876
Alfisols									
	Ustalfs	2 470 581	1 084 231	501 676	253 839	998 870	22 555	346 868	5 678 621
	Xeralfs	80 071	209 377	270 520	118 833	25 340	–	176 034	880 174
Ultisols									
	Ustults	1 649 310	824 575	101 301	–	1 091 366	58 903	131 048	3 856 502
	Xerults	933	2672	–	–	22	–	15 958	19 586
Inceptisols									
	Ustepts	1 675 118	832 200	257 463	377 338	611 443	83 255	331 262	4 168 080
	Xerepts	163 793	178 312	264	310 697	8342	–	9246	670 654
Entisols[d]	In ustic moisture regimes								4 620 113
Entisols	In xeric moisture regimes								840 021
Total semiarid soils									30 802 179
Total ice-free land area									130 268 185

[a] Courtesy of USDA–Natural Resources Conservation Service, Soil Survey Division, World Soil Resources, 2002.
[b] Entisols with the aridic (torric) moisture regime are designated at the great group level (e.g., Torripsamments and Torriorthents).
[c] The semiarid soils category contains some soils of subhumid climates (500–1000 mm annual precipitation) in cold regions and along coasts.
[d] Entisols with the ustic and xeric moisture regimes are designated at the great group level (e.g., Ustorthents and Xerofluvents).

Other diagnostic properties used to subdivide semiarid soils include redox, petroferric, vertic, andic, lithic, climatic, anthropogenic, particle-size, and groundwater characteristics. Classification of semiarid soils includes the Kastanozems, Chernozems, and Phaeozems in the WRB system; Meadow Chernozem-like, Meadow Chestnut, Semidesert Brown, and Semidesert Meadow Brown soils in the Russian system; and Isohumisols and Ferrallisols in the Chinese system. In the Soil Taxonomy system, semiarid soils are classified as, in order of abundance, Alfisols, Entisols, Mollisols, Inceptisols, Ultisols, Oxisols, Vertisols, and Andisols (Table 1).

Human Land Use of Arid and Semiarid Soils

Early civilizations arose in arid and semiarid Sumeria in the fourth millennium BC as irrigated agriculture on floodplain soils encouraged stable settlements, led to surplus food, and freed people to pursue specialized trades and to develop social order and cultural creativity. Similar cultural developments arose along rivers in other arid and semiarid climates, such as the Indus of ancient India and Hoang-Ho of ancient China. In the western hemisphere as well, societies developed in arid and semiarid climates, such as the Inca, Aztec, and Hohokan cultures.

Grazing of domesticated cattle, sheep, and goats has been the traditional land use of arid and semiarid soil away from the irrigated floodplains. Today, in addition to grazing, arid and semiarid soils have ecological and global biogeochemical properties important to humans. Arid and semiarid ecosystems have some of the highest biodiversity on Earth as grasses, cacti, shrubs, reptiles, birds, and mammals have adapted to wide and rapid shifts in temperature and moisture. Much of the carbon in the global carbon cycle, at least 800×10^{15} g, is stored as soil carbonate in arid and semiarid soils. The source of much of the local, regional, and global dust is from arid and semiarid soils.

Desertification, which is defined by the United Nations Convention on Desertification as 'land degradation resulting from climatic and human activities,' is a major issue of importance in arid and semiarid climates because desertification directly affects about one-sixth of the world's population in both developing and developed countries. Increasing human population and livestock pressures have accelerated desertification and caused shrubs to invade grasslands with subsequent erosion of exposed topsoil. Wind erosion is a major agent for desertification in arid and semiarid regions because it removes organic matter and fine mineral particles in A horizons that have important water and nutrient storage properties. Water erosion is also important, especially at the arid–semiarid boundary of about 250 mm of precipitation where erosion rates are commonly at their greatest.

Limited water supply is of paramount importance to expanding populations in arid and semiarid regions of North and South America, Australia, Africa, Asia, and the Middle East. In the western USA, for example, the drought did not move to the people, the people moved to the drought. While the lack of water threatens population growth in arid and semiarid regions, the lack of water is the natural state, and the state that imparted the unique properties to arid and semiarid soils. It is also the lack of water that makes arid and semiarid places some of the most open, least developed places on Earth, and some of the last 'wild' places for future generations to enjoy and study.

See also: **Classification of Soils**; **Desertification**; **Factors of Soil Formation:** Climate

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*. New York: Oxford University Press.

Brandt J and Thornes JB (1996) *Mediterranean Desertification and Land Use*. New York: John Wiley.

Cooke R, Warren A, and Goudie A (1993) *Desert Geomorphology*. London: UCL Press.

Dregne HE (1983) *Desertification of Arid Lands*. New York: Harwood Academic Press.

Durant W (1935) *Our Oriental Heritage*. New York: Simon and Schuster.

Eswaran H, Reich PF, Kimble JM *et al.* (2000) Global carbon stocks. In: Lal R, Kimble JM, Eswaran H, and Stewart BA (eds) *Global Climate Change and Pedogenic Carbonates*, pp. 15–61. Boca Raton, FL: CRC Press.

Gile LH, Peterson FF, and Grossman RB (1966) Morphology and genetic sequences of carbonate accumulation in desert soils. *Soil Science* 101: 347–360.

Harris D (1996) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. Washington, DC: Smithsonian Institution Press.

Krogh L (2002) Major classification systems. In: Lal R (ed.) *Encyclopedia of Soil Science*, pp. 176–182. New York: Marcel Dekker.

Schlesinger WH, Reynolds JF, Cunningham GL *et al.* (1990) Biological feedbacks in global desertification. *Science* 247: 1043–1048.

Schmidt RH (1979) A climatic delineation of the 'real' Chihuahuan desert. *Journal of Arid Environments* 2: 243–250.

Smith GD (1986) *The Guy Smith Interviews: Rationale for Concepts in Soil Taxonomy*. Soil Management Support Service monograph no. 11. Washington, DC: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy – A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. USDA Agriculture Handbook number 436. Washington, DC: US Government Printing Office.

Spaargaren OC (2000) Other systems of soil classification. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E137–E174. Boca Raton, FL: CRC Press.

Strahler AN and Strahler AH (1987) *Modern Physical Geography*. New York: Wiley.

Velasco MH (1983) *Uso y Manejo del Suelo*. Mexico: Editorial Limusa.

Wilding LP (2000) Introduction: general characteristics of soil orders and global distribution. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E175–E182. Boca Raton, FL: CRC Press.

Humid Tropical

S W Buol, North Carolina State University, Raleigh, NC, USA

Introduction

Soils are the physical, chemical, and biological media at the upper surface of the Earth's land areas capable of accepting plant roots. A wide range of geologic materials, soil moisture, and temperature conditions

that differ regionally and among adjacent soils on local landscape positions provide for diverse suites of contrasting soils that interact differently with biological communities and human attempts to sustain food production within the humid tropics.

Humid Tropical Setting

Many climatic classifications have been published. The parameters of soil temperature and moisture regimes used to classify soils do not conform to general concepts of humid tropics. The following criteria, adopted by the National Cooperative Soil Survey in the USA, are most universally used when classifying soil.

Soil-Temperature Regimes of the Tropics

Mean annual soil temperatures are 2–4°C warmer than mean annual air temperatures. Soil-temperature regimes (STRs) are defined by two criteria, mean annual soil temperature and seasonal temperature difference (Figure 1). Seasonal soil-temperature difference is determined as the mean soil temperature of June, July, and August compared with the mean soil temperature of December, January, and February. Almost all of the soils within the geographic tropics have soil temperatures that seasonally differ by less than 6°C and are identified by the prefix 'iso' placed before the name of the mean annual soil-temperature regime. Except for a few soils near the northern and southern extremities of the geographic tropics in Africa, this one characteristic is the only soil property that is nearly universal with the concept of tropical soil. The practical aspect of this soil property is that seasonal soil temperatures seldom have to be considered when planting food crops.

Soils with mean annual soil temperatures of 22°C or higher are classified as isohyperthermic; soils with mean annual soil temperatures of 15–22°C are identified as isothermic. Freezing conditions are seldom a problem in the isothermic and isohyperthermic STRs. Higher elevations in tropical areas have isomesic STRs, with mean annual soil temperatures less than 15°C; crop growth is slow, night-time freezing is common, and even cold-tolerant crops such as potatoes seldom grow well where mean annual soil temperature is below 10°C.

Soil-Moisture Regimes of the Tropics

Soil-moisture regimes (SMRs) are defined to classify a soil's ability to supply water to plants without irrigation (Figure 2). In soils where the groundwater table is not reached by the roots of most crop plants, the SMR is determined by the seasonal distribution of rainfall in 'normal' years. Normal years are ±1 standard deviation of long-term means. Most food crops require a reliable supply of water for at least 90 consecutive days. To calculate SMRs, the mean monthly precipitation is compared with the calculated potential evapotranspiration. A soil-water balance is then constructed for the mean rainfall in the area. The duration of time during which water is normally available either from average rainfall or as stored available water in the soil during a period when soil temperatures are warm enough to grow the crop determines the SMR.

The perudic SMR has precipitation that exceeds potential evapotranspiration every month of normal years. Although this may seem desirable, these areas present weed, insect, and disease problems, and the constantly humid conditions make it difficult to harvest mature grain crops.

The udic SMR has fewer than 90 cumulative days when water is not available in the rooting zone in normal years. It is possible to grow food crops any time of the year without irrigation when the temperature is warm enough for that crop. Available water is less reliable during some part of the year, and farmers often select more drought-tolerant crops or may choose not to plant during that period, but perennial plants are adequately supplied with water throughout most years in most of the udic SMR areas.

The ustic SMR is borderline to the common concept of humid. In normal years, soils with an ustic SMR have at least 90 consecutive days when moisture is available, but more than 90 cumulative days when water is not available in the rooting zone. Natural vegetation is either seasonal rain forest or savanna. At least one crop can be reliably grown each year, and it is possible to grow two crops per year on some soils, but there is a seasonal dry period of 90 days or more when crop production is not possible without irrigation. The reliable dry season of the ustic SMR is a distinct advantage for weed and disease control and grain-crop harvest in isohyperthermic and isothermic STRs.

Soils that have a reliable moisture supply for fewer than 90 consecutive days in normal years have an aridic SMR and are excluded from the concept of humid tropics.

The above SMRs are used to classify only the well-drained upland soils. Groundwater often saturates the rooting zone of some soils within all areas of the humid tropics. These soils are identified as having an 'aquic soil moisture condition' and are commonly referred to as poorly drained soils. Soils with aquic soil-moisture conditions most often occur in areas adjacent to rivers and lakes or in broad, level

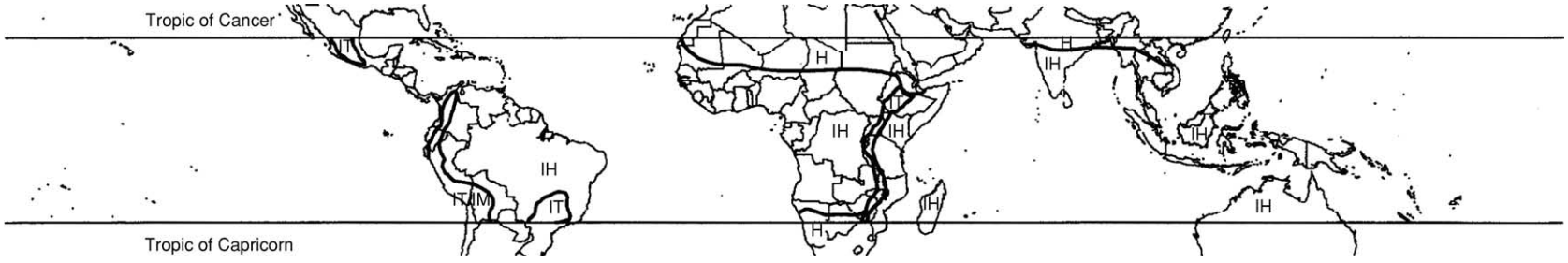


Figure 1 Generalized map of soil-temperature regimes in the tropics. IH, isohyperthermic; IT, isothermic; IT/IM, isohyperthermic and isothermic; H, hyperthermic.

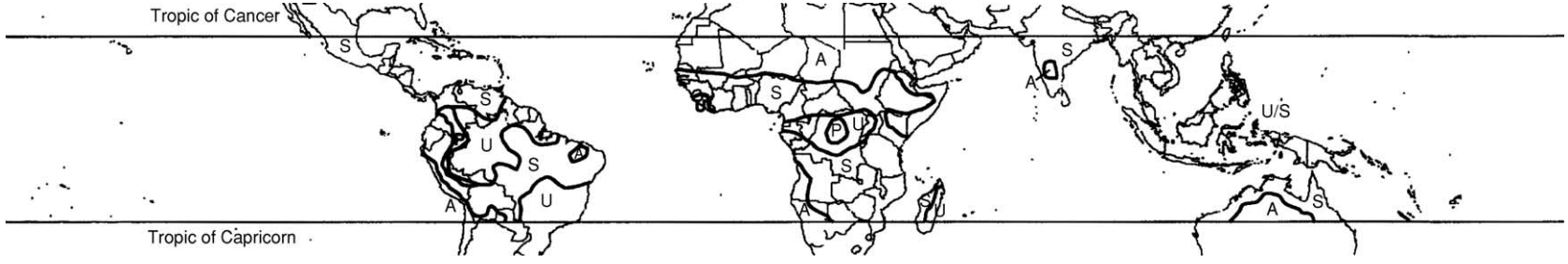


Figure 2 Generalized map of soil-moisture regimes in the tropics. P, perudic; U, udic; S, ustic; U/S, udic and ustic; A, aridic.

landscapes. Soils with aquic soil-moisture conditions are saturated with water only during some part of the year. They are often seasonally utilized for subsistence food crops, but most require engineered drainage systems before commercial agricultural production is possible.

Chemical and Mineralogical Composition of Soils

Of the chemical elements needed for plant and animal physiology, only carbon, oxygen, hydrogen, nitrogen, and to some extent sulfur are derived from air and water. The other essential elements are obtained from the minerals in the soil. Mineralogical properties of soils are derived from the geologic material within which the soil is formed. An inadequate supply of any essential element limits plant growth. The most frequent limitations result from insufficient plant-available nitrogen, phosphorus, potassium, calcium, or magnesium.

Practically no nitrogen is present in soil minerals. Nitrogen enters the soil as ammonium and nitrate dissolved in rainwater or via fixation from the air by nitrogen-fixing microbes. Some nitrogen-fixing microbes in the soil are symbiotic and the nitrogen they extract from the air is incorporated into their legume plant host. Other nitrogen-fixing microbes are not symbiotic, and the nitrogen they extract from the air is incorporated into their cells. Nitrogen is concentrated in organic residues in the surface layers of soil. As organic residues decompose, inorganic forms of nitrogen are released into the soil solution and become available to growing plants, leach into the groundwater during periods of excessive rainfall, or return to the air as nitrogen gas during periods when the soil is saturated with water. Plant-available nitrogen contents in soil are transient and closely related to supplies of organic residue.

Phosphorus is present in only a few minerals. Iron and aluminum phosphates are extremely insoluble and do not release phosphorus rapidly enough for rapid plant growth. The release rate is so slow that soils with high iron and aluminum contents tend to absorb phosphate applied as fertilizer and decrease its availability to plants. Apatite, a more soluble calcium phosphate mineral capable of supplying plant-available phosphorus, is a common source of phosphorus and often present in limestone.

Potassium is present in mica and feldspar minerals. These minerals are rather easily decomposed in the soil environment and consequently are seldom present in materials that have been repeatedly transported and deposited on the land surface.

Calcium and magnesium are most abundant in carbonate minerals associated with limestone and some sandstone. Carbonate minerals are also relatively unstable when subjected to weathering and therefore are present only in recent geologic sediments, limestone and some sandstone.

Soil pH is a measure of the acidity or alkalinity of water in the soil and has a direct effect on how rapidly many of the essential elements are available to growing plants. In the absence of carbonate minerals, soils in the humid tropics are acid in reaction and only limited quantities of essential elements present in the soil are available for plant growth. Acid soils with pH values less than approximately 5.2 also have a concentration of aluminum ions that is toxic to some but not all crop plants. Additions of lime (finely ground calcium and calcium:magnesium carbonates) are desirable and often necessary to reduce or eliminate aluminum toxicity and increase the availability of the essential elements to most crop plants growing in acid soils.

The rate at which essential elements in the soil are available to plants is critical to understanding soil fertility. Plants extract the elements they need from the soil as inorganic ions in the soil solution. The amount of each essential element in the soil that is available to plants changes rapidly as the moisture content of the soil changes and also depends on the rate at which organic compounds decompose to release organically bound elements as available inorganic ions. Less than approximately 1% of the total amount of most essential elements in the soil is present in an available form. Plant species differ greatly in the rate at which they need to acquire essential elements for adequate growth. The rate at which nutrients become available influences natural plant communities and is directly related to human food production. Most human food crops require 90–120 days to mature. Food crops must have a rate of nutrient availability many times faster than required by native ecosystems. A high-yielding grain crop of rice, wheat, or corn must acquire approximately as much phosphorus in 90 days as trees acquire from the same area of land in more than 20 years. In addition, tree roots usually penetrate more deeply and exploit a larger volume of soil than food crops. Therefore, the concentration of available nutrient elements near the soil surface must be considerably greater to supply adequately the needs of a food crop than to support tree growth.

Humans harvest and transport their food crops to a domicile some distance from the site where the crop was grown. Often the seed portion of the plant is consumed and only the less nutrient-rich stems, leaves, and roots of the crop plant are returned to the soil as organic residue. Considerable amounts of

organic residues are required to fertilize a crop plant, because these residues decompose slowly to release inorganic ions for crop growth. The common practice of burning residues facilitates rapid crop growth by releasing organically bound nutrients.

Historically humans have populated areas of soil with high levels of mineral fertility. These are commonly igneous or volcanic materials of basic mineral composition, sedimentary rocks such as limestone rich in calcium, magnesium, and phosphorus, and recent flood plains frequently renewed by depositions of material derived from fertile geologic materials and eroded surface soil. Where the mineral composition of the soils contains only small amounts of essential elements and a large amount of slow-growing natural biomass is present, a system of food production known as slash and burn is practiced. Although some essential elements are volatilized and lost, fire is the primary method of rapidly decomposing organic material and creating a short time period when nutrients contained are rapidly available as inorganic ions. If there is enough biomass, at least one crop can be successfully grown in the 90 days after burning. If correctly done, burning also assures that the surface temperature of the soil becomes high enough to reduce weed competition by killing most weed seeds near the soil surface. A second and third crop is often possible before the available supply of essential elements is exported from the field as human food and the rate of nutrient availability is reduced to a point where crop yields are low and/or weeds become a major problem. After the farmer abandons that land, a succession of native communities that are able to grow with lower rates of nutrient flow from the soil invade the site. After some years, the slow-growing native vegetation acquires enough nutrients in its biomass that it can again be cut, dried, and burned to obtain a site for another brief sequence of crop plants. This method of nutrient-availability management has numerous variations among different indigenous cultures. Only low human population densities are sustained by slash-and-burn agriculture because of the long periods of time (usually between 10 and 30 years and inversely related to the mineral fertility of the soil) that must be allowed for natural vegetation to accumulate sufficient quantities of essential elements needed to fertilize a food crop after burning. Where domestic animals are allowed to graze large areas of native vegetation, essential elements concentrated in their excrements are often collected and used to fertilize small areas of food crops. In areas where infrastructure enables crops to be exported and essential nutrients imported as concentrated fertilizer, continuous food-crop production is practiced on even the most chemically infertile soils. Many combinations and variations of these strategies presently exist throughout the humid tropics.

Mineral and Chemical Grouping of Soils

Broad geographic areas of soils with similar mineral and chemical composition are outlined in Figure 3. Many localized areas of soils that differ significantly from those described in the following discussions are present within each identified area.

Reworked Sediments of Low Fertility

Unlike many areas in temperate latitudes, where soils form from material that has been loosened and deposited by massive glaciations in rather recent geologic times, many areas in the tropics have land surfaces that have been exposed to weathering, soil formation, and erosion for many geologic ages. In these areas the present soils have formed from material almost devoid of minerals that contain essential elements, and soil acidity renders the low quantities present only slightly available to plants. Soils formed from such materials are not only deficient in available forms of the major plant essential elements, but also often deficient in plant-available forms of copper, zinc, boron, and sulfur. The nutrient-poor natural biomass growing on such areas often contains inadequate amounts of essential nutrients for slash-and-burn agriculture. Without the importation of essential nutrients, human habitation is often nomadic and limited to small, isolated sites of more fertile soil.

The watershed uplands between the Amazon and Parana rivers in Brazil are representative of infertile reworked material. The soils form in thick deposits of material that has been weathered through several cycles of erosion and deposition. Quartz, kaolinite, gibbsite, and iron oxides are the predominant minerals. Content of essential elements is very low, and the majority of the soils are Oxisols (Ferralsols) (*Nomenclature of Soil Taxonomy* is used to identify soils and, where applicable, the corresponding units of the *Soil Map of the World* follow in parentheses) and, in very sandy sediments, equally infertile Quartzipsamments (Ferralic Arenosols). The most infertile of these soils, Acrustox (Acric Ferrosols), support sparse native vegetation of grass and low shrubs called 'cerrado' in Brazil. Each year wildfires burn extensive areas during the dry season of the ustic SMR. Attempts to graze the native vegetation often result in nutrient deficiencies in cattle. Extensive areas of these soils are now being utilized for mechanized agriculture with external sources of lime and fertilizer. In addition to wheat, soybeans, and other food crops, sugar cane for ethanol fuel production is grown.

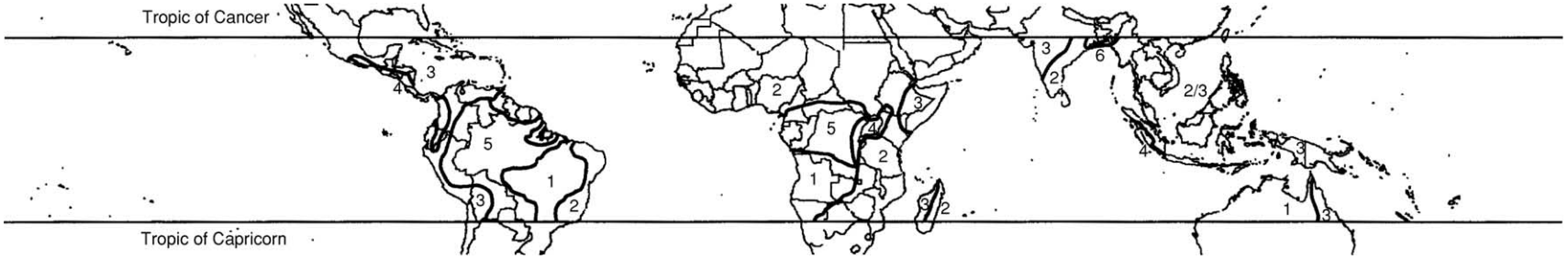


Figure 3 Generalized map of mineral and chemical soil regions in the tropics. 1, low-fertility sediments; 2, acid igneous rock; 3, limestone and other base-rich rock; 4, volcanic materials; 5, major river basins; 6, flood plains.

On inclusions of limestone, basalt, and sediments from these more fertile materials, Eutrustox (Rhodic Ferralsols) are present. These more fertile soils were naturally forested with seasonal deciduous forests that have been removed as farmers seek the more fertile soils for food-crop production.

Acid Igneous Rock

These areas are most extensive in Southeast Asia, parts of the eastern highlands in South America, West Africa, the central highlands of East Africa, and Eastern India. Ultisols (Acrisols) and Inceptisols (Cambisols) formed in these materials can be considered of moderate fertility but most commonly deficient in P and Ca. In most areas there are inclusions of more base-rich sedimentary rock where Alfisols (Luvisols) have formed. Most agriculture is limited to slash and burn unless infrastructure for markets and fertilizer exist.

Limestone and Other Base-Rich Rock

Where tectonic upheavals have exposed areas rich in carbonate and other nutrient-bearing rock, the soils are relatively rich in essential elements. Soils on the Deccan plateau of western India are fertile Vertisols (Vertisols) and Mollisols (Cambisols) formed in base-rich rock. The Andean chain along the west coast of South America and in Central America was formed by tectonic uplift and associated volcanic activity. Some of the uplifted materials are limestone, rich in carbonate, while other portions of the region are granite and schist, relatively poor in essential elements. Tectonic instability, steep slopes, and glaciers at the highest elevations have resulted in rapid erosion of surface materials and deposits of relatively fertile sediments in intermountain valleys of fertile Eutrudepts and Haplustepts (Eutric Cambisols).

Volcanic Material

A small proportion of the soils in humid tropical areas are formed in relatively nutrient-rich material derived from volcanic activity. Volcanic ash and basalt materials of various ages, including some recent volcanic activity, form Andisols (Andosols). Soils derived from volcanic ash and basalt historically have supported intense human settlement. Java, southern Sumatra, parts of the Philippines and other Pacific islands, Central America, the Andean mountains in South America and the Rift Valley in Africa are major areas of volcanic material. Many volcanic areas are mountainous and the higher elevations have isothermic or isomesic STRs where crop growth is reliable but cool temperatures slow crop growth at the higher elevations. Development and maintenance of roads are expensive due to the rugged relief, earthquakes, and volcanic activity. Transportation infrastructure is seldom reliable beyond local markets.

Major River Basins

The major river basins include the Amazon, Orinoco, and Paraguay river basins in South America, and the Congo basin in Africa (Figure 3). The basins are composed of sedimentary materials derived from surrounding uplands and mountains. The fertility of the soils reflects the mineral composition of the upland areas from which the sediments are derived.

The western portion of the Amazon basin and the central portion of the Congo basin have udic and perudic SMRs, while the eastern portion of the Amazon basin has an ustic SMR. Oxisols (Ferralsols) predominate in the Congo basin and eastern portion of the Amazon basin, where the sediments are derived from the Guyana and Brazilian shields. Ultisols (Nitosols) predominate in the western part of the Amazon basin, where the sediments are derived from the Andes and are slightly more nutrient-rich than those to the east. Both Ultisols and Oxisols are nutrient-poor, but inclusions of more nutrient-rich material are present. Alfisols (Luvisols and Eutric Nitosols) are formed in these more base-rich sediments in western Brazil and on the eastern boarders of the Congo basin. Many areas of these more chemically fertile soils are being cleared of forest vegetation for crop production and cattle grazing.

Some of the sediments in the basins are sandy. Soils formed in the sandy sediments are very infertile Quartzipsamments (Arenosols), where the water table is deep, or Spodosols (Podzols), formed where the water table is near the soil surface. It has been estimated that perhaps 10% of the Amazon basin has these kinds of soils. One major area of sandy soils is in the headwaters of the Rio Negro in the northwestern part of the basin, where it merges with the Orinoco watershed. Smaller areas of sandy soils are scattered throughout the basins, often detectable by coffee-colored 'black' waters of the rivers that drain these sandy watersheds.

Floodplains

Soils formed in river floodplain sediments and deltas are often fertile. The fertility is derived from eroded topsoil from the surrounding uplands. Where the mineral composition of the soil is relatively rich in essential elements, people are able to grow and harvest food with little or no attention to chemical fertilization. Crop yields are usually low but reliable. Most soils formed in floodplains are Entisols and Inceptisols, with aquic soil conditions (Gleysols). Despite the

physical dangers from flooding, human settlements are concentrated in these areas, attracted by the soil fertility and access to transportation afforded by the rivers. Road maintenance is extremely difficult where broad floodplains are prevalent. Most agriculture is subsistent unless the flooding hazards and periodic saturation can be controlled. Narrow areas of flood plains are present along coastlines and rivers throughout the humid tropics, but only a few areas are shown in Figure 3.

Organic Deposits

Organic remains accumulate in areas that remain saturated for most of the year, forming organic soils (Histosols). Histosols are present in many floodplain areas but contiguous areas are too small to outline in Figure 3. Cultivation is problematic and drainage systems are needed to aerate the soil and provide stable trafficking.

Human Utilization

The predominant soil limitation in humid tropical regions is low chemical fertility. Major areas of soil are formed from geologic material that contains very limited quantities of life-essential phosphorus, potassium, calcium, and magnesium. The most chemically infertile soils are present as uplands in the interiors of Africa and South America. Most of these soils are physically deep and have reliable moisture for one or two food-crops each year. The most infertile of these soils do not support enough natural biomass to sustain even slash-and-burn subsistence agriculture. Where rapid and reliable infrastructure has been developed, sustained commercial agriculture is now practiced. Reliable markets, fertilizer, lime, and fuel supplies are essential. A sufficient amount of phosphorus must be applied and mixed into the soil to saturate the iron and aluminum oxide surfaces to the extent that sufficient phosphorus becomes available before crops can be grown. Carbonate, in the form of crushed limestone, must be applied to raise the pH of the soil and inactivate the extractable aluminum. Nitrogen fertilizer is needed for nonlegume crops. Potassium is required for high yields, and small amounts of copper, zinc, boron, and molybdenum are needed in many areas. After an initial investment is made in altering the chemical conditions, fertilizer requirements are annually no greater than in other grain-growing soils around the world. Modern soil-testing technology is utilized to determine annual fertilizer formulations and rates. The physical stability of the inert soil minerals, a paucity of river systems that must be bridged, and the gentle topography aid in economical road construction in most parts of the region.

Somewhat less infertile soils with dense forest vegetation can be and are being used for slash-and-burn subsistence agriculture. The rate at which essential nutrients are released from the minerals is too slow to replenish those removed in rapidly growing food crops. The slow-growing forest vegetation that regenerates in abandoned fields can accumulate enough nutrients in its biomass that when cut and burned one or two food crops can be grown only every 10–30 years. These soils have the potential for more intense food crop production if chemically fertilized, which is seldom possible without a market infrastructure.

Human populations in the humid tropics are concentrated on soils formed from materials relatively rich in phosphorus, calcium, magnesium, and potassium. Such soils are primarily located in floodplains and deltas, in areas of volcanic activity, and in upland areas underlaid by limestone, base-rich igneous, and metamorphic rock.

Importing essential elements to compensate for those harvested in human food can increase crop yields from all soils. Applying animal and human wastes wherein the essential elements have been gathered from sites distant from the food-growing area is a partial solution in many cultures. However, the concentration of essential elements in organic residue and waste materials is low, transportation is difficult, and the rate at which the residues decompose to release inorganic ions for crop uptake is slow. Where commercial infrastructure is available, concentrated inorganic fertilizers have proven to be the most economical method of supplying nutrients harvested in human food.

Further Reading

Buol SW and Eswaran H (2000) Oxisols. *Advances in Agronomy* 68: 151–195.

Buol SW and Sanchez PA (1988) Soil characteristics and agronomic practices for sustainable dryland farming. In: Unger PW, Jordan WR, Snead TV, and Jensen RW (eds) *Challenges in Dryland Agriculture: A Global Perspective*. Proceedings of the International Conference on Dryland Farming, Bushland, TX: Amaillo.

Cochrane TT, Sanchez LG, de Azevedo LG, Porras JA, and Garver CL (1985) *Land in Tropical America*. Cali, Colombia: Centro Internacional de Agricultura Tropical (CIAT).

FAO–UNESCO (1971–1981) *Soil Map of the World*, vols I–IX. Paris, France: UNESCO.

Lepsch IF, Buol SW, and Daniels RB (1977) Soils–landscape relationships in the Occidental Plateau of Sao Paulo State, Brazil. I. Geomorphic surfaces and soil mapping units. *Soil Science Society of America Journal* 41: 104–109.

Lopes AS and Cox FR (1977) A survey of the fertility status of surface soils under "cerrado" vegetation in Brazil. *Soil Science Society of America Journal* 41: 742–747.

Lopes AS and Cox FR (1977) Cerrado vegetation in Brazil: an edaphic gradient. *Agronomy Journal* 69: 828–831.

Osher LJ and Buol SW (1998) Relationship of soil properties to parent material and landscape position in eastern Madre de Dios, Peru. *Geoderma* 83: 143–166.

Sanchez PA (1976) Properties and management of soils in the tropics. New York: John Wiley.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agricultural Handbook No. 436. Washington, DC: USDA–NRCS.

Tyler EJ, Buol SW, and Sanchez PA (1978) Genetic association of soils in the Upper Amazon Basin of Peru. *Soil Science Society of America Journal* 42: 771–776.

Van Wambeke A (1981) *Calculated Soil Moisture and Temperature Regimes: South America*. SMSS Technical Monograph No. 2. Ithaca, NY: Cornell University Press.

U

UNSTABLE FLOW

T S Steenhuis, J-Y Parlange, and Y-J Kim, Cornell University, Ithaca, NY, USA
D A DiCarlo, USDA Agricultural Research Service, Oxford, MS, USA
J S Selker, Oregon State University, Corvallis, OR, USA
P A Nektarios, Agricultural University of Athens, Athens, Greece
D A Barry, University of Edinburgh, Edinburgh, UK
F Stagnitti, Deakin University, Warrnambool, VIC, Australia

Introduction

The importance of understanding unstable preferential flow processes cannot be overstated considering its relevance to crop irrigation, groundwater recharge, and vadose-zone transport of nutrients and contaminants. An important aspect of preferential infiltration is the residence time of water in the vadose zone, where contaminants and nutrients are degraded or exchanged in the media. The residence time is shorter for soils with preferential flow than without. Preferential flow can also be caused by macropores. Both types of preferential flow are similar in that the gravity force dominates sorptive processes. This article is concerned mainly with the unstable finger flow formation.

Conditions for Unstable Flow

Although the existence of fingered flow has been known since 1945, the practical importance of instability in stratified soils was not recognized prior to 1972, when Parlange and coworkers demonstrated that fingering can occur in a fine-over coarse-textured profile. In the early years, investigators were concerned with instabilities caused by gradients of temperature or solute concentration in miscible liquids. A few noticed that 'narrow tongues' formed in hydrophobic sands.

Several attempts were made in the 1950s and 1960s by Saffman and other researchers to relate the fingering in petroleum engineering to vadose-zone fingers; but this was problematic, because the first laboratory experiments used a viscous oil, glycerin, which was displaced by compressed air in Hele–Shaw cells. This is not an obvious analog to water displacing air in a stratified medium. As a consequence, viscosity became initially a dominant feature in many theories, although this is now known to be incorrect. This does not mean that some of the equations derived by Hele–Shaw cells cannot be properly reinterpreted. A necessary condition for instability, for water and air in a porous media, neglecting both viscosity and density of air, becomes:

$$K_s > Q \qquad [1]$$

where Q is the water flux imposed by the upper layer of fine material, and K_s is the saturated conductivity of water in the coarse layer underneath. The condition in eqn [1] is obviously necessary, because if Q was greater than K_s, the whole area would have to be saturated to carry the water and no finger would be present. It is a fundamental contribution of Hillel and coworkers to have noticed that eqn [1] is not constraining enough and that water entry in the coarse layer is associated with a 'water entry' suction for the water to penetrate the coarse layer. They further suggest that, upon rewetting, the water entry suction will be higher, resulting in drier fingers, leading to interesting hysteresis phenomena. In all cases, hysteresis is crucial by limiting lateral capillary diffusion of water which would otherwise remove the presence of fingers, as often happens in Hele–Shaw cells.

Finger Diameter and Structure

In early work in Hele–Shaw cells, the optimal width of the fingers was found by balancing the destabilizing effects of gravity and the stabilizing effect of surface tension. Finger width for soils were initially derived 'Green and Ampt soils,' with discontinuous wetting fronts. In the mid-1970s, Parlange and coworkers derived an expression for the finger

width, d, based on the analysis of Richards' equation, i.e., a diffuse front, yielding:

$$d = \pi \frac{S^2}{K(\theta - \theta_i)} \frac{1}{1 - Q/K} \quad [2]$$

where S is the sorptivity given by:

$$S_c^2 = \int_{\theta_i}^{\theta} D\left[\theta + \bar{\theta} - 2\theta_i\right] d\theta \quad [3]$$

where D is the soil-water diffusivity and θ_i is the initial water content, assumed small enough that the soil-water conductivity, K, at θ_i is negligible compared with its value at θ_c. The value of θ_c is a strong function of the initial water content and is not predicted by the theory. Equation [2] is valid for a two-dimensional finger in a slab chamber. In the field, where the fingers are three-dimensional π is replaced by 4.8.

Initially, for a soil initially dry, Parlange and coworkers assumed that θ_i in eqn [2] corresponds to saturation. However, fingers are rarely saturated in soils (as they are in a Hele–Shaw cells) and water content varies along the fingers. As shown in **Figure 1** the finger tip is the wettest (the red color) and it dries behind the finger. The moisture content, θ, varies with depth (measured from the interface between layers) according to the equation:

$$z = \int_{\theta_o}^{\theta} \frac{D d\bar{\theta}}{K - v(\bar{\theta} - \theta_i)} \quad [4]$$

where v is the constant downward speed of the fingers obtained after a short time, i.e., after all mergers have taken place and the fingers have reached a steady configuration. θ_o is the value of θ at $z = 0$ and, if we assume that the maximum value of θ corresponds to a water entry value θ_e, then eqn [4] gives:

$$vt = \int_{\theta_o}^{\theta_e} \frac{D d\theta}{K - v(\theta - \theta_i)} \quad [5]$$

which gives $\theta_o(t)$ when v and θ_e are known. In particular, when $t \to \infty$, θ_o approaches an asymptotic value $\theta_{o\infty}$ with:

$$K(\theta = \theta_{o\infty}) = v(\theta_{o\infty} - \theta_i) \quad [6]$$

When Q/K_c is negligible in eqn [2], we obtain a simpler equation for d:

$$d = \pi S_c^2 / K_c(\theta_c - \theta_i) \quad [7]$$

Note that both S_c^2 and K_c are inversely proportional to the viscosity; accordingly d is independent of viscosity. The influence of viscosity can be felt only through $[1 - Q/K_c]$ in eqn [2] and, thus, is irrelevant when fingering is important and Q/K_c is small.

Finally, for coarse sands, and as long as θ is not too small where the description of the Gardner–Ritsema soil-water conductivity is realistic:

$$K = K_c \exp \alpha(h - h_c) \quad [8]$$

where α is more or less constant and h is the matric potential. Then, eqn [3] shows that for a coarse sand:

$$S^2 \simeq 2(\theta - \theta_i) \frac{K_c}{\alpha} \quad [9]$$

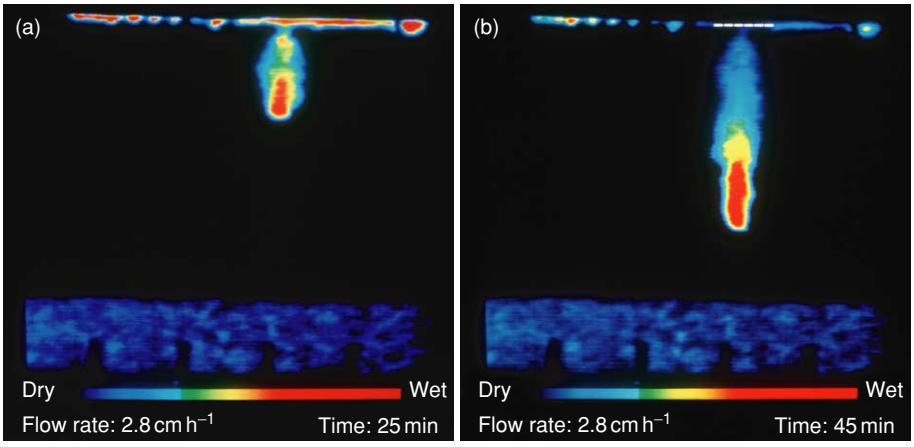


Figure 1 Finger formation in a coarse sand. Note that near the soil surface the distribution zone is visible. This zone carries the water from the rainfall that is uniformly applied at a rate of 2.8 cm h^{-1} to the one finger. Color indicates the moisture contents, with red the closest to saturation. Then the colors for decreasing moisture contents are: yellow, light green, light blue, dark blue, and black: (a) 25 min after water application; (b) 45 min after water application. Reproduced from Nektarios PA, Steenhuis TS, Petronic AM, and Parlange J-Y (1999) Fingered flow in laboratory golf putting greens. *Journal of Turfgrass Management* 3: 53–67.

Hence, we find the remarkably simple result:

$$d \simeq 2\pi/\alpha \qquad [10]$$

i.e., the dependence on θ_o and θ_i has disappeared. This explains why d is essentially constant in time and space and is inversely proportional to Gardner's α.

Note that, for water contents between θ_e and θ_o, all properties are measured on a drying curve of the matric potential. However, as the finger moves downwards within the sand, there is a very narrow zone at the finger tip where the water content increases rapidly, thus operating on a wetting curve, but reliable matric potential data are almost impossible to get in that region. Insight in this region was obtained by measuring reliable water contents down to the pore scale with neutron radiography. A 'dynamic' pressure, $h(\theta)$, was calculated based on Darcy's law by measuring the velocity of the finger, the distributed moisture content, and the unsaturated conductivity. The shape of the curve is similar to that which can be derived from earlier fingered flow experiments depicted in Figure 2 and is very much different from expected soil-water pressure relationships for the same soil. Although these effects are sometimes attributed to nonequilibrium effects, applying continuum results to the pore scale is problematic.

In fingered flow experiments where initial moisture content is varied, the finger size initially decreases when the moisture content increases from 0 to 0.005 $cm^3\,cm^{-3}$ and then later increases with increasing moisture content when it reaches the width of the chamber at 0.04 $cm^3\,cm^{-3}$ (Figure 3; note that the

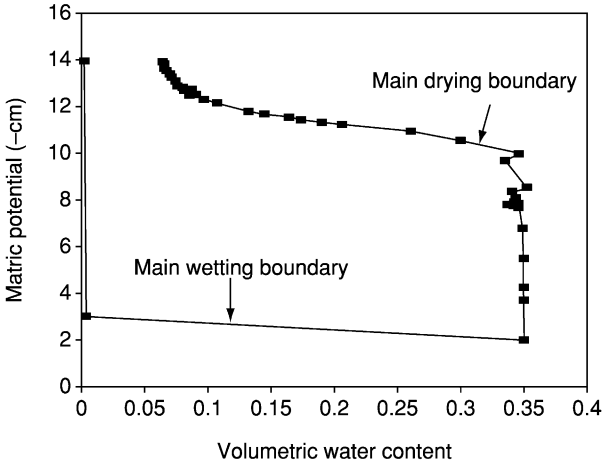


Figure 2 Nonequilibrium soil moisture curve obtained from fingered flow experiments. *h*, matric potential. Reproduced from Lin, Y, Steenhuis TS, and Parlange J-Y (1994) Closed-firm solution for finger width in sandy soils at different water contents. *Water Resource Research* 30: 951.

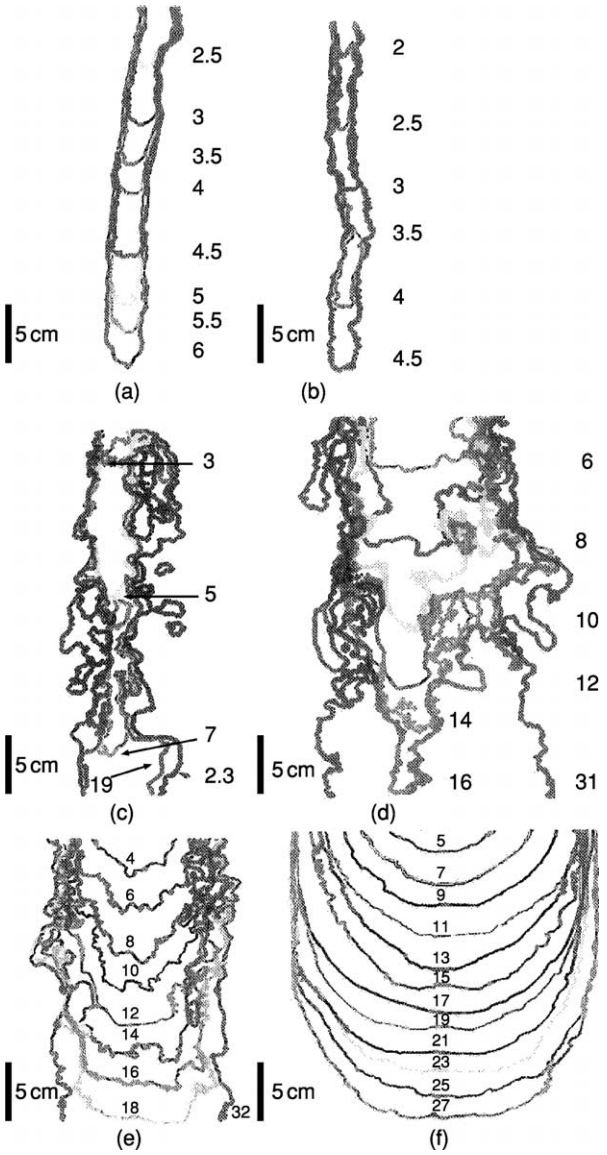


Figure 3 (a) Tracings of the totally dry sand. The sequence of tracings were taken every 30 s. The numbers accompanying the tracings are the minutes after infiltration started. A scale indicates 5 cm; (b) the advancing wetting front in the 0.01 $cm^3\,cm^{-3}$ moist sand. The infiltration front has a smaller width than in the totally dry case; (c) the advancing wetting front in the 0.02 $cm^3\,cm^{-3}$ initial moist sand pack. The tracings are depicted every 2 min. In addition, the latest tracing (after 23 min of infiltration) is shown; (d) advancing wetting fronts in the 0.03 $cm^3\,cm^{-3}$ moist sand. Two-minute tracings are shown until the front reached the bottom of the visible area. In addition, the tracing of the front is depicted when the infiltration was stopped (after 31 min); (e) advancing wetting fronts in the 0.04 $cm^3\,cm^{-3}$ initial moisture case. Two-minute tracings are shown until the front reached the bottom of the visible area. In addition, the front is depicted when the infiltration was stopped (after 32 min.); (f) the tracings of the wetting front when the chamber was imbibed from the bottom, or with a 0.04 $cm^3\,cm^{-3}$ (measured) initial moisture content. Tracings are depicted every 2 min. Reproduced with permission from Bauters TWJ, Dicarlo DA, Steenhuis TS, and Parlange J-Y (2000) Soil water content dependent wetting front characteristics in sands. *Journal of Hydrology* 231–232: 244–254.

surface induction zone is not shown). As θ_i increases, θ_c decreases rapidly and Q/K_c is not negligible in eqn [2] any more. Eqn [10] does not hold and d increases rapidly with θ_i.

The theory above applies only for a liquid, e.g., water or oil, and a gas, e.g., air. To extend the formula to the case of oil and water, it is necessary to obtain the dependence of the front velocity on its curvature. To obtain the dependence of the front velocity on curvature, the flow in the narrow diffuse region, ahead of it, must be analyzed, and to do so we must know the flow further ahead (here the pure oil). In the case of a gas, this presents no problem as it is assumed that the air can move freely ahead and does not affect the flow of water. Usually this will not be the case when oil is displaced by water, although it represents a limiting case.

The discussion has been limited, so far, to the movement of liquids. To include the transport of solutes, it is necessary to recall the physical configuration of the flow paths (**Figure 4**, where for clarity **Figure 1** has been redrawn schematically). The soil near the surface is more or less uniformly wet. This is the distribution zone in which water and solutes are funneled into the fingers of the conveyance zone below. The thickness of the distribution zone is that of the fine layer for layered soils and is of the order of the finger size for soils without such a layer. In the conveyance zone, all fingers move down with a velocity v. The number of fingers depends on the flow rate. For example, in **Figure 1** the flow rate is relatively low and only one finger is formed. In other experiments, where the flow rate is higher, several fingers form and, as a consequence, are spaced closer.

For the case when the initial concentration in the distribution zone is C_o, and the rainfall is solute-free, the concentration in the percolating water out of the distribution zone can be described as similar to a linear reservoir:

$$C = C_o \exp(-\lambda t) \quad [11]$$

where λ is the coefficient equal to q/w, q is the steady-state flow rate, w is the apparent water content of the distribution zone and equals $L(\rho k_d + \theta_s)$, d is the depth of the distribution zone, θ_s is the saturated moisture content, ρ is the bulk density of the soil, and k_d is the desorption partition coefficient. In case water is added with a solute concentration, C_o, the concentration in the water leaving the distribution zone is:

$$C = C_o(1 - \exp(-\lambda t)) \quad [12]$$

Equations [11] and [12] are equivalent to those used for sludge by the US Environmental Protection Agency in predicting the loss of metals from the incorporation zone.

It is reasonable to assume that the transport in the preferential flow paths of the conveyance zone can be described with the convective–dispersive equation, viz:

$$D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} = \frac{\partial C}{\partial t} \quad [13]$$

where D is the dispersion coefficient, and v is the velocity of the solute and equals approximately $q/(\beta(\rho k_d + \theta))$, where β is the wetted fraction in the conveyance zone by fingers and θ is the moisture content. Using Laplace transforms, eqn [13] can be integrated subject to the boundary condition described in eqn [11], with no solutes initially present in the column, and for $4D\lambda/v^2 < 1$ as:

$$C = \frac{1}{2}C_o \exp(-\lambda t)\left[\exp\left\{\frac{vx}{2D}(1-a)\right\}\operatorname{erfc}\left(\frac{x - vta}{2\sqrt{Dt}}\right) + \exp\left\{\frac{vx}{2D}(1+a)\right\}\operatorname{erfc}\left(\frac{x + vta}{2\sqrt{Dt}}\right)\right] \quad [14]$$

where

$$a = \sqrt{1 - \frac{4D\lambda}{v^2}}$$

The last term can usually be neglected when x or t is sufficiently large, i.e., $(x + vta)/(4Dt)_{1/2} > 3$. For the boundary condition (eqn [12]), we can find the solution by superposition.

Note that in eqn [14], as discussed above, the velocity is a function of the soil properties and not of the flow rate. The number of fingers in the conveyance zone adjusts itself to facilitate the different flow rates. This has been demonstrated with a fingered

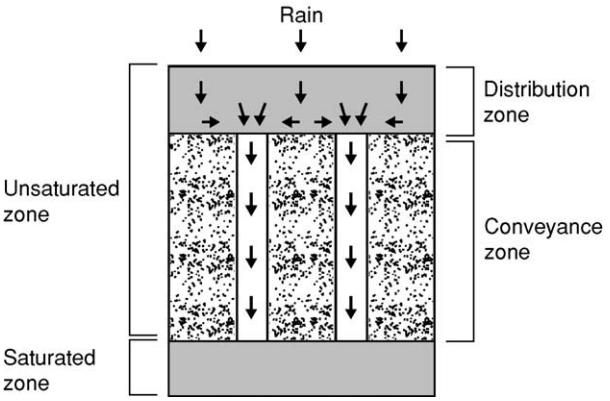


Figure 4 Schematic of the fingered flow process in the soil with preferential flow paths.

flow experiment that was carried out in the laboratory with 50-cm-long, 14-cm-diameter duplicate columns filled with coarse sand subjected to different intensities of steady-state rainfall. Chloride was applied at the surface as a pulse during each steady-state rainfall using different columns for different flow rates. The results are given as a function of cumulative flow (Figure 5a) rate and time (Figure 5b). In Figure 5a, the initial breakthrough of chloride takes more water for the high flow rate of 2 cm h^{-1} than for the low flow rate (0.3 cm h^{-1}). This is, obviously, a direct consequence of having a greater portion of the column wetted for the high flow rate. In Figure 5b, when chloride concentrations are plotted as a function of time, initial breakthrough for all three flow rates is approximately at the same time, clearly establishing that water and solute velocity is the same for the different flow rates. When the results are fitted to eqn [14], the solute velocity is, in all cases, approximately 35 cm h^{-1}, with no clear trend among the different flow rates, as expected.

In conclusion, it has become obvious in the last 25 years that, for soils in the field, preferential flow is the rule rather than the exception. Fingered flow is one of these types of preferential flow and can often be found above the primary coastal aquifers that are overlain by coarse-grained soils. Recognizing that these fingered flow paths can rapidly carry chemicals down to groundwater is a first step in keeping our aquifers clean for generations to come.

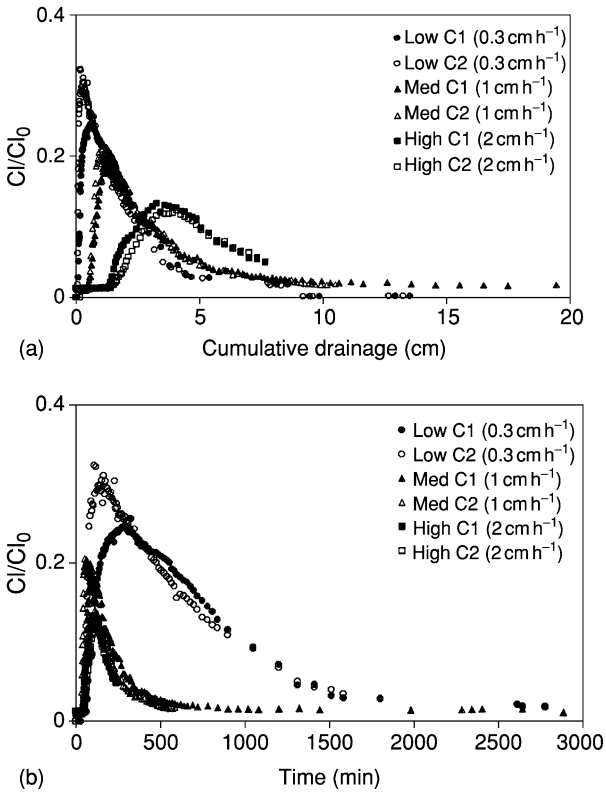


Figure 5 Chloride breakthrough curves (CBTs) for sand columns with a pulse application for three different flow rates of 0.3, 1, and 2 cm h^{-1} plotted: (a) as a function of cumulative drainage after application of the chloride pulse; (b) as a function of time after application of the chloride pulse. The experiment was carried out in duplicate with C1 and C2 indicating the two columns.

Further Reading

Bauters TWJ, DiCarlo DA, Steenhuis TS, and Parlange J-Y (2000) Soil water content dependent wetting front characteristics in sands. *Journal of Hydrology* 231–232: 244–254.

DeBano LF and Dekker LW (2000) Water repellency bibliography. *Journal of Hydrology* 231–232: 409–432.

Glass RJ, Steenhuis TS, and Parlange J-Y (1989) Mechanism for finger persistence in homogeneous, unsaturated, porous media: theory and verification. *Soil Science* 148: 60–70.

Hill DE and Parlange J-Y (1972) Wetting front instability in homogeneous soils. *Soil Science Society of America Proceedings* 36: 697–702.

Hillel D (1987) Unstable flow in layered soils: a review. *Hydrological Processes* 1: 143–147.

Parlange J-Y and Hill DE (1976) Theoretical analysis of wetting front instability in soils. *Soil Science* 122: 236–239.

Philip JR (1975) Stability analysis of infiltration. *Soil Science Society of America Proceedings* 39: 1042–1049.

Raats PAC (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Proceedings* 37: 681–685.

Saffman PG and Taylor GI (1958) The penetration of a fluid into a porous medium or Hele–Shaw cell containing a more viscous liquid. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 245: 312–331.

Selker J, Leclerq P, Parlange J-Y, and Steenhuis TS (1992a) Fingered flow in two dimensions. 1. Measurement of matric potential. *Water Resources Research* 28: 2513–2521.

Selker J, Parlange J-Y, and Steenhuis TS (1992b) Fingered flow in two dimensions. 2. Predicting finger moisture profile. *Water Resources Research* 28: 2523–2528.

Steenhuis TS, Boll J, Shalit G, and Merwin IA (1994) A simple equation for predicting preferential flow solute concentration. *Journal of Environmental Quality* 23: 1058–1064.

URBAN SOILS

J L Morel, C Schwartz, and L Florentin, Laboratoire Sols et Environnement ENSAIA-INPL/INRA, France
C de Kimpe, Agriculture and Agri-Food Canada, Ottawa, ON, Canada

Introduction

Soils in urban and suburban areas are transformed by human activities. For several decades, soil surveys and research were largely focused on agricultural and forest lands, and intensively managed and disturbed soils were not much investigated as revealed by the white (nonmapped) areas representing cities on most soil maps. Urban soils are used for many purposes, including urban and industrial activities, forestry, and agriculture. They are characterized by a strong spatial heterogeneity resulting from the various inputs of exogenous materials and the mixing of original soil material. The basic functions of natural and extensively modified soils are essentially the same. The evolution of urban soils is controlled by the same factors as natural soils, but the human factor imposes extremely rapid transformation cycles in comparison with those dominant under natural conditions. They often hold pollutants that may be a threat to human health. Anthropogenic soils can be investigated at least in part with traditional soil survey approaches; however these methods must be properly adapted and new methodology must still be developed. It is only through a multidisciplinary approach that urban soils will be better understood and their use optimized to protect human health and the quality of natural resources, e.g., preventing groundwater contamination.

Definition

'Urban soils' is a class of Anthropic soils, a term already used in several classification systems. Urban soils are soils extensively influenced by human activities, found mostly but not only in urban areas. They include: (1) soils that are composed of a mixture of materials differing from those in adjacent agricultural or forest areas, and that may present a surface layer greater than 50 cm, highly transformed by human activity through mixing, importing, and exporting material, and by contamination; (2) soils in parks and gardens that are closer to agricultural soils but offer different composition, use, and management than agricultural soils; and (3) soils that result from various construction activities in urban areas and that are often sealed. According to this definition, urban soils are essentially under strong human influence in urban and suburban environments; they may exert a strong effect on human health, on plants and soil organisms, and on water infiltration. They are differentiated from other strongly influenced soils such as those found in quarries, mines, and mine tailings, and airfields away from cities. However, it is sometimes difficult to set a clear boundary between urban soils and agricultural soils (Figure 1).

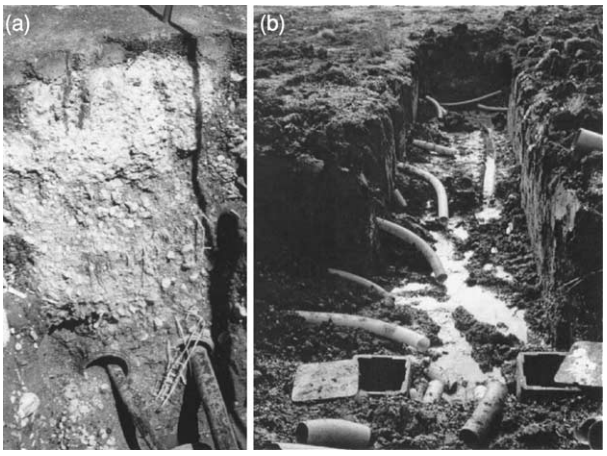


Figure 1 Urban soils and agricultural soils: (a) urban soil profile showing a sealed surface, imported soil materials, and electric wires; (b) agricultural soil profile showing strong perturbation induced by the incorporation of plastic drains.

Use of Soils in Urban Areas

In cities, soils provide support for infrastructure (buildings, roads, railways, parking lots, bridges), shelter for cables (electricity, telephone, television) and pipes of various size and composition (drinking water, wastewater, gas), and substrate for plants (isolated trees along streets, trees in public parks, and ornamental and edible plants in public and private gardens). They are also used for agricultural (horticulture, suburban agriculture, gardening), and industrial (buildings, mining, industrial waste disposal) production activities, and for recreation (stadiums, playgrounds). For centuries, the regions surrounding residential areas have been used for provision of construction materials and domestic waste disposal.

There are numerous soil types in urban areas as a result of these various uses, which differ in the intensity and duration of human impact. Human impact on urban soils may be light (e.g., urban forests) or, in contrast, maximal in developed built

areas (e.g., sealed and artificial soils). Therefore, soils in a city differ according to the degree of human transformation. In general, artificiality increases from the periphery to the center of the city, where original soils are often removed and replaced by anthropogenic materials. Older cities show the most modified soils, as they are generally constructed on their own waste materials that have accumulated over the years. For example, cities like Paris and Moscow are built on several meters of anthropogenic materials that hold remains of former human activities and materials.

Another main feature of urban soils is the high frequency of usage change with time. An example is the conversion of former industrial sites (now termed 'brownfield' sites) to new activities, including residential, public, and recreational activities. Such intensive changes are often dramatic and may cause problems for residents as a result of possible soil contamination. Landfills located at the periphery of cities are, knowingly or not, frequently converted to residential development as a result of the pressure of increasing urban population. This may result in increasing risks for residents from contamination through direct contact, inhalation, and/or consumption of garden produce.

Evolution of Soils in Urban Areas

As known from traditional soil science, a soil is the outcome of the evolution of a parent mineral material under the influence of climate, vegetation, topography, and time. Soils generally display a sequence of layers with an organically enriched surface horizon. Soil formation is generally a slow process that involves typically three main processes: weathering, transport, and accumulation. Differences in climatic conditions, parent rocks, and topography are at the origin of a wide range of soil types developed under natural conditions and, in general, the soil type can be deduced from knowledge of the conditions that have controlled its formation.

The three fundamental processes of formation (climate, parent rock, topography) and evolution of soils also apply to urban soils. However, urban soil formation is strongly influenced by the 'human factor' (Figure 2), which often creates a new 'parent rock' (e.g., debris from former buildings or industrial wastes), and modifies the conditions of its evolution through its influence on water circulation and composition in urban areas. The influence of the 'human factor' can also be interpreted in terms of soil formation and evolution: (1) weathering: transformation of the original material by mixing, compaction, or aeration of material layers; (2) transport: excavation of soil layers, leading to the partial or total elimination

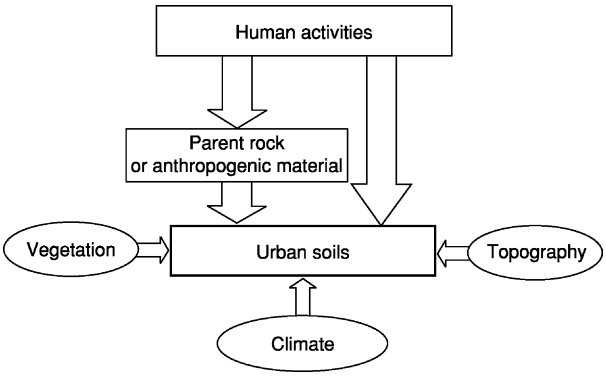


Figure 2 Urban soil formation and evolution.

of the original soil profile; and (3) accumulation: addition of exogenous materials from various origins (soil materials, minerals, technological compounds, and inert, organic, or toxic wastes). The kinetics of these processes are very rapid in comparison to natural processes, as a result of the increasing use of modern equipment, like tractors and bulldozers. A week or even a day is often the time scale required to modify completely the urban landscape. Heterogeneity in the landscape and in the soil may therefore change rapidly. Digging for new buildings, bringing in material from large distances for landscaping, disposing of rubbish, debris and topsoil for leveling and preparing the land for a new use often prevent any relation with the original parent material in the vicinity of the site being examined. Under such conditions, the natural weathering mechanism of soil formation does not play a great role if any, whereas transport and accumulation are predominant.

Another main feature of urban soils is the construction of barriers, e.g., sidewalk, which notably reduces the infiltration of water in soil, and increases its transport to streams, often causing flooding problems. Oxidizing conditions generally dominate in urban soils as a result of lack of water in the soil profile. However, soil compaction and leakage of drinking or wastewater from pipes may induce locally strong changes in redox and water flux conditions.

Composition and Heterogeneity of Urban Soils

The composition of soils in the cities depends largely on the nature of the materials in which they are developed. Coarse textures, with mostly sand and coarser material (rubble and gravel), generate great horizontal as well as vertical heterogeneity. This is a frequent characteristic of urban soils. Bulk density is generally low (<0.5) or very high (>1.60) as a function of the parent material. As an example, Figure 3 shows a soil

Figure 3 Soil profile in a twelfth-century urban area (Nancy, France).

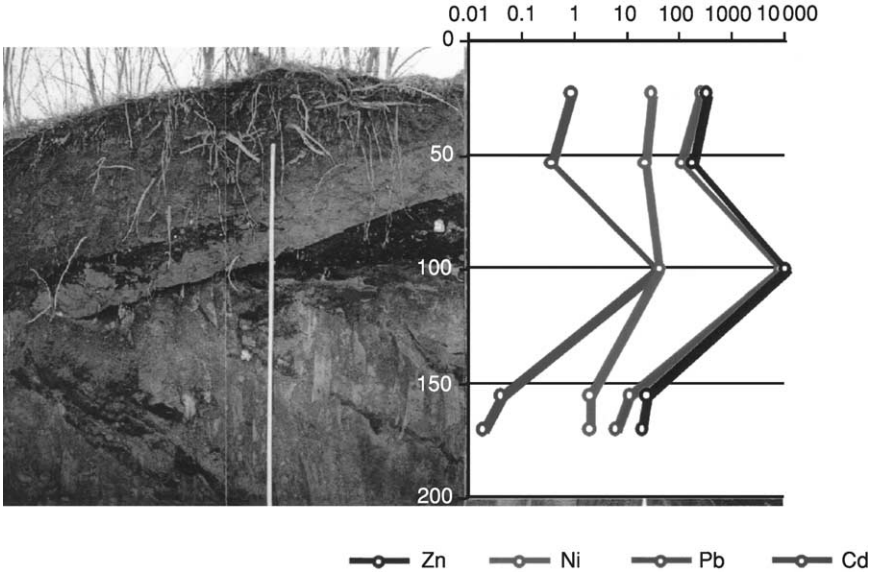


Figure 4 Soil profile and metal concentrations (mg kg^{-1}) in a former industrial site.

profile observed from a pit excavated for building construction in the city of Nancy, France. The soil exhibited a depth of more than 7 m and consisted of six distinct horizons. The three horizons at the bottom were the relic of a former agricultural soil. It presented a clear Ap horizon (15 cm) colored by organic matter, very compacted and weakly calcareous as a result of former agricultural practices or addition of calcareous construction material. The second horizon was not calcareous and offered a higher porosity with galleries of earthworms. The three upper horizons (0 to 1 m) were successive depositions of raw material. They were very heterogeneous and made of earthy material of sandy clay texture dominated by more than 80% of fine material. They were rich in organic matter and presented a brownish color. Numerous tree roots were present. The second horizon was strongly compacted and the third was characterized by the presence of building material of lighter density with fragments of dark gray schists coated with clayey sands.

Another example is given by a soil developed on a former industrial site (Figure 4). With the exception of the two upper horizons made of agricultural soil material brought in to facilitate the revegetation of the site, most of the material was of anthropogenic

origin, deposited during industrial activity and mixed and compacted at the closing of the factory. Many types of debris are present, including concrete, metal wires, slag, and industrial sand, each offering specific physical (density, porosity, mechanical resistance) and chemical properties (nutrients, metals, organic pollutants). In general, the analysis shows the high pH that is a common feature in anthropic soils, often as the result of alkalizing products mixed in the soil. Organic matter content may be high, and the C/N ratio is generally high because of the presence of organic material low in N, especially in industrial soil material; these high values are often due to contamination with oily wastes.

Soil Contamination in Urban Areas

Because of their diversified origin, urban soils may contain pollutants, the location, characteristics (including availability), and potential evolution of which must be established properly for future land use. Urban and suburban soils prompt a strong interest because of the growing public concern about the environment and human health. For example, urban horticulture provides a nonnegligible percentage of the food supply to large populations, especially in developing countries, and there is a need to improve this type of production while ensuring its safety, as well as addressing the issues related to a wide range of land uses. Also, in urban areas children are often in direct contact with soil material, and soil quality (pollutants, pathogens) in playgrounds may affect their health. Some urban soils, in particular those observed on brownfields where they were previously used as support for industrial production, may contain large amounts of mineral and organic pollutants that accumulated over time. The profile of the industrial site described above contains elevated amounts of some metals (Cd, Pb, and Zn), especially in the horizons made of industrial material, e.g., slag. Also organic contaminants (e.g., hydrocarbons, polychlorobiphenyls, herbicides) are commonly found in urban soils; polycyclic aromatic hydrocarbons (PAHs), a special class of hydrocarbons, some of which exhibit carcinogenic properties, are found in some former industrial sites (e.g., gasworks, coking plants). The soil shown in Figure 4 was also contaminated by lead and zinc (Pb and Zn), two metals widely present in urban areas (Table 1). An urban soil profile studied in the city of Nancy showed the influence of the material in place or brought in. These materials contributed to modifying the content of metals in the profile. The metal concentrations in the urban soil were lower than in the industrial site, but metals were present at significant concentrations in the upper horizon as a result of the various urban activities, including traffic and water runoff from zinc roofs.

Table 1 Concentration of metals in the profile of an urban soil (see Figure 3)

	Cd	Cu	Cr	Hg	Ni	Pb	Zn
Depth (cm)	($mg\,kg^{-1}$)						
20	0.67	66.2	65.0	2.00	28.3	415.1	325.7
175	0.08	11.8	39.5	0.04	24.5	28.4	50.4
199	0.75	41.0	59.7	0.06	50.0	21.4	119.3
215	0.26	20.2	44.7	0.04	28.6	26.6	70.4
294	0.11	15.0	54.0	0.02	32.6	24.3	63.4
725	0.03	5.8	25.7	0.02	13.4	26.0	32.1

Urban soils have different retention capacities for organic components, but their alkaline reaction often limits the mobility of heavy metals. In most urban soils, a pH shift to alkalinity constitutes an alkaline (carbonate) geochemical barrier in the topsoil. This barrier hinders the mobilization of heavy metals. Methods developed for agricultural soils (such as plant tests, selective extraction, microbiological and enzymatic tests) may be adapted for assessing the risk of transfer of pollutants to the food chain.

Garden soils

Gardens are a place of strong interactions between soils and human activities. In general, a very intensive form of agriculture is conducted, resulting in soils with high fertility and great diversity, deriving from the multiplicity of gardening practices. Soil quality (i.e., nutrient and pollutant content) in gardens is related to the quality of the parent materials, but through various inputs and modifications of the soil profile, the gardener is the most important factor in soil quality. Inputs range from traditional agricultural amendments and chemicals (manure, lime, pesticides, fertilizers) to domestic wastes and industrial amendments that may contain several inorganic and organic contaminants. In general, rates of amendment application are far higher than in traditional agriculture production, and garden soils tend to exhibit a deep upper horizon with a high concentration of organic matter and mineral nutrients, e.g., nitrogen and available phosphorus. In general, the organic matter content is directly related to the age of the garden. Therefore, the soil quality, hence fertility, of garden soils is as variable as gardening practices. This may be reflected by the biomass production of lettuce grown on various soil samples collected from a set of gardens (Figure 5). But pollutants may also accumulate in garden soils and be transferred to the food chain by direct consumption of vegetables.

Heavy metals tend to accumulate in garden soils and, in general, their concentration is on average twice that in agricultural soils, probably because of the input of various amendments to the garden soils (Figure 6). As for fertility, contamination is highly variable from one garden to another. This can be observed from the analysis of a set of soil samples collected from family gardens located in the Lorraine region (France) and in the nearby Saarland region (Germany) that showed a wide range of values, probably due to gardener practices and the proximity of industries. Soil physical properties and accumulation of heavy metals in the Ap horizon depend on parent material, substrates, and anthropogenic modifications. In Saarbrücken, only 18% of the garden soils were natural soils. Also metal concentration correlated well with soil fertility, as expressed by the total and available phosphorus contents (Figure 7). Finally, the content of heavy metals can be explained in decreasing order of importance by the age of the garden, previous use of the garden (including practices of the gardeners), natural concentration of metals in the soils, and atmospheric deposition.

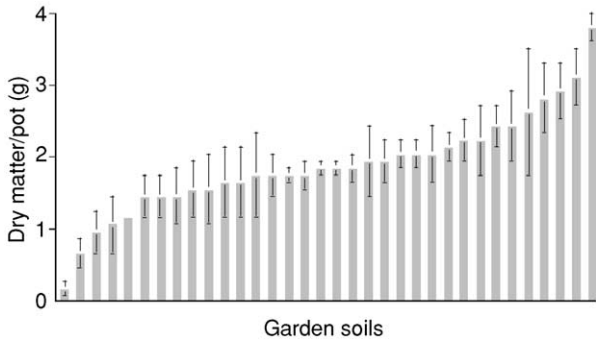


Figure 5 Biomass production of lettuce (*Lactuca sativa*) grown on a set of soil samples collected in family gardens in the Lorraine region (France).

Functions of Urban Soils

The link between soil functions and various soil properties such as substrate, texture, and humus content for soils at urban, industrial, and mining sites can be made using a set of indicators of soil quality such as those developed in Germany: rooting depth, wetting and aeration, nutrient status, and acid neutralization capacity. The high degree of surface sealing in urban soils limits the water partitioning that normally exists under natural conditions. Reduced seepage towards the subsurface causes high surface runoff and floods. Attempts are made to reduce this effect by increasing rainwater infiltration and groundwater recharge, and by reducing discharge by increasing water retention. Water and air supply in urban soils, an important feature for plant and tree growth, are controlled by bulk density, amount of medium, coarse, and very coarse mottles, organic matter content, and texture expressed as contents of clay, silt, and sand.

Urban soils are characterized by a great ecological heterogeneity, and show special distinctness of vegetation and fauna. They are habitats for plants and soil organisms, and for their filtering, buffering, and transforming of organic and inorganic pollutants. The root depth is, however, often limited due to abrupt horizon transitions, especially in the presence of a large percentage of coarse material (>2 mm). As a medium for plant growth, urban soil supports a large population of amenity vegetation in diversified habitats, including parks, gardens, roadsides, and turf areas. Much of the urban vegetation is cultivated, but there are also relics from natural vegetation or spontaneous infestation by opportunistic species.

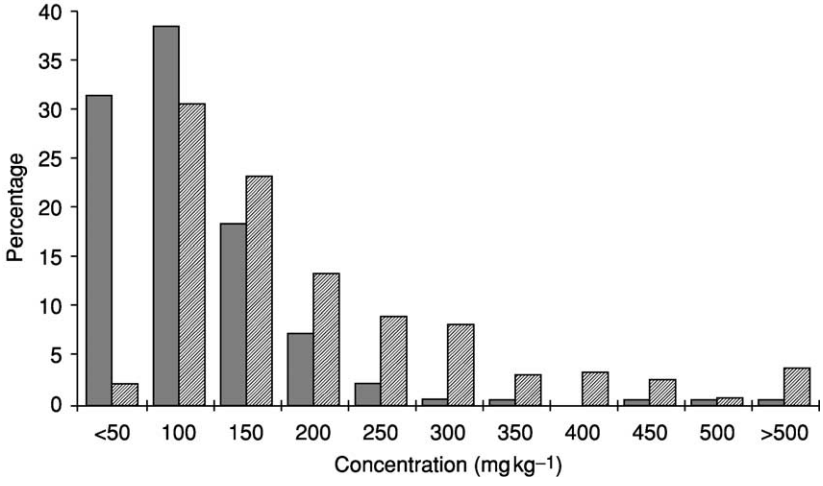


Figure 6 Frequency distribution of copper concentration in garden soils. ■ Regional soils ($n = 185$); ▨ garden soils ($n = 233$).

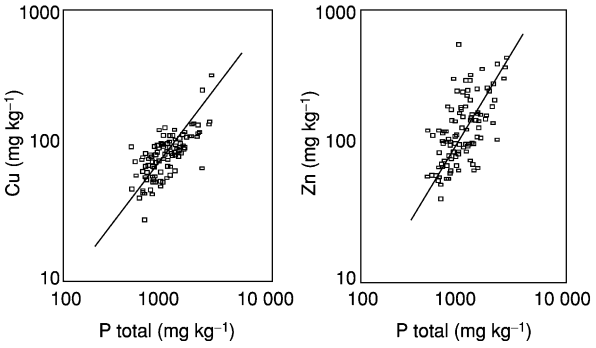


Figure 7 Relation between phosphorus and zinc and copper contents in garden soils.

The habitats may create stresses to vegetation survival: whereas normal root spread requires a circular soil disk preferably with a diameter equivalent to the crown spread and one meter deep, most urban tree roots are closeted in a narrow and shallow strip of substandard soil. In urbanized areas, roots of isolated trees are often sequestered in a restricted space; lateral room space is hindered, with a narrow strip, often less than 2 m, of poor-quality soil sandwiched between building foundations and highly compacted road material. The subsurface material quality is also often ignored. The presence of mortar, concrete, and asphalt attests to the widespread contamination of sites by construction debris that poses a physical hindrance to root development. Also the volume for water and nutrient storage is largely diminished.

Pedology and Archeology

Contrary to the situation in natural and slightly modified soils, e.g., agricultural or forest soils that most often display distinguishable horizons, a major challenge in urban soils, not only for their characterization, but also for determining their potential uses, is the heterogeneity of the layers associated with the fact that the latter are not always horizontal (Figure 3). Heterogeneity can also be very large within a single layer. This prompts a number of questions: what is an urban soil profile? What is an urban pedon? What area, both horizontally and vertically, is required for its description? What volume of soil is necessary for its characterization? How to describe the texture of urban soil, in particular the material coarser than sand? The presence of rubble may indeed confer special properties, and prevent some future uses of the soil.

Concepts must be properly developed in order to develop a suitable and useful classification scheme. Soil description involves the characterization of the successive layers with a specific composition. Urban soils are thus polycyclic soils composed by

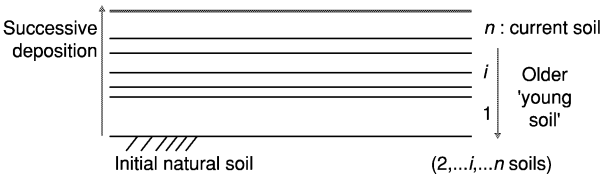


Figure 8 Urban soils in old cities are developed on waste material.

the superposition of several young soils (Figure 8). Using this approach, and as long as the human factor is clearly established, an urban soil survey is not strikingly different from a traditional soil survey: urban soils are the ultimate members of a continuum characterized by increasing human influence, and ranging from soils slightly or not affected by human influence (forests, rangeland) to agricultural lands and to urban soils where the human imprint is maximum, and where most natural features have often disappeared. For example, during early city development in previous centuries, when large machinery and transportation facilities were not available, rubble and debris were often disposed of at the site or at a short distance: cities were actually built on their own wastes. Therefore, the soils may contain layers or strata reflecting the processes that ruled the accumulation and such layers are the 'memory' of past activities: each layer corresponds to sedimentary materials representing an urban soil, generally poorly developed but significantly influenced by people.

Soils contain a large array of historical information, which has been proved very useful in understanding ecological and anthropological evolution. The materials brought about by human activities may also undergo pedological evolution leading to their significant transformation. In general, archeologists and pedologists have had only a few interactions in the past. Combining pedology and archeology by sharing research tools will increase our understanding of human evolution and help to predict better the problems connected to urban soils. Each layer of an urban soil contains products that are characteristics of the technological evolution of human societies. Combining archeology and pedology enables prediction of the location of some pollutants, such as heavy metals (Pb, Cu, Zn) and their fate in the profile.

Management of Urban Soils

The world is in the midst of a massive urban transition unlike that of any other time in history. In 1975, approximately one-third of the world's people lived in urban areas. By 2025, the proportion will have risen to almost two-thirds, which corresponds to more

than 5 billion people, mostly in developing countries. In developed countries, the population shift involves migration away from concentrated urban zones to large, sprawling metropolitan regions. Since the beginning of the twentieth century, the French population has increased considerably in cities, and it amounts to up to 70% of the population. In these large cities, as well as in smaller ones, green and open spaces play an important role as recreation areas, and they are the 'lungs' of urbanites. In the above context, characterizing the land that provides the infrastructure for these growing cities becomes particularly important. It is also a challenge: industrial and service businesses that are often located at the outskirts of the cities or in the suburbs near workers and employees are not particularly welcome amidst the new urban developments, and they are forced to move to a greater distance. This is favoring the emergence of brownfield sites, lands and buildings contaminated by previous industrial activities. Such areas are then recycled and often used for other human activities, that may present various risks for health. Waste disposal, especially the accumulation in landfills outside cities, creates similar concerns.

Within urban and suburban areas, soils are used as physical support for diverse anthropocentric activities and needs: infrastructure both at the soil surface and underground (e.g., roads, parking lots, railroads, pipes, sewers, ducts) and for buildings, sources (e.g., topsoil, landfill material from excavation) and sinks (e.g., dredged sediments, compost of organic urban wastes) of raw materials, food production, recreational activities (e.g., parks, landscaping), and memory and cultural heritage (e.g., cemeteries). Soil scientists have the expertise, as well as the social responsibility, to address issues related to these uses.

See also: **Applications of Soils Data**; **Land-Use Classification**

Further Reading

Arbeitskreis Stadtböden (1988) Substrate und Substratmerkmale von Böden der Stadt- und Industriegebiete. *Mitteilungen der Deutschen Bodenkundlichen Gesellschaft* 56: 311–316.

Bullock P and Gregory PJ (1991) *Soils in the Urban Environment*. Oxford: Blackwell Scientific Publications.

Burghardt W (1994) Soils in urban and industrial environments. *Zeitschrift für Pflanzenernährug und Bodenkunde* 157: 205–214.

Burghardt W, Zuzok A, and Heinen P (1987) Untersuchungen zur Kennzeichnung der Anreicherung und Verteilung von Schwermetallen in urbanen Böden. *Landschaft + Stadt* 30–38.

Craul PJ (1992) *Urban Soil in Landscape Design*. New York: John Wiley.

De Kimpe C and Morel JL (2000) Urban soils: a growing concern. *Soil Science* 165: 31–40.

FAO-UNESCO (1990) *Soil Map of the World – Revised Legend*. Rome: FAO.

Guillerme A (1999) Histoire du sol urbain. *Comptes Rendus de l'Académie d'Agriculture de France* 85: 129–140.

Hiller DA and Burghardt W (1993) Neues Leben im toten Boden. *Die Geowissenschaften* 1: 10–16.

Hollis JM (1992) Proposals for the classification, description and mapping of soils in urban areas. *English Nature Science* 2(4): 40.

United Nations (UN) Population Division (1995) *World Urbanization Prospects: The 1994 Revision*, pp. 132–139. New York: UN.

World Resources Institute (1996) Cities and the environment. In: *World Resources 1996–97*, pp. 1–30. New York: Oxford University Press.

V

VADOSE ZONE

Contents

Hydrologic Processes

J W Hopmans, University of California–Davis, Davis, CA, USA
M Th van Genuchten, US Salinity Laboratory, Riverside, CA, USA

Introduction

The distinction between groundwater and the unsaturated zone is usually made within a hydrologic context that views water as the agent of change of the subsurface and the main driver for transport of chemicals between the atmosphere and groundwater. This region between the soil surface and groundwater table known as the vadose zone, involves a complex array of time-dependent, nonlinear physical, chemical, and biological processes, including interactions with groundwater and the atmosphere. The soil is the uppermost part of the vadose zone, subject to fluctuations in water and chemical content by infiltration and leaching, water uptake by plant roots, and evaporation from the soil surface. It is the most dynamic part of the subsurface, as changes occur at increasingly smaller time and spatial scales when moving from the groundwater toward the soil surface. Soil depth is usually controlled by the maximum rooting depth of plants (generally within a few meters of the soil surface). By contrast, the vadose zone may extend much deeper than the surficial soil layer and includes unsaturated rock formations and alluvial materials to depths of 100 m or more.

In the last few decades of the twentieth century, research interests in the deeper vadose zone have increased dramatically, instigated by a need to sustain quality of groundwater and maintain adequate resources for drinking water and ecologic purposes. Unquestionably, our society has negatively impacted the quantity and quality of its soil, water, and air resources. Chemical pollution generated by past agricultural, industrial, and municipal activities has contaminated soil and groundwater and surface-water systems worldwide. Unfortunately it continues to do so up to the present. Scientists and others are now increasingly aware that soil is a critically important component of the Earth's biosphere, not only because of its food-production function, but also as a safekeeper of local, regional, and global environmental quality. For example, management strategies in the vadose zone will offer the best opportunities for preventing or limiting pollution, and for remediation of ongoing pollution problems. Because chemical residence times in groundwater aquifers can range from a few years to thousands of years, their pollution is often essentially irreversible. Prevention or remediation of soil and groundwater contamination starts, therefore, with proper management of the vadose zone. Understanding the intricate processes in the vadose zone is a challenge because of the many complex nonlinear physical, chemical, and biological interactions that simultaneously control the transfer of heat, mass, and momentum between the atmosphere and the groundwater table.

Physical Processes

The physical characteristics of the vadose zone control such processes as natural or artificial recharge to deep groundwater and surface water–groundwater interactions. When considering the water budget of the vadose zone as a whole, many terms are similar and equally as important as those of a soil-water budget, with differences in measurement techniques mostly

predicated by differences in the spatial scale. Whereas typical soil-infiltration measurements are conducted over areas smaller than 1 m^2, estimates of infiltration rates over much larger areas and depths are required in the vadose zone, for which the assumption of one-dimensional flow is generally inadequate. Whereas the soil-water budget requires quantification of drainage, i.e., flow of water beneath the rooting zone (*See* **Drainage, Surface and Subsurface**), the water budget of the vadose zone includes net infiltration (*See* **Infiltration**), percolation, and recharge rates at large depths. 'Net infiltration' is generally defined as the water flux below the root zone that is not further influenced by evaporation or plant transpiration. 'Percolation rate' is defined as the net infiltration rate at any depth within the vadose zone, whereas 'recharge rate' defines the water flux into the groundwater across the groundwater table. Depth variations in percolation rates are caused by lateral flow and temporal variations in precipitation and evapotranspiration (*See* **Evapotranspiration**), whereas estimated rates are affected by measurement type. Specific vadose zone methods to estimate percolation and recharge rates include physically based methods that assume darcian water flow (*See* **Darcy's Law**) throughout the vadose zone, and tracer methods. The darcian methods generally require dedicated and highly specialized instrumention for large-depth measurements, but neglect fractured flow. Environmental tracer methods estimate percolation rate or age of water at a given depth based on *in situ* concentrations of natural tracers such as tritium, chloride, chlorine-36, and nonradioactive isotopes such as deuterium and oxygen-18, assuming that these are mass-conservative.

Accurate estimation of vadose zone water and solute fluxes is especially challenging in arid climates, where they are orders-of-magnitude smaller than in humid agricultural settings. However, because of these small flow rates, the relevant time scales of recharge rate can be orders-of-magnitude larger. A variety of methods have been reviewed, including detailed numerical modeling, to estimate percolation and recharge rates at Yucca Mountain, Nevada. The different methods were compared to study the potential and performance of Yucca Mountain as a repository site for high-level radioactive waste over time scales of 1000 years and longer (Figure 1). The mean water table depth is approximately 500 m, whereas the potential repository location is at a mean depth of 300 m below the land surface, within a densely welded and fractured tuff horizon. In February 2002, President George W. Bush endorsed a formal recommendation by the Department of Energy (DOE) for the Yucca Mountain to accept a total of approximately 85 000 metric tons of radioactive wastes.

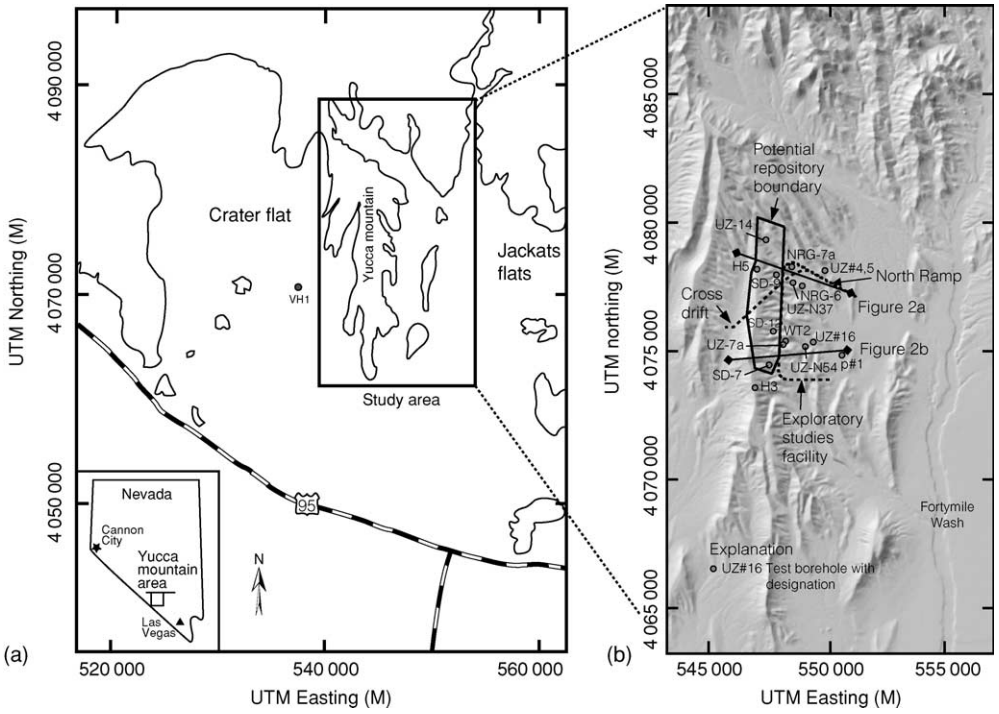


Figure 1 Location of (a) the Yucca Mountain area, Nevada, and (b) the study area. (Reproduced with permission from Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvatsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.)

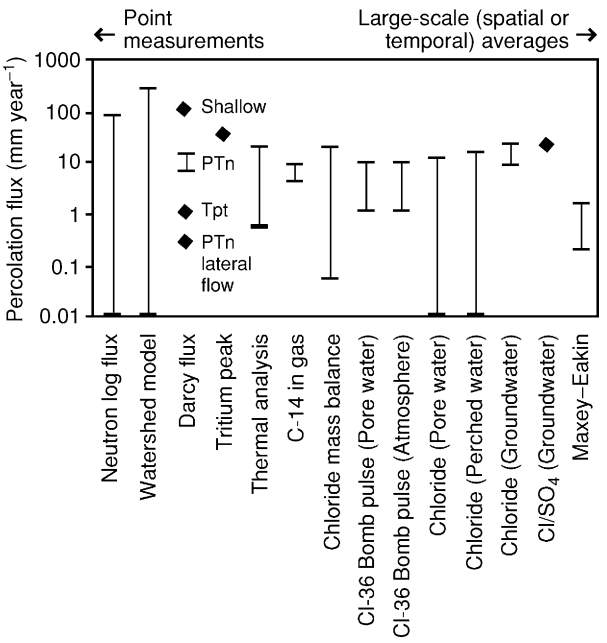


Figure 2 Comparison of percolation fluxes estimated by various methods. Bars represent ranges of estimates; points represent single estimates. PTn and Tpt indicate different hydrogeologic units of the Yucca Mountain site. (Reproduced with permission from Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvarsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.)

Estimated percolation rates at Yucca Mountain vary between 0.5 and 20 mm year^{-1} (Figure 2), with differences in magnitudes controlled by variable precipitation, topography, soil depth, and highly variable, physical vadose zone properties, including fractures and faults. The different methods each yield percolation values typical of their specific space and time scales, with results complementing and partially overlapping each other. For example, shallow methods are representative for relatively small time scales, whereas rates estimated from the deep vadose zone and recharge measurement techniques are an integration of spatial mixing over much longer time scales.

Chemical Processes

The justification of vadose zone characterization and monitoring of contaminant transport arises from the simple observation that, for contaminants and microorganisms to reach the groundwater table, they must pass through the vadose zone first. Therefore, monitoring of the vadose zone and appropriate management practices using vadose zone concepts is a prerequisite for understanding and successfully preventing groundwater contamination. Much of the *Handbook of Vadose Zone Characterization and Monitoring* is devoted to current and emerging techniques for vadose zone measurement and monitoring of chemical and microbial pollutants. The general theories and concepts of chemical fate and transport in the vadose zone are largely identical to those occurring in soils. Therefore, the focus here is on the added complexities to which constituents are subjected as they move through the vadose zone. Most additional complications in vadose zone transport arise because of the much wider range in relevant space and time scales and the presence of preferential flow mechanisms (fractures, sediment, and rock beddings) that make it difficult to delineate transport networks for the larger vadose zone. Moreover, the increasing physical heterogeneity (*See* **Spatial Variation, Soil Properties**) and associated spatially variable soil moisture conditions and preferential flow mechanisms of the vadose zone affect geochemical and microbial reactions, resulting in spatial and temporal variations of concentrations and transport of which the implications have yet to be fully understood.

As dissolved solutes move through the vadose zone, various physical, chemical, and biological properties control their fate. In addition to diffusion and dispersion, the fate and transport of chemicals in the subsurface are influenced by sorption on to the solid phase and biological transformations. Diffusion and dispersion of the transported chemical are a function of both pore size distribution and water content. Mechanical or hydrodynamic dispersion is the result of water mixing within and between pores as a result of variations in the pore-water velocity. Increasing dispersivity values causes greater spreading of chemicals, thereby often also decreasing their peak concentration. Sorbed chemicals move through the vadose zone slower than noninteracting chemicals, with the degree of sorption largely depending on mineral type, specific surface area of the solid phase, and organic matter fraction. In addition, biogeochemical processes and radioactive decay affect contaminant concentration, for example by cation exchange, mineral precipitation and dissolution (*See* **Precipitation–Dissolution Processes**), complexation, oxidation–reduction reactions (*See* **Oxidation–Reduction of Contaminants**), and microbial biodegradation and transformations. Moreover, all of these processes depend on such environmental conditions as temperature, pH, water saturation, and redox status (*See* **Redox Potential**), including, their spatial variations.

The fate and transport of microorganisms (including pathogenic bacteria, viruses, and protozoan parasites) follow many similar processes to chemical compounds. For example, many microbes exibit

enhanced transport relative to the water solution as a result of electrostatic repulsion from negatively charged solid surfaces, as well as owing to size exclusion from smaller areas of the pore space. There is also evidence that biologically reactive solutes and bacteria preferentially adsorb to gas–liquid interfaces and may be subject to inferfacial adsorption–desorption processes. In addition, transport of microorganisms is controlled by their retention (such as filtration) in the porous matrix, which is a function of the size of the specific microorganisms, the water-filled pore size distribution, and pore-water velocity as well as concentration and ionic composition of the aqueous phase. As such, microbial biomass accumulation may reduce the porosity and alter the hydraulic properties of vadose zone soils. Additionally, inorganic, organic, and microbiologically active colloids with diameters between 0.01 and 10 μm can strongly adsorb otherwise immobile chemicals or microbes and thus facilitate their transport by acting as a mobile solid phase.

No better example of the complexities of vadose-zone transport can be illustrated than by highlighting progress made at the US DOE Hanford site in Washington State, toward measurement, characterization, modeling, and remediation of vadose zone contamination from leaking, high-level radioactive waste storage tanks. The subsurface fate of the nuclear waste is increasingly complex because of various potential preferential flow mechanisms occurring, such as fingered flow, tunnel flow, and flow associated with poorly sealed boreholes (**Figure 3**). Soon after World War II, the US Atomic Energy Commission built many 2.0- to 3.8-million-liter, carbon-steel single-shell tanks at the Hanford Site to store high-level waste generated from production of plutonium and uranium. Inventory studies show that more than 4 million liters of waste leaked from these tank farms into the surrounding vadose zone. The vadose zone at Hanford consists of permeable and poorly consolidated sands and gravels that are bedded but highly discontinuous in the lateral direction. Also, the vadose zone includes many sharp contrasts in texture between layers. The regional groundwater table at depths of 60–90 m below the surface discharges to the Columbia River, which flows adjacent to the Hanford site. Gamma logging data from approximately 140 dry wells (boreholes) surrounding the tanks shows contamination to depths of at least 40 m, probably mostly caused by preferential flow. The different preferential flow processes have caused widespread contamination of ^{137}Cs and other radionuclides, which otherwise would have been highly retarded. Fingered flow occurs because of density-driven flow, whereas tunnel flow is caused by zones of increased permeability around the tanks. The geologic complexities of the Hanford site, combined with man-made changes, have prevented adequate prediction of the transport of the various radioactive nuclides, even with the use of sophisticated multidimensional flow and transport models.

Whereas this example applies to point-source pollution, even more complications arise when responding to questions on nonpoint pollution of groundwater such as occurs through crop production by application of fertilizers (e.g., nitrates) and irrigation (e.g., salinity and toxic trace elements). Typically, using distributed modeling techniques, flow and transport

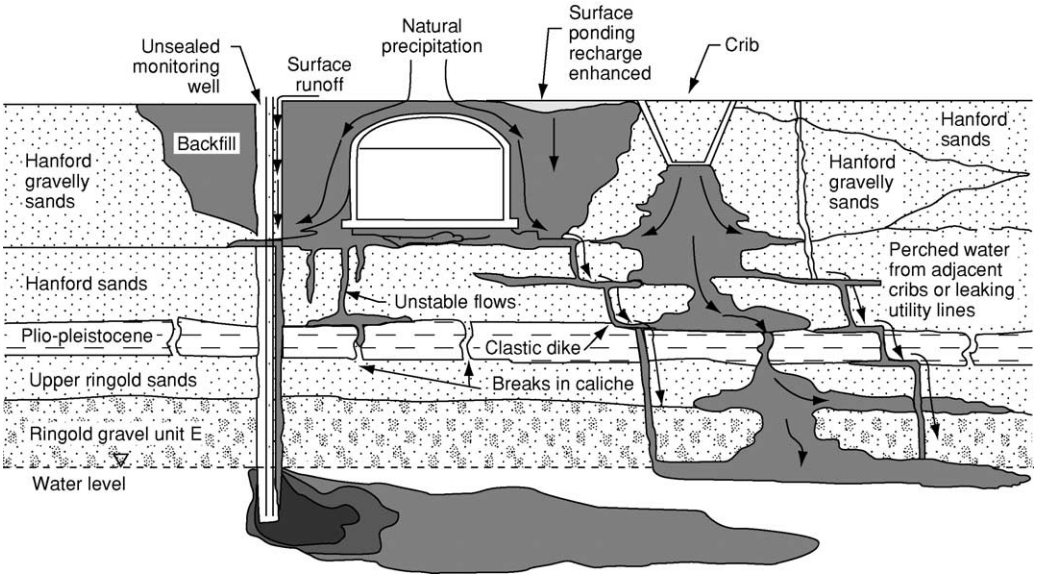


Figure 3 Conceptual model of fluid blow beneath single-shell tanks at the US Department of Energy's Hanford site.

are simulated for major characteristic units within a watershed or landscape (e.g., using soil map units, vegetation type, and slope). However, the connection and integration of the resulting simulation units within the watershed beg many questions of scale and associated nonlinearity effects. These nonlinear effects may be exacerberated when including chemical processes that are highly variable in space and time because of local soil variations in soil chemical composition, saturation, aeration, pH, redox status, and other factors. There is a need for interdisciplinary and multifaceted research approaches to improve the understanding of biological effects on soil chemical reactions and processes.

Biological Processes

In the soil environment, bioremediation and phytoremediation can be used to reduce vadose zone and groundwater contamination, by means of biological processes sustained by selected plants or microbes. Favorable environmental conditions of plant and microbial growth in soils and the vadose zone allow for bioremediation of contamination by inorganic and organic chemicals. Microbial processes mainly transform the parent chemical, thereby reducing its concentration, and by the formation of less-toxic metabolites, whereas plants can accumulate specific chemical species, e.g., through bioaccumulation of heavy metals in plant tissues. In addition, root water uptake by plants and trees can reduce leaching (*See* **Leaching Processes**) of water and salts below the rooting zone, thereby controlling percolation rates and contaminant transport through the vadose zone.

Differences between the microbiology of the vadose zone and the saturated zone are mostly caused by differences in water saturation, which affects the prevailing fluid fluxes and the availability of oxygen and other nutrients. With oxygen being the oxidizing agent for aerobic microorganisms, its supply may be limited in the deep vadose zones and/or in locally saturated (parched) areas. For most other nutrients, fluxes are controlled by the degree of water saturation and associated percolation rates and by organic matter supply rates. Consequently, nutrient availability is relatively abundant in the root zone and in humid climates, while their fluxes generally are limited in the vadose zones of arid climates. Although the vadose zone is generally unsaturated, local water-saturated inclusions may occur, promoting anaerobic microbiological processes, using electron acceptors other than oxygen, such as nitrate, Mn^{4+}, Fe^{3+}, or sulfate, depending on the redox status of the system. Irrespective of climate, the local nutrient supply in the vadose zone is at times predominantly controlled by diffusion to or from regions where transport is mainly by preferential flow through macropores (*See* **Macropores and Macropore Flow, Kinematic Wave Approach**) and cracks. Whereas, in general, microbial population density and activities are low in the vadose zone below the soil-rooting zone, their densities and activities can be orders-of-magnitudes higher at contaminated sites that facilitate microbial colonization. Microbes are generally located at air–water and water–solid interfaces; however, microbial heterogeneity is generally unpredictable, as it is conditioned by local variations in nutrient availability.

Enhanced microbial degradation (*See* **Pollutants:** Biodegradation) is generally achieved by gaseous nutrient deliveries of electron donors and acceptors to the contaminated sites, e.g., by bioventing through injection of air to simulate aerobic biodegradation of petroleum hydrocarbons (*See* **Hydrocarbons**). Microbial degradation of recalcitrant organic contaminants may require a secondary carbon supply, whereas other nutrients such as N and P are sometimes required. Inorganic contaminant concentrations of heavy metals or radionuclides can be reduced by microbes through their transformation (*See* **Metals and Metalloids, Transformation by Microorganisms**) to less-toxic states or by changing their mobility by using them as electron acceptors. A specific form of bioremediation is phytoremediation (defined as remediation through plants), which is mostly effective in near-surface soils. Many processes may contribute to phytoremediation (**Figure 4**), including specific root uptake followed by bioaccumulation and/or volatilization, and biodegradation in the rhizosphere (*See* **Rhizosphere**) sustained by root exudates and organic matter of decaying roots.

Certain plants have been adapted genetically to grow in soils containing toxic levels of metals. A classic example of this is the use of specific plants for the bioaccumulation of selenium, for example crop and grassland species. Research in selenium remediation methods accelerated in the 1980s after the discovery of high concentrations of Se in agricultural drainage water, followed by high mortality rates of grazing waterfowl at Kesterson Reservoir in Merced County, California. The specific advantage of crop or grassland plant species to remediate Se-laden soils is that these can be harvested and subsequently used as a Se supplement for Se-deficient forage or as an amendment for Se-deficient rangelands. Studies have demonstrated that grasses such as tall fescue can effectively take up Se if soils are supplemented with organic matter. Other successful examples include bioaccumulation of lead and other toxic metals, and radionuclides such as Cs, U, Cd, and Cr.

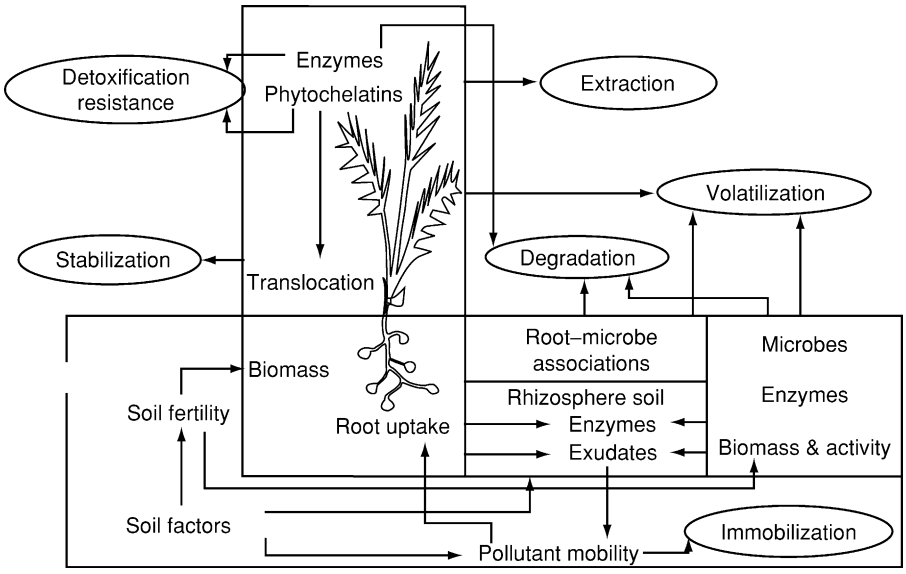


Figure 4 Simplified overview of soil–plant–microbe processes and interactions involved in phytoremediation. (Reproduced from Wenzel WW, Adriano DC, Salt DI, and Smith R (1999) Phytoremediation: a plant–microbe-based remediation system. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 457–508. Agronomy Series No. 37. Madison, WI: American Society of Agronomy, Inc.)

There are many questions that need to be resolved before bioremediation techniques can be applied universally to a wide range of contaminant classes. Major issues revolve around the complex interactions between microbiological, geochemical, and hydrogeologic processes, the unpredictable local distribution of microsites, and the difficulty in supplying nutrients to these spatially distributed, contaminant-transforming microorganisms (*See* **Spatial Patterns**). In addition, successful bioremediation is controlled by the physical (e.g., soil moisture, temperature, oxygen diffusion) and chemical (e.g., organic matter type and content, soil adsorptive properties, availability of micronutrients) soil environment.

An additional example of mostly unknown territory occurs in the mixing zone between surface and subsurface water underneath streams, known as the hyporheic zone (HZ). This region of increased biochemical activity within the upper few centimeters of stream sediment affects the type and rate of material transformation as water moves downstream, thereby significantly changing stream-water composition and the stream ecosystem, as well as groundwater chemistry. Specific examples include the influence of oxygen supply to fish eggs buried in the HZ, the influence of particulate and dissolved organic matter on microbial activity and resulting stream-water chemistry, and the function of the HZ to denitrify, thereby ameliorating high-N stream load. There is a definite need to understand better the hydrology of the HZ and the coupling of hydrologic with biogeochemical processes, through dedicated interdisciplinary experiments and numerical modeling.

Scale Issues

For the past few decades, soil scientists have applied soil physical data to characterize flow and transport processes in large-scale, heterogeneous vadose zones, using measurement scales that are typically much smaller. For example, prediction of soil-water dynamics at the field scale is usually derived from the measurement of soil hydraulic properties from laboratory cores, collected from a limited number of sampling sites across large spatial extents. Soil parameters obtained from these small-scale measurements are included in numerical models with a grid or element size that is many times larger, with the numerical results extrapolated to predict large-scale flow and transport behavior. Because of the typical nonlinearity of physical properties, their use across spatial scales is inherently problematic. Specifically, the averaging of processes determined from discrete, small-scale samples may not describe the true soil behavior involving larger spatial structures. Moreover, the dominant physical flow processes may vary between spatial scales. Considering that soil physical, chemical, and biological measurements are typically conducted for small measurement volumes and that the natural variability of soils is enormous, the main question asked is how

small-scale measurements can provide information about large-scale flow and transport behavior. Answers to this question may require the estimation of appropriate, effective soil parameters for use in describing the behavior of pollutant plumes at the field or landscape scales.

A conceptual solution to the problem of scale issues of vadose-zone modeling might lie in considering the controlling effect of small-scale processes on larger-scale flow behavior. Hence, vadose zone properties are nonunique and scale-dependent, resulting in effective properties that vary across spatial scales and merely serve as calibration parameters in simulation models. Therefore, their accurate prediction in heterogeneous materials can only be accomplished using scale-appropriate measurements, including those that measure at the landscape scale.

Opportunities and Challenges

There are a number of opportunities that have come about in the past decade through experimental innovations and increasing environmental awareness. These opportunities include improved physical characterization of the vadose zone at larger spatial scales, which needs to be extended to chemical and biological measurements also.

First, it is becoming increasingly clear that there is a pressing need for subsurface observations and property-measurement techniques at spatial scales much larger than the usual laboratory or field-plot scale. The scale problem is extremely complex because of the general presence of large spatial and temporal variabilities of the soil physical, chemical, and biological properties in question. For example, at the heart of many hydrologic projects lies the need to understand better the flow and transport of water and associated chemical constituents into and through the vadose zone above an aquifer, and within a watershed. Hence, developmental work is needed on fundamental concepts and measurement technologies to establish appropriate soil parameters for use in theories or models describing the behavior of vadose zone water flow and pollutant plumes across spatial scales. In addition, appropriate measurement techniques and field experiments are needed to characterize effective field-scape and landscape-scale soil properties. In particular, we note the potential of using inverse methodologies perhaps in combination with rapidly improving invasive and noninvasive geophysical techniques, to infer *in situ* dynamic soil physical characteristics and the development of instruments that combine multiple measurements within a single device, thereby minimizing soil heterogeneity effects.

Second, although it is evident that large-scale characterization is needed, there is also increasing awareness within the scientific community that the physical, chemical, and biological processes in the vadose zone are controlled by mechanisms operating at the pore-size scale. Improved predictions of subsurface flow and transport will probably be a function of the development and application of innovative pore-scale modeling approaches (e.g., Lattice–Boltzmann, percolation, and related methods), and associated measurement techniques that operate at the microscopic level. Examples of the latter are nuclear magnetic resonance (NMR), computed tomography (CT), and spectromicroscopy.

Third, improved characterization and interpretation of subsurface processes will require increasing efforts toward an interdisciplinary partnership, integrating physical with chemical and biological processes. The soil physicist and/or vadose zone hydrologist must seek collaborations in other disciplines to ensure that different measurement types are collected and integrated to study more effectively and determine relationships between flow and transport processes at the microscale and in the laboratory, with ultimate application to the watershed scale. To achieve this goal, soil physics must be taught within the broader context of hydrology and the environmental sciences.

See also: **Darcy's Law**; **Drainage, Surface and Subsurface**; **Evapotranspiration**; **Hydrocarbons**; **Infiltration**; **Leaching Processes**; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Metals and Metalloids, Transformation by Microorganisms**; **Oxidation–Reduction of Contaminants**; **Pollutants:** Biodegradation; **Precipitation–Dissolution Processes**; **Redox Potential**; **Rhizosphere**; **Spatial Patterns**; **Spatial Variation, Soil Properties**

Further Reading

Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (1999) *Bioremediation of Contaminated Soils*. Agronomy Series No. 37. Madison, WI: American Society of Agronomy.

Ainsworth CC, Brockman FJ, and Jardine PM (2000) Biogeochemical considerations and complexities. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 829–947. Columbus, OH: Battelle Press.

Cheng HH and Mulla DJ (1999) The soil environment. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 1–13. Agronomy Series Number 37. Madison, WI: American Society of Agronomy.

Dane JH and Topp GC (2002) *Methods of Soil Analysis,* part 4, *Physical Methods*. SSSA Book Series No. 5. Madison, WI: Soil Science Society of America, Inc.

Ferré TPA and Kluitenberg GJ (eds) (2003) Advances in measurement and monitoring methods, special section. *Vadose Zone Journal* 2: 443–654.

Findlay S (1995) Importance of surface–subsurface exchange in stream ecosystems: the hyporheic zone. *Limnology and Oceanography* 40: 159–164.

Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvarsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.

Hopmans JW and Bristow KL (2001) Current capabilities and future needs of root water and nutrient uptake modeling. *Advances in Agronomy* 77: 103–183.

Hopmans JW, Nielsen DR, and Bristow KL (2002) How useful are small-scale soil hydraulic property measurements for large-scale vadose zone modeling. In: Smiles D, Raats PAC, and Warrick A (eds) *Heat and Mass Transfer in the Natural Environment*, pp. 247–258. Geophysical Monograph Series no. 129. Washington, DC: American Geophysical Union Publications.

Parlange MB and Hopmans JW (1999) *Vadose Zone Hydrology: Cutting across Disciplines*. Oxford, UK: Oxford University Press.

Shafer DS, Bertsch JF, Koizumi CJ, and Fredenburg EA (2000) Gamma borehole logging for vadose zone characterization around the Hanford high-level waste tanks. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 445–457. Columbus, OH: Battelle Press.

Sparks DL (2001) Elucidating the fundamental chemistry of soils: past and recent achievements and future frontiers. *Geoderma* 100: 303–319.

Stephens DB (1996) *Vadose Zone Hydrology*. Boca Raton, FL: CRC Press.

Tyler SW, Scanlon BR, Gee GW, and Allison GB (1998) Water and solute transport in arid vadose zones. In: Parlange MB and Hopmans JW (eds) *Vadose Zone Hydrology: Cutting Across Disciplines*, pp. 334–373. Oxford, UK: Oxford University Press.

Wenzel WW, Adriano DC, Salt DI, and Smith R (1999) Phytoremediation: a plant–microbe-based remediation system. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 457–508. Agronomy Series No. 37. Madison, WI: American Society of Agronomy, Inc.

Wilson LG, Everett LG, and Cullen SJ (1995) *Handbook of Vadose Zone Characterization and Monitoring*. Boca Raton, FL: Lewis.

Wood TI and Faybishenko B (2000) Large-scale field investigations in fractured basalt in Idaho: lessons learned. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 396–405. Columbus, OH: Battelle Press.

Wu LW (2000) Selenium accumulation and uptake by crop and grassland plant species. In: Frankenberger WT Jr and Engberg RA (eds) *Environmental Chemistry of Selenium*, pp. 657–686. New York: Marcel Dekker.

Microbial Ecology

P A Holden and N Fierer, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

The vadose zone consists of unsaturated porous media and rock extending from the surface soil to the groundwater table. This zone is characterized by a lack of water relative to other major Earth compartments such as the ocean, fresh water, aquatic sediments, and groundwater. Low nutrient availability, low water content, and low potential energy per unit volume of water, termed 'water potential,' constrain life in the deeper vadose-zone relative to the surface. However, the vadose zone harbors numerous and diverse microbes, including bacteria, fungi, protozoa, and viruses. Vadose-zone microbial ecology refers to the study of interactions between vadose-zone microbes and their environment. Generally, microbial population density, diversity, and total activity, as well as available nutrients, and moisture and temperature fluctuations, decline sharply with depth below surface soil. For example, relative to surface soil, culturable heterotrophic bacteria in the unsaturated subsurface are only one-tenth as numerous. Regardless, the integral mass of microbes along the depth profile is large, which accounts for the overall importance of vadose-zone microbes as catalysts in terrestrial nutrient-cycling, including pollutant biodegradation.

Definition of the Vadose Zone

The vadose zone is the Earth's terrestrial subsurface that extends from the surface to the regional groundwater table. As shown in Figure 1, the vadose zone includes surface soil, unsaturated subsurface materials, and a transiently inundated capillary fringe. The subsurface materials include partially weathered soils and unweathered parent material. The vadose zone may be very shallow (less than 1 m) or very deep (extending hundreds of meters or more), depending on the depth to the water table.

The vadose zone has low water content relative to the saturated zone below the water table and is therefore commonly referred to as the unsaturated zone. Above the capillary fringe, vadose-zone pore spaces are generally air-filled, with thin water films coating solid particles. Pore spaces become water-filled when rainfall percolates, followed by drainage and gradual drying. It is the relative lack of water and its

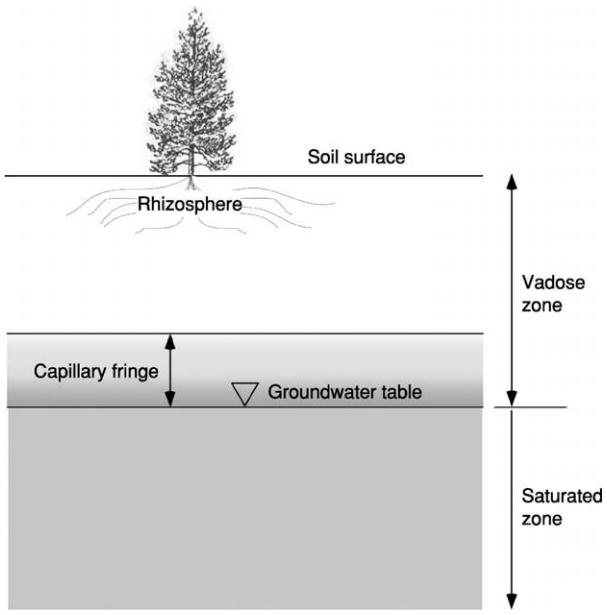


Figure 1 The vadose-zone profile.

transience, the complete lack of sunlight, and complex physical and chemical gradients that make the vadose zone an interesting and unique compartment for microbes.

Vadose-zone microbial ecology is the study of the microbes residing in the vadose zone, their interactions with each other, and their interactions with their surrounding environment. Compared with soil microbiology, a field for which a significant amount of agriculturally oriented research has been conducted, vadose-zone microbial ecology is a relatively young science. However, questions regarding the presence of microbes in the vadose zone, where they live, what they do, and how they do it are integral to the work of biogeochemists, environmental scientists, and engineers concerned with the fate of nutrients and pollutants.

Abundance and Distribution of Microbes in the Vadose Zone

Quantity Along a Depth Gradient

The abundance and distribution of microbes in the vadose zone are key parameters that need to be measured in order to understand vadose-zone microbial ecology. Abundance is studied by sampling (typically by aseptically coring through the face of a trench or vertically from the surface) and laboratory quantification at each depth interval sampled. Laboratory quantification of microbial abundance is performed using a variety of methods, including: directly counting eluted bacteria, culturing, measuring substrate uptake and/or mineralization, measuring total phospholipid fatty acids (PLFAs), measuring biomass carbon (C), quantifying extractable deoxyribonucleic acid (DNA), or counting using a most-probable-number (MPN) technique. Specific phylogenetic or functional groups of microbes can also be sensitively quantified using molecular biological approaches such as either fluorescence *in situ* hybridization (FISH) coupled with directly counting cells, or quantitative polymerase chain reaction (Q-PCR). Typically, each of these methods provides estimates of microbial abundance that are in general agreement with one another. Along a depth gradient (Figure 2), microbes tend to be most numerous in the surface soil, with their abundance declining sharply below the surface then reaching a nonzero plateau in the subsurface. The initial rate of decline with depth is more rapid in some vadose-zone systems than in others. Microbial population densities tend to increase in the capillary fringe and at the water table. In shallow vadose zones, the majority of the microbial biomass is in the surface material. In deeper vadose zones, microbial biomass densities are highest in the surface regions, but the total microbial biomass in deeper materials can be high due to the integration of sparser population densities across a large spatial area.

Generally, the number of culturable cells decreases more rapidly with depth than the number of total extractable cells. For example, by culturable counts, subsurface populations of heterotrophic bacteria may be one-tenth the size of surface populations, but, by direct counts, subsurface populations may be one-fifth the size of surface populations. In general, the

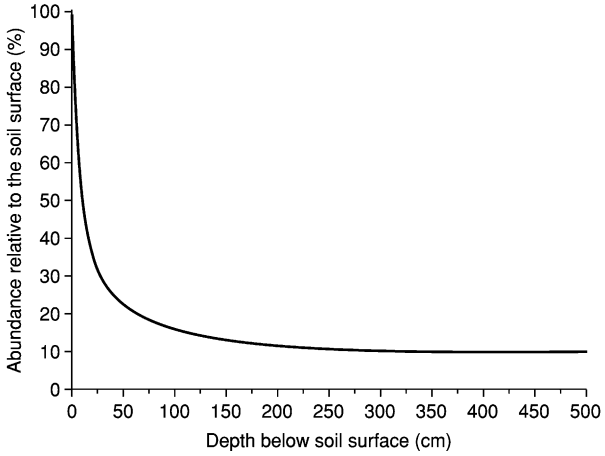


Figure 2 Typical depth trend of the vadose-zone constituents that are relevant to microbial ecology.

total numbers of culturable bacteria are higher in polluted vadose materials.

Protozoan population densities decline rapidly with depth below surface soil, which may be attributable to the concomitant decline in abundance of bacteria, their primary food source. Direct counts of fungal populations also decline with depth, but at a more dramatic rate than bacteria. Overall, total nonbacterial microbial biomass, including protozoa, algae, diatoms, and fungi, is lower than bacterial biomass in surface soils and decreases more dramatically with depth.

While the majority of soil microbial biomass is found in the surface horizons (Figure 1), where nutrients are most abundant (Figure 2), these horizons also experience extreme fluctuations in moisture and temperature which may result in significant temporal variability in population sizes. Microbial population sizes in the subsurface tend to experience a minimal degree of seasonal variation.

Distribution at the Meso- and Microscales

The distribution of microbes with depth through the vadose zone, as described above, is useful to know if we want to build large-scale mathematical models of microbially mediated processes in the vadose zone. However, there is also a secondary nature to microbial distribution which occurs at the meso-, and micro-, or perhaps 'microbe-' scales in soil. As depicted in Figure 3a, mesoscale variation in microbial abundances depends on the spatial distribution of bulk soil properties, including flow paths that preferentially conduct water through both surface and subsurface soils. Different microbial communities have been found along preferential flow paths as compared to the surrounding vadose material. In surface soils, preferential flow paths harbor more carbon and therefore more microbes; in the subsurface, this phenomenon is attenuated. At the microscale, the scale of individual pores, microbes are less active in the smallest pores where the diffusional resupply of nutrients is restricted. In contrast, larger, interconnected pores tend to contain more metabolically active microbes.

At the scale of individual microbes, which in some cases may be at the pore scale, there are at least three known possible configurations for microbial growth habits (Figure 3b). These configurations are mainly applicable to bacteria, the most abundant microbes in the vadose zone, but may also apply to other microbial particles, including fungal spores, viral particles, and protozoa. The distribution of filamentous fungi is generally a mesoscale phenomenon, because fungal hyphae can structurally, and physicochemically, bridge pores and particles. As shown in Figure 3b,

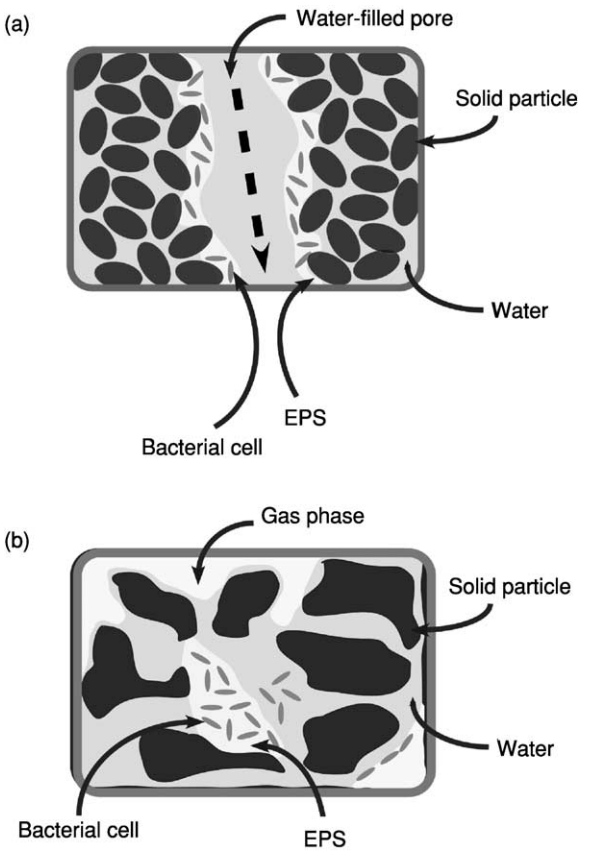


Figure 3 (a) Mesoscale and (b) microscale orientations of vadose-zone bacteria relative to other constituents, including extracellular polymeric substances (EPS) surrounding bacterial cells.

microbes can colonize the air–water interface, exist planktonically in water, or attach to surfaces and grow in biofilms consisting of cells and their associated expolymeric substances (EPS). The growth habit dictates movement: microbes at the air–water interface and in bulk water are mobile, but microbes embedded in EPS are relatively immobile. The growth habit also affects the availability of nutrients and the susceptibility to predation: microbes in EPS tend to be more protected from predation but often have reduced access to nutrients. Compared with what we know regarding gross microbial distributions at the macro- or depth-gradient scale, we know very little about microbial distributions at smaller scales and how specific configurations affect larger-scale vadose zone processes.

Diversity of Vadose Microorganisms

The microbes inhabiting the vadose zone are diverse taxonomically, metabolically, and morphologically. There are thousands of bacterial species (species richness) in a gram of surface soil. The proportional

representation of each taxonomic group (evenness) varies with depth within a vadose zone and between different types of vadose zones. For example, the overall abundance of Gram-positive bacteria has been shown to increase relative to Gram-negative bacteria along a vadose-zone depth profile, and there are lower fungal-to-bacterial ratios at depth than at the surface.

Taxonomic diversity is assessed with a range of techniques, including: the identification of culturable microbes, PLFA fingerprinting, fingerprinting based on 16S ribosomal DNA (rDNA), and the enumeration of different 16S rDNA sequences in a clone library. Metabolic diversity can be assessed by exposing serial enrichment cultures or microcosms to a variety of substrates and measuring the potential for substrate metabolism. Both taxonomic and metabolic diversity decline with depth through the vadose-zone profile, following the general trend depicted in Figure 2. In a related pattern, the number of dead and viable-but-nonculturable cells (VBNC) tends to increase with vadose-zone depth. Morphological diversity can be assessed by the microscopy of cultivated cells, but, as with any diversity assessment based on conventional laboratory culturing, the method may yield biased estimates of diversity because so few soil microbes (less than 1%) can be cultivated.

A fundamental question in vadose-zone microbial ecology is: How did the microbes, particularly those found at greater depths, get there? In part, diversity studies address this question by looking for similarities between the microbial community composition of the surface soil and that of the deeper vadose zone. If identical taxonomic groups are found at the surface and at depth, this suggests that microbes are transported downward through the vadose zone. If there are microbial groups found at depth that are unique from those at the surface, this suggests that microbes may have resided in the initial parent materials before the formation of the vadose zone. Generally, while there is overlap in the taxonomies of surface and deep vadose microbial communities, there are also unique groups.

Water is the main agent responsible for the inoculation of the subsurface by the surface. Thin water films in the vadose zone typically limit the transport of microbes through the profile. However, water and gas transiently advect and diffuse through soil pores. Preferential flow of water through certain pores causes these paths to have differing chemical and biological characteristics when compared with surrounding soil. Further, recharge of water to the vadose zone varies and, in turn, this leads to additional variations in pore chemistry and microbial communities along preferential flow paths. Overall, site hydrology is a key determinant of the abundance and types of microbes inhabiting the vadose-zone profile. However, at any given depth, the variability in microbial community composition and abundances can be high, and this heterogeneous distribution contributes greatly to microsite-level variation in microbial processes.

Activity of Vadose-Zone Microbes

Microbial metabolic activity declines with depth in the vadose zone, roughly paralleling the curve in Figure 2. The observed decline in activity with depth is related to the previously mentioned declines in microbial viability and abundance with depth. However, *in situ* activity is strongly affected by available nutrients. For example, *in situ* denitrification rates at depth in the vadose zone are generally much lower than the rates measured for surface soils (top 10 cm). The decrease in denitrification rates is partly a result of the low abundance of denitrifying bacteria, but is primarily a result of the low carbon availability in the deeper depths of the vadose zone. When subsurface vadose-zone samples are amended with nutrients, the activity per viable microbe in the subsurface is approximately the same as that in the surface.

Physiological markers of stress indicate that vadose-zone microbes are subject to starvation with increasing depth. Starvation is enhanced when water availability is low, because both the intrinsic physiology of microbes and the substrate availability to microbes are affected. Microbial biodegradation activity declines with depth, as shown with simple analysis of CO_2 evolution patterns. However, degradation rates can be as high or higher when either water or aqueous-phase nutrients are added to the vadose zone.

Measurements of enzyme activity are used to assess enzyme abundance in vadose-zone samples. Dehydrogenase, a useful measure of oxidative activity in all microbes, positively correlates with bacterial abundance, with dehydrogenase activity generally decreasing with depth below the surface. Other enzymes related to carbon, nitrogen, phosphorus, and sulfur metabolism show similar patterns. The low enzyme activity of deeper soils appears to be related to the low numbers of microbes and the low quantities of available nutrients.

Physical and Chemical Characteristics Affecting Vadose-Zone Microbial Ecology

The vadose zone is an open, multiphase system (Figure 3b) containing dissolved and undissolved solids (organomineral complexes and biomass), liquids

(water and nonaqueous-phase liquid pollutants), and gases (also known as soil gases). Open systems are inherently dynamic which, in the vadose zone, means that water, gases, and solutes are continuously being redistributed. The concept of chemical equilibrium in the vadose zone is only valid over short distances and short time scales. The ramification for microbes is that their local environment is in a state of continuous flux, due to either microbial metabolism or external environmental factors. Six key environmental factors are commonly cited as important to vadose zone microbial ecology: water availability, temperature, nutrient availability, oxygen, redox potential, and pH. Nutrient availability, redox potential, and oxygenation change with alterations in soil-water content: drying soil increases air-filled soil pore space, while wetting inhibits oxygenation and lowers redox potentials. Water is therefore considered to be a unifying environmental factor in the vadose zone.

Vadose-zone microbial ecology is affected by both soil-water potential and soil-water content (θ). Total soil-water potential (ψ) is defined as the potential energy per unit volume of soil and has two main components in the near subsurface: matric potential (ψ_m) and solute potential (ψ_s). In nonsaline soils, ψ_m is the dominant component of ψ. Total soil-water potential is related to the relative humidity of surrounding soil gas by:

$$\psi = \frac{RT}{V_w} \ln\left(\frac{\text{rh}}{100}\right) \qquad [1]$$

where R is the gas constant ($P - L^3/\text{mol} - T$), T is temperature (T), V_w is molal volume of water (moles per cubic liter), and rh is relative humidity (percentage).

As per eqn [1], ψ is negative when the rh is less than 100%. Alternatively, ψ is zero when the rh is equal to 100% (i.e., when water activity or $a_w = 1$). The vadose-zone environment may have a ψ equal to zero over a range of water contents that qualitatively span moist to fully saturated. For all $\psi = 0$ conditions, soil-water content indicates the degree to which pores are air-filled and thus oxygen-bearing. As soil dries, both θ and ψ decrease, but ψ is most relevant to microbial ecology. At approximately -0.25 MPa, vadose-zone bacteria become slightly nutrient-stressed, because the diffusional resupply of solutes is restricted. A total soil-water potential of -1.5 MPa is approximately the lower limit of what is physiologically tolerable to many surface soil bacteria, but filamentous fungi and actinomycetes are often more resilient to even lower water potential conditions. Vadose-zone microbes can adapt to be extremely desiccation-tolerant, as evidenced by the presence of viable microbes in deep, arid vadose zones where the total water potential is *c.*-50 MPa or less.

Water is an important physical environmental factor governing vadose-zone microbial ecology; temperature is another. For both water and temperature, average conditions vary from surface to subsurface. However, perhaps more importantly, both water content (and potential) and temperature are more variable in surface soils as compared to deep in the vadose zone. For example, temperature extremes are greatest at the soil surface but either increase or decrease (depending on the season) with depth to a constant temperature of around 20°C (*See* Figure 4 in **Thermal Properties and Processes**). Similarly, surface soils wet and dry frequently depending on the climate and season, yet subsurface soils are more stably moist or dry. In general, more stable temperature and moisture regimes at depth will select for vadose-zone microbial communities with narrower temperature and moisture optima and with lower tolerance for extreme or rapid fluctuations in environmental conditions.

The availability of nutrients to microbes in the vadose zone declines steeply with depth below the surface to low equilibrium values (**Figure 2**). In many vadose-zone environments, nutrients are more abundant in the surface soils because of plants and their contributions to soil nutrition. However, the textural composition of the vadose zone also changes with depth, affecting nutrient availability. For example, the percentage of clay, which controls the availability of exchangeable ions, typically decreases with depth. Further, the percentage of organic matter, which controls the abundance and availability of dissolved and available sorbed organic nutrients, also decreases with depth.

Soil gas composition affects redox status and aeration, and it often changes with depth through the vadose zone. For example, the oxygen content of soil gas in the subsurface is generally below the atmospheric concentration of 21%, often dropping to 15% or less. Also, CO_2 concentrations often range from 0.033% at the surface to as high as 4% or more at soil depths below 2 m. The CO_2 increase is due to the combined effects of microbial respiration and low rates of diffusional exchange between gases above- and below-ground. The high rates of CO_2 production and limited diffusional transport tend to increase alkalinity at depth, resulting in higher pHs in the subsurface compared with the soil surface. The unique gas conditions of the deeper soil depths facilitate the selection of microaerophiles and carboxyphiles.

The physical morphology of the vadose-zone particles is complex near the soil surface and more homogeneous in the subsurface (**Figure 4**). Undoubtedly,

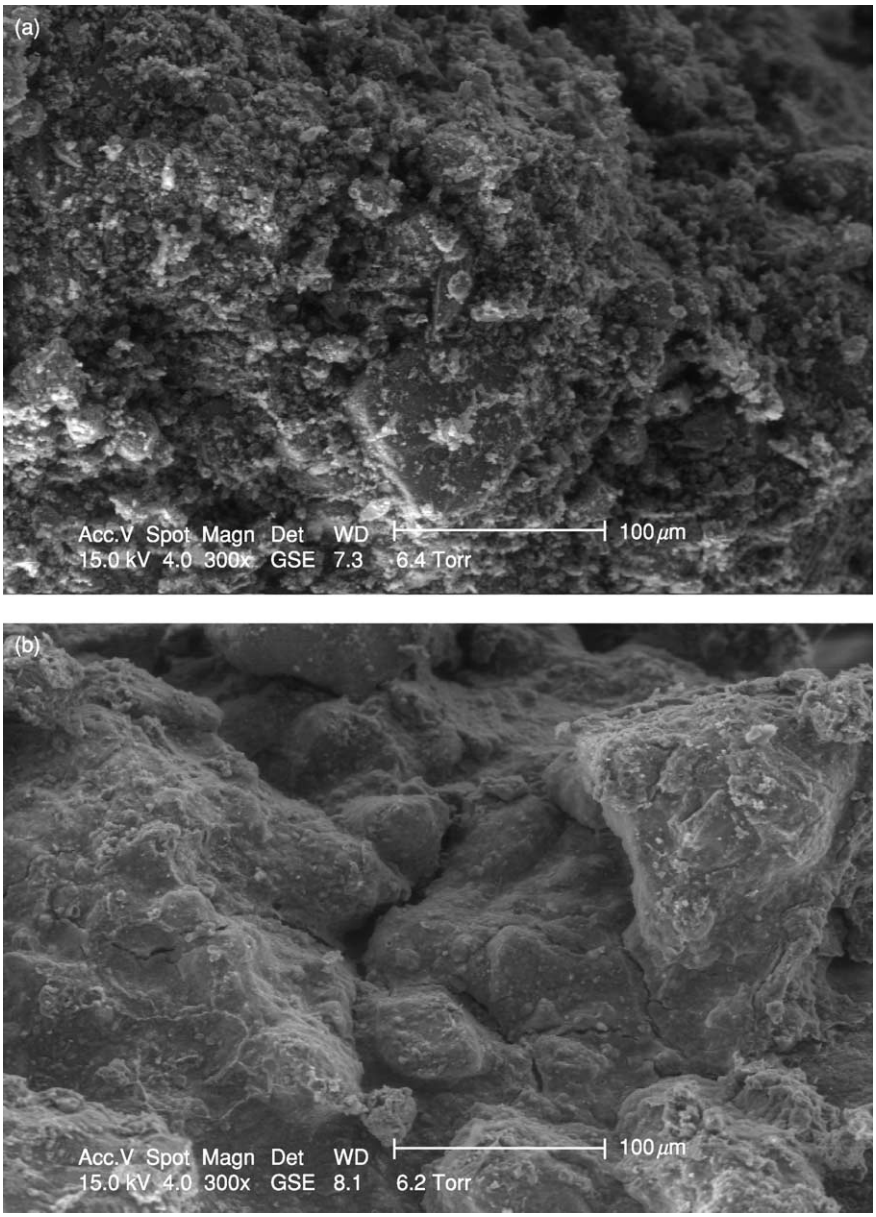


Figure 4 Environmental scanning electron micrographs of (a) surface and (b) subsurface vadose-zone materials at 300× magnification.

microbial growth and activity shape the physical appearance of vadose-zone particles and surfaces, but the relationship between gross surface textural morphology and microbial ecology is not yet quantified. Certainly, the physical composition of the vadose zone can influence the distribution and abundance of microbes: low-porosity rock such as basalt and higher-porosity sands and silts with low carbon content harbor fewer microbes as compared to buried soils (paleosols), which contain more organic carbon. Soil texture also influences the distribution of microbes, since higher clay contents are often associated with greater substrate availability due to the sorption of nutrients on to clay surfaces. However, at a given water content, soils with higher clay contents have lower soil-water potentials, restricting the rates of nutrient diffusion.

Modeling Vadose-Zone Microbial Processes

Mathematical models are useful for predicting the fate of nutrients and pollutants in the vadose zone. Although the vadose zone is not a well-mixed system, because of its heterogeneity it is sometimes modeled as a complete mix reactor where microbes are not

limited by diffusional or advective resupply of nutrients. In such cases, X, the concentration of growing microbes (n per cubic liter) increases according to a simple first-order relationship with respect to time:

$$\frac{dX}{dt} = \mu X \qquad [2]$$

where μ is specific growth rate (t^{-1}), where t is true.

Specific growth rate, for a particular microbe and for a given set of environmental conditions, varies with substrate concentration, as per the Monod equation:

$$\mu = \frac{\mu_{max} S}{K_s + S} \qquad [3]$$

where μ_{max} is maximum specific growth rate (t^{-1}), S is substrate concentration (per cubic milliliter), and K_s is substrate concentration at one-half μ_{max} (per cubic milliliter).

Equation [2] can be related to a change in substrate concentration over time by a yield coefficient, Y, where Y is mass of substrate utilized per mass of microbes. In that case, the rate of substrate depletion is:

$$\frac{dS}{dt} = -\mu X Y \qquad [4]$$

An equation that is analogous to eqn [3] but more applicable to substrate depletion by nongrowing microbes is the Michaelis–Menten equation:

$$v = \frac{v_{max} S}{K_m + S} \qquad [5]$$

where v is substrate depletion rate ($ml^{-3}\,t^{-1}$), v_{max} is maximum substrate depletion rate ($ml^{-3}\,t^{-1}$), S is substrate concentration (per cubic milliliter), and K_m is substrate concentration at one-half v_{max} (per cubic milliliter). Note that eqn [5] describes the variable v, whose units are different from the Monod specific growth rate, μ, because v describes the mass of substrate depleted per unit volume of catalysts (free enzymes or nongrowing microbes). When cells are nongrowing and do not experience mass transfer limitations:

$$\frac{dS}{dt} = -\frac{vX}{\rho_c} \qquad [6]$$

where ρ_c is microbial cellular density (per cubic milliliter).

Transport processes may limit the observed rate of microbial reaction in the vadose zone. At the scale of microbial biofilms (**Figure 5a**), the composite biofilm EPS and cells (**Figure 3a**) can significantly impede the diffusion of nutrients to biofilm bacteria. If we assume a steady state at that spatial scale, the

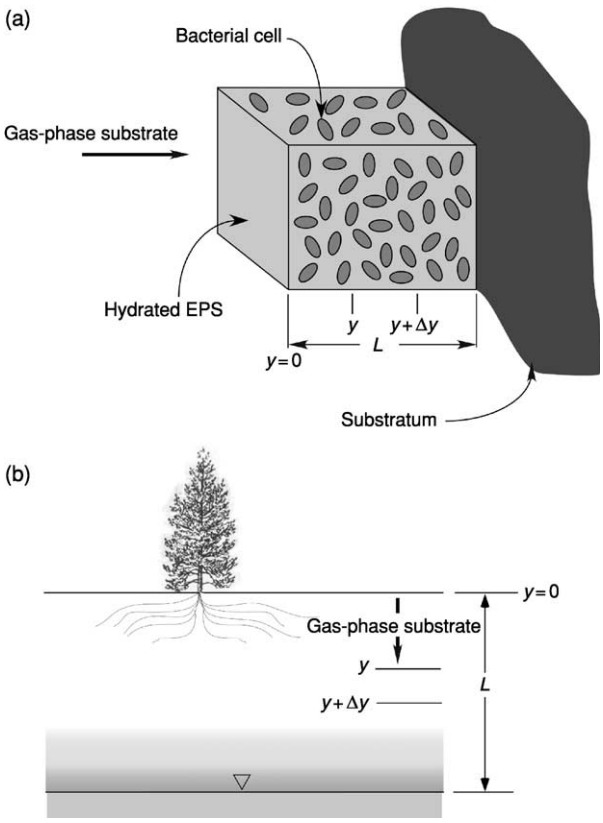


Figure 5 Idealized control volumes depicting (a) biodegrading bacterial biofilm and (b) the vadose-zone profile where, for modeling purposes, L is total depth, y is the distance into the system, and Δy is the change in distance. EPS, extracellular polymeric substances.

applicable differential equation that accounts for both diffusional substrate resupply and microbial reaction is:

$$D\frac{d^2S}{dy^2} = -\mu X Y \eta_i \qquad [7]$$

where D is effective diffusivity of limiting nutrient ($l^2\,t^{-1}$), y is distance along the axis of diffusion (L), and η_i is internal effectiveness factor, a dimensionless parameter that reduces microbial reaction to the rate constrained by diffusional rate limitation.

The boundary conditions for eqn [7] are: for $y = 0$, S is a maximum concentration occurring at the outside edge of the biofilm; and, for $y = L$, $dS/dy = 0$.

Equation [7] can be used to describe gas-phase substrate mass transfer and microbially mediated substrate reactions along a depth profile (**Figure 5b**), as long as the advection of gas through soil pores is relatively rapid and there are no rate-limiting sorption processes. Where mass flow through soil pores and/or sorption processes limit the rate of biogeochemical processes in the vadose zone, the

mathematical model includes explicit mass flow terms for the moving phase. Equations for describing mass flow of solution phase nutrients are typically based on the Darcy equation. Models of the latter kind will also include an applicable reaction term, e.g., eqn [4] or [6], if the purpose of the model is to describe the transport and fate of nutrients or pollutants.

Applications of Vadose-Zone Microbial Ecology

Pollutant Remediation

The vadose zone is frequently subject to pollution from either activities on the ground surface or below-ground leaking storage tanks. Depending on the physicochemical properties of the particular pollutant, the local environmental conditions, and the length of time following release, pollutants can be distributed through various regions of the vadose zone, as depicted in Figure 6. The vadose zone, either passively or actively, is frequently relied upon to attenuate pollutants that would otherwise migrate into groundwater. Active uses include leach fields for disposal of septic tank effluent and bioventing for the remediation of volatile organic pollutants. Passive uses rely upon the natural degradation capabilities of the vadose zone rather than an engineered system. Table 1 summarizes the various technologies that use the vadose zone to attenuate and remediate environmental pollution.

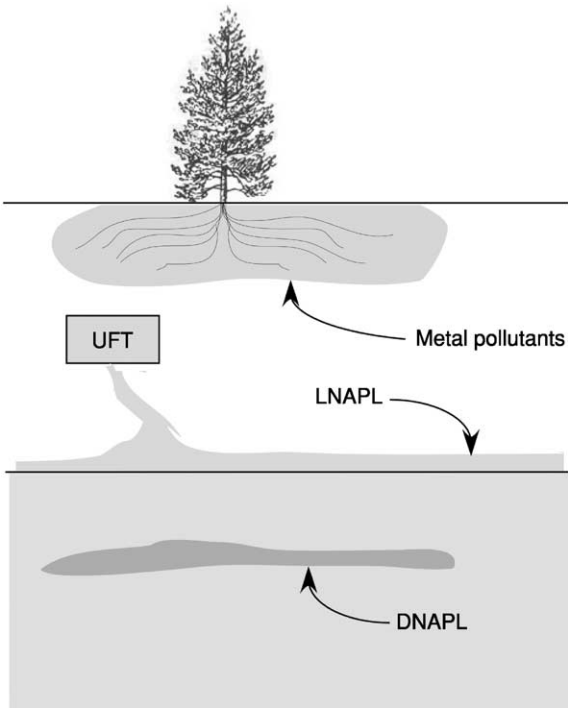


Figure 6 Profile of the polluted vadose zone depicting the orientation of a leaking underground fuel tank (UFT), light non-aqueous-phase liquid (LNAPL) and dense nonaqueous-phase liquid (DNAPL).

Pollution can alter vadose-zone microbial ecology. For example, sites used for waste disposal often have higher populations of culturable microbes and increased microbial activity relative to unimpacted sites. The effects of pollution on vadose-zone microbes and the effectiveness of microbial remediation depend on the characteristics of the specific pollutant and the nature of the microbial communities. Pollutant toxicity as well as bioavailability, i.e., the integrated mass transfer and biodegradation characteristics of the pollutant, are important determinants of the rate of pollutant bioattenuation in the vadose zone. Toxicity can be quantified in the laboratory using bioassays, but pollutants often occur in mixtures which have different toxicity characteristics to individual chemicals. Bioavailability largely depends on the physicochemical characteristics (e.g., solubility, partitioning characteristics, volatility) of the pollutant *in situ*.

The Vadose Zone as an Analog for Other Extreme Habitats

From the near surface to the extremely deep subsurface (several kilometers below the surface), the varying regions of subterranean Earth may physically and chemically mimic the harsh environments that exist on other planets. Although mostly beyond the scope of this chapter, a significant body of new knowledge has been gathered in the last two decades to confirm that microbes of varying taxonomies exist in the deep subsurface, living in high-temperature, high-pressure, and anoxic conditions. Much of the work that has generated the current knowledge of microbial ecology in the very deep subsurface has been funded by the US Department of Energy Subsurface Science Program. While much of that work has been focused on the microbiology of saturated zones, some research has been conducted on vadose-zone samples. Many questions in deep subsurface microbial ecology have been partly answered for specific sites, yet generalizable answers are still needed. How and when did microbes first inhabit deep vadose environments? How do they survive in environments that appear decoupled from aboveground primary productivity? What limits their proliferation: is it the lack of carbon or other nutrients, desiccation, unfavorable pH, and/or lack of terminal electron acceptors? Can subsurface microbiology reveal clues about the origins of life on Earth? Continued interest in these questions still fuels

Table 1 Selected waste treatment technologies and applications requiring vadose-zone microbes

Technology	*Application*	*Description*	*Comments*
Leach field	Septic tank effluent disposal	Organic and nutrient removal; effluent percolates from near-surface distribution pipes through underlying subsoil	Used frequently for rural domestic wastewater treatment; depending on percolation rates, nitrate and microbial pollutants may enter the groundwater
Phytoremediation	*In situ* bioremediation of metal and organic pollutants	Selected plants are cultivated in contaminated soil; plants sequester metals and enhance organic pollutant degradation in the rhizosphere	Low cost and high aesthetic value as compared to mechanical systems. Metal-enriched plant material requires special disposal; efficacy depends on the plant, pollutant, rooting depth, and extent
Bioventing	*In situ* bioremediation of volatile organic carbon pollutants	Air is piped belowground to enhance soil aeration, pollutant mobilization, and aerobic biodegradation	Used frequently for the remediation of refined petroleum that has leaked from underground storage tanks
Natural bioattenuation	Organic (volatile and semivolatile) and inorganic pollutant remediation	Natural physical, chemical, and biological processes distribute and possibly destroy waste; monitoring programs track performance	Used frequently, either intentionally or not; rate of remediation and pollutant fates are uncertain
Bioswale	Storm water	Storm water runoff collects in a constructed depression aboveground and percolates belowground, where dissolved and particulate pollutants are filtered	Used frequently in urban settings; depending on percolation characteristics, pollutants may enter groundwater directly. Sorption and biotransformation capacity for metals is uncertain
Land application	Disposal of treated wastewater	Effluent is applied by irrigation or spraying on to the soil surface and allowed to infiltrate for the removal of organic compounds and nutrients	Used for some domestic and livestock wastewater disposal; less effective in cold or wet climates. Pathogen distribution a potential problem when irrigating food crops

research in subsurface microbial ecology. Given that most of the biomass on Earth resides in the subsurface, it is imperative that we reach a better understanding of microbial life in the subsurface and the role of subsurface microbes in the Earth's biogeochemistry.

See also: **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics; **Mineral–Organic–Microbial Interactions**; **Thermal Properties and Processes**

Further Reading

Fyfe WS (1996) The biosphere is going deep. *Science* 273: 448.

Holden PA (2001) Biofilms in unsaturated environments. In: Doyle R (ed.) *Methods of Enzymology* 337: 125–143

Kieft TL and Brockman FJ (2001) Vadose zone microbiology. In: Fredrickson JK and Fletcher M (eds) *Subsurface Microbiology and Biogeochemistry*, pp. 141–169. New York: Wiley-Liss.

Konopka A and Turco R (1991) Biodegradation of organic compounds in the vadose zone and aquifer sediments. *Applied and Environmental Microbiology* 57: 2260–2268.

Madsen EL (1998) Epistemology of environmental microbiology. *Environmental Science & Technology* 32: 429–439.

Papendick RI and Campbell GS (1981) Theory and measurement of water potential. In: Parr JF, Gardner WR, and Elliot LF (eds) *Water Potential Relations in Soil Microbiology*, pp. 1–22. SSSA Special Publication No. 9. Madison, WI: Soil Science Society of America.

Richter DD and Markewitz D (1995) How deep is soil? *Bioscience* 45: 600–609.

Stephens DB (1996) *Vadose Zone Hydrology*. Boca Raton, FL: CRC Lewis Publishers.

Wan J, Wilson JL, and Kieft TL (1994) Influence of the gas–water interface on transport of microorganisms through unsaturated porous media. *Applied and Environmental Microbiology* 60: 509–516.

Whitman WB, Coleman DC, and Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences* 95: 6578–6583.

Viruses *See* Bacteriophage

VOLCANIC SOILS

G Uehara, University of Hawaii, Honolulu, HI, USA

Introduction

Molten magma extruded as liquid lava or ejected as cinders or ash from deep within the Earth adds to the variety of parent materials and soils we find on the landscape. Unlike their plutonic counterparts that cool and solidify slowly underground and are subjected to weathering only after uplift and erosion, volcanic rocks cool and solidify immediately upon exposure to the atmosphere. The faster cooling rate results in formation of glassy and finer rock textures in volcanic rocks so that plutonic and volcanic rocks with identical chemical composition can lead to formation of different soils. This is especially true for finely divided pyroclastic ejecta that cool even before they land. The porosity and texture of such cindery or ashy materials differ so greatly from those of other parent materials that glassy volcanic ash develops into soils that fall in the exclusive Andisol soil order.

In addition to rock texture and porosity, volcanic materials also vary in chemical and mineral composition. In fact it is the chemical composition of the magma that determines, to a large extent, the texture and porosity of the extruded material. Highly glassy and porous pyroclastic rocks are more common in high-silica magma that form acidic rocks than the less viscous, basic magma that erupt less explosively and produce basalitc lavas.

The higher viscosity of silicic magma results in the formation of steep-sided, cone-shaped volcanoes exemplified by Mount Fuji in Japan, whereas the basic, low-viscosity magma form broader, shield volcanoes exemplified by Mauna Loa in the island of Hawaii. Both types of volcanoes erupt pyroclastic materials and lava, but the silicic magma have more frequent explosive eruptions and produce more ash than the shield volcanoes. The 1980 eruption of Mount St. Helens in the USA illustrates the explosive nature of intermediate and silicic eruptions. Silicic eruptions often force nearby residents to evacuate the area, whereas sightseers flock to view molten lava pouring out from vents and fissures along rift zones of shield volcanoes.

The range of silica content of volcanic rocks is shown in Table 1. As silica decreases, the content of iron, magnesium, calcium, and sodium increases. These differences in chemical composition influence the rocks' susceptibility to weathering. In general, the basic rocks weather more rapidly owing to the ease with which sodium-, calcium-, magnesium-, and iron-bearing silicate minerals undergo hydrolysis and dissolution. The rate of weathering is even faster if the erupted material is ash.

Table 1 Chemical composition of volcanic rocks

	Alkali rhyolite	*Dacite*	*Andesite*	*Tholeiitic olivine basalt*	*Nephelinite*
SiO_2	74.57	63.58	54.2	47.9	39.07
TiO_2	0.17	0.64	1.31	1.65	3.86
Al_2O_3	12.58	16.67	17.17	11.84	12.82
Fe_2O_3	1.30	2.24	3.48	2.32	8.75
FeO	1.02	3.00	5.49	9.8	6.39
MnO	0.05	0.11	0.15	0.15	0.26
MgO	0.11	2.12	4.36	14.07	6.14
CaO	0.61	5.53	7.92	9.29	14.20
Na_2O	4.13	3.98	3.67	1.60	4.09
K_2O	4.73	1.40	1.11	0.54	2.07

Source: Nockolds SR (1954) Average chemical compositions of some igneous rocks. *Bulletin of the Geological Society of America* 65: 1007–1032.

Rock Weathering

As a rule, the weathering rate of volcanic rocks is influenced by its chemical composition, glass content, and particle-size distribution. Volcanic ash, for example, weathers very rapidly owing to its high glass content and high specific surface associated with smaller particle size. Volcanic glass, however, varies in chemical composition. Felsic glass is high in silicon, potassium, and sodium and is more resistant to weathering than mafic glass high in iron and magnesium.

Artifacts such as arrowheads made of obsidian show little sign of weathering even after centuries of burial, primarily because they are felsic in composition and have low surface-area-to-mass ratios. It is a different matter with ash particles, called shards, which geologists and soil scientists often use as stratigraphic markers and evidence of ancient volcanic eruptions. Shards have high surface-area-to-mass ratios so that volcanic ash particles still identifiable as shards after deposition and buried for thousands of years are more likely to be felsic, as mafic shards would long have been weathered to clay minerals.

Soil Formation

Any student of soil science who has walked over a lava flow or freshly deposited volcanic ash can sense

that soil formation on such materials would be rapid. Owing to the high porosity and large pores of most lavas and ash deposits, very little runoff occurs, and plants are able to establish themselves with relative ease. Nitrogen-fixing lichens are the first to colonize lava flows, and, under warm and humid climates, dense forests can cover lava fields in less than a century. In humid areas, the soils that first form are Folists of the Histosol order. Folists are well-drained Histosols that consist primarily of an organic surface horizon resting on rocks or fragmental materials. If the underlying rock is A'a lava the soil is often a Typic Udifolist, but if the rock is Pahoehoe lava the soil is commonly a Lithic Udifolist. This distinction between a Typic and Lithic Udifolist is crucial from an agricultural standpoint. A Typic Udifolist with its clinkery A'a substrate is ideal for use as orchards, but its Lithic counterpart, with an unbroken and smooth rock substrate, is not. Figure 1 shows papaya production on a Typic Udifolist.

In the case of volcanic ash, plants readily send roots deep on to the loose ash deposit, and rapid buildup of organic matter occurs throughout the rooting depth. The first recognizable soil would most likely be a Vitrand. As the name implies, the soil is an Andisol consisting mainly of unweathered glass fragments.

Folists and Vitrands do not remain unchanged for very long. While soils change little in a human lifespan, one can, given the right circumstances, understand how volcanic soils change with time. Such circumstances exist in the Hawaiian Islands which formed from the movement of the Pacific plate over a stationary hotspot. Islands that form from outpouring of lavas from the hotspot are displaced in a northeasterly direction, providing a chain of islands that is now the Hawaiian islands. The age of the rocks in the island chain ranges from 0.4 million years on the island of Hawaii to 5 million years on the island of Kauai. Ten of the 12 soil orders of Soil Taxonomy occur in the island chain. Unlike Folists and Vitrands that can form in less than a century, and are common on the youngest island of Hawaii, Oxisols and Ulitisols that require prolonged weathering and leaching to develop do not exist there. On Kauai, the oldest island, the opposite is true, wherein Oxisols and Ultisols dominate and Folists and Vitrands are nowhere to be found. The greatest diversity of soils is found on the island of Oahu, midway between Hawaii and Kauai. Here soils found on the old islands dominate, but young, posterosional volcanic eruptions have rejuvenated part of the landscape. A soil order that does not occur on the young islands, but peaks in the intermediate islands and slightly diminishes in abundance on the oldest island, is the Vertisol. The Vertisols form in valley floors on the dry, leeward side of the older islands. On the leeward side, rainfall is lower and the bases that accumulate in the valley floor provide the right conditions for smectite formation. Vertisols do not occur on the young islands because the valleys have not yet formed. In the oldest islands, Vertisols in the wetter zones lose their bases and eventually become kaolinitic.

Figure 1 Papaya production on a young volcanic soil, which was originally a Udifolist.

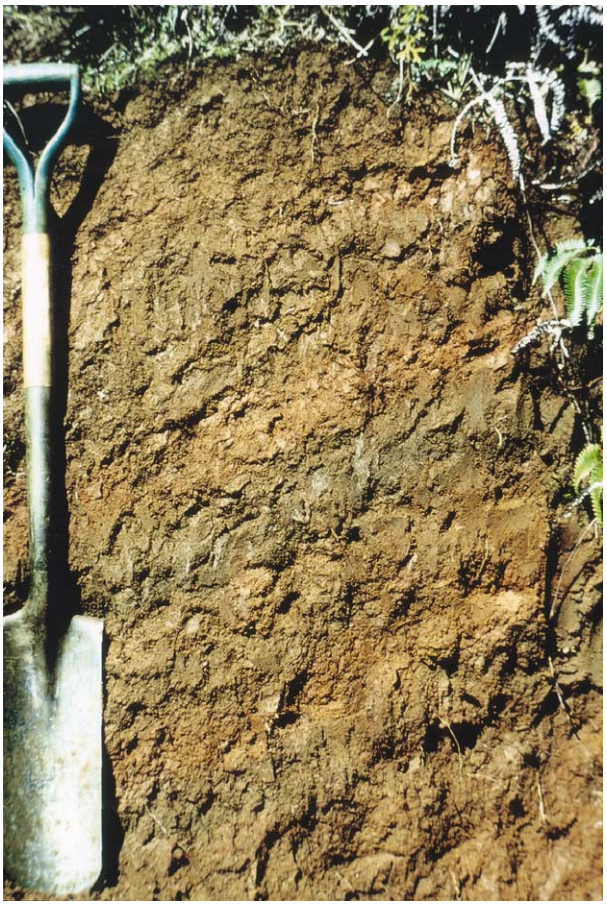


Figure 2 A highly weathered Andisol, an Acrudoxic Hydrudand.

The soils developed from volcanic ash age much faster. Vitrands quickly turn to Haplustands in the dryer zone and to Hydrudands in the wetter areas. The Hydrudands are some of the lightest soils in the world, with bulk densities less than $0.5\,cm^{-3}\,g$ and field water content by weight generally exceeding 100%. Owing to their porous nature, they are readily leached of bases and silica, leaving a residue high in hydrated iron and aluminum oxides. These highly weathered, prematurely old soils occur on the youngest islands (Figure 2).

Vertisols

A combination of basic parent material and low-leaching environment often serves as a necessary condition for Vertisol formation. Sediment deposited by streams and rivers cutting through basaltic rock often provides the first condition, and the semiarid tropics that circle the globe north and south of the equatorial humid tropics provide the second. The Vertisols of Sudan's Gezira region, situated between the White and Blue Nile, is such an example. There are no volcanoes in the Gezira, but the soils there owe their origin to the Blue Nile, which transports basaltic sediments from the highlands of Ethiopia to the Gezira. A similar situation exists in the experimental fields of the International Crop Research Institute for the Semi-Arid Tropics located near Hyderabad, India. The black Vertisols in the Hyderabad area sit side by side with red Alfisols to form a red-and-black soil complex. Here again, there are no signs of volcanoes nearby, for the Vertisols are products of sediment transported by water from basalts of the Deccan Flats.

In Table 2, the chemical composition and chemical properties of a Vertisol from Hawaii are shown. The Vertisols of the semiarid tropics are remarkably alike. Often referred to as tropical black earths, they have low organic carbon content and their dark color has been attributed to dark iron minerals and manganese coatings. Their chemical and physical properties are mainly determined by the high clay and smectite content.

Oxisols

A key feature of Oxisols is the low cation exchange capacity (CEC) of their clay. This condition is most likely to occur when the clay fraction is low in layered silicate clay and rich in iron oxides. Volcanic rocks low in silica and rich in ferromagnesium minerals weather rapidly to Oxisols under warm and humid conditions. While time and intense leaching can produce Oxisols from a variety of materials, volcanic materials, due to their porous nature, produce Oxisols with greater ease. The chemical composition and selected properties of an Oxisol formed from basic volcanic rock are shown in Table 3. The extreme weathering of the soil is indicated by its placement in the Acrudoxic subgroup and also by the rise in the $1\,mol\,l^{-1}$ KCl pH in the oxic horizons midway in the profile. In the oxic horizons the pH measured in $1\,mol\,l^{-1}$ KCl is higher than the pH measured in water, indicating that the net charge of the soil material is positive. Soils with these characteristics are infertile, but can be made productive with phosphorus fertilization, as well as liming with calcium silicate rather than calcium carbonate.

Ultisols

From a taxonomic standpoint, 'volcanic soil' is not a very useful term as a great variety of soils can have volcanic origins. Ultisols can form from volcanic rocks where rainfall is sufficiently high to leach bases, and conditions exist that favor activation and

Table 2 Chemical composition and properties of a Vertisol developed from basaltic alluvium (Lualualei series: fine, smectitic, isohyperthermic, Typic Gypsitorrerts)

												Chemical properties								
	Chemical composition (% of whole soil)													*Exchangeable bases (cmol kg^{-1})*				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*(cmol kg^{-1})*	*H_2O*	*1 mol l^{-1} KCl*
0–3	30.54	7.71	19.29	26.26	0.30	1.50	1.57	0.26	0.07	0.50	11.63	0.66	0.08	17.1	15.2	0.8	1.4	34.1	7.1	5.9
3–25	31.74	7.59	20.34	25.81	0.33	1.07	1.72	0.28	0.08	0.48	10.81	0.42	0.06	15.1	15.1	1.3	0.4	32.9	7.2	5.4
25–35	33.91	7.52	19.04	26.28	0.45	1.57	0.62	0.29	0.05	0.51	10.87	0.21	0.05	14.6	13.4	2.5	0.2	32.9	7.2	5.4
35–75	33.54	7.59	19.57	25.82	0.43	1.72	0.34	0.37	0.06	0.36	10.54	0.17	0.04	16.0	10.5	3.9	0.2	32.1	6.8	5.2
75–123	32.25	6.18	20.02	24.24	0.33	1.58	2.56	0.44	0.04	0.38	11.39	0.17	–	73.1	9.3	7.2	0.2	29.8	5.6	4.9
>123	31.52	6.57	20.39	22.37	0.38	1.50	4.39	0.49	0.07	0.29	11.56	0.16	–	73.8	10.7	8.4	0.3	30.6	5.8	5.0

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 3 Chemical composition and properties of an Oxisol developed from nepheline basalt (Kapaa series: very fine, sesquic, isohyperthrmic, Anionic Acrudox)

												Chemical properties								
	Chemical composition (% of whole soil)													*Exchangeable bases (cmol kg^{-1})*				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*(cmol kg^{-1})*	*H_2O*	*1 mol l^{-1} KCl*
0–30	6.9	7.6	24.6	38.5	0.06	0.32	–	0.03	0.28	0.63	21.7	3.92	0.22	0.7	0.4	0.20	0.40	15.9	5.0	4.5
30–40	4.4	7.2	28.1	39.3	0.09	0.25	–	0.03	0.11	0.55	19.9	1.46	0.07	0.2	0.0	0.10	0.20	3.7	5.5	5.7
40–63	4.1	8.1	28.8	38.5	0.08	0.08	–	0.03	0.07	0.58	19.7	1.09	0.04	0.4	0.1	0.20	0.10	2.6	5.5	5.7
63–90	4.3	7.5	32.9	34.3	0.17	0.13	–	0.03	0.07	0.63	20.0	0.64	0.02	0.1	0.1	0.20	0.10	1.9	5.8	5.7
90–123	7.1	7.5	30.9	35.2	0.15	0.11	–	0.03	0.07	0.57	18.9	0.46	0.02	0.8	0.1	0.20	0.10	3.7	5.3	5.4
123–150	13.2	8.4	24.8	36.8	0.15	0.38	–	0.03	0.07	0.61	15.5	0.48	0.02	1.0	0.1	0.30	0.10	6.1	5.5	4.9

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Service Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

illuviation of clay. It is not unusual to find two soil orders, in close proximity, forming from the same parent rock. Ultisols and Oxisols, for example, have developed synchronously on the same geomorphic surface on the volcanic island of Kauai in the Hawaiian Islands. The Ultisols, however, always occur on the knick points of the landscape, where soil creep activates and mobilizes clay along sheer planes in the subsoil. The chemical composition and properties of such an Ultisol are given in Table 4. The same shearing action leads to formation of Alfisols in drier locations, where leaching is less intense and basic cations are retained in the profile.

In climates where Ultisols dominate, water is plentiful but nutrients are in short supply, whereas, in semiarid regions where Alfisols predominate, the land is fertile but short of water.

Volcanic soils may have special meaning to people living in humid regions where soils that under normal circumstances would be Ultisols are maintained in an Alfisol-like condition by periodic additions of nutrient-rich ash from a neighboring volcano. The advantage of having an adequate supply of water and nutrients in contrast to having one or the other was, for our ancestors, the difference between being able to form creative, sedentary communities in contrast to simply surviving by hunting and gathering. Thus it may not be the soil itself but the presence of an active volcano nearby that makes volcanic soils what they are.

Andisols

If one had to choose a soil to represent the class of soils we call volcanic, it would be an Andisol. This is because all Andisols are volcanic soils, but the reverse is not true. Andisols have two characteristics which make them unique. They contain high amounts of short-range-order and noncrystalline materials, and have low bulk densities. Two short-range-order minerals commonly found in highly leached and weathered Andisols are allophane and imogolite. These minerals have high specific surfaces (approximately equal to $1000\,m^2g^{-1}$) and pH-dependent charge. The high specific surface of the inorganic colloids enables the soil to sequester large quantities of organic matter. The prefix 'Ando' on Andisol is the Japanese word for 'dark,' used to describe the low chroma of organic-rich volcanic ash soils of Japan. In Japan, where the ash is generally high in silica and low in iron, organic matter imparts a nearly black color to the soil. In Andisols developed from low-silica, high-iron and -magnesium ash, the iron oxide masks the black color of organic matter to produce a dark-brown soil (Figure 2).

The chemical composition and selected chemical properties of two Andisols, one developed in a perudic, and the other in an ustic moisture regime, are shown in Tables 5 and 6. Both soils are excellent agricultural soils, with the wetter soil requiring careful nutrient management and the drier soil requiring irrigation for profitable farming.

Human Interactions with Volcanic Soils

Our human ancestors may have avoided extinction and evolved by adapting to the diverse and ever-changing volcanic environment of the African Rift. The rich soil, constantly rejuvenated by sprinkling of volcanic ash, even today supports large herds of herbivores and their predators, and some of the highest nonurban human population densities in Africa.

When modern humans exited Africa through the Rift zone, then crossed Asia and the Bering Straits, they encountered the North, Central, and South American segment of the 'Ring of Fire' that circles the Pacific. Like the rich alluvial soils of the Nile, Tigris, Euphrates, and Indus rivers that enabled humans to shift their energies from hunting and gathering to creative endeavors, the fertile and well-watered volcanic soils of Central and South America provided the new settlers with the means to do the same.

Even today, the influence of volcanic soils on human society is evident. In Indonesia, for example, nearly half of its 200 million people live on the tiny volcanic island of Java. Efforts to move whole communities from Java to the larger and less densely populated neighboring islands have not been easy. Villagers accustomed to farming Andisols with characteristics similar to the one shown in Table 5 found it difficult to survive on Ultisols with features similar to those shown in Table 4.

Today science and technology enable farmers to understand and overcome soil constraints, but only a century ago soil fertility determined to a large degree outcomes of human interactions with the environment. Unfortunately this is still true in many developing countries, where science-based soil management practices have yet to be adopted. The volcanic soils of Africa, Asia, and Latin America that helped people in the past now need our help to perform and contribute to human well-being as they have in the past.

Table 4 Chemical composition and properties of an Ultisol developed from basalt (Haiku Series: very fine, isohyperthermic, Ustic Palehumult)

	Chemical composition (% of whole soil)											Chemical properties								
														Exchangeable bases ($cmol\,kg^{-1}$)				CEC	pH (1:5)	
Depth (cm)	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO_2	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	C (%)	N (%)	Ca	Mg	Na	K	($cmol\,kg^{-1}$)	H_2O	1 $mol\,l^{-1}$ KCl
0–18	16.8	11.6	7.6	46.3	0.15	0.97	–	0.06	1.15	0.55	15.9	3.08	0.26	–	0.6	0.1	0.3	15.9	5.1	4.1
18–33	16.8	11.2	11.1	43.0	0.12	0.77	–	0.07	1.23	0.51	15.5	2.79	0.23	–	0.6	0.1	0.2	14.3	5.0	4.0
33–45	14.1	7.9	15.2	42.7	0.07	0.66	–	–	0.99	0.57	17.6	1.97	0.16	0.2	0.8	0.2	0.1	15.5	4.9	4.1
45–70	11.9	6.5	23.2	36.9	0.09	0.78	–	0.02	0.65	0.52	19.1	1.78	0.13	0.6	0.7	0.4	0.1	12.2	5.2	4.4
70–98	18.9	5.2	27.4	28.4	0.12	0.98	–	–	0.31	0.39	18.1	1.08	0.08	1.0	0.8	1.0	0.1	12.7	5.1	4.0
98–155	17.3	5.7	30.1	26.4	0.12	1.62	–	–	0.11	0.42	17.6	0.91	0.06	0.6	0.2	0.8	0.1	12.0	5.0	4.0
155–175	17.0	5.2	32.9	24.7	0.11	1.54	–	–	–	0.32	18.0	0.74	0.04	0.4	0.3	>0.8	0.1	12.4	4.9	4.0

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii.* Soil Survey Investigations Report No. 29. USDA Soil Conservation Services/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 5 Chemical composition and properties of a Haplotorrand developed from basaltic ash (Waikaloa series: medial, amorphic, isothermic, Typic Haplotorrands)

	Chemical composition (% of whole soil)											Chemical properties								
														Exchangeable bases ($cmol\,kg^{-1}$)				CEC	pH (1:5)	
Depth (cm)	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO_2	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	C (%)	N (%)	Ca	Mg	Na	K	($cmol\,kg^{-1}$)	H_2O	1 $mol\,l^{-1}$ KCl
0–13	32.3	3.7	21.1	14.5	0.44	3.30	3.44	1.76	0.80	0.63	18.3	7.53	0.32	27.9	7.8	0.40	4.80	81.0	6.6	5.7
13–25	32.9	3.5	23.9	15.0	0.42	2.86	3.55	2.24	0.87	0.61	14.7	3.69	0.16	28.0	6.4	0.40	5.60	78.0	7.1	6.2
25–50	36.1	3.2	24.6	15.4	0.43	2.03	2.45	2.26	0.88	0.44	11.7	1.88	0.11	32.6	7.1	0.80	4.80	88.0	7.3	6.2
50–63	38.6	3.4	24.0	15.1	0.47	2.07	3.44	2.25	0.82	0.38	10.0	1.02	0.09	34.6	9.8	1.10	3.00	97.0	7.5	6.3
63–78	40.4	4.3	22.5	14.8	0.43	2.04	3.17	2.41	0.92	0.36	9.2	0.77	–	33.3	12.8	1.60	0.60	93.0	7.6	6.4
78–98	44.8	3.2	20.9	13.9	0.40	2.08	2.11	2.42	1.42	0.26	8.6	0.31	–	33.0	15.9	2.20	0.30	>100	7.8	6.5
98–125	51.9	2.5	20.4	6.7	0.29	1.00	2.61	4.20	2.70	0.09	8.1	0.15	–	34.0	12.4	5.40	0.50	>100	8.1	6.8
125–163	55.7	1.6	21.1	4.8	0.31	0.77	1.68	4.19	2.66	0.09	7.6	0.11	–	33.3	14.4	6.10	1.30	>100	8.2	6.9

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii.* Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 6 Chemical composition and properties of a Hydrudand developed from basaltic ash (Hilo series: medial over hydrous, ferrihydritic, isohyperthermic Acrudoxic Hydrudands)

	Chemical composition (% of whole soil)											*Chemical properties*								
														Exchangeable bases ($cmol\,kg^{-1}$)				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*($cmol\,kg^{-1}$)*	*H_2O*	*$1\,mol\,l^{-1}$ KCl*
0–40	12.90	5.24	24.32	27.56	0.15	–	–	–	–	–	27.30	5.30	0.41	2.0	1.8	0.1	0.1	67.6	5.8	5.6
40–53	8.78	5.58	34.40	26.26	0.25	–	–	–	–	–	25.82	3.06	0.22	2.2	0.6	–	0.1	68.4	6.1	6.2
53–65	8.58	5.13	35.92	26.10	0.34	–	–	–	–	–	24.90	2.61	0.17	2.1	0.3	–	–	60.4	6.3	6.5
65–80	8.54	5.30	35.04	26.62	0.37	–	–	–	–	–	25.50	2.67	0.19	1.2	0.6	–	–	62.9	6.3	6.5
80–123	10.82	5.38	33.78	26.38	0.36	–	–	–	–	–	24.46	2.80	0.20	1.8	0.4	–	–	60.5	6.3	6.4
123–140	10.73	5.60	32.56	25.97	0.25	–	–	–	–	–	24.13	2.57	0.16	1.8	0.5	–	–	65.5	6.4	6.4
140–168	10.68	5.68	30.04	28.95	0.25	–	–	–	–	–	24.43	2.24	0.15	2.5	0.5	–	–	71.0	6.3	6.4

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

List of Technical Nomenclature

Acrodoxic Hydrudands — These are highly weathered soils derived from volcanic ash that occur in high-rainfall areas. The subsoil material contains more that 100% water on a weight basis and dries irreversibly when exposed to the atmosphere

Udifolists — These are well-drained organic soils that occur mainly in the high-rainfall areas of the Hawaiian Islands on forested lava flows

Further Reading

Beinroth FH, Uehara G, and Ikawa H (1974) Geomorphic relationships of Oxisols and Ultisols on Kauai, Hawaii. *Soil Science Society of America Proceedings* 38: 128–131.

Bleeker P and Parfitt RL (1974) Volcanic ash and its clay mineralogy at Cape Hoskins, New Britian, Papua New Guinea. *Geoderma* 11: 123–135.

Christiansen RL and Peterson DW (1981) Chronology of the 1980 eruption activity. In: Lipman PW and Mullineaux DR (eds) *The 1980 Eruptions of Mount St. Helens*, pp. 17–30. Geological Survey Professional Paper 1250. Washington, DC: US Government Printing Office.

Clague DA, Dalrymple GB, Wright TL *et al.* (1989) The Hawaiian/Emperor chain. In: Winterer EL, Hussong DM, and Decker RW (eds) *The Geology of North America*, vol. N. *The Eastern Pacific Ocean and Hawaii*, pp. 187–287. Geological Society of America.

Dethier DP, Pevear DR, and Frank D (1981) Alteration of new volcanic deposits. In: Lipman PW and Mullineaux DR (eds) *The 1980 Eruptions of Mount St. Helens*, pp. 649–665. Geological Survey Professional Paper 1250. Washington, DC: US Government Printing Office.

Jones RC, Babcock CJ, and Knowlton WB (2000) Estimation of the total amorphous contact of Hawaii soils by the Rietveld method. *Soil Science Society of America Journal* 64: 1100–1108.

Lowe DJ (1986) Controls on the rate of weathering and clay mineral genesis in airfall tephras: a review and New Zealand case study. In: Coleman SM and Dethier DP (eds) *Rates of Chemical Weathering of Rocks and Minerals*, pp. 265–330. New York: Academic Press.

Nockolds SR (1954) Average chemical compositions of some igneous rocks. *Bulletin of the Geological Society of America* 65: 1007–1032.

Parfitt RL (1975) Clay minerals in recent volcanic ash soils from Papua New Guinea. *Bulletin of the Royal Society of New Zealand* 13: 241–245.

Plucknett DL (1971) The use of soluble silicates in Hawaiian agriculture. *University of Queensland Papers* 1: 203–223.

Raymundo ME (1965) *The Properties of the Black Earths of Hawaii*. PhD Dissertation. Honolulu, HI: Department of Agronomy and Soil Science, University of Hawaii.

Saigusa M, Shoji S, and Kato T (1978) Origin and nature of halloysite in Ando soils from Towada tephra, Japan. *Geoderma* 20: 115–129.

Shoji S, Fujiwara Y, Yamada I, and Saigusa M (1982) Chemistry and clay mineralogy of Ando soils, Brown forest soils, and Podzolic soils formed from recent Towada ashes, northeastern Japan. *Soil Science* 133: 69–86.

Stevens KF and Vucetich CG (1984) Pedogenic weathering of Upper Quaternary tephras in New Zealand. Isovolumetric geochemical evidence of cation movement. *Chemical Geology* 47: 285–302.

Uehara G and Gillman GD (1980) Change characteristics of soils with variable and permanent change minerals. I. Theory. *Soil Science Society of America Journal* 44: 250–252.

USDA–SCS (1978) *Soil Survey Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

USDA–SCS (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Washington, DC: Natural Resource Conservation Service.

Wada K (1985) The distinctive properties of Andosols. In: Stewart BA (ed.) *Advances in Soil Science*, vol. 2, pp. 173–229. New York: Springer-Verlag.

W

WAKSMAN, SELMAN A.

H B Woodruff, Soil Microbiology Associates, Watchung, NJ, USA

Selman A. Waksman, professor and research specialist at Rutgers University, the State University of New Jersey, USA, was a highly creative soil scientist. From the time of preparation of his senior college thesis in 1915 until his death in 1973 he specialized in the study of the microbes of the soil. During that 58-year period, he published 447 scientific reports. They covered the full range of the early development of soil microbiology, extending from the descriptive phases of the microorganisms of the soil to details of their physiology, from investigations of their influence on soil fertility to studies of their interactions together within soils. His research on the microbial associations led to a demonstration that soil microorganisms produce antagonistic agents. He proposed the name 'antibiotic' for such products. Certain of the antibiotics derived from soil microbes, for example streptomycin, discovered in his laboratory, have proven useful in the treatment of infections of humans, animals, and plants.

Selman Waksman was a prolific writer, publishing papers in a wide range of scientific journals, in several languages. He was author or co-author of 28 books. His *Principles of Soil Microbiology*, an 897-page volume, the first edition published in 1927, for years became the standard textbook of his field. He guided 78 students to graduate degrees. He was an inspiring lecturer. His presentations were filled with stories of his relationships with leaders of his field and were avidly followed by his students and by general audiences. He was beloved by his students, who spoke of him familiarly as 'The Old Maestro.' Waksman's name is generally included with Winogradsky, Omeliansky, Beijerinck, and the Americans Lipman and Thom in lists of the pioneers of soil microbiology.

Illustrative of the breadth of Selman Waksman's interests and the scope of his accomplishments is the quantitative record of his publications, written in association with his students. Fifty-eight were descriptive reports of the microbes of the soil, 53 of the enzymology and biochemistry of soil microbes, 11 on sulfur oxidation, 35 on peat, composts, and humus, 18 on microbes of the sea, 20 on taxonomy and methodology, eight on organic acids and fermentation products of molds, 18 on general physiology, and 31 were biographies and discussions on the philosophy of science. Finally, 195 were on antibiotics, including the announcement of 17 new antibiotics discovered in his laboratory.

Waksman was greatly honored, based on his productive career. By the time of his 80th birthday, he had received 22 honorary degrees, 12 from universities of Europe, Asia, and South America. He had been granted 63 prizes, awards, and medals. He was elected to the prestigious National Academy of Sciences, served as President of the American Society for Microbiology and was awarded the 1952 Nobel Prize in Physiology or Medicine.

What were the background, family, experiences, and education that led to such a creative career?

Selman Abraham Waksman was born 8 July, 1888 according to the old Russian calendar (now 22 July) in Novaia-Priluki, a small village in the Russian Ukraine. The villages and small towns of western Russia and Ukraine were populated by recently liberated serfs, who subsisted by small-scale farming and livestock husbandry, and by Jewish artisans and tradesmen who engaged in crafts, shop-keeping, and the processing and marketing of farm and forest products. These activities influenced the young child, and provided a motivation for his later decision to specialize in unraveling the chemistry of living things.

The inhabitants of Novaia-Priluki were poor. There was a rigid separation in culture and approaches to life between two groups, the Russian and the ethnic Jewish populations. The Jews stressed education, but a specialized religious education, directed by the synagogue Rabbinate, devoted almost exclusively to the Bible and the Talmud. The adult men spent much of their free time in study, discussion, and

arguments in their religious halls. Yiddish was their universal language, only a minority learning the intricacies of Hebrew and developing the ability to write in Russian. Waksman's mother was more accomplished than most. She vowed that her son Selman would become even more so, therefore she strove constantly to see that objective fulfilled.

No formal state schools existed in Novaia-Priluki. The sole opportunity for a broad education was private tutors, who either taught small groups or, if financial circumstances permitted, individual students. Waksman was taught privately. He learned the rudiments of language and mathematics and limited geography. He had an exceptional ability to memorize almost instantly anything read, thus progressed more rapidly than many of his student associates. Somewhat chagrined by the advantages offered him, compared to his friends, he, with some associates, organized a volunteer school, in which they passed on to economically deprived young folks the results of their studies. It was a most fortunate experience. It not only honed the lecturing skills Waksman exemplified later as a college professor, but provided the means whereby he could remove himself from the drudgery of mere subsistence.

Once he had sampled the full extent of educational opportunities available in his home town, Waksman moved to a larger town, Zhitomir, capital of a nearby prefecture. It had a Gymnasium for advanced studies, so accomplished teachers were available who would accept students for private study. From Zhitomir he progressed to an even larger city, Odessa, where highly qualified professors were located. In these locations at a distance from home, his prior experience as a volunteer instructor became significant. He offered himself as an instructor in the daytime, thus earning funds to cover living costs and tutors' fees, taking his own lessons during evening hours.

After years of study, at 22 years of age, Waksman felt qualified to take the 2-week extensive written and oral exams whose passage was required to obtain permission to enter a university. He did so, and passed.

Now that the essential qualification had been earned, entry into a university became possible. For Jews in Russia, however, that was very difficult to accomplish because of a quota system. Waksman realized that, graduating in Russia as a specialist in his chosen area, he would face few opportunities for employment. Therefore, he gave serious consideration to emigrating to the USA. Close relatives had done so and their letters contained descriptions of the opportunities for advancement that existed in America. The turning point came with the death of his mother. His father quickly remarried. Feeling that he no longer had a home to return to and with little opportunity available to advance in his objective for life work, he emigrated.

His port of entry to the USA was Philadelphia. Relatives, the Kornblatts, had advanced from tenant farmers to ownership of a farm specializing in poultry. They lived nearby in Metuchen, New Jersey. He joined them, working on the farm to earn his board. For the first time he became directly responsible for agricultural practices, obtaining best yields of vegetables and enhancing egg production.

The experience aiding the Kornblatts on their farm strengthened Waksman's intent to study the application of chemistry to biology. He applied to Columbia University in New York City to enter a premedical program, was accepted, but had no means to cover the costs involved. He visited nearby Rutgers College, at that time a private school that had associated with it federally supported Schools of Engineering and of Agriculture, supported financially through the Federal Morrill Land Grant Act, and the site of an agricultural experiment station supported by the state of New Jersey. There he met Jacob G. Lipman, professor of soil chemistry and microbiology, an excellent experimentalist and skilled leader, who was slated to become Director of the Agricultural Experiment Station. In Lipman, Waksman found his mentor. In soil microbiology he discovered the topic he had been striving toward all his life, providing him the opportunity to study the relationship of chemistry to living systems. He applied for a scholarship at Rutgers College, received one and became an undergraduate student, the sole person in his class majoring in soil microbiology.

Waksman was highly successful in his studies. Despite his limited knowledge of the English language, he received high grades. As he entered his senior year, he had already fulfilled most of the requirements for graduation. Therefore, a senior thesis became his prime objective for the year. His life as an experimentalist in the field of soil microbiology had begun.

The prior investigations in soil microbiology had been directed primarily toward individual organisms, the nitrogen-fixing rhizobia and azotobacter, or soil organisms that caused plant diseases. Waksman chose to apply a holistic approach to the topic. He enumerated the types of organisms existing in soils, defined the changes in numbers and types that occurred at different soil depths, among differing soil types, at different temperatures and moisture levels. From trenches he had dug on the college farm, using good aseptic techniques, he isolated and classified hundreds of organisms. At the end of his senior year, he tabulated his results. They were presented orally by his professor, Jacob Lipman, at a Society of American Bacteriologists meeting. An abstract of the talk was

published in the first issue of the newly established *Journal of Bacteriology*.

Highly appreciative of Waksman's ability to perform science independently, Lipman encouraged him to remain at Rutgers as a graduate student, working towards an MS degree in soil microbiology. He provided the financial wherewithal to make it possible. Waksman had become fascinated with the small, hard, slowly growing microbes that appeared almost universally in his agar platings of soils at high dilutions, often by count a dominant proportion of the soil population. They were known to others, but were very little studied. They had been generally ignored because of the smallness of their colonies and slowness of their growth. These microorganisms were the actinomycetes, the cells of which are typical of bacteria, approximately the same size in diameter, but in morphology filamentous, resembling the molds. Waksman decided to concentrate on them as an MS thesis objective and with an associate he isolated many, studying each in detail. When the isolates were typical of past descriptions, he applied the established species names, but often he found cultures that differed distinctly from previously known actinomycetes. He created names for these, often related to his personal experiences: *lipmanii* and *halstedii*, his favorite teachers; *rutgersensis* and *californicus*, his university and choice for his PhD education; *bobili*, the nickname for his wife.

Upon completion of his MS program, Waksman realized the need to enhance his knowledge of biochemistry, a special branch of chemistry, and to do so he accepted a student appointment to the Department of Biochemistry at the University of California in Berkeley. As subjects for his investigations there, he took with him a selection of the cultures isolated at Rutgers and with them concentrated on defining the enzymes they produced.

Upon receiving his PhD degree at Berkeley, Waksman was offered a position at Rutgers as a college instructor in soil microbiology, with simultaneous appointment as researcher in the Agricultural Experiment Station, at a small salary. He returned to New Brunswick and designed a graduate-level lecture program for soil microbiology. He continued his research studies on the actinomycetes and extended them to include the fungi and protozoa of the soil. For a few years he also worked part-time at a commercial concern to supplement his salary.

Assertive in his approaches, he came into conflict with recognized specialists in the field of soil microbiology. H.J. Conn of Cornell University was convinced that the fungi observed in soil platings were of little significance, that they were derived from wind-blown fungal spores. E.J. Russell in England, a member of the prestigious Rothamsted Experiment Station, had proposed that soil protozoa are controlling members of soil populations because they consume soil microorganisms as a source of nutrition. Waksman objected to both claims, did research to prove the correctness of his ideas, and, through his publications on soil microorganisms he became well-known in the field. He, with the aid of his students, greatly expanded the understanding of microbes of the soil, especially the actinomycetes. New students were attracted to his laboratory, as well as many non-degree-seeking visitors who desired to work several weeks or months conducting studies in association with him.

By 1924 Waksman had completed studies defining the nature of the common members of the soil population and he faced the question of an objective for his future program. To search for the answer, he chose to tour the well-known laboratories of Europe, to spend hours in conversation with specialists in his field and allied subjects. For 5 months he did so, visiting a variety of persons. Mostly, he was received with enthusiasm by those who appreciated his scientific record. Sometimes he had to negotiate an entry where unknown, but always he entered into detailed discussions, sometimes somewhat heated if he disagreed with accepted dogma. He had firmly fixed ideas about many chemical and biochemical processes, a few not in tune with current thinking, and he avidly promoted them.

Waksman came home thoroughly refreshed, convinced that his holistic approach to the study of soil microbes was correct. He had chosen the new objective of defining what the grand mixture of organisms present in soils was accomplishing. What products did they produce in the soil? How did they affect soil fertility? He became convinced that a significant factor in defining fertility was humus, the resistant organic matter resulting from the microbial attack on plant residues, composts, other organic substances that come into contact with soils. Using a system of analysis, dubbed proximate analysis, he followed the changes in organic matter in the soil, then followed the slower attack on humus itself by other microbes, causing its gradual breakdown. He became convinced that the resistant organic complexes in soils have fertility-defining significance.

He extended his studies to follow details of the microbial processes that occur during the composting of waste agricultural products. He also investigated the special geographic situations in which humus is broken down less rapidly than it is formed, resulting in the formation of peat bogs. Here, likewise, he stated his opinions assertively, which eventually brought him into conflict with members of the

chemistry section of his department. In contrast to his belief that humus is a significant substance in soil fertility, some chemists at Rutgers, led by the excellent experimentalist Jacob Joffe, differed. Joffe had previously been Waksman's associate in research on sulfur-oxidizing organisms and was co-discoverer with him of *Thiobacillus thiooxidans*. He and fellow chemists considered soils to be the equivalent of ion exchange substances, binding essential plant nutrients as they entered the soil, gradually releasing them to support plant growth, with the soil organic fraction being relatively insignificant for fertility. In many aspects, the arguments advanced by members of both sections of the department had validity, but their inability to come to an agreement led to a permanent subdivision of the Soils Department into two separate departments, Soil Chemistry and Soil Microbiology. Waksman was appointed chairman of the latter.

In his new capacity, Waksman had one professional associate, Robert Starkey, one graduate student position, and one technician funded permanently by the University and its associated Experiment Station. He greatly expanded his research program by receiving financial support from private funding agencies and occasionally by government grants. His work was also supplemented by accepting commercial support for projects with practical objectives that had relation to his ongoing programs. One that had special significance was support offered by Merck & Co., Inc., at that time a commercial concern in New Jersey especially interested in producing bulk chemicals for pharmaceutical and industrial uses. Merck directors felt the need to expand into organic acid manufacture and provided a fellowship to Waksman to investigate organic acid production by soil microbes. Initially, the effort was directed at fumaric acid, a product useful in the dry-cleaning industry. It was soon extended to citric acid, which was then being produced for the beverage industry by culturing *Aspergillus niger* in pure culture in large trays. Waksman and his students isolated from soils new highly efficient organisms that produced the desired acids. Because of close contacts with A.J. Kluyver of the Technische Hoogschool in Delft (Waksman's associate Robert Starkey had recently spent a sabbatical year in the Netherlands with him), Waksman became familiar with the value of cultivating molds aerobically for citric acid production by submerged culture on rotary shaking machines. With adequate oxygen available at all times, mold growth was far more rapid and it was much more dense in the shaken cultures than that obtained by the prior stationary culture techniques. This was an advanced method of culture for microbes, not only valuable in organic acid production but highly significant when Waksman extended his program utilizing submerged culture procedures into the antibiotic field.

Supplementing his broad attack on understanding the activities of microbes within the soil, Waksman became involved with many related activities. Much effort was spent in developing a classification system for the filamentous bacteria, commonly termed the actinomycetes. In 1943, he, with A.T. Henrici, proposed a new generic name *Streptomyces* for the frequently found spore-forming members of the order Actinomycetales. The new genus name became official and was included in Bergey's *Manual of Determinative Bacteriology.*

Waksman was consulted concerning the role of bacteria in initiating the fouling of ship bottoms with barnacles and other marine life. Because of the special wartime needs that developed soon thereafter, the project rapidly expanded and Waksman started summer residences at the Oceanographic Institution in Woods Hole, Massachusetts, supervising students there. For several years he applied the basic procedures of soil microbiology to investigation of the microbes of the sea, until the demands of the program required full-time supervisory effort, whereupon he resigned. Eighteen publications resulted from this scientific side excursion.

Waksman spoke frequently of threefold objectives for research in soil microbiology: first, defining the nature of the organisms present in soils; second, defining the microbiological processes that occur in soils that influence soil fertility; and, finally, elucidation of the interactions, both growth-promoting and growth-inhibiting, that occur among the multitudes of organisms that exist and interact together in soils.

Waksman's approach to the latter objective was based on observations he had made throughout his career. As early as 1923, Starkey and Waksman, in a paper describing results of their study of the microbiological changes that occur during recovery from partial sterilizations of soils, had stated: "Certain actinomycetes produce substances toxic to bacteria – around an actinomycete colony upon a plate a zone is found free from bacterial growth." They had seen such inhibitions as well as growth enhancements repeatedly when they plated soils. In 1938, Waksman decided it was appropriate to initiate a formal study of the interaction of microbes upon one another in soils. Three companion publications resulted. The first was a historical review of microbial interactions. The second, prepared with a student, covered antagonistic actions among microbes grown in the laboratory in artificial substrates. The third, with another student, covered interactions during the decomposition of plant residues. Waksman was fully prepared, therefore, to appreciate an experiment

reported the following year by Rene Dubos. Dubos, as a former Waksman student, had utilized the technique of soil enrichment at Rutgers to isolate cellulose-degrading organisms from the soil. Waksman had recommended him to Oswald Avery of the Rockefeller Institute for a project to find an enzyme capable of hydrolyzing the resistant mucoid capsule of the pneumococcus. Dubos had not only been successful in that endeavor, using the soil-enrichment procedure learned under Waksman, but had extended his studies, enriching soil with living cocci and from that enriched soil isolating a bacillus that killed cocci, including the pneumococci. The antibacterial agent responsible for the killing, which he named tyrothricin, was described in a 1939 publication. Waksman immediately encouraged his cooperators at Merck to aid Dubos by growing the tyrothricin producer in large quantities.

Waksman also saw the potential of the method as a means of discovering other important antibacterial agents. He decided to organize a project aimed at their discovery as products of actinomycetes. He assigned his university-supported graduate student to full-time study of microbial associations. Quickly, a new extremely active inhibitory substance was obtained from an actinomycete. It was crystallized in association with a Merck chemist. It was named actinomycin (Figure 1). Additional graduate students were enlisted to the task, supplemented by short-time visitors who arrived at the laboratory. The scientific publication describing actinomycin had resulted in intense newspaper publicity, attracting an even greater number of visitors than usual. Thus was organized an expanded program on antibacterial substances derived from soil microbes. When requested by an editor of an abstract journal to coin a name under which reports of such substances could be grouped, Waksman proposed the term 'antibiotic' and it became accepted worldwide.

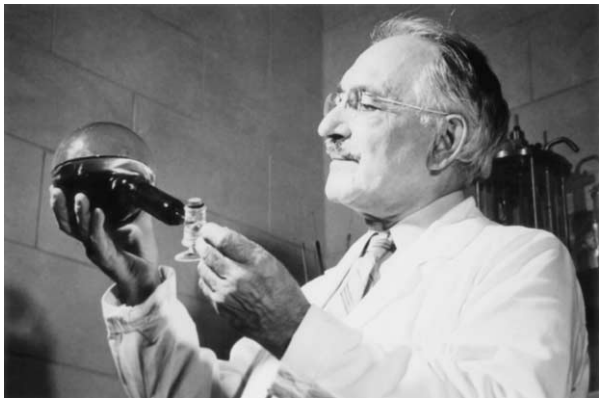


Figure 1 S.A. Waksman purifying actinomycin, his first antibiotic. (From the public files of Rutgers University, photographer unknown.)

The antibiotic discovered in Waksman's laboratory that has had greatest commercial significance is streptomycin, a broad-spectrum inhibitor, that investigators at the Mayo Clinic showed to be active against experimental tuberculosis infections in guinea pigs. Clinical trials of streptomycin soon followed. No effective therapeutic agent existed for human tuberculosis, therefore streptomycin quickly became the dominant approach for treatment of the disease, resulting in many striking cures, in spite of frequently expressed side-effects. Commercial production of streptomycin was initiated in America, initially by Merck & Co., Inc., but with conversion of the streptomycin product patent to a nonexclusive basis at Waksman's and Rutgers University's request, production was undertaken by several additional companies in the USA as well as abroad.

The commercial success of steptomycin led to an intensive search for improved antibiotic producers among the soil population, especially among actinomycetes, both at Rutgers and elsewhere. Many useful agents were found and actinomycete antibiotics have achieved great commercial value.

As the result of Waksman's prominence in the field of antibiotics and the success stories concerning streptomycin, his presence was sought constantly for advice and for the presentation of lectures, especially after he was awarded the Nobel Prize for Physiology or Medicine in 1952 for his studies on the microbes of the soil leading to the discovery of streptomycin. Students flocked to his laboratory, far more than budgets could support or for whom space was available. Waksman recommended, and the University assented, that the royalties earned from the sale of streptomycin, as well as other discoveries from his laboratory, be applied for the construction and support of a specialized Institute of Microbiology located on the Rutgers University campus. Its objective was not to be limited to research on microbes of the soil but to be applied to research on microbes broadly. Through the financial support of the royalties, the Institute of Microbiology was built and became fully staffed. Selman Waksman served as its first Director. In the Institute, important fundamental discoveries have been made and basic investigative work on microorganisms is continuing.

Supplementing this expanded activity, a research and teaching program in agricultural microbiology has been continued in Cook College, which is the successor name for the Rutgers University Agricultural College. Present in that college is a museum devoted to the development of the science of soil microbiology. It celebrates the organization in 1901 of a specialized Department of Soil Microbiology, the first such university department organized in the USA. Special

emphasis in the museum is given to Selman A. Waksman's creative leadership and to the honors accorded, including award of a Nobel Prize in 1952.

To honor Selman Waksman, in 1973 the name of the Institute for Microbiology at Rutgers University was changed to the Waksman Institute for Microbiology. For Selman A. Waksman, a dedicated research scientist for more than 50 years, and this author's mentor during the period when the first Rutgers antibiotics were discovered, the honor provided by the revised Institute name assures permanent recognition for his creative leadership in the field of soil microbiology.

See also: **Bacteria:** Soil; **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics

Further Reading

Waksman SA (1927) *Principles of Soil Microbiology*. Baltimore: Williams and Wilkins.

Waksman SA (1938) *Humus. Origin, Chemical Composition, and Importance in Nature*, 2nd edn. Baltimore: Williams and Wilkins.

Waksman SA (1949) *Streptomycin – Its Nature and Practical Application*. Baltimore: Williams and Wilkins.

Waksman SA (1954) *My Life with the Microbes*. New York: Simon and Schuster.

Waksman SA (1959) *The Actinomycetes*. Vol. I. *Nature, Occurrence and Activities*, Baltimore: Williams and Wilkins.

Waksman SA (1961) *The Actinomycetes*. Vol. 2. *Classification, Identification and Descriptions of Genera and Species*, Baltimore: Williams and Wilkins.

Waksman SA and Curtis RE (1916) The actinomycetes of the soil. *Soil Science* 1: 99–134.

Waksman SA and Lechevalier HA (1963) *The Actinomycetes*. Vol. 3. *The Antibiotics of Actinomycetes*, Baltimore: Williams and Wilkins.

Waksman SA and Starkey RL (1931) *The Soil and the Microbe*. New York: John Wiley.

Woodruff HB (1968) *Scientific Contributions of Selman A. Waksman*. New Brunswick, NJ: Rutgers University Press.

WASTE DISPOSAL ON LAND

Contents
Liquid
Municipal

Liquid

C P Gerba, University of Arizona, Tucson, AZ, USA

Introduction

The land application of human wastes has been practiced for centuries. The benefit of the nutrients provided in these wastes has been recognized as a major benefit in crop production. However, because of the presence of pathogenic organisms in such wastes, the use of untreated wastes is no longer practiced in developed countries. The main objectives of land application of wastewater today are further effluent treatment, groundwater recharge, and provision of nutrients for agricultural crops. Land treatment is defined as the controlled application of wastewater onto the land surface to achieve a planned degree of treatment through natural physical, chemical, and biological processes within the plant–soil–water matrix. During land treatment of wastewater effluents, biological and chemical pollutants are removed by physical (settling, filtration), chemical (adsorption, precipitation), and biological (e.g., plant uptake, microbial transformation) processes.

Land-treatment systems are capable of reducing the levels of pathogenic microorganisms, biochemical oxygen demand (BOD), suspended solids, nutrients (nitrogen and phosphorus), toxic metals, and trace organics. Suspended solids are removed by filtration and sedimentation. Soluble organics are removed by microbial action in the soil. Nitrogen is removed by sedimentation-filtration (e.g., particle-associated organic nitrogen), adsorption to soil, volatilization (e.g., NH_4), uptake by crops, and biological denitrification. Phosphorus is removed by adsorption to soil particles, chemical precipitation, and uptake by vegetation.

There are three major land-treatment processes: slow rate, rapid infiltration, and overland flow. A comparison of the design features of these land-treatment

Table 1 General characteristics of the three methods used for land application of sewage effluent

	Application method		
Factor	*Low-rate irrigation*	*Overland flow*	*High-rate infiltration*
Main objectives	Reuse of nutrients and water Wastewater treatment	Wastewater treatment	Wastewater treatment Groundwater recharge
Soil permeability	Moderate (sandy to clay soils)	Slow (clay soils)	Rapid (sandy soils)
Need for vegetation	Required	Required	Optional
Loading rate	1.3–10 cm week^{-1}	5–14 cm week^{-1}	>50 cm
Application technique	Spray, surface	Usually spray	Surface flooding
Land required for flow of 10^6 l day^{-1}	8–66 hectares	5–16 ha	0.25–7 ha
Required depth to groundwater	Approx. 2 m	Undetermined	5 m or more
BOD and suspended solid removal	90–99%	90–99%	90–99%
N removal	85–90%	70–90%	0–80%
P removal	80–90%	50–60%	75–90%

BOD, biochemical oxygen demand.

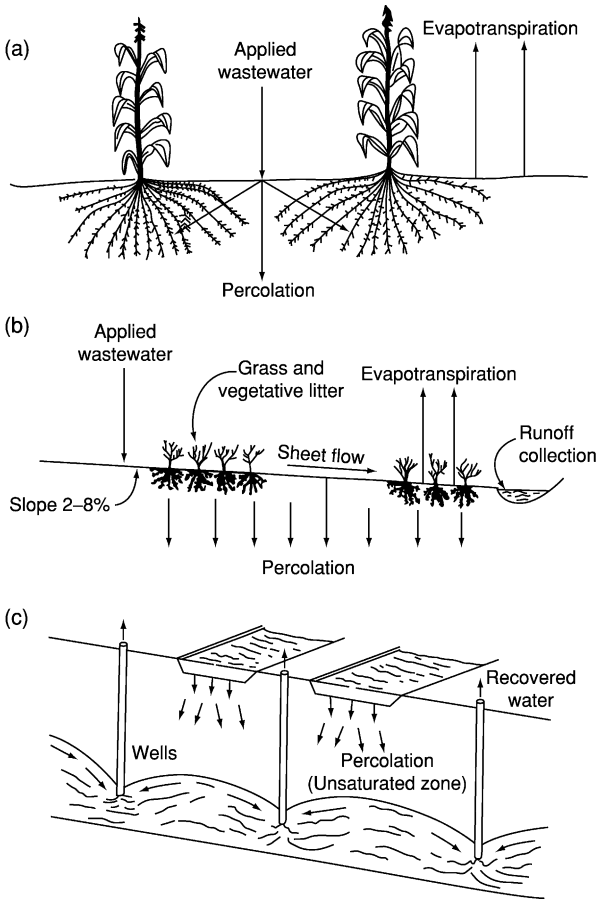


Figure 1 (a) Slow-rate irrigation system; (b) overland-flow system; (c) rapid-infiltration system. (Reproduced from the US Environmental Protection Agency (1981) *Technology Transfer Process Design Manual for Land Treatment of Municipal Wastewater*. EPA/1-81-013. Cincinnati, OH: USEPA.)

processes is shown in Table 1 and Figure 1. Expected removals of BOD, nitrogen, and phosphorus are also shown in Table 1.

In recent years, rapid-infiltration and slow-rate systems have been studied most. Rapid-infiltration systems are seen as important in the recharge of drinking-water aquifers, and slow-rate systems as a means of crop production in the arid regions of the world. Because of the potential impact of contaminants on underground drinking-water supplies, the fate of trace organics and pathogens have received increasing attention where rapid infiltration is used to supplement groundwater supplies. Because of the transport of food crops around the world today, the fate of pathogens in slow-rate systems has also received more attention.

Types of Land-Treatment Processes

Slow-Rate Process

Slow-rate irrigation systems are the most frequently used land-treatment system. There are approximately 1200 systems in the USA. Slow-rate land treatment is the application of wastewater to a vegetated land surface, with the applied wastewater being treated as it flows through the plant–soil matrix. Some of the flow percolates to the groundwater and some is used by the vegetation. Off-site surface runoff of the applied wastewater is generally avoided in the design. Nitrogen is often the limiting factor in these systems because of strict standards for levels of nitrate in drinking water. In arid lands, salts at acceptable levels for crop production may be limiting. Usually sandy loams or clay loams are the preferred soil type.

In slow-rate systems, the soil can be as shallow as 0.3 m for grass crops, but 1.5 m is preferred for complete wastewater treatment. Retention of contaminants is a function of residence time of the wastewater in the soil and the degree of contact with the soil colloids.

BOD in most slow-rate treatment systems is applied at rates of less than 11 kg ha^{-1} day^{-1}, which is

an order of magnitude lower than the capacity of the soil. High-strength food-processing wastewater systems are loaded as high as $9500\,kg\,ha^{-1}\,day^{-1}$. The BOD reductions in most systems exceed 98%. Filtration through the soil results in removal of 99% or more of total suspended solids.

Nitrogen removal is achieved using a nitrogen balance that matches the expected removal plus a percolate nitrate nitrogen of less than $10\,mg\,l^{-1}$, which is the drinking-water standard in the USA. Slow-rate systems use forage crops to remove much of the applied nitrogen. To achieve the nitrogen uptake rates expected for the crop, excess nitrogen must be applied so that the crop can compete effectively with soil microorganisms for the available nitrogen. Biological nitrogen removal occurs by nitrification–denitrification. The loss due to denitrification depends on the BOD-to-N ratio and the soil temperature, pH, and moisture. Intermittent application, which is characteristic of slow-rate systems, serves to enhance nitrification followed by denitrification. Ammonia volatilization losses of 10% can be expected if the soil pH is greater than 7.8 and the cation exchange capacity is low (low absorption of ammonium by the soil). For soils with less than 2% organic matter, nitrogen storage in the soil can be a significant loss for the first 3–4 years of operation of the system. Eventually, equilibrium is reached and the net storage of nitrogen stops.

Phosphorus removal has been found to be generally around 98–99%. Removal is by formation of insoluble phosphates with metals at the soil surface.

The concern with pathogens in slow-rate systems is related to their potential for reaching the groundwater, contamination of adjacent surface waters, and contamination of the crop. The low rate of wastewater application usually limits the potential for groundwater contamination, and contamination of surface waters can be controlled by proper management and site selection. However, enteric viruses have been detected in groundwater beneath slow-rate systems when high rates of wastewater are applied to areas with high water tables. Food-crop contamination with pathogens easily occurs during furrow irrigation, but can be limited by drip irrigation and covering the soil surface with plastic sheeting.

Rapid-Infiltration Process

In rapid infiltration, most of the applied wastewater percolates through the soil and the treated effluent percolates through to the groundwater, where it may be collected in recovery wells for drinking water or irrigation. The wastewater is applied intermittently at high loading rates ($6–125\,m\,year^{-1}$) onto a permeable soil (e.g., sands or loamy sands). Application is usually in basins and vegetation is usually not planted. Recovery wells may be used to reclaim the treated water for irrigation or as a drinking-water source. The treatment potential of rapid-infiltration systems is lower than in slow-rate systems. Removal of nitrogen is generally low but may be enhanced by encouraging denitrification. Denitrification requires adequate carbon (as found in primary wastewater effluents) and low oxygen levels, necessitating flooding periods as long as 9 days, followed by drying periods of approximately 2 weeks. Climate and season will affect the intervals of flooding and drying due to their influence on microbial activity.

Typical BOD loadings for rapid-infiltration systems using municipal wastewater range from 27 to $175\,kg\,ha^{-1}\,day^{-1}$ with removals ranging from 74 to 96%. BOD loadings of industrial systems range from 112 to $676\,kg\,ha^{-1}\,day^{-1}$. Suspended-solids loadings of $112–224\,kg\,ha^{-1}\,day^{-1}$ or more require more frequent disking or scarifying of the infiltration basin surface to avoid plugging of the soil.

Nitrification–denitrification is the principal mechanism of ammonia and nitrogen removal from wastewater in rapid-infiltration systems. Ammonia adsorption also plays an important role in retaining ammonia in the soil long enough for biological conversion. Nitrification and denitrification are affected by temperature and the amount of organic carbon available. It has been found that nitrification rates of up to $67\,kg\,ha^{-1}\,day^{-1}$ can be achieved under favorable moisture and temperature conditions. Nitrogen removal is a function of detention time, BOD-to-N ratio, and anoxic conditions. Detention time is related to hydraulic loading rates through the soil profile. For effective nitrogen removal (80% or more), the loading rate should not exceed 3–4 cm. The BOD-to-N ratio needs to be 3:1 or more to ensure adequate carbon to drive the denitrification reaction. Secondary wastewater effluent will have a BOD-to-N ratio of approximately 1:1, while primary effluent usually has a BOD-to-N ratio of 3:1. To overcome the low BOD-to-N ratio in secondary effluent, a longer period of application (7–9 days) is necessary. Typical removal of total nitrogen ranges from 38 to 93%.

Phosphorus removal is accomplished by adsorption and chemical precipitation. The adsorption occurs quickly and the chemical precipitation occurs more slowly; removal typically ranges from 40 to 97%.

Removal of bacteria, helminths, and protozoan parasites is largely accomplished by filtration. Almost all are removed within a meter of the soil surface under unsaturated conditions. Virus removal is dependent upon adsorption to the soil surface; viruses have the potential to travel long distances under the

right conditions (50 m or more). The depth of the unsaturated or vadose zone is important in the degree of removal of viruses before the wastewater reaches the aquifer. Expected removals of virus should be 99% or more.

Overland-Flow Process

In overland-flow treatment, wastewater is applied at the upper reaches of grass-covered slopes of 2–10% grade and allowed to flow over the vegetated surface to runoff collection ditches. Typical slopes are 2–4% in grade and 36–45 m long. The overland-flow process is best suited to sites having relatively impermeable soils. The most suitable soils are clay or clay-loamy soils with a permeability equal to or less than $0.5\,cm\,h^{-1}$. About 50 overland-flow systems have been built in the USA. The wastewater is treated by a thin biofilm and plants as it flows down the slope. Removal of nitrogen is due to nitrification followed by denitrification, and uptake by the grass or other plants. Phosphorus removal is via adsorption and precipitation.

In municipal systems, the BOD loading rate typically ranges from 12.3 to $4\,kg\,ha^{-1}$. Biological oxidation accounts for the 90–95% removal of BOD normally found in overland-flow systems. A typical BOD concentration in treated runoff water is approximately $10\,mg\,l^{-1}$. Overland flow is effective in removing most suspended solids, with effluent total suspended solids between 10 and $15\,mg\,l^{-1}$.

The removal of nitrogen depends on nitrification–denitrification and the crop uptake of nitrogen. Denitrification can account for 60–90% of the nitrogen removed, with denitrification rates of $160\,kg\,ha^{-1}$. Phosphorus removal in overland-flow systems is limited to approximately 40–50% because of the lack of soil–wastewater contact.

Overland flow is not very effective in removing microorganisms. Fecal coliforms can be reduced by approximately 90% when raw or primary effluent is applied; however, minimal removal occurs when secondary effluents are applied. Enteric virus removals of up to 85% have been observed with overland flow.

Treatment Mechanisms

In the land application of wastewater, the soil is used as a treatment medium. Some substances pass through the soil and into the groundwater, some are utilized by growing plants, while others are retained almost indefinitely within the soil. Proper design of land-application facilities must relate the fate of pollutants to the properties of soil with which they may interact and minimize the fraction of contaminants passing through to groundwater. There are several separate-unit processes that can be adopted for removal of suspended solids, total dissolved solids, biochemical oxygen demand, nitrogen, phosphorus, and toxic substances. Waste-treatment mechanisms that occur in soils can conveniently be categorized as physical, chemical, and biological. Within each category, various processes act to remove or alter specific pollutants.

Filtration

As wastewater moves into and through soil pores, suspended solids are removed by mechanical filtration. The actual depth at which removal occurs varies with the size of suspended particles, soil texture, and the rate of application. The larger the hydraulic loading rate and the coarser the soil, the greater the distance required. However, when loading rates are such that much of the applied water is held in the soil, additional removal can occur as suspended materials settle or adhere to the surfaces of the soil particles. Some organic particles and some organisms such as worms and protozoa are large enough so that removal by simple blockage occurs. Both bacteria and viruses, however, are small enough that they can move through the soil pores. While filtration in the soil matrix limits the movement of most bacteria to a meter or less, viruses are too small to be affected, and filtration is not believed to play a significant role in their removal. In practice an organic mat or 'Schmutsdecke' forms at the soil surface, enhancing the removal of particulates. Suspended matter can evidently clog soil pores and thereby severely reduce the infiltration rate. In rapid-infiltration operations, flooding and drying cycles help restore infiltration rates. During flooding the infiltration rate is reduced over time. Drying of the Scmutszdecke results in its cracking and removal during the next infiltration event.

Adsorption and Precipitation

Chemical reactions among dissolved ions or compounds and interactions with the soil solid-phase alter the mobility of waste pollutants. Some dissolved constituents are retained within the profile indefinitely, while the movement of others is only temporarily restricted or unaffected. Two processes, adsorption and precipitation, account for most of the retention. Adsorption refers to the net effect of interactions between dissolved and particulate (i.e., viruses) matter of the soil and humus particles. Precipitation is the formation of an insoluble form of a substance that was originally in solution. Ion exchange is one type

of adsorption process. Anion exchange is of very minor importance in land-application systems. Phosphorus is the only anion appreciably retained in soils, though the primary mechanism is not anion exchange. The predominant ion exchange reactions in most soils in the USA involve cations and are related to the cation exchange capacity. In wastewater-treatment systems, the retention of dissolved cations depends largely on their concentration in solution entering the soil. Large concentrations overcome a higher degree of attraction of other cations toward the soil matrix. Many of the heavy metal cations such as zinc, copper, nickel, cadmium, mercury, lead, and chromium are present in wastewater at concentrations too low to be appreciably affected by cation exchange reactions. These cations seem to be incorporated into the soil solid-phase in a nonexchangeable form.

Chemical Reactions

Phosphorus and nitrogen are at least partially controlled by chemical mechanisms in the soil. The most important chemical mechanism for nitrogen removal is the reaction of positively charged ammonium ions with the soil cation-exchange complex.

Volatilization

Volatilization refers to the evaporation of chemical vapor from soil or water and its subsequent loss to the atmosphere. Many organic compounds are volatile in water, as are some nitrogen compounds (e.g., ammonia, nitrous oxide) generated from biological transformations. In addition, some inorganic chemicals (e.g., selenium compounds) may be rendered volatile through biological reactions. For a given chemical, the extent of volatilization is very dependent on the soil and atmospheric conditions. In general, volatilization is greatly reduced in soil compared with water, because the soil solid-phase retains the chemical mass, thereby reducing its vapor pressure. In addition, the soil can offer substantial resistance to the transport of the chemical from the soil profile to the surface, particularly if the soil is wet and little upward flow of water is occurring. In rapid-infiltration processes where the wastewater is ponded over the surface for prolonged periods of time, the primary route of volatilization loss is from the surface of the standing water. During the drainage cycle, when the soil becomes unsaturated, volatile constituents in solution near the surface can evaporate and escape to the atmosphere.

Ammonia present in wastewater is very volatile and vaporizes from the anhydrous form immediately upon exposure to air. Nitrite may be transformed anaerobically by biological denitrification to several gas species (primarily N_2O and N_2) when soil-water content is high and a source of organic carbon is present.

Biological Mechanisms

Soil microorganisms alter the waste constituents through organic matter decomposition, inorganic transformations, and nutrient assimilation. These processes are largely restricted to the upper meter of soil. The ability of soil microorganisms to decompose organic matter is a function of their population complexity. The diversity of microorganisms enhances the capability of soil to degrade a wide variety of organic substances; some prey on pathogenic bacteria and reduce the survival of viruses. The rate at which organic matter decomposition occurs and the exact nature of the intermediate and end products depend in part on the composition of the added organic matter in the wastewater. Soil factors, however, exert considerable control as well. The presence or absence of oxygen, more than any other single factor, determines the rate and end products. The oxygen status of the soil is a function of soil porosity, and properties that favor rapid transmission of water also favor oxygen movement unless the soil is completely saturated.

Decomposition proceeds most rapidly under aerobic conditions. Under aerobic conditions, up to 60% of the organic carbon may be respired by organisms as carbon dioxide in the initial stage of decomposition. Much of the rest is incorporated into microbial cells, and some of this is subsequently respired when the population declines. Through organic matter decomposition, nitrogen, phosphorus, sulfur, and a host of other trace elements are converted from organic to inorganic forms, i.e., mineralization. Many of the elements that are mineralized during organic matter decomposition are then subject to inorganic transformation in the soil and are assimilated by plants in overland-flow and irrigation-treatment systems. Vegetation thus functions as a sink where waste nutrients, most notably nitrogen and phosphorus, can be effectively immobilized. When harvested, vegetation plays an integral part in the renovation or 'treatment' of the applied wastewater.

Fate of Specific Contaminants

Metals and Trace Elements

Factors that affect the retention of trace elements by soils include soil texture, pH, soil organic matter, and contents of amorphous oxides of Fe, Al, and Mn. Most studies have a high capacity to attenuate the concentration of copper and lead. The retention of

other trace elements studied is best correlated with the clay content of free iron oxides of the soil. Capacities for attenuating the cationic elements (Cu, Pb, Be, Zn, Cd, Ni, and Hg) in soils tends to increase with their clay and iron oxide contents. For the anionic trace elements (SeO_3, VO_3, AsO_4, CrO_4), retention in the soil also increases with increasing clay and iron oxide content of the soils. In general, the solubility of the cationic trace-element species increases as the pH of the soil decreases. By contrast, the solubilities of anionic trace-element species in the soil tend to increase as the pH of the soil increases.

Metals and trace elements may be removed by a number of mechanisms during land application of wastewater. Those associated with suspended matter are retained at or near the soil surface. In rapid-infiltration systems, they accumulate in the colloid material in the surface-clogging layer, which eventually must be removed to restore infiltration. Smaller suspended particulates that can move through soil pores without becoming trapped are also attenuated by sorption to mineral surfaces in the soil matrix. For dissolved elements, ion exchange, precipitation, surface adsorption, and complexing with organic compounds are important. The principal mechanisms that immobilize dissolved trace elements in the wastewater within the soil appear to be absorption and precipitation. Most soils appear to have a high capacity to remove trace elements during land treatment of wastewater. The amounts of trace elements removed by crops are small compared with the amounts applied to the soils though application of wastewater (Table 2).

Because soil sorption may be finite for most trace elements, there may be long-term accumulation of metals in the soil. Groundwater concentrations of silver, barium, cadmium, cobalt, and chromium below the rapid-infiltration site at Hollister, California, have been unaffected by the additions of wastewater to the overlaying soil. However, manganese, nickel, iron, zinc, lead, and copper are above background levels. Soil samples taken at the Whittier Narrows infiltration facility after more than 20 years of operation have shown elevated levels of cadmium, chromium, copper, nickel, lead, and zinc in the top 60 cm, but not below that depth, suggesting that the soil has the capacity to remove metals for many more years of operation before groundwater is affected.

Organic Compounds

Organic compounds vary greatly in their mobility, volatility, and persistence in soil. In land-application systems, volatile compounds volatilize prior to application, and only soluble organic compounds enter the soil. The fate of the soluble compound depends on the degree to which it is chemically or biologically transformed during its passage through the system. Organic compounds degrade by hydrolysis, photodecomposition, or redox reactions. Microbial conversion occurs chiefly at or near the soil surface, where bacterial populations and organic carbon levels are high. Low organic carbon levels limit microbial action in the deeper regions of the vadose zone during treatment in rapid-infiltration systems. The travel time of an organic compound may be roughly estimated by its retardation factor in a given soil. The overall action of the chemical and microbiological processes transforming an organic compound moving through the soil may be crudely expressed as the half-life or degradation rate constant. The length of the half-life compared with the travel time can be used as an index of the potential for the compound to survive its contact and passage through the soil.

The dissolved organic carbon (DOC) in domestic, secondary-treated wastewater during rapid infiltration is reduced by 50% after passage of a few meters through the soil. DOC is primarily removed by microbial action during passage through the soil, with more than half being removed during passage through the top 8 cm of soil. With long residence

Table 2 Expected trace element removal by vegetation from wastewater-irrigated soils

Element	*Typical concentration in wastewater ($g\,l^{-1}$)*	*Annual input ($g\,ha^{-1}$)*	*Typical concentration in vegetation ($g\,g^{-1}$)*	*Annual removal ($g\,ha^{-1}$)*	*Removal (%)*
As	<0.005	<60	1	5	8.3
Cd	0.005	60	0.5	2.5	4.2
Cr	0.025	300	0.5	2.5	0.8
Cu	0.10	1200	15	75	6.3
Hg	0.0009	11	0.02	0.1	0.9
Pb	0.05	600	2	10	1.7
Zn	0.15	1800	50	250	13.9

At an application rate of 1.2 m $year^{-1}$. Assuming annual dry matter yield of 5 t ha^{-1} e.g., potatoes.
Adapted from Page AL and Chang AC (1984). Fate of wastewater constituents in soil and groundwater: trace elements. pp. 31-1 through 13-16. In: Pettygrove GS and Asano T (eds) *Irrigation with Reclaimed Muncipal Wastewater. A Guidance Manual.* Report Number 84-1. Sacramento, CA: California State Water Resources Control Board.

times in the soil, aquifer DOC is reduced to 1 mg l^{-1} after 12–24 months.

Disinfection By-products

In most cases, wastewater is disinfected before land application with chlorine, chloramine, ozone, or ultraviolet light. In the case of chemical oxidizers, this results in the formation of disinfection by-products which may potentially be harmful to human health. Some of these compounds are listed in Table 3 along with their probable fate during land application of wastewater. In addition to known disinfection by-products, residual halogen is present when chlorine is used. This material may then appear as adsorbable organic halogen (AOX). Sorption processes do not play a significant role in their removal. Under anoxic conditions removal is probably based on co-metabolism and is related to the DOC that is needed as a substrate. Ozone disinfection of wastewater improves the biodegradability of refractory organic compounds and co-substrate concentration for AOX co-metabolism in soil columns. AOX removal in 1-m soil columns to which secondary effluent is applied averages 30%.

Endocrine Disruptors

Many endocrine-disrupting compounds are present in wastewater at trace quantities (i.e., estrogens, pesticides, polychlorinated biphenyls, phthalates, pharmaceuticals). Estrogens are probably the largest contributors to the endocrine-disruption activity in domestic wastewater, since daily urination adds estrogen to household wastewater. Humans are known to excrete between 10 000 and 100 000 ng l^{-1} of 17beta-estradiol per day. Synthetic estrogens such as 17alpha-ethinyl estradiol have been detected in wastewater and in surface waters affected by effluent discharge. Measurements of endocrine-disrupting compounds displaced by a fluorescent estrogen assay or cell proliferation assay have shown the reduction of endocrine disruptors during rapid infiltration. By cell proliferation assay, the estradiol equivalent value of 71 ng l^{-1} was found to be reduced to 13.7 ng l^{-1} after passage of 63 m through the vadose zone. The equivalent 17beta-estradiol activity measured by a binding assay was reduced by 97% after travel through the vadose zone to reach the aquifer. Since the study site had been in operation for more than 10 years, it seems unlikely that endocrine disruptors would ever break through the vadose zone.

Table 3 Fate of disinfection by-products (DBP) during rapid infiltration

DBP	*Disinfectant*	*Estimated fate during recharge*
Chloroform	Chlorine	Volatilization
	Chloramines	Sorption
Bromodichloromethane	Chlorine	Volatilization
	Ozone	Sorption
Trichloroacetic acid	Chlorine	Mineral sorption
MX	Chlorine	Sorption
		Degradation
Bromate	Ozone	Reaction with soil

MX, 3-chloro-4(*d*-chloromethyl-5-hydroxyl-2(*H*)-furanone.
Adapted from NRC (1994) *Groundwater Recharge Using Waters of Impaired Quality*. Washington, DC: US Government Printing Office.

Pharmaceuticals

There has been increasing concern about the fate of micropollutants originating from pharmaceuticals and active ingredients in personal-care products (e.g., soaps, deodorants) that are introduced into domestic wastewater. There is evidence that substances of pharmaceutical origin are not completely eliminated during wastewater treatment or biodegraded in the environment. At two rapid-infiltration sites in Arizona, neither ibuprofen nor naproxen or any other acidic drug were detected in groundwater wells downgradient of the infiltration basins (Table 4). These results indicate a high potential for degradation of anti-inflammatory and lipid-regulator drugs. However, antiepileptic, carbamazepine, and primidone drugs have been detected in all of the wells, suggesting that they are able to persist in soil-treatment systems.

Microorganisms

From the standpoint of acute diseases, microbial contaminants are of greatest concern during the land application of wastewater. Microorganisms are responsible for more than 90% of all waterborne outbreaks reported each year in the USA. Hundreds of different types of pathogenic microorganisms (i.e., bacteria, viruses, protozoa, helminths) are excreted in fecal material of infected hosts and these can find their way into municipal wastewater. The number of types of pathogenic microorganisms present in wastewater varies by location and over time at a given location. A variety of factors influence pathogen content of wastewater, including the incidence of the disease in the population, the season of the year, the economic status of the population, and water-use patterns. Diseases caused by waterborne organisms range from mild gastroenteritis to severe and life-threatening illness such as hepatitis, cholera, typhoid, paralysis, meningitis, heart disease, etc. New enteric pathogens are discovered almost yearly and the significance of new ones changes over time. *Cyclospora*

Table 4 Fate of pharmaceuticals after rapid infiltration

Use/origin	*Compound*	*Secondary-treated wastewater (mg l^{-1})*	*Detection in groundwater (ng l^{-1})*
Lipid regulator	Gemfibrozil	1235	Not detected
Antiepileptic	Carbamazepine	Not detected	455
	Primidone	110	115
Analgesic/anti-inflammatory	Ibuprofen	3380	Not detected
	Naproxen	6280	20
	Fenoprofen	35	Not detected
	Propyphenazone	20	15

Adapted from Drewes JE, Heberer T, and Reddersen K (2002) Fate of pharmaceuticals during indirect potable reuse. *Water Science and Technology* 46: 73–80.

and TT hepatitis virus are examples of new pathogens that have been recognized only recently, and the significance of caliciviruses (Norovirus) as a major cause of food-borne illness has only recently been recognized.

During land application of wastewater, pathogens are removed from the system by a combination of die-off (inactivation in the case of viruses) and physical filtration or adsorption by the soil or plant material. Pathogen removal by overland flow systems has been little-studied, but it is not expected to result in large reductions of enteric bacteria or viruses. Irrigation systems and rapid-infiltration systems have been studied more because of the potential for crop and aquifer contamination.

Transmission of food-borne illness by enteric pathogens due to irrigation with untreated wastewater has been well-established for more than 100 years. For this reason, irrigation with untreated wastewater for food, crops is usually forbidden. Wastewater to be used for food crop production should meet the same standards as drinking water and be intensely monitored. Thus almost all irrigation with wastewater is for the production of nonfood crops or fruit crops. However, recent outbreaks of disease associated with apples and raspberries caused by the parasites *Cryptosporidium* and *Cyclospora* suggest that only wastewater treated to potable standards should be used in any food crop production.

Factors controlling the survival of pathogens are shown in Table 5. Table 6 shows the range of reported survival times of selected enteric pathogens in the environment. Temperature is probably the most important factor controlling persistence of pathogens on plant material and soil. The die-off of pathogens increases as the temperature increases; however, under proper conditions (high humidity), some growth of bacterial pathogens may occur on the plants in the field. Rotavirus and enteroviruses appear to be inactivated very rapidly on irrigated grass under the conditions found in the arid southwestern USA. Poliovirus type one is inactivated at a rate of $0.06\,\log_{10}\,h^{-1}$ and rotavirus SA-11 at $0.04\,\log_{10}\,h^{-1}$ during the winter (4–10°C) in Arizona (USA). The rates of die-off during the summer are $0.37\,\log_{10}\,h^{-1}$ and $0.2\,\log_{10}\,h^{-1}$. It would appear that 8–10 h is needed during the summer and 16–24 h during the winter before inactivation of these viruses. *Salmonella* and other enteric bacteria can survive for several weeks on grass if sufficient organic matter and moisture is available. Helminth eggs such as *Ascaris* are

Table 5 Factors influencing enteric pathogen survival in soils

Factor	*Comments*
Temperature	Lower temperatures promote increased survival
Moisture content	Survival is decreased in drying soils. Increases may result in bacterial growth
pH	Survival is decreased pH extremes
Soil type	Survival influenced by chemical, textural, and mineralogical properties such as pH, exchangeable ion content, and water-retention capacity
Organic matter	Survival enhanced and growth of bacteria possible
Soil microflora	Survival is decreased in nonsterile soil
Salt species and concentration	Cationic type and ionic strength may influence survival
Soil surfaces	Survival may be enhanced by adsorption to soil particles
Saturation	Air–water interface under unsaturated conditions may enhance viral inactivation

Table 6 Survival of pathogens in soils

Organism	*Survival time (days)*
Coliforms	38
Fecal streptococci	26–77
Salmonellae	15–90
Salmonella typhi	1–120
Entamoeba histolytica cysts	6–8
Enteroviruses	8–175
Ascaris ova	Up to 2 years

believed to survive for 30–60 days, although they may survive many months in the soil itself.

Temperature also significantly influences the survival of enteric pathogens in soil. Poliovirus persists for 8 days in saturated sand and sandy loam soils at 4°C, whereas more than 175 days are required at 8°C. Survival of bacteria in soils is also favored by cold temperatures. However, freeze–thawing increases mortality of enteric bacteria and protozoa. *Cryptosporidium* may persist in soils for several weeks, although its survival time is reduced at temperatures above 35°C and if multiple freeze–thaw cycles in the soil take place. Extreme acidic or alkaline conditions (pH <6.0 or >8.0) tend to adversely affect the survival of both enteric bacteria and viruses in soil. Greater moisture also enhances survival of most enteric organisms in soil as does greater concentrations of organic matter in the wastewater.

Transport of microorganisms through the soil is influenced by many of the same factors that affect survival of microorganisms (Table 7). Because of their large size, bacteria, protozoan parasites, and helminths are usually removed within a meter or less of the soil surface. Because of their small size (20–200 nm), viruses have the potential to travel the greatest distances through soil. Removal largely occurs by adsorption to the soil surface and is influenced by the surface properties of the virus and soil. Electrostatic and hydrophobic interactions control the degree of adsorption. Transport is favored in sandy soils with a low clay content. Adsorption largely occurs near the soil surface because of the presence of iron oxides to which the negatively charged viruses readily adsorb. Transport is more limited under unsaturated conditions and at pH levels greater than 8.0. Organic matter in wastewater reduces adsorption because of competition for adsorption sites on the soil surface.

Table 7 Factors influencing microbial transport

Factor	*Comments*
Size of microorganism	Transport of bacteria and parasites is limited because of size in most agricultural soils
Adsorption	Major factor in limiting virus transport through soils. Also plays a role in bacterial transport
pH	Lower pH enhances adsorption
Salt species and concentration	Increased cation valency and ion concentration increase retention; low ionic strength promotes desorption and transport
Organic matter	Some organics (e.g., humic and fulvic acids) interfere with adsorption and cause desorption
Metal oxides	Virus adsorption to iron oxides occurs readily in soils
Hydraulic conditions and moisture content	Increase flow rates and saturated flow decrease adsorption

See also: **Groundwater and Aquifers**; **Pollutants:** Persistent Organic (POPs); **Pollution:** Groundwater; **Waste Disposal on Land:** Municipal

Further Reading

Amy G, Wilson LG, Conroy A *et al.* (1993) Fate of chlorination by-products and nitrogen species during effluent recharge and soil aquifer treatment (SAT). *Water Environment Research* 65: 726–734.

Asano T (1985) *Artificial Recharge of Groundwater.* Boston, MA: Butterworth.

Asano T (1998) *Wastewater Reclamation and Reuse*. Lancaster, PA: Technomic.

Bitton G (1999) *Wastewater Microbiology.* New York: Wiley-Liss.

Crites R and Tchobanoglous G (1998) *Small and Decentralized Wastewater Management Systems*. New York: McGraw-Hill.

Drewes JE and Jekel M (1998) Behavior of DOC and AOX using advanced treated wastewater for groundwater recharge. *Water Research* 32: 3125–3133.

Drewes JE, Heberer T, and Reddersen K (2002) Fate of pharmaceuticals during indirect potable reuse. *Water Science and Technology* 46: 73–80.

Fox P, Narayanaswamy K, Genz A, and Drewes DE (2001) Water quality transformation during soil aquifer treatment at the Mesa Northwest Water Reclamation Plant, USA. *Water Science and Technology* 43: 343–350.

Hurst CJ, Gerba CP, and Cech I (1980) Effects of environmental variables and soil characteristics on virus survival in soil. *Applied and Environmental Microbiology* 40: 1067–1079.

Page AL and Chang AC (1984) Fate of wastewater constituents in soil and groundwater: trace elements. pp. 31-1 through 13-16. In: Pettygrove GS and Asano T (eds) *Irrigation with Reclaimed Muncipal Wastewater. A Guidance Manual*. Report Number 84-1. Sacramento, CA: California State Water Resources Control Board.

Quanrud DM, Arnold RG, Wilson LG *et al.* (1996) Fate of organics during column studies of soil aquifer treatment. *Journal of Environmental Engineering-ASCE* 122: 314–321.

Reed SC and Crites RW (1984) *Handbook of Land Treatment Systems for Industrial and Municipal Wastes*. Park Ridge, NJ: Noyes Publications.

Reed SC, Crites RW, and Middlebrooks EJ (1995) *Natural Systems for Waste Management and Treatment*, 2nd edn. New York: McGraw-Hill.

US Environmental Protection Agency (1981) *Technology Transfer Process Design Manual for Land Treatment of Municipal Wastewater.* EPA/1-81-013. Cincinnati, OH: EPA.

US National Research Council (1994) *Ground Water Recharge Using Waters of Impaired Quality*. Washington, DC: US Government Printing Office.

Yates MV and Gerba CP (1998) Microbial considerations in wastewater reclamation and reuse. In: Asano T (ed.) *Wastewater Reclamation and Reuse*, pp. 437–488. Lancaster, PA: Technomic.

Municipal

D A C Manning, University of Newcastle, Newcastle upon Tyne, UK

Introduction

Waste materials are produced by all societies. The nature of waste and the quantities requiring management vary according to both geographic and historical factors, largely because the definition of waste is governed by both political and economic considerations. Given that at the present time it is required that waste is managed to minimize possible pollution of water resources (drinking water and fisheries and/or recreation) and the atmosphere, disposal to land is the dominant procedure used in North America and Europe for the disposal of the residues left after recovery of potentially useful materials and energy. In other parts of the world, disposal to land is widespread, depending on local political, economic, and environmental factors.

The artificial process of disposal to land by landfill, land-raising, or land-spreading produces accumulations that may be as much as 30 m thick of organic matter derived from plants and animals, mineral matter (including glass), metals, and plastics. Once deposited, these materials are subjected to natural processes of decomposition, and in good practice the site is restored to a new use, through appropriate capping and revegetation. Disposal to land thus generates an artificial 'soil' profile that can affect root systems and that influences pore water and gas compositions in the unsaturated and saturated zones. To understand the consequences of land disposal on soil systems, in the broadest sense, it is necessary to understand the chemical processes that take place within landfill and how they change with time.

Municipal solid waste (MSW) is waste generated by urban populations and collected for disposal by municipal authorities. It is typically composed of putrescible waste from food production, distribution, and consumption, together with discarded consumer goods, packaging, and papers. It differs from industrial wastes that are specific to particular factories or processes and so have a focused impact requiring attention on a case-by-case basis. It differs from mine waste and demolition wastes, which are 'inert' through their lack of putrescible material (although important because of, for example, reactions involving sulfides or sulfates).

The amounts of MSW generated by the UK, the USA, and Japan are summarized in Table 1. This shows variation in the relative proportions of key waste types. Japan has little land available for landfill disposal, whereas land is abundant in the USA. Like all European countries, the UK has a very tightly regulated waste management system, coupled with shortage of land close to urban centers where there is considerable popular opposition to waste incineration.

MSW Composition and Degradation

The composition of municipal solid waste (Figure 1) is dominated by paper (33%) and putrescible waste (20%). Putrescible waste includes raw and cooked

Table 1 Amounts of waste produced annually in the USA, UK, and Japan, and relative proportions of methods of disposal (1995)

	USA	*UK*	*Japan*
Household waste (t year^{-1})	210	20	50
Landfill (%)	63	90	20
Incinerated (%)	16	5	75
Recycled/reused (%)	17	5	5

Based on data in Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.

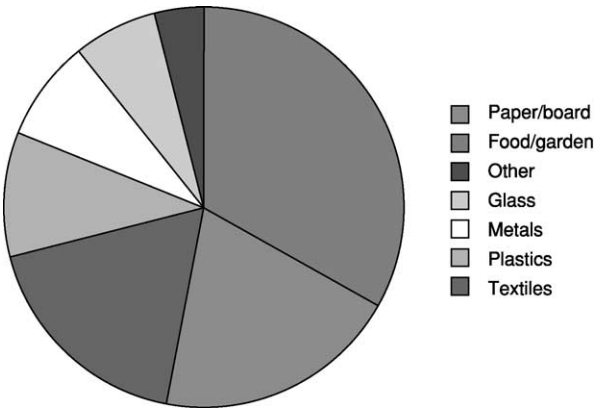


Figure 1 Approximate composition (percentage weight) of municipal solid waste (UK, 1992). (Data from Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.)

food waste and waste from domestic gardens or yards. It is composed of fats, carbohydrates and proteins, all of which decompose readily through microbial-mediated putrefaction. Paper waste is dominated by cellulose, with a significant mineral component (up to 25%) through the use of kaolin and calcite as paper-coating and -filling agents. Compared with putrescible waste, paper is relatively inert within a landfill and may not degrade for many years.

Within a landfill, waste is deposited ideally under dry conditions and is rapidly compacted and buried beneath a daily cover (often a local clay or an inert waste stream). Anaerobic conditions become established rapidly, and microbial degradation is dominated by communities of anaerobic bacteria. Although moisture ingress is minimized as far as possible, the inherent moisture content of the waste and some input from rainfall is sufficient to produce a liquid, 'landfill leachate,' that drains to the bottom of the waste pile. At the same time, microbial reactions within the leachate and on surfaces within the waste produce methane gas, which rises through the waste pile. Landfill management revolves around the control of both leachate and gas, to prevent harm that might arise from their escape, and to recover energy from landfill gas.

Typically, landfill gas is composed of approximately 66% methane and 34% carbon dioxide, with varying proportions of other gas species (Table 2). This 'end-member' composition is diluted with nitrogen from the atmosphere (oxygen having been consumed prior to the onset of anaerobic conditions). Reported landfill gas compositions may contain oxygen due to ingress of air to the sample prior to analysis. High levels of hydrogen sulfide are unusual, but may arise in circumstances where a site has received gypsum waste (especially plaster-based building products) or other sulfate-bearing industrial waste.

Landfill leachate composition varies according to the age of a landfill and the types of waste that have been deposited within it. Typical compositions are given in Table 3. The dominant cationic species

Table 2 Typical composition of landfill gas

Component	*Typical value (vol %)*	*Observed maximum (vol %)*
Methane	63.8	88.0
Carbon dioxide	33.6	89.3
Oxygen	0.16	20.9
Nitrogen	2.4	87.0
Hydrogen	0.05	21.1
Carbon monoxide	0.01	0.09
Hydrogen sulfide	0.00002	35.0
Others	Less than 0.02	NA

Based on data in UK Department of the Environment (1989) Waste Management Paper 27. *Landfill Gas*. London, UK: The Stationery Office.

Table 3 Typical contents of major constituents of landfill leachate (derived from Department of the Environment, 1995)

		Acetogenic leachate		*Methanogenic leachate*	
Component	*Units*	*Minimum*	*Maximum*	*Minimum*	*Maximum*
pH	pH units	5.12	7.8	6.8	8.2
Conductivity	$\mu S\,cm^{-1}$	5800	52 000	5990	19 300
COD[a]	$mg\,l^{-1}$	2740	152 000	622	8000
BOD_5[b]	$mg\,l^{-1}$	2000	68 000	97	1770
TOC[c]	$mg\,l^{-1}$	1010	29 000	184	2270
VFA[d] (as C)	$mg\,l^{-1}$	963	22 414	<5	146
Alkalinity	$mg\,l^{-1}$ $CaCO_3$	2720	15 870	3000	9130
Chloride	$mg\,l^{-1}$	659	4670	570	4710
Nitrate	$mg\,l^{-1}$	<0.9	79.7	0.9	9.3
Sulfate	$mg\,l^{-1}$	<5	1560	<5	322
Phosphate	$mg\,l^{-1}$	0.6	22.6	0.3	18.4
Ammonium	$mg\,l^{-1}$	249	4641	364	2623
Sodium	$mg\,l^{-1}$	474	2400	474	3650
Potassium	$mg\,l^{-1}$	350	3100	100	1580
Magnesium	$mg\,l^{-1}$	25	820	40	1580
Calcium	$mg\,l^{-1}$	270	6240	23	501
Iron	$mg\,l^{-1}$	48.3	2300	1.6	160
Manganese	$mg\,l^{-1}$	1.40	164.0	0.04	3.59

[a]Chemical oxygen demand.
[b]Biochemical oxygen demand.
[c]Total organic carbon.
[d]Total volatile fatty acids, as C.
Adapted from UK Department of the Environment (1995) Waste Management Paper 26B. *Landfill Design, Construction and Operational Practice*. London, UK: The Stationery Office.

are ammonium, Na, K, Mg, Ca, and Fe. The dominant anionic species include chloride, bicarbonate, and organic acid anions. Leachates differ from surface waters and from most natural shallow groundwaters in having high ammonium and organic acid anion contents, sometimes as the dominant cation or anion, respectively.

Variation in landfill gas and leachate composition with time is shown in Figure 2. Atmospheric oxygen is consumed rapidly, and anaerobic conditions lead initially to the formation of hydrogen (during acetogenesis, which is the stage when ethanoate (acetate) is a dominant product of decomposition) and then methane (methanogenesis; which is the stage when methane is the dominant product of decomposition). Gas production peaks and then declines over periods of months to years (depending on local conditions). Over a similar period, leachate composition shows a major change as acetogenic conditions are replaced by methanogenesis. As organic acid anions are consumed quantitatively by microbial communities that produce methane, inorganic solutes also reduce to very low levels. Ammonium is not affected greatly by this change, remaining at 100- to 1000-mg l^{-1} levels.

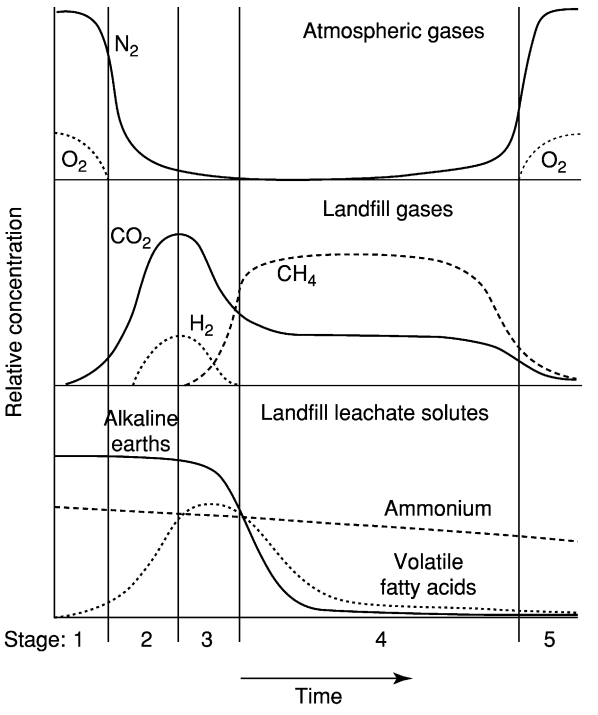


Figure 2 Changes with time in (a) atmospheric gases, (b) landfill gas, and (c) leachate composition (expressed relative to chloride). Stages 1 and 5 are dominated by aerobic conditions. Stages 2, 3, and 4 are anaerobic acidogenesis, acetogenesis, and methanogenesis, respectively. (Adapted from Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175; Robinson HD (1995) *A Review of the Composition of Leachates from Domestic Wastes in Landfill Sites*. Report CWM/072/95. Department of the Environment. London, UK: The Stationery Office.)

Microbial Reactions and Waste Degradation

The decomposition of the constituents of the putrescible fraction (proteins, lipids, and carbohydrates) is summarized in Figure 3. Initially, relative molecular masses are reduced through aerobic hydrolysis and anaerobic fermentation reactions that reduce chain length largely without destroying the characteristic functional groups, and that produce intermediate products that are readily water-soluble. Thus proteins degrade to give amino acids, carbohydrates yield simple sugars, and lipids form glycerol and long-chain fatty acids.

During the acetogenic stage, acetate is produced as an end-member of the degradation of amino acids (which also yield the ammonium ion) and long-chain fatty acids. The microbial communities responsible for the degradation of the fatty acids typically include two components in an obligate syntrophic relationship – one population oxidizes the volatile fatty acid anion, yielding hydrogen, and another reduces hydrogen or an inorganic species (sulfate, nitrate, or iron, all acting as electron donors). Examples of reactions known to occur for specific bacterial communities (e.g., *Syntrophobacter wolinii*

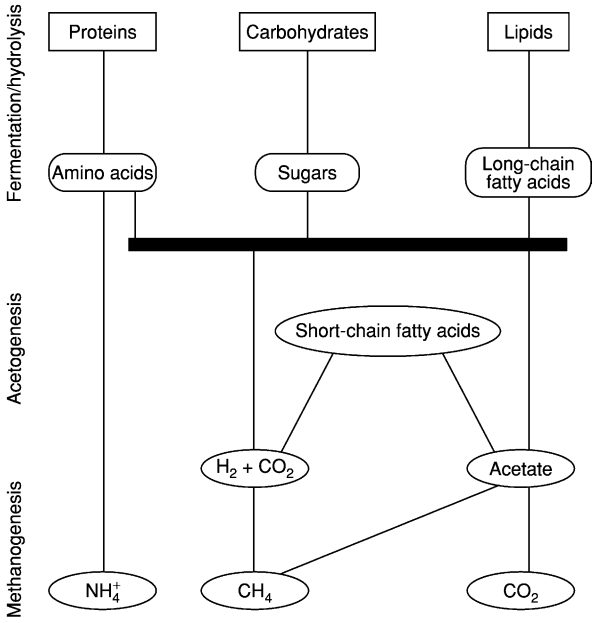


Figure 3 Schematic to summarize the decomposition of putrescible matter. (Adapted from Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175.)

Table 4 Examples of microbially mediated acetogenic reactions within landfill

	Microorganism	*Reaction*	ΔG_r^o *(kJ)*
Propionate-acetate	*Syntrophobacter wolinii*	$CH_3CH_2COO^- + 3H_2O = CH_3COO^- + H^+ + 3H_2 + HCO_3^-$	+76.1
Butanate-acetate	*Syntrophomonas wolfei*	$CH_3CH_2CH_2COO^- + 2H_2O = 2CH_3COO^- + H^+ + 2H_2$	+48.1
Sulfate reduction[a] *with oxidation of:*			
Hydrogen	*Desulfovibrio vulgaris*	$4H_2 + SO_4^{2-} + H^+ = 4H_2O + HS^-$	−152.2
Acetate	*Desulfobacter postgatei*	$CH_3COO^- + SO_4^{2-} = 2HCO_3^- + HS^-$	−47.6
Propionate	*Desulfobulbus* sp. or *S. wolinii* + *D. vulgaris*	$4CH_3CH_2COO^- + 3SO_4^{2-} = 4CH_3COO^- + 4HCO_3^- + 3HS^- + H^+$	−150.6
Methane	Species unknown in 1988; still controversial in 2003	$CH_4 + SO_4^{2-} = HCO_3^- + HS^- + H_2O$	−16.6

[a]Sulfate reducers can reduce nitrate to ammonium.

+ *Desulfovibrio vulgaris*) are summarized in Table 4. The 1:1 relationship for acetate and propionate that is observed in landfill leachates (Figure 4) appears to be characteristic of anaerobic systems of this type, as it is also observed for sedimentary waters (oil-field waters).

During the methanogenic stage, methane is produced in two ways once sulfate, nitrate, and iron in solution have been exhausted. First, decomposition of acetate in solution is facilitated by acetoclastic bacterial communities (e.g., *Methanosarcina barkeri*, *Methanosarcina mazei*, *Methanothrix soehngenii*). Secondly, methanogenic bacteria produce methane directly from hydrogen and carbon dioxide (or bicarbonate):

- General reaction (all methanogens except acetoclasts; syntrophic):

$$4H_2 + HCO_3^- + H^+ = CH_4 + 3H_2O$$
$$\Delta G_r^o = -135.6\,\text{kJ};$$

- Acetoclastic methanogenesis (*Methanosarcina barkeri*, *Methanosarcina mazei*, *Methanothrix soehngenii*):

$$CH_3COO^- + H_2O = CH_4 + HCO_3^-$$
$$\Delta G_r^o = -31.0\,\text{kJ}.$$

In all landfill leachates, the nature of the microbial community will be complex and variable in time as well as spatially. The production of methane is rarely homogeneously distributed throughout a landfill site, partly as a consequence of leachate-management procedures that mix leachates of different ages in separate parts of a landfill site.

Mineralogical Reactions Within Waste

The changes in the inorganic solute constituents of landfill leachates that are observed relate directly to mineralogical reactions that take place within the waste. These reactions can be predicted through use of the technique of geochemical modeling of landfill leachate compositions, in which a chemical analysis is recalculated to give thermodynamic concentrations of dissolved species, and then this information is used to calculate saturation indices (SI) for selected minerals. Those mineral species that have $SI > 0$ are predicted to precipitate, and those with $SI < 0$ are predicted to dissolve. Evidence that predicted mineral precipitation reactions take place can be obtained by determining the mineralogical composition of the suspended solids within leachate, or of scale that forms within leachate drainage and pumping systems.

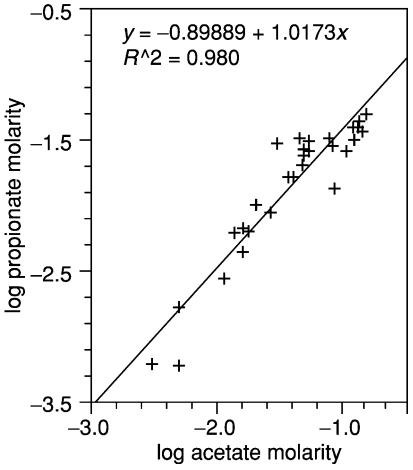


Figure 4 Relationship between acetate and propionate within landfill leachates. (Data from Manning DAC (1997) Acetate and propionate in landfill leachates: implications for the recognition of microbiological influences on the composition of waters in sedimentary systems. *Geology* 25: 279–281.)

Typically, landfill leachates have saturation indices for calcite and carbonate minerals that are greater than zero and so precipitation of calcite is predicted. Figure 5 shows the change in calcite SI with time, showing a general decrease with age. Also, calcite

saturation is observed over a wide range of pH values (Figure 6). Calcite precipitation is evident from the common observation that calcite forms scale within landfill leachate-management systems, and from the presence of calcite in a number of morphologies within leachate-suspended solids.

Silicate minerals typically show saturation indices greater than zero and so are predicted to precipitate. However, the ubiquitous presence of quartz and clays within waste- and landfill-containment systems means that newly precipitated silicate minerals cannot be recognized with confidence. Additionally, the very slow reaction kinetics of silicate mineral precipitation suggest that these minerals will not form within landfill systems under typical temperature conditions.

The recalculation of leachate compositions as activities allows mineralogical controls on leachate chemistry to be elucidated. Cation exchange reactions involving leachate and clay minerals clearly explain the distribution of K and ammonium in solution (Figure 7). Silica activities are tightly constrained to values consistent with a control by equilibrium between illitic and smectitic clays (Figure 8). These patterns are seen irrespective of the type of containment system used to hold the waste and instead relate

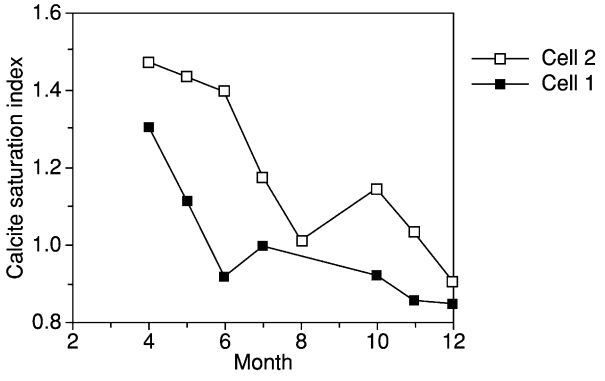


Figure 5 Variation in calcite saturation index calculated from leachate compositions for two landfill cells over time.

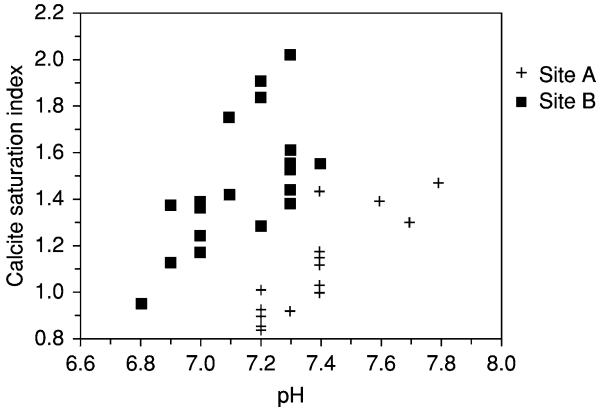


Figure 6 Variation in calcite saturation index calculated for leachate compositions with varying pH values.

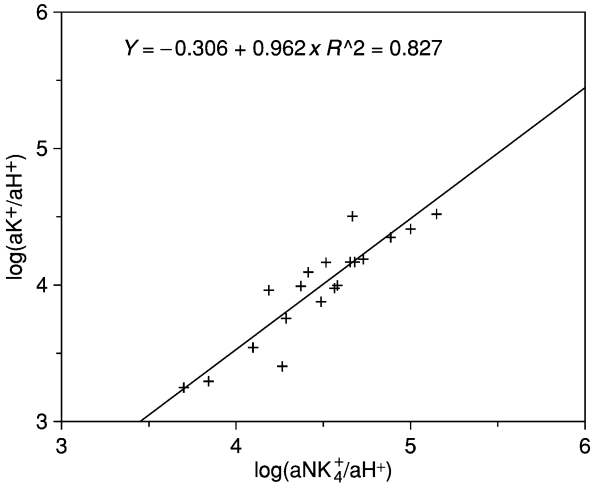


Figure 7 Ion activity diagram showing ammonium-to-hydrogen ion activity ratios $\log(aNH_4^+/aH^+)$, versus potassium-to-hydrogen ion activity ratios ($\log(aK^+/aH^+)$) for leachate data reported by Owen and Manning. A slope of 1 is consistent with mineral-controlled ammonium–potassium cation exchange. The term 'activity' refers to the thermodynamic concentration of the chemical species. (Adapted from Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.)

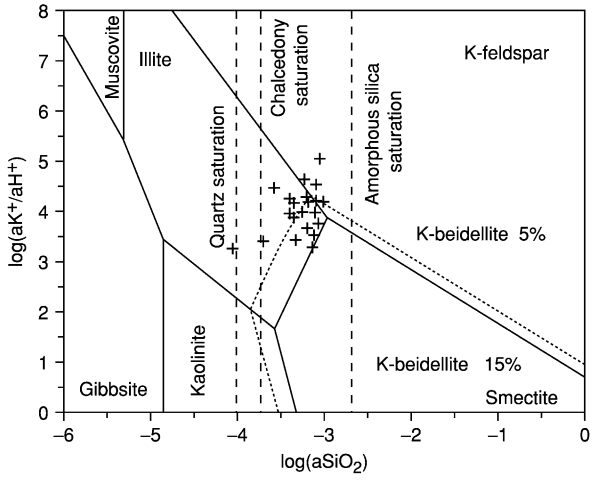


Figure 8 Ion activity diagram showing mineral stability fields calculated for soil minerals (including two smectite compositions: K-beidellite end-member, 5% (dashed) and 15% (solid)) as a function of potassium-to-hydrogen ion activity ratio ($\log(aK^+/aH^+)$) versus dissolved silica activity ($\log(aSiO_2)$), with activity ratios for landfill leachates. (Adapted from Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.)

to the nature of the minerals that are present within municipal waste as constituents of consumer goods. Sorbents such as animal litters are based on smectite–illite mixtures; clays within paper are kaolinite–illite mixtures.

Summary

Landfill disposal of municipal waste involves the artificial creation of an anaerobic environment in which degradation of putrescible matter and mineral reactions produce solutions and gases whose composition is controlled by natural processes analogous to those that take place in soils. Degradation products include methane and carbon dioxide, and calcium carbonate mineral precipitate. Clay minerals within waste control solution composition through cation exchange, and through the illite-smectite reaction.

List of Technical Nomenclature

–	pH (pH units)
ΔG_r^o	Gibbs free energy of reaction (kilojoules)
$\mu S\,cm^{-1}$	Conductivity (micro-Siemens per centimeter)
a	Activity
BOD_5	biochemical oxygen demand (5 days' incubation)
COD	chemical oxygen demand
kJ	Energy (kilojoules)
m	Molarity (moles per liter)
$mg\,l^{-1}$	Concentration (milligrams per litre)
TOC	total organic carbon
VFA	volatile fatty acids

See also: **Groundwater and Aquifers**; **Manure Management**; **Pollutants:** Persistent Organic (POPs); **Pollution:** Groundwater; **Waste Disposal on Land:** Liquid

Further Reading

Chistensen TH, Kjeldsen P, Bjerg PL *et al.* (2001) Biogeochemistry of landfill leachate plumes. *Applied Geochemistry* 16: 659–718.

Dearlove J (1998) *An Introduction to Landfill*. London, UK: Arnold.

Maliva RG, Missimer TM, Leo KC *et al.* (2000) Unusual calcite stromatolites and pisoids from a landfill leachate collection system. *Geology* 28: 931–934.

Manning DAC (1997) Acetate and propionate in landfill leachates: implications for the recognition of microbiological influences on the composition of waters in sedimentary systems. *Geology* 25: 279–281.

Manning DAC (2001) Calcite precipitation in landfills: an essential product of waste stabilization. *Mineralogical Magazine* 65: 603–610.

Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.

Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175.

Robinson HD (1995) *A Review of the Composition of Leachates from Domestic Wastes in Landfill Sites*. Report CWM/072/95. Department of the Environment. London, UK: The Stationery Office.

UK Department of the Environment (1989) Waste Management Paper 27. *Landfill Gas*. London, UK: The Stationery Office.

UK Department of the Environment (1995) Waste Management Paper 26B. *Landfill Design, Construction and Operational Practice*. London, UK: The Stationery Office.

Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.

Zehnder AJB (1988) *Biology of Anaerobic Microorganisms*. New York: John Wiley & Sons, Inc.

Water Availability *See* **Plant–Soil–Water Relations**

WATER CONTENT AND POTENTIAL, MEASUREMENT

G S Campbell and C S Campbell, Decagon Devices, Inc., Pullman, WA, USA

Introduction

Soil-water content is a measure of the amount of water (volume or mass) contained in a unit volume or mass of soil. If the measure is the volume of water per unit volume of soil, the water content is called the 'volume wetness.' If it is the mass of water per unit mass of soil it is called the 'mass wetness.' Mass wetness in soils is typically defined as the mass of water per unit mass of dry soil. Volume and mass wetness are related by:

$$\theta = \frac{\rho_b}{\rho_w} w \qquad [1]$$

where θ (cubic meter per cubic meter) is the volume wetness, w (kilogram per kilogram) is the mass wetness, ρ_b (kilogram per cubic meter) is the bulk density of the soil (mass of dry soil per unit volume) and ρ_w (kilogram per cubic meter) is the density of water (10^3 kg m^{-3}).

Soil-water potential is a measure of the potential energy per unit mass, volume, or weight of soil water, compared with that of pure, free water. It is the work required, per unit quantity of water, to remove an infinitesimal quantity of water from the soil to a pool of pure, free water. The energy of soil water responds to the hydrostatic or pneumatic pressure on the water, to the concentration of solutes in the water, to the forces that adsorb the water to the surfaces of soil particles, and to the position of the water in the gravitational field. The components of the water potential arising from these interactions are, respectively: ψ_p, the pressure potential; ψ_o, the osmotic potential; ψ_m, the matric potential; and ψ_g, the gravitational potential. The pressure and gravitational components can be positive or negative. The osmotic and matric components can only be negative.

Both water content and water potential must be known to characterize the status of water in soil. Water content tells how much water is there, but gives no information about the availability of the water for plant uptake or microbial activity, and no information about the direction of movement of the water. Water tends to move from regions of high potential to regions of lower potential. A gradient in water potential is the driving force for water flow. A plant or microbe is therefore able to obtain water from the soil as long as it can maintain an internal water potential below that of the soil. Since there is a limit to the lowest potentials attainable by living organisms, organisms may cease to be able to obtain water from soil when the soil still contains a substantial quantity of water.

The gravitational and pressure potentials are important for determining rates and directions of water flow when the soil is saturated or near saturation. Both are proportional to the distance from a reference plane to the soil location in question. They are therefore measured with a ruler. Our focus here will be on the other components of the water potential, the osmotic and matric, which are more difficult to measure.

Measurement of Water Content

Gravimetric Methods

Mass wetness is typically measured by oven-drying a sample. The mass wetness is the mass loss divided by the dry mass of the sample. Samples are typically dried for approximately 24 h at 105°C. Aside from errors in weighing, the two most serious errors come from loss of mass of nonaqueous volatiles in the sample and uncertainty in the point at which the sample is really 'dry.' Increasing or decreasing the drying temperature changes the mass loss, and this change can be substantial in some soils. Variation in oven humidity can also lead to some uncertainty. Typical errors are on the order of 0.005 kg kg^{-1}. There are a number of faster drying methods which give quicker results at the expense of some accuracy. Drying of samples in a microwave oven can decrease the time required for a measurement to approximately 20 min, while still giving acceptable accuracy for many purposes.

Mass wetness is frequently used in laboratory studies, but has little relevance in the field, where one is mainly interested in the ability of the soil to store water for later transpiration and evaporation. Here volume wetness is the relevant measurement. The only direct method for measuring volume wetness is to find the mass wetness, through oven-drying, of a sample of known volume. From the known volume and the soil dry mass, the bulk density is computed. The volume wetness is then obtained using Eqn [1]. Since this is a tedious and time-consuming

measurement and is subject to large uncertainty owing to the large spatial variation found in the field (*See* **Spatial Patterns**), volume wetness is often inferred from measurements of other soil properties. These include dielectric properties, thermal properties, and interactions with nuclear radiation.

Dielectric Properties

The relative dielectric permittivity of water is approximately 80 at microwave frequencies, while the permittivity of soil minerals, organic matter, and ice ranges from 3 to 5. A measure of soil permittivity is therefore strongly influenced by the amount of water present in the soil. One of two methods is typically used to measure the permittivity of the soil: time-domain reflectrometry (TDR) and capacitance. For the TDR method, parallel conductors are placed in the soil forming a transmission line, and one measures the time required for an electromagnetic pulse to traverse the transmission line, reflect from the distal end, and return along the same line. The speed of propagation of an electromagnetic disturbance in the transmission line is determined by the dielectric permittivity of the surrounding medium. It is computed from:

$$\epsilon = \left(\frac{ct}{2L}\right)^2 \qquad [2]$$

where c is the speed of light ($3 \times 10^8\,\mathrm{m\,s^{-1}}$), t is the time to traverse the transmission line, and L is the length of the transmission line. (*See* **Time-Domain Reflectometry**.)

TDR measurements obviously require very fast response circuitry. The time resolution must be on the order of 100 ps. Equipment with this kind of time resolution is necessarily expensive. When slower circuitry is connected to the transmission line and a voltage pulse applied through a resistor of value R, the transmission line acts like a capacitor, with the soil as the dielectric. The time constant for charging the capacitor (time to charge to 63% of its final voltage) is RC, where C is the capacitance of the probe. The capacitance is related to geometric factors and is directly proportional to the dielectric permittivity of the medium surrounding the probe. The charging time is therefore directly related to the soil permittivity.

The dielectric constant of the soil can therefore be measured by either method. TDR, however, due to its higher-frequency excitation, is less influenced by electrical conductivity of the soil. Since ice has a dielectric permittivity near that of soil minerals, dielectric methods measure unfrozen water content in frozen soil.

A number of sensors are available commercially, both for TDR and capacitance measurement of soil moisture. An example of a capacitance sensor is shown in Figure 1. Sensors of this type have been used extensively for the past 20 years or so. Since they measure the properties of the dielectric within the electromagnetic field of the probe or transmission line, it is important that a representative sample of the soil lie within the field of the sensor. Unfortunately, the field lies close to the surface of the sensor, so any air gaps or compaction around the sensor may adversely affect the measurement.

Thermal Properties

Both thermal conductivity and heat capacity of soil vary with water content. (*See* **Thermal Properties and Processes**.) The relationship with heat capacity is an easy one to use for determining water content. The water content can be computed from the volumetric heat capacity of soil (*See* Eqn [13] in **Thermal Properties and Processes**):

$$\theta = \frac{C - C_s\rho_b/\rho_s}{C_w} \qquad [3]$$

Here C, C_s, and C_w (joules per cubic meter per kelvin) are the volumetric heat capacities of the bulk soil, the soil minerals, and water; and ρ_b and ρ_s are the bulk and particle densities. Since C_w is typically about twice as large as C_s, changes in water content have a fairly large effect on C. If the bulk density of the soil is known and C is measured, all values on the right-hand side of Eqn [3] are known, so water content can be computed. Typically C is measured using a dual-needle thermal properties sensor. Figure 2 shows a typical dual-needle sensor.

Nuclear Methods

Gamma rays and neutrons both interact strongly with soil water. The methods used most widely for water-content measurement involve attenuation

Figure 1 An example of a capacitance soil-water content sensor. The circuitry to measure the capacitance is in the overmolded head of the probe. The capacitance sensor is in the blade and consists of parallel conductors sandwiched between insulating layers.

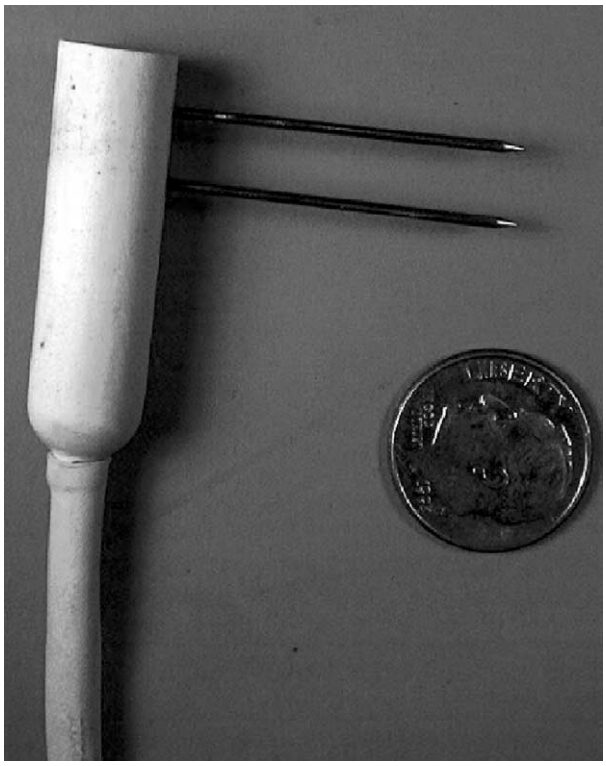


Figure 2 Dual-needle probe for measuring soil heat capacity. One needle contains a heater, the other contains a temperature sensor. The temperature rise in the temperature sensor resulting from a heat pulse to the heater is measured.

of transmitted gamma radiation and backscattering of neutrons, though neutron transmission and gamma backscatter have also been used. Gamma attenuation is typically a laboratory method. A soil column, 5–20 cm thick, is irridiated by a collimated beam of gamma radiation. Sources with appropriate energy levels are ^{137}Cs and ^{241}Am. The volume wetness is computed from:

$$\theta = \frac{\ln\left(\frac{I_d}{I_w}\right)}{\mu_w S} \qquad [4]$$

where I_d and I_w are the numbers of gamma photons passing through the dry and wet soil in unit time, respectively, μ_w is the mass attenuation coefficient for water, and S is the soil-column thickness. It is possible to determine the bulk density and the water content of a sample without measuring the gamma attenuation in the dry soil column by measuring the attenuation of two gamma beams having different energies.

Neutron scattering is a method for measuring volume wetness in the field. In a typical application, a neutron source such as ^{241}Am/Be is lowered into an access tube in the soil. The neutrons from the source initially have high energy, but, as they collide with low-atomic-mass nuclei (mainly hydrogen) in the soil, they slow and become 'thermal' neutrons, with energies typical of atoms at room temperature. These slow neutrons can be counted with a relatively simple detector. Since the main source of hydrogen nuclei in the soil is water, and since the number of scattered neutrons is proportional to the number of collisions with these hydrogen nuclei, there is a relationship between neutron count and water content:

$$\theta = a + b\frac{I}{I_{std}} \qquad [5]$$

where a and b are soil-specific empirical constants, I is the count in soil, and I_{std} is the count in a standard scattering medium. Neutron scattering is much less affected by air gaps around the sensor than are dielectric methods. The scattering volume for the neutrons varies with soil-water content from approximately 15 cm diameter in saturated soil to, e.g., 70 cm in dry soil.

Measurement of Water Potential

While the water potential concept is straightforward, no method exists for directly measuring the energy status of water in soil. All of the methods currently available equilibrate an external phase of some sort with the water in the soil and then measure the potential of the equilibrated phase. The sensing phase can be solid, liquid, or gas. Typical solid-phase sensors are heat-dissipation matric-potential sensors, gypsum blocks, dielectric matric-potential sensors, and filter paper. The tensiometer is a liquid-phase sensor. Thermocouple psychrometers and dew-point potentiometers are vapor-phase sensors.

Solid-Phase Sensors

For each of the solid-phase sensors, a porous matrix of ceramic, gypsum, or cellulose equilibrates with the soil matric potential. The water content of the matrix is then determined by weighing and drying (filter paper), measuring thermal conductivity (heat-dissipation sensor), electrical conductivity (gypsum block), or dielectric constant (dielectric sensor). From predetermined relationships between these measures of water content and matric potential, the measurements are converted to water potential.

Heat-Dissipation Matric-Potential Sensors

A heat-dissipation sensor is shown in Figure 3. It has a cylindrical, ceramic outer matrix that fully encases a stainless-steel needle. Inside the needle is an electrical resistor that runs the length of the probe

Figure 3 Heat-dissipation matric-potential sensor. The darker, upper part of the cylinder is the overmold, containing wires and connections. The lighter, lower part is the ceramic matrix, which equilibrates with the soil.

(typically 3 cm) and a temperature sensor, either a thermocouple or thermistor. To measure thermal conductivity, a known amount of current is passed through the resistor for a preset amount of time and data are collected on the temperature change (δT) over time (t). The thermal conductivity (λ_h) can be calculated from the slope (m) of the δT versus logarithm of time curve using:

$$\lambda_h \cong \frac{q}{4\pi m} \qquad [6]$$

where q is the power applied to the internal resistor (watts per meter). Because the ceramic matrix properties remain fixed, the thermal conductivity is a measure of the water content of the matrix, which, in turn, is related to the water potential of the soil. These sensors must be calibrated and, at midrange water potentials, have a strong temperature dependence. The heat-dissipation probe functions over an extremely wide range of water potentials (field capacity to air dryness).

Electrical Resistance Matric Potential Sensors

Electrical resistance sensors measure the electrical resistance between two electrodes packed in a granular (sand) or gypsum matrix to obtain water potential. As the sensor matrix comes into equilibrium with the surrounding soil, water fills the voids in the matrix. The increase in water causes the electrical resistance between the electrodes to decrease. The resistance of the sensor is related to the water content, and a value of water potential can be calculated using its predetermined relationship with water content. Electrical resistance sensors are the most inexpensive sensors available for measuring water potential, and this may be why they are also the most popular. They are easy to install and simple to automate for data-logging purposes; however, they are sensitive to changes in salinity and temperature. The largest pores in a gypsum matrix are still fairly small, so these sensors have very low sensitivity in wet soil.

Dielectric Matric-Potential Sensors

Dielectric matric potential sensors differ very little from the other solid-phase sensors discussed above. Instead of a heated needle or two electrodes, two or more wave guides are separated by a porous matrix. The wave guides measure the dielectric permittivity of the porous matrix. An empirical relationship between the matrix permittivity and matric potential is used to infer matric potential of the matrix, which is equal to the matric potential of the soil at equilibrium. The high accuracy of the dielectric sensor provides excellent resolution of water potential; in addition, careful selection of a matrix allows relatively quick equilibration in a wide range of soils. However, although dielectric sensors are not sensitive to temperature and salinity to the same degree as the electrical resistance sensor, they are not completely insensitive either.

Filter-Paper Method

The filter-paper method is a simple method for measuring water potential. A single disk of Whatman No. 42 filter paper is brought to equilibrium with a soil sample, which generally takes 2–3 days. The filter paper is then weighed and dried. The water content of the paper is used to calculate water potential using the relation:

$$\psi_m = -11w^{-3.68} \qquad [7]$$

where the relationship between the water content of the filter paper and its water potential in joules per kilogram is defined by the two constants, which are specific for the specified filter paper.

Tensiometer

The tensiometer equilibrates water inside a tube with the water in the soil matrix. The soil is separated from the water in the tube by a porous ceramic which is permeable to water and solutes, but not to soil colloids and air. When the water in the tube equilibrates with the soil water, the suction on

Table 1 Water-potential units

	$J\,kg^{-1}$	MPa	$m\ H_2O$	Relative humidity	Freezing point (°C)	pF	Pore diameter (μm)
	−1	−0.001	−0.1	0.999993	−0.0008	1.01	290.08
	−10	−0.01	−1.02	0.999926	−0.0076	2.01	29.01
FC	−33	−0.033	−3.37	0.999756	−0.0252	2.53	8.79
	−100	−0.1	−10.2	0.999261	−0.0764	3.01	2.9
	−1000	−1	−102.04	0.992638	−0.7635	4.01	0.29
PWP	−1500	−1.5	−153.06	0.988977	−1.1453	4.18	0.19
	−10 000	−10	−1020.41	0.928772	−7.6352	5.01	0.03
Air dry	−100 000	−100	−10 204	0.477632		6.01	
Oven dry	−1 000 000	−1000	−102 041	0.000618		7.01	

Relative humidity computed assuming a temperature of 20°C.
FC, a typical value for the drained upper limit of water potential in soil; PWP, a typical value for the lower limit of plant-available water in soil; air dry, typical value; varies with atmospheric humidity; oven dry, typical value; varies with oven humidity and oven temperature; pF, base 10 logarithm of the water potential in centimeters of water; pore diameter, bubble pressure or diameter of the largest water-filled pore at the indicated potential.

Figure 4 Dew-point potentiometer for measuring water potential of soil samples using vapor equilibration. The sample is placed in the cup, sealed inside the instrument, and the dew point and sample temperature are measured.

the water is equal to the soil matric potential. (*See* **Tensiometry**.)

Vapor-Pressure Methods

If soil water is allowed to come to temperature and vapor equilibrium with a vapor phase, the potential in the vapor phase becomes the same as that in the liquid phase. Since a vapor barrier is impermeable to solutes as well as soil colloids, the potential in the vapor phase is equal to the sum of the osmotic and matric potentials in the soil. The potential in the vapor phase can be determined by measuring its relative humidity. The Kelvin equation relates humidity and water potential:

$$\psi = \frac{RT}{M_w} \ln h \qquad [8]$$

where R is the gas constant, T is temperature (Kelvin) M_w is the molecular mass of water, and h is the relative humidity (a fraction, not a percentage). It is possible to measure h by measuring either the wet-bulb depression or the dew-point depression of the headspace air. For soil wet enough to sustain plant growth, the relative humidity is greater than approximately 0.99, so the psychrometer or dew-point meter must be extremely accurate and sensitive. A number of such devices are available which resolve changes in humidity of 0.0001 at these high humidities. Figure 4 shows a dew-point meter for water-potential measurement.

Water-Potential Units

Water potential has been expressed in a variety of ways and with a variety of units. As stated earlier, water potential is the potential energy per unit quantity of water, where the quantity is sometimes mass, sometimes volume, and sometimes weight. Potential energy per unit volume is dimensionally equivalent to pressure, so pressure units are often used; equivalent concentrations of ideal solute, relative humidity, and freezing point have also been used. Table 1 shows some of these units.

See also: **Hydrodynamics in Soils**; **Neutron Scattering**; **Spatial Patterns**; **Tensiometry**; **Thermal Properties and Processes**; **Time-Domain Reflectometry**; **Water Potential**

Further Reading

Campbell GS, Calissendorff C, and Williams J (1991) Probe for measuring soil specific heat using a heat pulse method. *Soil Science Society of America Journal* 55: 291–293.

Flint AL, Campbell GS, Ellett KM, and Calissendorff C (2002) Calibration and temperature correction of heat dissipation matric potential sensors. *Soil Science Society of America Journal* 66: 1439–1445.

Gardner WH (1986) Water content. In: Klute A (ed.) *Methods of Soil Analysis*, 2nd edn, part 1. *Physical and Mineralogical Methods*. Agronomy Monographs No. 9, pp. 493–544. Madison, WI: American Society of Agronomy.

WATER CYCLE

D K Cassel and B B Thapa, North Carolina State University, Raleigh, NC, USA

Introduction

The water cycle is an endless process of water circulation on the planet Earth. Water is an essential constituent of all plants and animals, most of which contain more than 60% water, and many contain more than 95%. The planet Earth is dominated by the hydrosphere, which contains all of the Earth's water. Approximately 1.39 billion km^3 of liquid water is stored in depressions, primarily oceans, which occupy over 70% of the Earth's surface. Water is unique in many ways: it has a high heat capacity, high heats of vaporization and fusion, and can exist in solid, liquid, and gaseous phases at temperatures commonly encountered on Earth. The ease with which water can move from one phase to another in response to additions or losses of heat or energy allows water to move or cycle from one storage field to another.

Overview

Approximately 97% of the Earth's water is salty and is stored in oceans (Table 1). The remaining 3% is fresh water and, if uncontaminated, is considered to be potable and drinkable by animals and utilized by nonsalt-tolerant plants. Fresh water is stored primarily as ice in glaciers and icecaps, but also in groundwater, lakes, streams, and rivers. Water retained as soil moisture is a small fraction of the total, but is extremely important for the production of food and fiber. The amount of water existing in the gaseous state is relatively constant and is estimated to be 12 900 km^3.

Table 1 Estimate of global water distribution

Source	*Volume × 1000 km³*	*Total water (%)*	*Fresh water (%)*
Oceans	1 338 000	96.5	–
Ice, glaciers, snow	24 064	1.74	68.7
Groundwater			
Fresh	10 530	0.76	30.1
Saline	12 870	0.94	–
Soil moisture	16.5	0.001	0.05
Ground ice and permafrost	300	0.022	0.86
Lakes			
Fresh	91.0	0.007	0.26
Saline	85.4	0.006	–
Atmosphere	12.9	0.001	0.04
Swamps	11.47	0.0008	0.03
Rivers	2.12	0.0002	0.006
Biological water	1.12	0.0001	0.003
Total fresh and saline	1 385 984	100.0	100.0

Reproduced with permission from Shilkomanov IA (1993) World fresh water resources. In: Gleick PH (ed.) *Water in Crisis: A Guide to the World's Fresh Water Resources*. New York: Oxford University Press.

The water cycle, commonly called the 'hydrologic cycle,' is a conceptual model that relates to the transport of water to and from the various storage fields or pools on the Earth. Movement of water from one storage field to another depends on the energy sources available for the processes of evaporation, condensation, precipitation or deposition, runoff, infiltration, melting, and groundwater flow. It is estimated that 380 000 km^3 of water are recycled each year. Human recognition of this water-recycling process has led to the utilization of renewable (recycling) flowing water to generate hydroelectricity in many parts of the world. A given water molecule could follow an infinite number of pathways as it progresses through the various processes in the water cycle.

At the global scale (macro scale), the water-cycle model appears to be a simple process of evaporation and precipitation and is typically viewed as transport of water from oceans and lakes to land masses and back to oceans and lakes. The water, once on land, is eventually returned to the oceans in runoff water or as precipitation fed by water evaporated from the soil or transpired by plants. When considered at the regional scale, human intervention in the water cycle often exists, and the water-cycle model becomes more complicated. For example, in arid regions, in addition to precipitation, water previously stored in the groundwater might be 'mined' for domestic purposes, or water falling in another region might be transported to the arid region for domestic or commercial uses or to satisfy the transpiration demand to grow crops. Conversely, water in humid regions might be impounded in dams and transported by canal or pipeline to another region or watersheds.

The water cycle becomes even more complex at the local or field scale, and we often view it as the water balance. The individual mechanisms of water transport become more important at smaller scales. For example, at the local scale, a high percentage of the water might be transferred to the land by irrigation canals rather than by precipitation. Likewise, lateral subsoil transport of water from sloping fields might be the major mode of transport rather than runoff. In

many agricultural fields, a conscious effort is made to reduce or at least control runoff to allow more water to infiltrate and to control soil erosion. The soil moisture component of the water cycle may appear to be insignificant at the global scale (Table 1), but at the field scale it is elevated to primary importance in growing plants.

The quality of waters transferred by runoff, groundwater flow, and precipitation is an important consideration. Pollutant-laden runoff has significant negative impacts on aquatic habitats and decreases the usefulness of water for further use. As the population increases, not only is more water needed, but human activities tend to impair water quality as it continues to travel in the water cycle. Examples of water-quality impairment are the contamination of groundwater by organic and inorganic chemicals, and the fouling of lakes, streams, and oceans by rubbish and human and animal wastes. Acid rain due to sulfur, a by-product of some industrial operations, has damaged plant and animal habitats. These impacts on water quality can lead to health hazards and drastic environmental and economic impacts at the regional and local scales. Fortunately, the evaporation process in the water cycle purifies polluted water, allowing it eventually to reappear as fresh precipitation to be reused to support plant and animal life.

Water Cycle Processes

The water cycle is a conceptual model to describe the transfer and temporary storage of water among various storage fields or reservoirs. These reservoirs include lakes, streams, oceans, groundwater, soil moisture, glaciers, snowfields, the atmosphere, and the biosphere. Energy balance and the water cycle are closely linked. Water is transferred from one storage field to another by processes of evaporation, transpiration, sublimation, condensation, precipitation, infiltration, runoff, and groundwater flow (Figure 1). An estimate of the approximate residence time in various water reservoirs is given in Table 2.

Evaporation converts liquid water to the gaseous phase. Evaporation is a distillation process that leaves all impurities and contaminants behind in the soil or water body from which it evaporates. Evaporation of water requires a large quantity of energy, most of which comes from the Sun. The heat required to vaporize water is $540\,\mathrm{MJ\,kg^{-1}}$ at 373 K (101 kPa) and increases to $2.26\,\mathrm{MJ\,kg^{-1}}$ at 273 K. Evaporation occurs when the vapor pressure of the atmosphere is less than the vapor pressure of liquid water at the evaporating surface. The evaporation rate is controlled by the energy available, temperature, humidity, wind speed, and salinity of the water. On a global

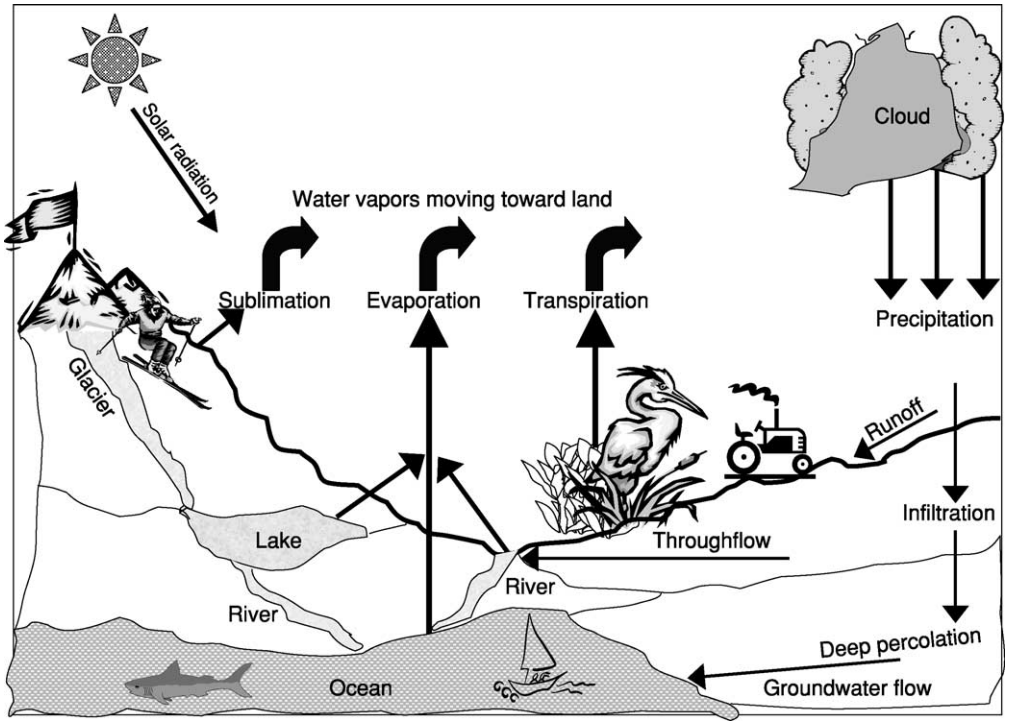


Figure 1 The water cycle.

Table 2 Approximate residence time of water in various reservoirs

Reservoir	*Residence time (years)*
Glaciers	40
Seasonal snow	0.4
Soil moisture	0.2
Groundwater, shallow	200
Groundwater, deep	10 000
Lakes	100
Rivers	0.04

Adapted from Wetzel RG (1983) *Limnology*, 2nd edn. New York: Saunders College Publishing.

scale, evaporation of water is most intense in the subtropical oceans, where solar radiation is relatively constant throughout the year and provides a constant energy source to vaporize water.

Sublimation is conversion of solid-phase water (ice) directly to the vapor phase without passing through the liquid phase. This process requires approximately 2.85 MJ kg^{-1}.

Transpiration is the conversion of water from liquid to gas as it passes through plant stomata, small openings on the undersides of leaves of vascular plants. The stomata of some plants have the ability to open and close. Transpiration is a passive process, being controlled by several factors. When soil water content is high (absolute value of soil water pressure is small), the transpiration rate of plants with extensive root systems tends to be controlled by the solar energy available. Wind tends to increase transpiration rate. In general, 99% of the water entering plant roots passes through the plant and is lost to transpiration.

The processes of evaporation and transpiration on a given area of land often occur simultaneously, and it is difficult to differentiate between the evaporative loss of water from the land surface and the transpiration loss by plants. In such cases the combined loss of water by these two processes is referred to as evapotranspiration. Generally, potential evapotranspiration exceeds precipitation from landmasses during the summer at middle and high latitudes.

Condensation is the process whereby water is converted from the gaseous phase to the liquid phase. This process releases a vast amount of heat. Condensation occurs when the atmospheric temperature cools below the point at which the relative humidity approaches 100%. When this condition is reached, condensation nuclei such as dust, smoke particles, or salts provide sites for water vapor to condense to form water droplets. These condensed water droplets form clouds or fog. The water droplets in the clouds continue to increase in size when the right conditions are encountered and eventually coalesce into ice crystals or a liquid. Condensation nuclei such as salts, e.g., NaCl, are hygroscopic, thus allowing water to condense at a relative humidity of less than 100%. As clouds form, air currents move them, thus dispersing the water vapor around the Earth. When clouds eventually encounter environmental conditions where they can no longer hold the moisture, they release some or part of it as precipitation. Another result of the condensation process is the formation of dew during summer and autumn nights, when the air temperature cools to the point where the relative humidity at the plant or land surfaces rises to 100%.

Precipitation is any aqueous deposit, whether in liquid or solid form, that develops in a water-vapor-saturated atmosphere and falls to land or water surfaces. In general, precipitation over continents tends to exceed evaporation, whereas over oceans, evaporation tends to exceed precipitation. Most precipitation forms from natural clouds, but most clouds exist without yielding precipitation. On the other hand, small precipitation events derived from water vapor emitted from smoke stacks are not uncommon. Precipitable water is the amount of water potentially available in the atmosphere for precipitation, usually measured in a vertical column that extends from the Earth's surface to the upper edge of the troposphere. In many clouds, water droplets and ice crystals are too small to overcome natural updrafts found in the atmosphere. Hailstones are formed by the continual thickening of ice crystals until they become heavy enough to overcome the updraft.

The distribution of rainfall reaching the soil surface can be modified by vegetation covering all or part of the soil surface. Deforestation in some regions of the world has changed the rainfall pattern. The amount of rainfall intercepted by vegetative cover depends on leaf density and the branching structure of the plants. Some of the water is evaporated directly from the vegetative surfaces, while some of it might drip from leaves or move by 'stemflow' down the leaves and stems to the soil surface.

The intensity, duration, and amount of rainfall events control to some degree the pathway water takes once it arrives at the soil surface. Rainfall intensity varies during a storm event. If the intensity is high, or if the duration is long, the likelihood is great that some of the water reaching the soil surface will run off. The energy associated with high-intensity rainfall increases the opportunity for soil erosion, a process that degrades productive soil as well as the quality of the runoff water.

Irrigation is the application of water to the soil, usually for the purpose of growing food or fiber. This very important process in the water cycle is human-imposed, but has tremendous impact on other processes in the water cycle. Approximately

40% of the world's food comes from irrigated land. One common source of irrigation is water diverted from large, constructed reservoirs. These reservoirs are often hundreds of kilometers away from the land being irrigated. Other sources of irrigation water are ponds, streams and rivers, and groundwater. Irrigation accounts for over 65% of the fresh water use, but, because of inefficiencies at various points in the irrigation process, less than half of that water reaches the plant roots. The remaining half is lost to evaporation during transport of water to the land to be irrigated, deep percolation, and runoff. Runoff water leaving the lower boundary or an irrigated field (irrigation return flow) often has a high salt content, thus reducing water quality of streams into which it drains. Irrigation water is applied to the soil by flooding, spraying, sprinkling, furrowing, and trickle methods. Groundwater resources continue to be depleted in some regions of the world due to pumping more water than is being recharged.

Infiltration is the process of water entry into the soil, generally by downward flow through the soil surface. Liquid water falling on the soil surface is temporarily stored on the surface, runs off, or infiltrates. The infiltrability, that is, the rate at which water enters the soil when water at atmospheric pressure is available at the soil surface, decreases with time. The infiltration rate of unsaturated soils is controlled by a number of factors, including the condition and cover of the soil surface, soil texture, pore-size distribution and porosity, soil water pressure or soil suction, soil structure, the rate at which water arrives at the soil surface, antecedent soil water content, and the permeability of underlying soil horizons. If a soil is saturated, water will not infiltrate because there is no pore space for it to enter. Coarse-textured soils usually have high infiltration rates unless they have high antecedent water contents. Fine-textured soils tend to have low infiltration rates, increasing the chances for water to run off. For some soils, water infiltrates along preferred pathways due to the presence of cracks and fissures, soil 'pipes,' and large pores created by animal or plant roots.

Runoff is excess water that does not infiltrate into the soil. When liquid water as precipitation or irrigation arrives at the soil surface at a rate that exceeds the soil's infiltration rate, the excess water is temporarily stored in small depressions on the soil surface. Once these depressions on sloping land are filled, additional water arriving at the soil surface that does not infiltrate is transported across the soil surface as runoff. Gravity provides the energy to transport runoff, or 'overland flow,' from higher to lower elevations. Overland flow occurs as sheet flow on the soil surface. Runoff water concentrates in rivulets or rills, waterways, and gullies, eventually emptying into streams and rivers, and finally into lakes and oceans. This runoff water provides energy that can transport detached soil materials in the processes of sheet, rill, and gully erosion. Flooding can occur if too much water runs off in a short time period. In some arid regions, land is set aside and managed to increase the amount of runoff water that is usually collected in reservoirs at lower elevations for various consumptive uses.

Drainage When soils become saturated or nearly saturated, the possibility that some of the water will continue to drain or percolate deeper into the soil or into the underlying water table increases. This process recharges ground water supplies and also provides 'base flow' water for streams and rivers. When irrigation is practiced, it is imperative that some water drains below the root zone to remove salts that are added to the soil in the irrigation water. Open ditches or buried drains carry the drainage water to a discharge outlet, where it is returned to a river or stream. Mobile chemicals present in the soil move in the percolating and drainage waters. Transport of chemicals has led to the contamination of groundwater, streams, rivers, and reservoirs throughout the world and is a continuing problem.

Environmental and Economic Implications of Water-Cycle Processes

Tremendous environmental impacts are associated with the processes in the water cycle. At the global scale, the geography of the entire world along with its many beautiful geologic features has been shaped primarily by the various components of the water cycle, without human intervention, acting over extended periods of time. For millions of years the water-cycle processes, driven and interacting with the energy-distribution patterns on the Earth's surface, have shaped Earth's surface. Geologic erosion, deposition of eroded and transported solids, the formation of rivers, lakes, and streams by runoff water, and the formation of wetlands in low areas are results of water-cycle processes. All of these processes result from a combination of ordinary and episodic events associated with the water cycle.

At the regional scale, one can only imagine the flooding that occurred as huge amounts of runoff water loosened and scoured minerals from higher elevations, transporting them to lower elevations to settle and eventually develop into sedimentary and metamorphic rocks. Phosphates, carbonates, and placer gold are but a few of the numerous minerals and chemicals that are mined that have been concentrated in certain regions throughout the world by

water-cycle processes. Similarly, very fertile alluvial flood plains, such as the Indo-Gangetic Plain in Asia, have developed.

Glaciation, a process that occurred in some regions of the Earth in the past and presently is ongoing in some regions, transports vast amounts of ice that scour and shape the surface features of the Earth. As glaciers melt, rivers, lakes, prairie potholes (closed depressions), wetlands, and numerous other features are formed. At the local scale, for example, the magnitudes of the individual processes of the water cycle for a prairie pothole in eastern North Dakota differ from those for land just a few hundred meters away. Surface runoff from land in the small, depressional watershed surrounding the pothole contributes to the water stored in the pothole, where it will be evaporated or percolate downward to the groundwater. On the other hand, runoff water from the nearby land outside the depressional watershed contributes its runoff to streams. This seemingly small difference creates great diversity in ecology within a distance of 100 m or less.

Extreme variation in the magnitude of one or more components of the water cycle, for example, precipitation, causes catastrophic events such as flooding or drought, although these two terms are defined in terms of norms as perceived by humans today. Yet flooding as a catastrophic event has occurred for billions of years and has been a key process in the formation of sedimentary rocks.

On a shorter time scale, vagaries in various processes of the water cycle have caused disasters for humans. In fact, humans have altered the water cycle at various scales for millennia, and currently the process of global warming, which is still a highly debated topic, is intimately linked with the water cycle. Impacts of global warming that have been predicted using models are shifts in temperature, rainfall distribution, and food production. These shifts are predicted to have greater impacts in some regions of the world than in others. Acid precipitation is another man-induced problem that affects the quality of water and is changing the ecology of large regions. Nitrogen and sulfur oxides released into the air as a result of industrial processes and fuel combustion are further oxidized and converted into nitric and sulfuric acids. These acids are washed from the air by precipitation. Acid precipitation refers to wet forms of acid pollution contained in rain, sleet, snow, fog, and vapor transport. Still another problem affecting the quality of water is the input of chemicals and heat that is discharged into stream waters. Excess nitrogen and phosphorus can lead to excessive algal growth. Siltation of lakes and rivers and accompanying fish kills are common occurrences.

Water in lakes and streams has a large impact on quality of life and the economy. Water for transportation led to the development of ancient and modern civilization. Water for industrial and domestic use is diverted from lakes, streams, and groundwater supplies. After the liquid water is used, much of it continues its travel in the water cycle as 'wastewater' discharged into streams and lakes. This water is reused further downstream for additional domestic and industrial applications. In some locations energy in the form of hydroelectricity is 'harvested' from the flowing water. Without this source of energy, many areas of the world could not have been developed. The use and reuse of water continue until the water eventually evaporates or reaches the ocean and then evaporates. The development of cities, industries, agriculture, and recreational parks are dependent on water as it passes along the various stages in the water cycle.

Land-Management Impacts

Human interventions in the 'natural' water cycle at many scales have occurred as a result of implemented land- and water-management (or mismanagement) practices. Several examples of the effects of management practices involving the water cycle at different scales, either directly or indirectly, follow. Intervention of the water cycle at the regional scale began in ancient times with the construction of irrigation canals that diverted water from one region to another, thus increasing the amounts of water applied to irrigated areas. The anticipated response of increasing food production by applying greater amounts of water to the lands was obtained. However, an unanticipated regional-scale response was that, in time, the soils became saline and were eventually abandoned for crop production. Failure to provide soil drainage led to the gradual increase in salt content of irrigated soils. Eventually, even salt-tolerant crops failed to grow. This undesirable salinization process was a direct effect of inadequate land and water management. Unfortunately, today we are still losing arable land due to salinization caused by inadequate drainage of irrigation water.

Another type of salinization problem, caused by modification of the water cycle as a result of changes in land management at the field or landscape scale, arose in the subarid Northern Great Plains (USA) due to the adoption of 'summer fallow.' Dryland small grain production was, and still is, common in this area even though precipitation in a typical year is often insufficient to produce a grain crop. The practice of summer fallow was developed to accumulate water in the soil profile, thereby using the precipitation occurring over a 2-year period to produce one

grain crop every 2 years. Typically, the soil was kept bare the year before planting the grain crop (year 1) so that water falling as precipitation during that year would infiltrate, thus increasing the amount of stored soil moisture. The grain crop planted in year 2 would derive benefit from the stored moisture in year 1 in addition to precipitation occurring in year 2. Use of this practice over decades gave rise to 'saline seeps.' In essence, summer fallow was inefficient in conserving water for the following crop. Much of the water evaporated from the soil surface in year 1. However, some of the water in higher-rainfall years percolated through the soil. This percolating water carried salts leached from underlying strata having high salt contents. When the drainage water encountered an aquatard, it moved obliquely downslope and eventually seeped from the hillside, causing excessively wet, salty nonarable soils at lower elevations.

As the human population increases, more and more land is being altered and subjected to different land uses. The land use changes have impacts on the processes of the water cycle, the scale of impact being dictated by the extent and nature of land management. Compaction of soils by farm and construction machinery, as well as human and animal footprints, decreases infiltration with a resultant increase in runoff. The eroded soil material from farmland and construction sites may include adsorbed hazardous chemicals. The construction of sedimentation basins to collect runoff water to allow solids removed in runoff water to settle out is a management practice that has arisen to help control the environmental degradation.

The construction of dams and ponds occurs at many scales and, regardless of scale, they cause some alteration of the magnitudes of the processes in the water cycle. The water that fills ponds and lakes impounded by dams arises from runoff waters, groundwater, seeps, and springs. The impounded water in farm ponds is often used for irrigation. At a regional scale, for example, some of the large dams constructed in California impound water that is transported hundreds of miles and used for domestic consumption, industry, and irrigation. In some regions of the world, so much water is diverted for irrigation and other human uses that rivers run dry for parts of the year. Rivers affected in this manner include the Yellow River in China, the Ganges in South Asia, and the Colorado in the United States. In addition, in many regions the groundwater supplies are being mined for irrigation and other uses at rates exceeding their recharge rates, resulting in dropping water tables.

Urbanization has drastically altered the water cycle at various scales. Precipitation falling on roofs of houses and buildings, on streets and highways, and paved parking lots near industries, schools, and shopping centers does not infiltrate the soil, but increases the runoff component of the water cycle. This stormwater runoff carries suspended solids and numerous chemicals, such as spilled petroleum products, and nitrogen and phosphorus fertilizers that degrade water quality and cause eutrophication. In addition, it is common for the increase in amount of runoff water to initiate or exacerbate downstream flooding. Following Hurricane Floyd in 1999, the state of North Carolina began a program to develop new flood plain maps for five river basins which incorporates changes in flood plain elevation caused by two decades of development.

Summary

Human life cannot exist without water: life originated in water, and civilization began near water; water is the shelter for many marine creatures; water fulfills human recreational needs. The importance of water in the world today cannot be overstated. In addition to being the major constituent of human life forms, water is a required input to produce food and clothing, manufacture products, transport people and products around the world, and produce a significant percentage of the world's energy.

Water molecules are in continuous motion and keep moving from one storage field to another. Liquid water in rivers moves downslope to the ocean, where solar radiation converts it to the gaseous phase, which in turn transports the water vapor back to the landmasses, where it condenses to form precipitation. The precipitation is divided into water that infiltrates the soil and runoff water that carries with it suspended solids, chemicals, pesticides, and organic and inorganic wastes. This endless cycle is called the water cycle or hydrologic cycle.

The total amount of water on Earth remains constant, but it is separated into different processes or pathways depending on environmental factors and human interventions. Thus processes such as evaporation, condensation, precipitation, infiltration, and runoff and subsurface water movement induce the transfer of the three phases of water. The evaporation processes can be viewed as a distillation process that purifies water, leaving impurities in the ocean, whereas runoff and percolation are water-polluting processes of the water cycle. The runoff process affects the quality of surface waters. Fish kills as a result of algal growth and eutrophication are primarily due to runoff generated in response to inappropriate land management, deforestation, and misuse of heavy machinery, chemical fertilizers, and pesticides.

Percolated water can leach undesirable chemical compounds below the rooting zone of plants and into drainage water or through the vadose zone into the groundwater. Even though the ocean serves as a huge sink for salts and pollutants, the contamination of streams, rivers, lakes, and groundwater resources poses a great threat to plant and animal health, the environment, the economy, and civilization.

Further Reading

Berner EK and Berner RA (1996) *Global Environment: Water, Air, and Geochemical Cycles*. Upper Saddle River, NJ: Prentice-Hall.

Browning KA and Gurney RJ (1999) *Global Energy and Water Cycles*. Cambridge, UK: Cambridge University Press.

Carpenter SR, Fisher SG, Grimm NB, and Kitchell JF (1992) Global change and freshwater ecosystems. *Annual Review of Ecology and Systematics* 23: 119–139.

Gleick PH (1993) An introduction to global fresh water issues. In: Gleick PH (ed.) *Water in Crisis: A Guide to the World's Fresh Water Resources*, pp. 3–12. New York: Oxford University Press.

Guymon GL (1994) *Unsaturated Zone Hydrology*. Englewood Cliffs, NJ: Prentice-Hall.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Howells G (1995) *Acid Rain and Acid Waters*, 2nd edn. New York: Ellis Harwood.

Postel S (1999) *Pillar of Sand: Can the Miracle Last?* Saddle Brook, NJ: WW Norton.

Vorosmarty CJ, Green P, Salisbury J, and Lammers RB (2000) Global water resources: vulnerability from climate change and population growth. *Science* 289: 284–288.

Wetzel RG (1983) *Limnology*, 2nd edn. New York: Saunders College.

Water Erosion *See* **Erosion:** Water-Induced

WATER HARVESTING

D Hillel, Columbia University, New York, NY, USA

Definition

The term 'water harvesting' generally refers to the collection of rainstorm-generated runoff from a particular area (a catchment) in order to provide water for human, animal, or crop use. The water thus collected can either be utilized immediately, as for irrigation, or be stored in aboveground ponds or in subsurface reservoirs, such as cisterns or shallow aquifers, for subsequent utilization. As such, water harvesting is an ancient practice that has enabled some societies to subsist in semiarid and arid areas where other sources of fresh water (e.g., rivers, lakes, or aquifers) are scant or unavailable.

Surface Runoff

Whenever the rate at which rainwater is applied to the soil surface exceeds the rate of infiltration into the soil, the excess tends to accumulate over the surface. Where the surface is not perfectly flat and smooth, the excess water collects in depressions, forming puddles. The total volume of water thus held, per unit area, is called 'surface-storage capacity.' It depends on the geometric irregularities (roughness) of the surface as well as on the overall slope of the land (Figure 1).

Only when the surface storage is filled and the puddles begin to overflow can actual runoff begin. The term 'surface runoff' thus represents the portion of the water supply to the surface that is neither absorbed by the soil nor accumulates on its surface, but that runs downslope.

Surface runoff typically begins as sheet flow but, as it accelerates and gains in erosive power, it eventually scours the soil surface to create channels. There exists

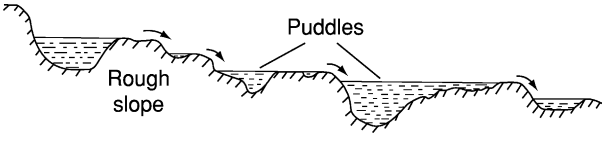


Figure 1 Effect of surface roughness and slope on surface storage of rainfall excess. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

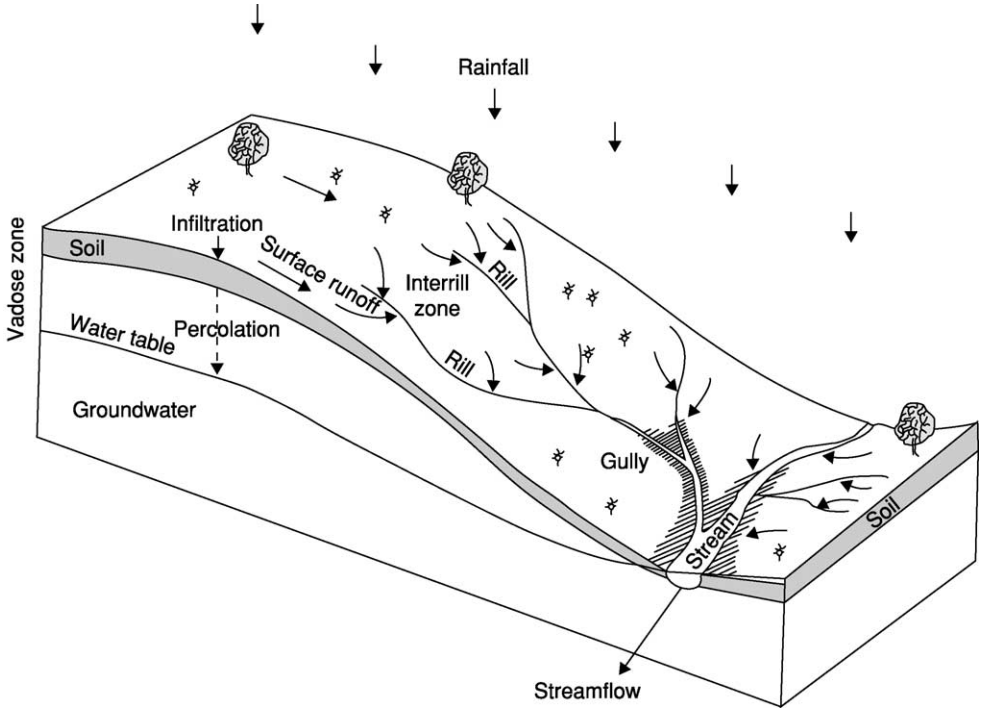


Figure 2 A sloping area exhibiting overland flow as well as runoff through rills and gullies. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

a wide spectrum of channel geometries and flow patterns. On the one extreme is the thin, sheet-like runoff called 'overland flow.' It is likely to be the primary form of surface runoff from small areas or fields having little topographic relief. The next distinctive form of flow takes place in small channels called rills. The latter gather the overland flow in a continuous fashion along their length to form the lowest order of stream flow. As these small streamlets merge with one another, they form higher-order channels, called gullies, which collect concentrated tributaries as well as distributed (lateral) overland flows (Figure 2).

Runoff Control and Utilization

Uncontrolled runoff is never desirable, as it is likely to cause soil erosion on slopes as well as flooding and silting in bottomlands. In humid regions, where rainfall may be excessive, measures may be needed to ensure the safe routing and conveyance of the runoff. Such measures, called 'surface drainage,' include shaping the land and treating it so as to direct the runoff via protected (grassed or even concrete-lined) channels. In semiarid regions, by way of contrast, natural rainfall is barely sufficient for crops, hence farmers typically strive – by such means as terracing and mulching – to cause as much of the rainfall as possible to infiltrate the soil, and thereby to minimize runoff.

The situation is fundamentally different in arid (as distinct from semiarid) regions. In many arid regions, large tracts of land are basically unsuitable for conventional rainfed farming, owing to the paucity and instability of rainfall, the nature of the soil (too shallow, stony, or saline to permit cultivation), or the rough topography. From the point of view of farmers (though not of ecologists, who are concerned about an area's natural biota rather than about crop production per se), rain falling on such lands is almost totally lost – being insufficient either to recharge groundwater or to support an economically viable crop.

Most of the meager rainwater generally infiltrates the soil to a shallow depth only, and it is quickly returned to the atmosphere – either by direct evaporation from the soil or by transpiration from native vegetation. Occasional intense rainstorms may none the less generate runoff and cause sudden flash floods. Although runoff from any particular rainstorm may be high, under natural conditions the total seasonal runoff seldom exceeds 10% of the annual precipitation.

The possibility of controlling and even increasing the amount of surface runoff obtainable from such lands can be of great importance. Particularly where no other dependable water source is available, the runoff thus obtained may constitute the major supply of an inhabited area.

Ancient Methods

The art of inducing, collecting, and utilizing runoff has been practiced by desert-dwelling communities since antiquity, in such disparate regions as southwestern Asia, northern Africa, and southwestern North America. Remnants of extensive water-harvesting systems are found in these regions.

Highly noteworthy are the works of the ancient Nabateans, who inhabited the Negev Desert of southern Israel some 2000 years ago. They began as caravan drivers, conveying aromatic, spicy, and medicinal plant-products, as well as other precious objects of trade, from distant sources across the desert to the major population centers along the shores of the Mediterranean Sea. The Nabateans established stations along their trading routes, and these grew in time into permanent settlements and even into cities. To sustain their population in the desert environment, the Nabateans built extensive terraces and channels, and hewed out numerous cisterns. Remains of their cities (most notable among them being the city of Petra, in southern Jordan) can be visited today.

The first imperative of desert settlement was the provision of potable water for humans and livestock. This was done by means of cisterns, which are artificially constructed reservoirs filled with surface flows during infrequent rains. The early cisterns were undoubtedly leaky and inefficient. Building efficient cisterns became possible only with the advent of watertight plaster, made of burned and slaked lime. Also crucial was the ability to recognize suitable rock formations, such as soft marly chalk, which could be hewed out readily and was not as fissured and leaky as the hard limestone that is also prevalent in the same region. The Nabateans were also skilled at choosing appropriate sites for their cisterns and at ensuring that they could be filled with water annually.

Where cisterns could be located alongside natural streams, they were filled directly by flash floods. However, most cisterns in the Negev were built along the lower reaches of hillsides and depended on the direct collection of runoff from the higher slopes. Many hundreds of such cisterns were built in the Negev, and they are clearly discernible landmarks even today. A typical one resembles a giant necklace, with the glistening white pile of excavated rock appearing to hang like a pendant from the two collection channels that ring the hill and curve down its sides from opposite directions (Figure 3). To parched travelers through the desert, to whom these cisterns would beckon from afar, no sight could be more gladdening.

The application of water-harvesting techniques to enable farming in desert areas (that is to say, collecting runoff from sloping areas and using it to irrigate plots or tracts of flat land, where crops could be grown) has been called 'runoff farming.' In ancient times, stone dikes were constructed across tributary streambeds, thus forming a series of terraced fields. These fields were watered by flows arriving from the upper watershed of the stream, as well as from the adjacent hillslopes. The slopes were often divided into sections, the runoff from each of which was led by means of constructed channels to specific fields (Figure 4). The water retained and infiltrated in each field was then utilized by annual crops (e.g., wheat or barley) or by perennial crops (e.g., grapevines, olive trees, or other fruit trees).

Water Spreading

Another form of water harvesting, called 'water spreading,' consists of diverting flash floods from intermittent streams (known as *wadis* in the Middle East, *arroyos* in the American Southwest, and *dongas* in parts of Africa) on to adjacent tracts of land. It is a simple form of flood irrigation, controlled by dikes, check dams, or channels designed to direct and spread the expected flow. It is generally used to irrigate pastures and rangelands, but in places also to sustain groves of trees in arid areas.

Water spreading is typically practiced on small watersheds of a few square kilometers. In larger watersheds, the floods may occasionally be too violent or torrential to be controlled by simple diversion structures. In such cases, a more complex system may be required: the construction of dams designed not to retain the floods but merely to detain and regulate them, so as to provide farm units located downstream from the dam with controlled flows. Such dams, called 'detention dams,' are built across a stream in order temporarily to impound the flood. A large-diameter open pipe is laid through the dam to permit downstream flow at a predetermined rate, as through the drain of a bathtub. Thus, a flash flood that would normally last just a few hours is made to flow through the pipe and on to a series of fields for perhaps several days. The field dikes (made of compacted earth or stone) can then be built economically and safely to withstand floods of a known maximal intensity, so farming operations can be planned accordingly (Figure 5).

Runoff Inducement

The builders of runoff utilization systems in the Negev in ancient times had to contend with the paucity of natural runoff, and there is evidence that they actually strove to augment it. They developed techniques for inducing a greater portion of the rainfall to trickle downslope as runoff.

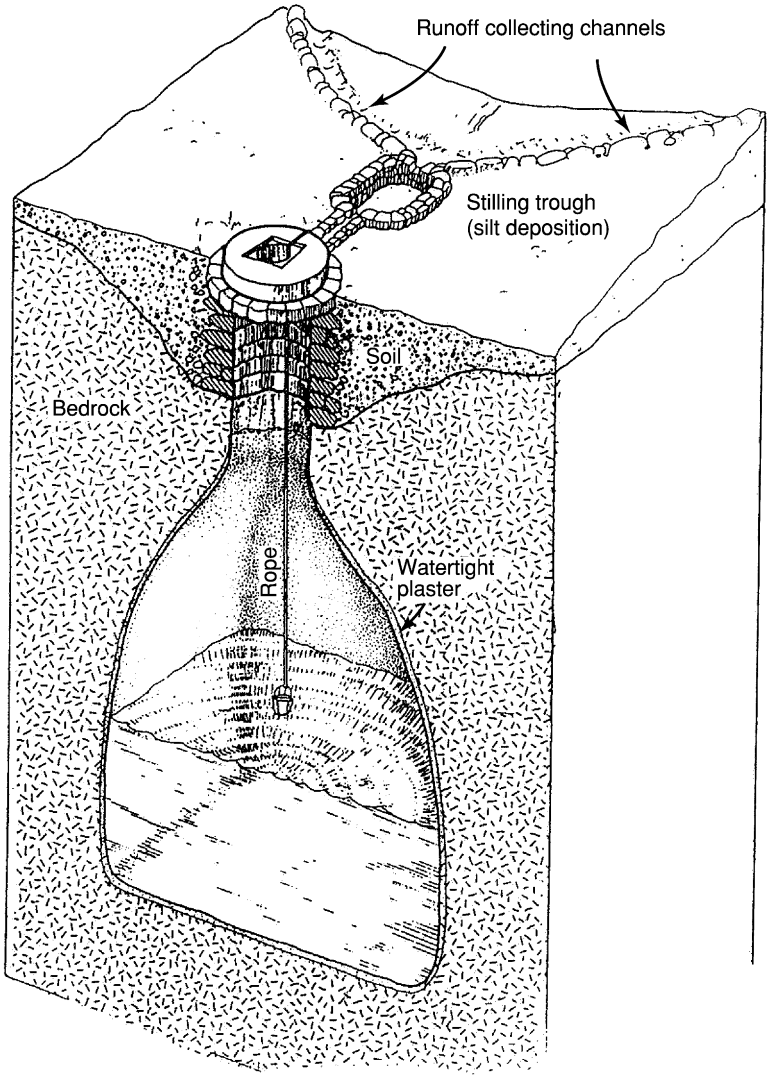


Figure 3 A typical hillside cistern excavated in the bedrock and filled by runoff from a sloping catchment. Reproduced from Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment.* New York: Praeger Publishers. © D Hillel.

Noticing that the soil had a natural tendency to crust, which was obstructed, however, by the desert's natural cover of loose gravel (commonly known as 'desert pavement'), the ancient inhabitants of the Negev region in southern Israel raked the stones off the surface in order to expose the finer soil material and to induce the formation of a surface seal. Even so, the ancient runoff farmers needed a water-contributing area approximately 20 times larger than the area to which water was directed for crop production. Similar techniques were used to collect runoff in cisterns for subsequent use as drinking water for humans and domestic animals (goats, sheep, and camels).

The importance of runoff inducement is greater than the mere increase in runoff yield that it may produce. The practice can also lower the 'runoff threshold' of a rainstorm, i.e., the minimal rainstorm (in terms of intensity, duration, and total amount of rain) needed to initiate runoff. This decrease in the threshold correspondingly increases the probability of obtaining runoff a sufficient number of times during the season to provide for the needs of domestic human use, as well as for agricultural or industrial purposes. This is especially important in view of the fact that most of the storms in an arid region result in light rains only, and that only a few storms (typically no more than two or three per season) are of sufficient intensity and quantity to yield runoff under natural conditions.

Still another climatic feature of arid regions that adds to the importance of runoff inducement is the interannual variability of rainfall. Most years provide less than the average (or mean) rainfall, and only a few anomalously rainy years skew the mean. (In statistical terms, mode and median rainfall tend to be less than the long-term mean rainfall.) In such regions, droughts are relatively frequent and may be very severe. In some years, there may be practically no

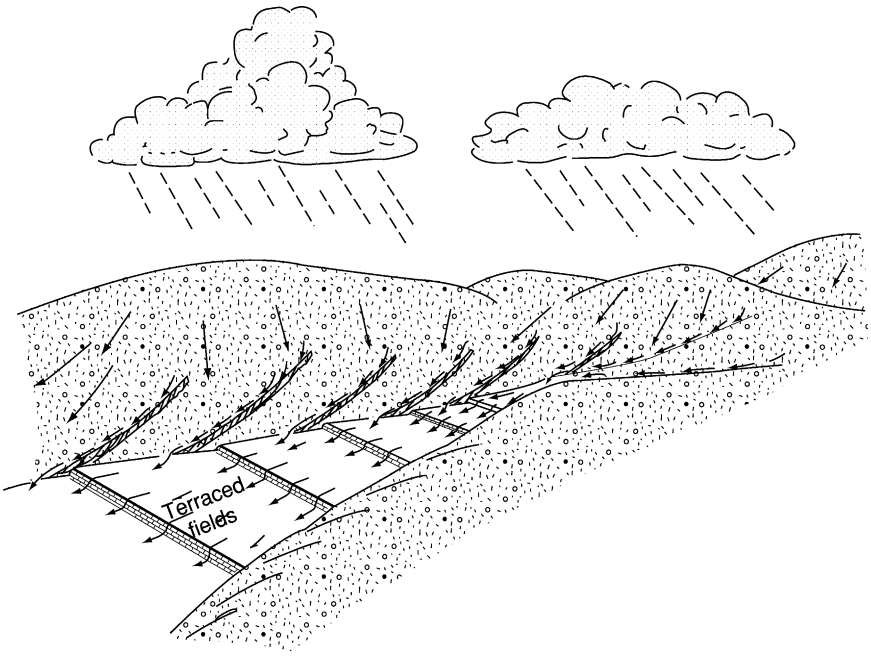


Figure 4 Ancient runoff farming in the Negev: water trickling off barren slopes was directed to terraced fields in the *wadi* bed. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

intense rainstorms and therefore no substantial natural runoff. Human enterprises dependent on the collection and utilization of runoff must therefore ensure, first of all, that there be a minimally sufficient water supply even in years of drought. Methods of runoff inducement can greatly contribute to this end.

Modern Methods

Modern technology holds the promise of more efficient runoff inducement than was possible in ancient times. The simplest, though generally most expensive, way to induce runoff is to cover the surface with an impervious apron of such materials as plastic, rubber, or aluminum sheeting, or by asphalt or concrete paving. A possibly more economical approach is to cause the soil itself to shed, rather than absorb, the rain. Runoff from natural surfaces can be increased several-fold by means of mechanical treatments (stone clearing, smoothing, and compaction), as well as by a variety of chemical treatments to encrust and stabilize the surface so as to prevent erosion of the runoff-contributing areas. Accordingly, the soil surface can be made water-repellent and relatively impermeable by treating it with sprayable clay-dispersants, sealers, and hydrophobic agents. The following series of treatments can be applied:

1. Eradication of vegetation and removal of surface stones, to reduce interception of rain and obstruction of overland flow, and to permit the formation of a continuous surface crust;
2. Smoothing of land surface, to obliterate surface depressions and prevent the retention of water in puddles;
3. Compaction of the soil top-layer to reduce its permeability. This can be done by means of a roller at optimal soil moisture content;
4. Dispersion of soil colloids to cause crusting, by means of sprayable solutions of sodic salts. This treatment pertains to soils that contain sufficient clay to be dispersible, but not so much as to exhibit marked shrinkage and cracking;
5. Impregnation of the surface with a sealing and binding substance such as an emulsion of asphalt that can form a water-repellent and stable coating.

With such methods, it is possible not only to increase the total yield per unit area of the watershed, but also to decrease the threshold of rain needed to form runoff and thus to increase the frequency of runoff supply and contribute to the efficiency and economic feasibility of agricultural and engineering systems designed to utilize runoff. In a desert with a seasonal rainfall of 250 mm, for instance, yields as high as 200 000 m^3 of water may be obtainable per square kilometer of treated area per season.

Several systems have been tried with respect to the size and arrangement of the contributing area in relation to water-receiving areas or reservoirs. A small watershed may be treated in its entirety so as to provide the maximal amount of water at the outflow

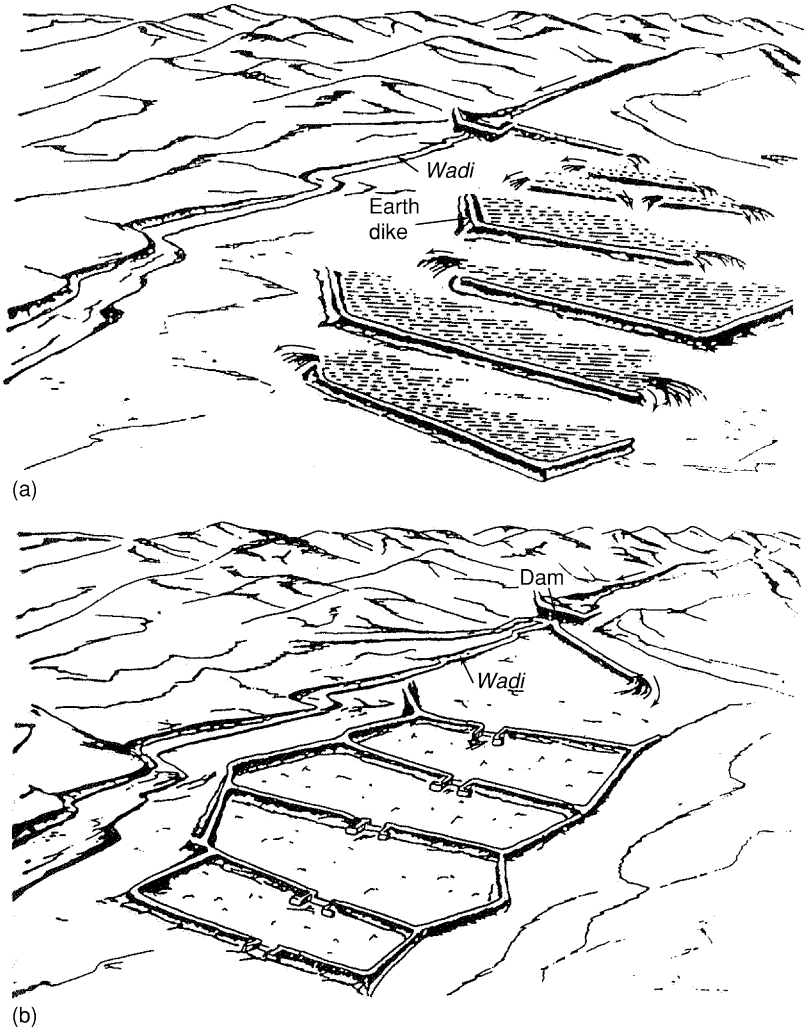


Figure 5 Two general methods of water spreading by diversion from an intermittent stream (a *wadi*) on to adjacent land: (a) diversion of an uncontrolled flow over unleveled land by means of a zigzag series of earthen dikes; and (b) diversion of a controlled flow from a detention dam to a series of level basins with concrete- or stone-lined spillways. Reproduced from Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment*. New York: Praeger Publishers. © D Hillel.

of the basin for conveyance to a pond or a series of irrigated fields. Another system of runoff farming may consist of strips treated to shed runoff, alternating with basins or areas treated to receive and absorb the runoff (Figure 6). A third possibility is to form microwatersheds, wherein each contributing area serves a single tree or row of plants.

In the ancient runoff-farming systems found in the Negev of Israel, the measured ratio of runoff-contributing areas to the runoff-receiving areas is generally between 30:1 and 20:1. In a region with a mean annual rainfall of 100 mm and a mean runoff yield of 10%, a 'runoff-to-runon' area ratio of 10:1 would double the effective water supply to an agricultural plot, a 20:1 area ratio would triple it, and a 30:1 ratio would quadruple it. The latter ratio would provide 400 mm (100 mm of rain plus 300 mm of runon) to a plot of arable land that would otherwise receive only 100 mm. (These data are hypothetical.) If using methods of runoff inducement can increase the mean runoff yield to approximately 50%, then the runoff-to-runon area ratio may be reduced to about

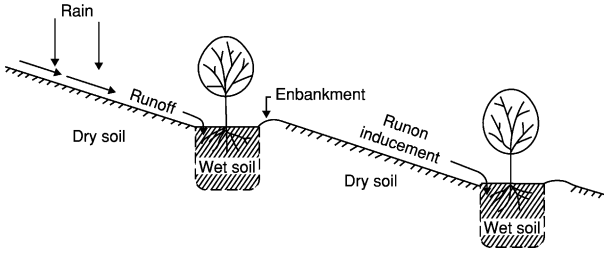


Figure 6 Schematic of a modern runoff-farming system. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

6:1 or even less. Moreover, the number of months or seasons without a minimally sufficient water supply (owing to the paucity of rain during a drought) would diminish and the entire system could thereby operate with a greater probability of success.

See also: **Infiltration**; **Overland Flow**

Further Reading

Bruins HM, Evenari M, and Nessler U (1986) Rainwater-harvesting agriculture for food production in arid zones. *Applied Geography* 6: 13–33.

Critchey W and Siegert K (1991) *Water Harvesting Manual*. Rome: Food and Agriculture Organization.

Dutt GR, Hutchinson CF, and Anaya Garduno M (eds) (1981) *Rainfall Collection for Agriculture in Arid and Semiarid Regions*. Farnham Royal, UK: Commonwealth Agricultural Bureaux.

Evenari M, Shanan L, and Tadmor N (1971) *The Negev: The Challenge of a Desert*. Cambridge, MA: Harvard University Press.

Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment*. New York: Praeger Publishers.

Hillel D (1994) *Rivers of Eden: The Struggle for Water and the Quest for Peace in the Middle East*. New York: Oxford University Press.

Water Management *See* **Crop Water Requirements**

WATER POTENTIAL

D Or, University of Connecticut, Storrs, CN, USA
M Tuller, University of Idaho, Moscow, ID, USA
J M Wraith, Montana State University, Bozeman, MT, USA

Introduction

Water status in soils is characterized by both the amount of water present and its energy state. Soil water is subjected to forces of variable origin and intensity, thereby acquiring different quantities and forms of energy. The two primary forms of energy of interest here are kinetic and potential. Kinetic energy is acquired by virtue of motion and is proportional to velocity squared. However, because the movement of water in soils is relatively slow (usually less than 0.1 $m\,h^{-1}$) its kinetic energy is negligible. Potential energy, which is defined by the position of soil water within a soil body and by internal conditions, is largely responsible for determining soil water status under isothermal conditions.

Like all other matter, soil water tends to move from where the potential energy is higher to where it is lower, in pursuit of equilibrium with its surroundings. The magnitude of the driving force behind such spontaneous motion is a difference in potential energy across a distance between two points of interest. At a macroscopic scale, we can define potential energy relative to a reference state. The standard state for soil water is defined as pure and free water (no solutes and no external forces other than gravity) at a reference pressure, temperature, and elevation, and is arbitrarily given the value of zero.

The 'Total' Soil Water Potential and its Components

Soil water is subject to several force fields, the combined effects of which result in a deviation in potential energy relative to the reference state, called the 'total soil water potential' (ψ_T) defined as: "The amount of work that an infinitesimal unit quantity of water at equilibrium is capable of doing when it moves (isothermally and reversibly) to a pool of water at similar standard (reference) state, i.e., similar pressure, elevation, temperature and chemical composition." It should be emphasized that there are alternative definitions of soil water potential using concepts of chemical potential or specific free energy of the chemical species water (which is different from the soil solution termed 'soil water'). Recognizing that these fundamental concepts are subject to ongoing debate, presented here are simple and widely accepted definitions which are applicable at macroscopic scales and which yield an appropriate framework for practical applications.

The primary forces acting on soil water held within a rigid soil matrix under isothermal conditions can be conveniently grouped as: (1) matric forces

resulting from interactions of the solid phase with the liquid and gaseous phases; (2) osmotic forces owing to differences in chemical composition of soil solution; and (3) body forces induced by gravitational and other (e.g., centrifugal) inertial force fields.

The thermodynamic approach, whereby potential energy rather than forces are used, is particularly useful for equilibrium and flow considerations. Equilibrium would require the vector sum of these different forces acting on a body of water in different directions to be zero; this is an extremely difficult criterion to deal with in soils. On the other hand, potential energy, mathematically defined as the negative integral of the force over the path taken by an infinitesimal amount of water when it moves from a reference location to the point under consideration, is a scalar (not a vector) quantity. Subsequently, we can express the total potential as the algebraic sum of the component potentials corresponding to the different fields acting on soil water as:

$$\psi_T = \psi_m + \psi_s + \psi_p + \psi_z \qquad [1]$$

where the component potentials ψ_i are discussed below.

ψ_m is the matric potential resulting from the combined effects of capillarity and adsorptive forces within the soil matrix. The primary mechanisms for these effects include: (1) capillarity caused by liquid–gas interfaces forming and interacting within the irregular soil-pore geometry; (2) adhesion of water molecules to solid surfaces due to short-range London–van der Waals forces and extension of these effects by cohesion through hydrogen bonds formed in the liquid; and (3) ion hydration and water participating in diffuse double layers (particularly near clay surfaces). There is some disagreement regarding the practical definition of this component of the total potential. Some consider all contributions other than gravity and solute interactions (at a reference atmospheric pressure). Others use a device known as a tensiometer to measure and provide a practical definition of the matric potential in a soil volume of interest in contact with a tensiometer's porous cup. The value of ψ_m ranges from zero when the soil is saturated to increasingly negative values as the soil becomes drier (note that $\psi_m = 0$ mm is greater than $\psi_m = -1000$ mm; in analogy, a temperature of 0°C is greater than −10°C). (*See* **Capillarity.**)

Applied theories for flow and transport in unsaturated porous media, particularly at low water content, commonly lump capillary and adsorptive forces without distinguishing individual contributions to the matric potential. Advanced frameworks simultaneously consider individual contributions of capillary and adsorptive forces for calculation of liquid–vapor interfacial configurations in angular pore spaces. The liquid–vapor interface is considered as a surface of constant, partial specific Gibbs free-energy (or matric potential) made up of an adsorptive component (A) and a capillary component (C):

$$\psi_m = A(h) + C(\kappa) \qquad [2]$$

with κ as the mean curvature of the liquid–vapor interface, and h as the distance from the solid to the liquid–vapor interface, taken normal to the solid surface (thickness of the adsorbed film). The capillary component C is given by the classic Young–Laplace equation:

$$C(\kappa) = \frac{-2 \cdot \sigma \cdot \kappa}{\rho} \qquad [3]$$

where κ is positive for an interface concave outward from the liquid, σ is the surface tension at the interface, and ρ is the density of the liquid. (*See* **Capillarity.**)

The adsorptive component in eqn [2] is attributed to two types of surface forces. The first kind includes long-range (>500 Å) electrostatic forces (e.g., diffuse double layer, DDL), and short-range (<100 Å) van der Waals and hydration forces, responsible for molecular interactions and structural changes in water molecules near the solid surface. The second kind is comprised of long-range forces due to the overlapping of two interfacial regions (e.g., mutual attraction between two clay platelets across a slit-shaped pore space). The combined effect of interfacial interactions results in a difference in chemical potentials between the liquid in the adsorbed film and the bulk liquid phase. This difference in chemical potentials may be expressed as an equivalent interfacial force per unit area of the interface, termed the 'disjoining pressure' (Π). The disjoining pressure is a function of liquid film thickness (h), and it can also be viewed as the difference between a normal component of film pressure, P_N (in equilibrium with the gaseous phase $P_N = P_G$), and the pressure in the bulk liquid phase, P_L:

$$\Pi(h) = P_N(h) - P_L = P_G - P_L \qquad [4]$$

The disjoining pressure is related to more conventional thermodynamic quantities such as Gibbs free energy. Gibbs free energy (G) per unit area of the interface may be defined on the basis of $\Pi(h)$ isotherms for constant pressure P_L, temperature T, and chemical (μ) and electric potentials of the liquid–gaseous and the liquid–solid interfaces as:

$$G(h) = -\int_{\infty}^{h} \Pi(h)\mathrm{d}h \qquad [5]$$

The value of $G(h)$ is equal to the work of thinning the film in a reversible isobaric–isothermal process from ∞ to a finite thickness h, with $\Pi(h) = -(\partial G/\partial h)_{T,\, P_L,\, \mu,\, \psi_g,\, \psi_s}$. The use of $\Pi(h)$ as the basic thermodynamic property is not a mere change of notation, but $\Pi(h)$ has advantages in cases where Gibbs thermodynamic theory is difficult to define, such as when interfacial zones overlap to the extent that the film does not retain the intensive properties of the bulk phase. The use of the disjoining pressure is advantageous from an experimental point of view because of the relative ease in accounting for different contributions (e.g., electrostatic effects).

The disjoining pressure is a sum of several components, similar to the concept of total soil water potential discussed above. The primary components of $\Pi(h)$ in porous media are: molecular, $\Pi_m(h)$; electrostatic, $\Pi_e(h)$; structural, $\Pi_s(h)$; and adsorptive $\Pi_a(h)$:

$$\Pi(h) = \Pi_m(h) + \Pi_e(h) + \Pi_s(h) + \Pi_a(h) \qquad [6]$$

$\Pi_m(h)$

The molecular component originates from van der Waals interaction between macro-objects (e.g., parallel clay plates). Various expressions, with $\Pi_m(h)$ often proportional to h^{-3}, were derived by Paunov *et al.* and Iwamatsu and Horii.

$\Pi_e(h)$

The electrostatic component of the disjoining pressure is calculated from the solution of the Poisson–Boltzmann equation for the DDL with appropriate boundary conditions. Approximate solutions are adequate for many applications, often with $\Pi_e(h) \propto h^{-2}$.

$\Pi_s(h)$

Some controversy exists regarding the origin of the structural component; some attribute it to changes in the structure (density) of water adjacent to solid surfaces and deformation of hydrated shells, while others attribute this force to the presence of a layer with a lower dielectric constant near the surface. Regardless of its exact origin, this component is responsible for the so-called hydration repulsion which stabilizes dispersion and prevents coagulation of some colloidal particles, even at high electrolyte concentrations: $\Pi_s(h) \propto h^{-1}$.

$\Pi_a(h)$

The adsorptive component of the disjoining pressure results from nonuniform concentrations in the water film due to unequal interaction energies of solute and solvent with interfaces in nonionic solutions. This is different to the nonuniform distribution of charged ions. This component of the disjoining pressure is likely to become very important for interactions between nonpolar molecules (e.g., NAPLs) which give rise to repulsive forces in the liquid film.

The form of the disjoining pressure isotherm $\Pi(h)$ is determined by the nature of surface forces. While the molecular component $\Pi_m(h)$ is always present, the influence of other components depends on surface properties, liquid polarity and its composition, and adsorption of dissolved components. The range of the electrostatic forces in dilute solutions of a 1:1 electrolyte (10^{-6}–10^{-7} mol l^{-1}) is in the range of 0.3–1.0 μm. Consequently, thick films of water and aqueous electrolyte solutions ($h > 500$ Å) are stable mainly through the $\Pi_e(h)$ component of disjoining pressure. The magnitude and contribution of $\Pi_e(h)$ primarily depend on the charges of the film and substrate surfaces. Dispersion forces become appreciable in the range $h < 500$ Å, and their influence is enhanced by large differences between the permittivity of the liquid and the solid. The forces of structural repulsion may come to play when the film thickness is less than 100 Å.

ψ_s is the solute or osmotic potential determined by the presence of solutes in soil water, which lower its potential energy and its vapor pressure. The effects of ψ_s are important when: (1) there are appreciable amounts of solutes in the soil; and (2) in the presence of a selectively permeable membrane or a diffusion barrier which transmits water more readily than salts. The effects of ψ_s are otherwise generally negligible when only liquid water flow is considered and no diffusion barrier exists. The two most important diffusion barriers in the soil are: (1) soil–plant root interfaces (cell membranes are selectively permeable); and (2) air–water interfaces; thus when water evaporates, salts are left behind. In dilute solutions the solute potential, also called the osmotic pressure, is proportional to the concentration and temperature according to:

$$\psi_s = -RT\,C_s \qquad [7]$$

where ψ_s is in kilopascals, R is the universal gas constant (8.314×10^{-3} kPa m^3/(mol K)), T is absolute temperature (Kelvin), and C_s is solute concentration (moles per cubic meter). A useful approximation which may be used to estimate ψ_s in kilopascals from the electrical conductivity of the soil solution at saturation (EC_s) in deciSiemens per meter (dS m^{-1}) is:

$$\psi_s \approx -36\,EC_s \qquad [8]$$

ψ_p is the pressure potential, defined as the hydrostatic pressure exerted by unsupported water that

saturates the soil and overlays a point of interest. Using units of energy per unit weight provides a simple and practical definition of ψ_p as the vertical distance from the point of interest to the free water surface (unconfined water table elevation). The convention used here is that ψ_p is always positive below a water table, or zero if the point of interest is at or above the water table. In this sense, nonzero magnitudes of ψ_p and ψ_m are mutually exclusive: either ψ_p is positive and ψ_m is zero (saturated conditions), or ψ_m is negative and ψ_p is zero (unsaturated conditions), or $\psi_p = \psi_m = 0$ at the free water table elevation. Note that some prefer to combine the pressure and matric components into a single term, which assumes positive values under saturated conditions and negative values under unsaturated conditions. Based on operational and explanatory considerations, we prefer to adopt the more commonly used separate components protocol.

ψ_z is the gravitational potential, which is determined solely by the elevation of a point relative to some arbitrary reference point, and is equal to the work needed to raise a body against the Earth's gravitational pull from a reference level to its present position. When expressed as energy per unit weight, the gravitational potential is simply the vertical distance from a reference level to the point of interest. The numerical value of ψ_z itself is thus not important (it is defined with respect to an arbitrary reference level) – what is important is the difference (or gradient) in ψ_z between any two points of interest. This value is invariant of the reference level location.

Soil water is at equilibrium when the net force on an infinitestimal body of water equals zero everywhere, or when the total potential is constant in the system. Though the last statement is a logical consequence of the definitions above, it is not strictly true. Constant total potential is a necessary but not a sufficient condition, and, for thermodynamic equilibrium to prevail, three conditions must be met simultaneously: (1) thermal equilibrium or uniform temperature; (2) mechanical equilibrium, meaning no net convection-producing force; and (3) chemical equilibrium, meaning no net diffusional transport or chemical reaction. In most practical applications, however, the macroscopic definition of the total potential and equilibrium conditions based on it are completely adequate.

The difference in chemical and mechanical potentials between soil water and pure water at the same temperature is known as the soil water potential (ψ_w):

$$\psi_w = \psi_m + \psi_s + \psi_p \qquad [9]$$

Note that the gravitational component (ψ_z) is absent in this definition. Soil water potential is thus the result of inherent properties of soil water itself, and of its physical and chemical interactions with its surroundings; whereas the total potential includes the effects of gravity (an 'external' and ubiquitous force field).

Total soil water potential and its components may be expressed in several ways depending on the definition of a 'unit quantity of water.' Potential may be expressed as (1) energy per unit of mass; (2) energy per unit of volume; or (3) energy per unit of weight. A summary of the resulting dimensions, common symbols, and units are presented in Table 1.

Only μ has actual units of potential; ψ has units of pressure, and h of head of water. However, the above terminology (i.e., potential energy expressions rather than units of potential per se) is widely used in a generic sense in the soil and plant sciences. The various expressions of soil water energy status are equivalent, with:

$$\mu = \frac{\psi}{\rho_w} = gh \qquad [10]$$

where ρ_w is density of water ($1000\,kg\,m^{-3}$ at 20°C) and g is gravitational acceleration ($9.81\,m\,s^{-2}$).

Measurement of Potential Components

Water Potential

A psychrometer (Figure 1) is commonly used for measurement of total water potential (ψ_w) in soils. The potential of the soil solution is in thermodynamic equilibrium with its ambient water vapor. Taking the vapor pressure above pure water at reference state ($\psi_w = 0$) as e_o, the vapor pressure (e) over a salt

Table 1 Units, dimensions, and common symbols for potential energy of soil water

Units	*Symbol*	*Name*	*Dimensions*	*SI units*	*CGS units*
Energy/mass	μ	Chemical potential	$L^2 t^{-2}$	$J\,kg^{-1}$	$erg\,g^{-1}$
Energy/volume	ψ	Soil water potential, suction, or tension	$ML^{-1}t^{-2}$	$N\,m^{-2}$(Pa)	$dyn\,cm^{-2}$
Energy/weight	h	Pressure head	L	m	cm

L, length; *M*, mass; *t*, time.

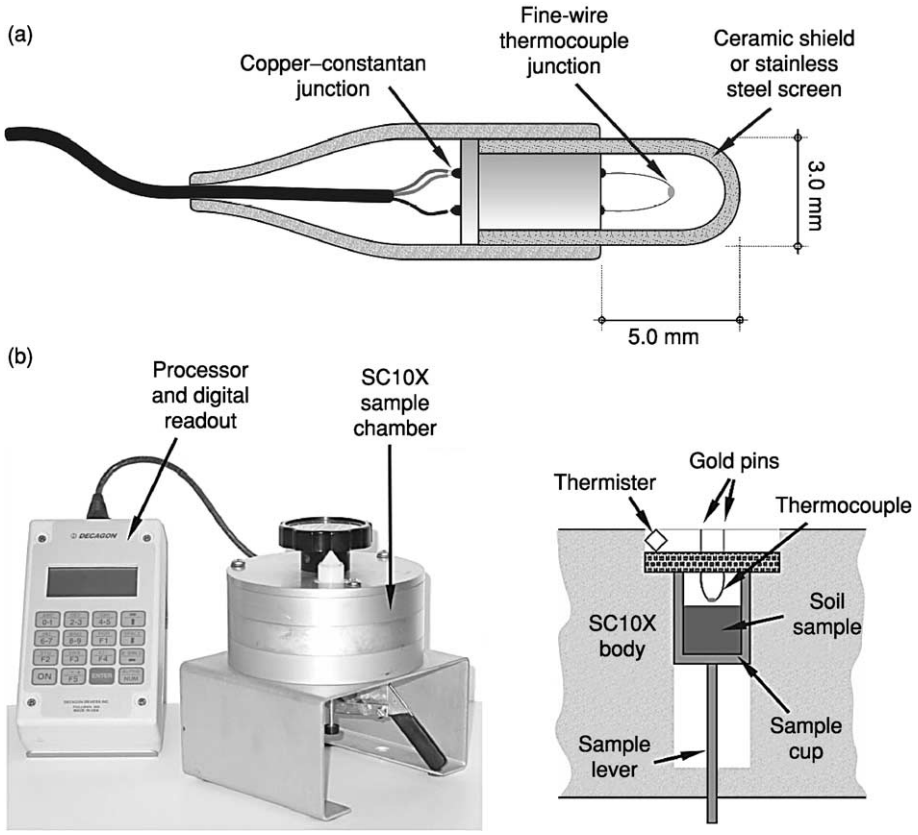


Figure 1 (a) A field psychrometer with porous ceramic shield (Wescor Inc., Logan, Utah, USA); and (b) SC10X sample chamber for psychrometric laboratory measurements of soil water potential (Decagon Devices Inc., Pullman, Washington, USA).

solution or soil water held in soil pores by matric forces is depressed relative to the reference state, i.e., $e < e_o$. A convenient measure obtained by the psychrometer is the relative vapor pressure of the ambient soil atmosphere, which is related to the water potential (ψ_w) of soil water through the well-known Kelvin equation:

$$\mathrm{RH} = \frac{e}{e_o} = \exp\left[\frac{M_w \psi_w}{\rho_w R T}\right] \qquad [11]$$

where e is water vapor pressure (kilopascals), e_o is saturated vapor pressure at the same temperature, M_w is the molecular weight of water ($0.018\,\mathrm{kg\,mol^{-1}}$), R is the ideal gas constant ($8.31\,\mathrm{J\,K^{-1}\,mol^{-1}}$ or $0.008314\,\mathrm{kPa\,m^3\,mol^{-1}\,K^{-1}}$), T is absolute temperature (Kelvin), and ρ_w is the density of water ($1000\,\mathrm{kg\,m^{-3}}$ at 20°C). Rearranging and taking a log-transformation of eqn [11] yields an expression for water potential ψ_w:

$$\psi_w = \frac{R T \rho_w}{M_w} \ln\left(\frac{e}{e_o}\right) \qquad [12]$$

The water potential in drier soils is lower such that fewer water molecules 'escape' into the ambient atmosphere, resulting in lower relative humidity (lower relative vapor pressure). Concentrated soil solutions having lower osmotic potentials have similar effect on reducing vapor pressure, as more water molecules are associated with hydrated salt molecules and are less free to 'escape' the liquid state. The inability to distinguish between matric and osmotic effects limits psychrometric measurements to soil water potential only. In some cases where the osmotic potential is negligible, psychrometric measurements are used to infer the matric potential.

Pressure Potential

Piezometers are commonly applied to measure ψ_p. A piezometer (Figure 2) is a tube that is placed in the soil to depths below the water table and that extends to the soil surface and is open to the atmosphere. The bottom of the piezometer is perforated to allow soil water under positive hydrostatic pressure to enter the tube. Water enters the tube and rises to a height equal to that of the unconfined water table. The elevation of the free water table is measured relative to the soil surface using a steel tape with bell sounder, or other electro-optic devices that indicate water table depth. The value of pressure potential

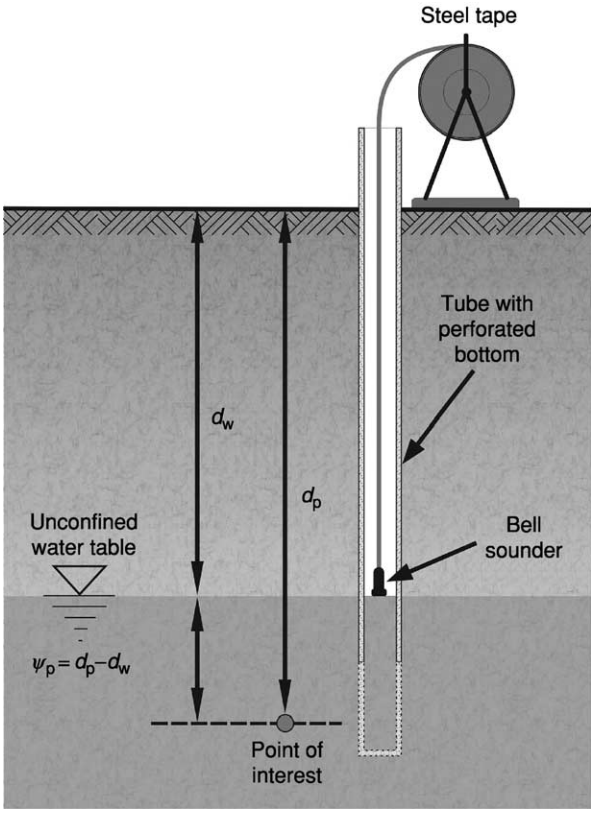


Figure 2 The concept of piezometer measurements.

expressed as energy per weight is simply the vertical distance from a point of interest to the surface of the free water table. Pressure potentials above the water table surface are always zero (nonzero pressure and matric potentials are mutually exclusive).

Matric Potential

Tensiometers or heat-dissipation sensors are commonly applied to measure soil matric potential. A tensiometer consists of a porous cup, usually made of ceramic and having very fine pores, connected to a vacuum gauge through a water-filled tube (Figure 3). The porous cup is placed in intimate contact with the bulk soil at the depth of measurement. When the matric potential of the soil is lower (more negative) than inside the tensiometer, water moves from the tensiometer along a potential energy gradient to the soil through the saturated porous cup, thereby creating suction sensed by the gauge. Water flow into the soil continues until equilibrium is reached and the suction inside the tensiometer equals the soil matric potential. When the soil is wetted, flow may occur in the reverse direction, i.e., soil water enters the tensiometer until a new equilibrium is attained. The tensiometer equation is:

$$\psi_m = \psi_{gauge} + (z_{gauge} - z_{cup}) \quad [13]$$

with ψ_{gauge} the reading at the vacuum gauge location and z indicating depth. The vertical distance from the gauge plane to the cup, expressed as a negative quantity, must be added to the matric potential measured by the gauge (ψ_{gauge}) in order to obtain the matric potential at the depth of the cup. This accounts for the positive head exerted by the overlying tensiometer water column at the depth of the ceramic cup. Note that using the difference in vertical elevation is appropriate only when potentials are expressed per unit of weight. Electronic sensors called pressure transducers often replace mechanical vacuum gauges. The transducers convert mechanical pressure into an electric signal which can be more easily and more precisely measured. In practice, pressure transducers can provide more accurate readings than other gauges, and in combination with data-logging equipment are able to supply continuous measurements of matric potential.

The tensiometer range is limited to suctions (absolute values of the matric potential) of less than 100 kPa, i.e., 1 bar, 10 m head of water, or ~1 atmosphere. Therefore other means are needed for matric potential measurement under drier conditions.

Heat-dissipation sensors may be applied for a matric potential range from −10 to −1000 kPa. The rate of heat dissipation in a porous medium is dependent on the medium's specific heat capacity, thermal conductivity, and density. The heat capacity and thermal conductivity of a porous matrix are affected by its water content. Heat dissipation sensors contain heating elements in line or point source configurations embedded in a rigid porous matrix with fixed pore space. The measurement is based on application of a heat pulse by applying a constant current through the heating element for specified time period, and analysis of the temperature response measured by a thermocouple placed at a fixed distance from the heating source. With the heat dissipation sensor buried in the soil, changes in soil water matric potential result in a gradient between the soil and the porous ceramic matrix, inducing water flow between the two materials until a new equilibrium is established. The water flow changes the water content of the ceramic matrix which, in turn, changes the thermal conductivity and heat capacity of the sensor, and hence the measured temperature response to the applied heat pulse.

As already mentioned above, for cases where the osmotic potential is negligible, psychrometric measurements can be used to infer the matric potential. A typical range for psychrometers is −800 to −10 000 kPa.

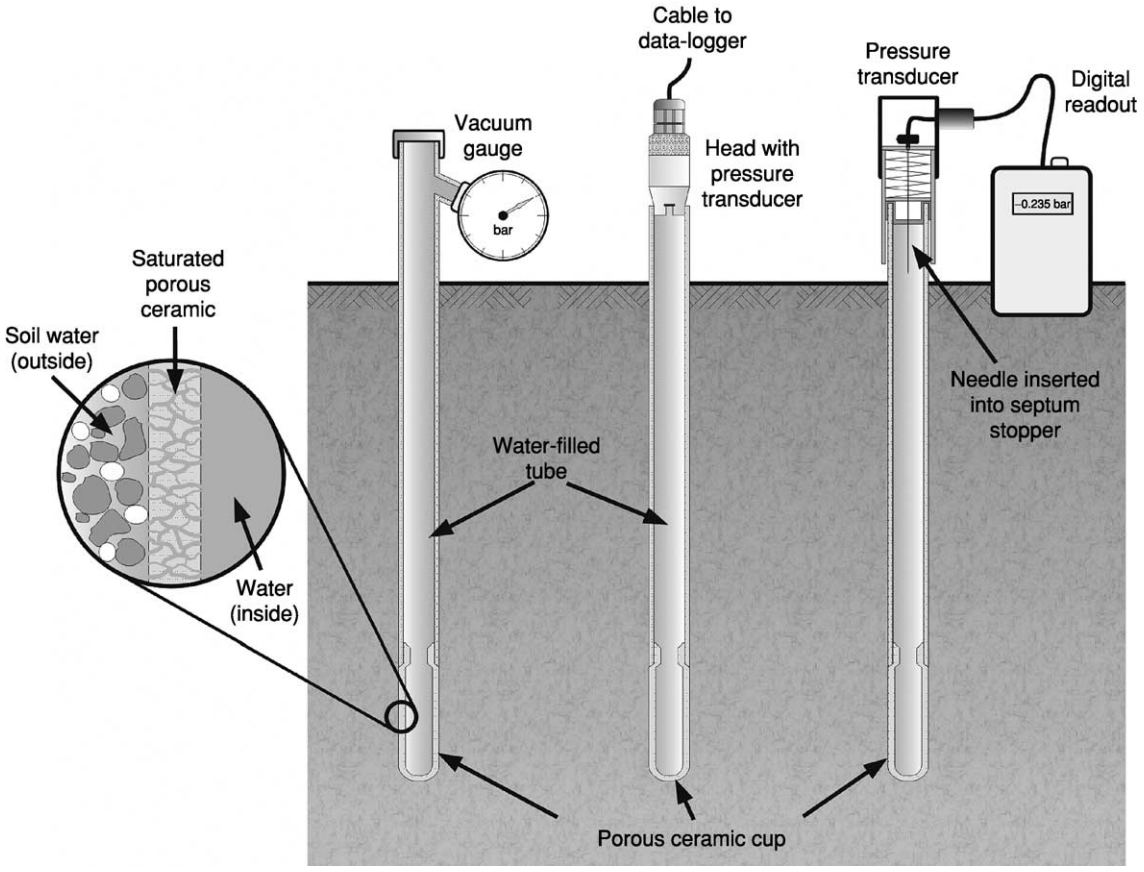


Figure 3 Tensiometers for matric potential measurement using vacuum gauges and electronic pressure transducers.

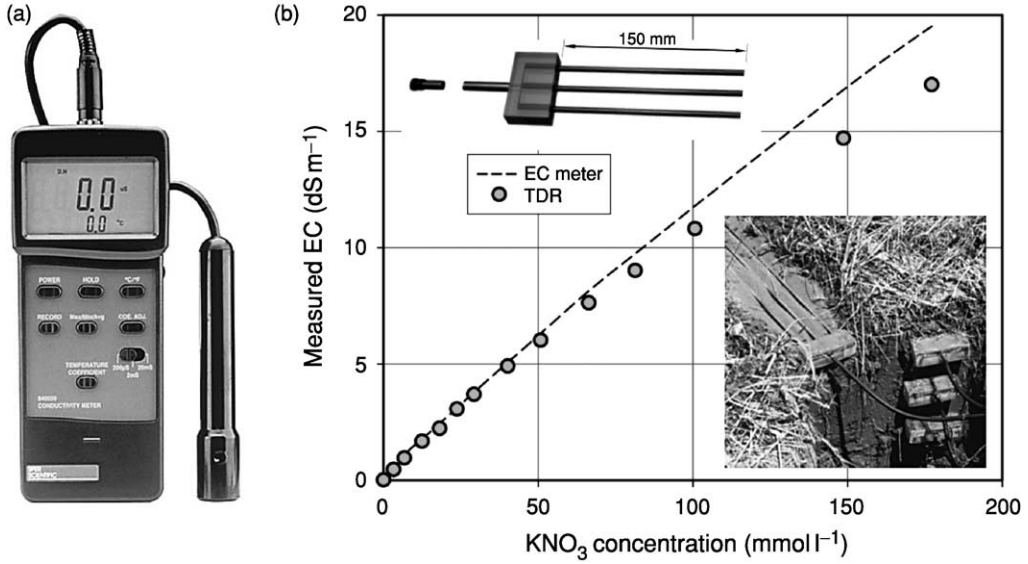


Figure 4 (a) Handheld electrical conductivity (EC) meter; (b) time-domain reflectometry (TDR) probes and solution EC versus concentration measured with TDR and EC meter. (Reproduced from Mmolawa KB and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.)

Osmotic Potential

Soil water solutions contain varied quantities and compositions of dissolved salts. The relationships between the salt concentration and ψ_s, and the possibility for estimating ψ_s from the electrical conductivity (EC) of the soil solution, were discussed above. Conventional measurement of soil solution EC involves solution extraction from saturated soil samples and measuring the EC using an electrical conductivity meter (**Figure 4a**).

Electrical conductivity meters rely on Ohm's law:

$$E = I \cdot R \qquad [14]$$

with E the electromotive force (volts), I the current flow (amperes), and R the resistance (ohms). For constant voltage, the current flowing through the solution is inversely proportional to the electrical resistance, or directly proportional to the electrical conductance. The solution EC is thus determined from known voltage and electrode geometry and measurement of the electric current. More recently, a variety of *in situ* methods such as time-domain reflectometry (TDR) have been used to deduce soil bulk EC from electromagnetic signal attenuation, hence enabling simultaneous measurements of water content and soil EC in the same undisturbed soil volume. Concurrent knowledge of θ and EC can be used to infer the soil solution EC, and hence to estimate ψ_s.

See also: **Capillarity**

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.

Bolt GH (1976) Soil physics terminology. *International Society Soil Science Bulletin* 49: 16–22.

Bristow KL, Campbell GS, and Calissendroff K (1993) Test of a heat-pulse probe for measuring changes in soil water content. *Soil Science Society of America Journal* 57: 930–934.

Corey AT and Klute A (1985) Application of the potential concept to soil water equilibrium and transport. *Soil Science Society of America Journal* 49: 3–11.

Dalton FN, Herkelrath WN, Rawlins DS, and Rhoades JD (1984) Time-domain reflectometry: simultaneous measurement of soil water content and electrical conductivity with a single probe. *Science* 224: 989–990.

Day PR, Bolt GH, and Anderson DM (1967) Nature of soil water. In: Hagan RM, Haise HR, and Edminster TW (eds) *Irrigation of Agricultural Lands*, pp. 193–208. Madison, WI: American Society of Agronomy.

Derjaguin BV, Churaev NV, and Muller VM (1987) *Surface Forces*. New York: Plenum.

Edlefsen NE and Anderson ABC (1943) Thermodynamics of soil moisture. *Hilgardia* 15: 31–298.

Hanks RJ (1992) *Applied Soil Physics*, 2nd edn. New York: Springer-Verlag.

Hendrickx JMH, Wraith JM, Corwin DL, and Kachanoski RG (2002) Solute content and concentration. Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, pp. 1253–1322. Madison, WI: American Society of Agronomy.

Iwamatsu MI and Horii K (1996) Capillary condensation and adhesion of two wetter surfaces. *Journal of Colloid Interface Science* 182: 400–406.

Iwata S, Tabuchi T, and Warkentin BP (1988) *Soil Water Interactions*. New York: Marcel Dekker.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Cremlingen-Destedt, Germany: Catena-Verlag.

Mitlin VS and Sharma MM (1993) A local gradient theory for structural forces in thin fluid films. *Journal of Colloid Interface Science* 157: 447–464.

Mmolawa KB and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.

Nitao JJ and Bear J (1996) Potentials and their role in transport in porous media. *Water Resources Research* 32(2): 225–250.

Novy RA, Toledo PG, Davis HT, and Scriven LE (1989) Capillary dispersion in porous media at low wetting phase saturations. *Chemical Engineering Science* 44(9): 1785–1797.

Paunov VN, Dimova RI, Kralchevsky PA, Broze G, and Mehreteab A (1996) The hydration repulsion between charged surfaces as an interplay of volume exclusion and dielectric saturation effects. *Journal of Colloid Interface Science* 182: 239–248.

Phene CJ, Hoffman GJ, and Rawlins SL (1971) Measuring soil matric potential *in situ* by sensing heat dissipation within a porous body. 1. Theory and sensor construction. *Soil Science Society of America Proceedings* 35: 27–33.

Rhoades JD and Oster JD (1986) Solute content. Klute A (ed.) *Methods of Soil Analysis*, part 1, *Physical and Mineralogical Methods*, 2nd edn, pp. 985–1006. Madison, WI: American Society of Agronomy.

Philip JR (1977) Unitary approach to capillary condensation and adsorption. *Journal of Chemical Physics* 66(11): 5069–5075.

Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Water Requirements *See* **Crop Water Requirements**

WATER RETENTION AND CHARACTERISTIC CURVE

M Tuller, University of Idaho, Moscow, ID, USA
D Or, University of Connecticut, Storrs, CT, USA

Introduction

A soil-water characteristic (SWC) curve describes the amount of water retained in a soil (expressed as mass or volume water content, θ_m or θ_v) under equilibrium at a given matric potential. An SWC is an important hydraulic property, related to size and connectedness of pore spaces, hence strongly affected by soil texture and structure, and by other constituents, including organic matter. Modeling water distribution and flow in partially saturated soils requires knowledge of the SWC, therefore plays a critical role in water management and in prediction of solute and contaminant transport in the environment. Typically a SWC is highly nonlinear and is relatively difficult to obtain accurately. Because the matric potential extends over several orders of magnitude for the range of water contents commonly encountered in practical applications, the matric potential is often plotted on a logarithmic scale. Figure 1 depicts representative SWC curves for soils of different textures, demonstrating the effects of porosity (saturated water content) and the varied slopes of the relationships resulting from variable pore-size distributions.

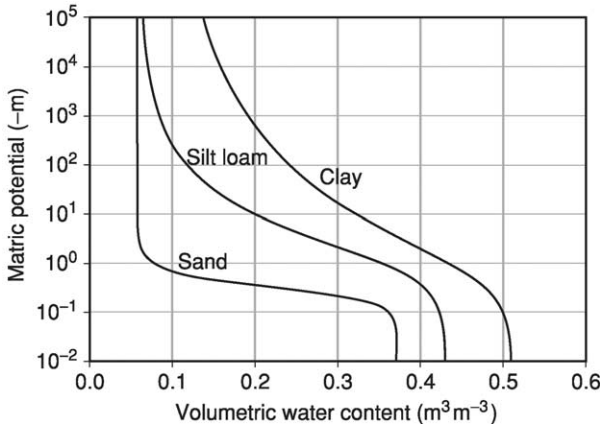


Figure 1 Typical soil-water characteristic curves for soils of different texture.

The Matric Potential

The matric potential is attributed to capillary and adsorptive forces acting between liquid, gaseous, and solid phases. Capillarity results from the surface tension of water and its contact angle with the solid particles. Under partially saturated conditions (i.e., in the presence of the nonwetting air phase), curved liquid–vapor interfaces (menisci) form within the porous soil system. Menisci radii of curvature (R) are a function of capillary pressure (P_c) and are calculated according to the Young–Laplace equation:

$$P_0 - P_c = \Delta P = \frac{2\sigma}{R} \qquad [1]$$

where P_0 is the atmospheric pressure (conventionally referenced as zero), P_c is the pressure of the soil water, and σ is the surface tension of the liquid–vapor interface. If soil pores were behaving as a bundle of capillary tubes, capillarity would be sufficient to describe the relationships between matric potential and soil pore radii. However, in addition to capillarity, the soil also exhibits adsorption, which forms hydration envelopes over the particle surfaces. The presence of water in films is most important in clayey soils with large surface area and is influenced by the electric double layer and the exchangeable cations present. In sandy soils, adsorption is relatively insignificant and the capillary effect predominates. In general, however, matric potential results from the combined effect of capillarity and surface adsorption. The capillary 'wedges' are at a state of internal equilibrium with adsorption 'films,' and the two cannot be considered separately. (*See* **Water Potential; Capillarity.**)

The Bundle of Cylindrical Capillaries Model

Early conceptual models for the SWC and liquid distribution in partially saturated porous media are based on the 'bundle of cylindrical capillaries' (BCC)

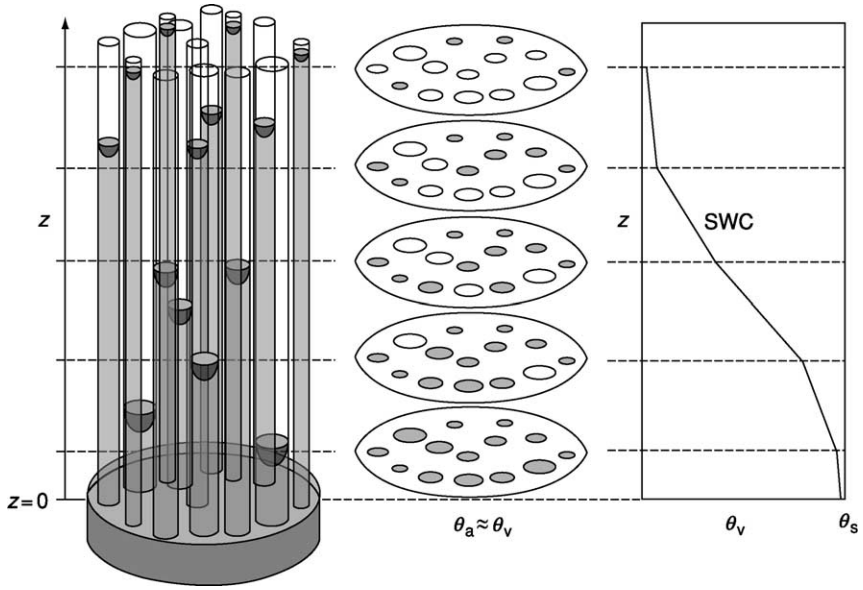


Figure 2 Relationship between the pore space, as represented by the bundle of cylindrical capillaries model, and the soil-water characteristic (SWC). Greater capillary rise occurs in smaller pores, which have smaller radii of meniscus curvature. (Note that z indicates the elevation above free water, θ_a stands for pore volume, θ_v for volumetric water content, and θ_s for volumetric water content at saturation.)

representation of pore-space geometry. The BCC representation postulates that at a given matric potential a portion of interconnected cylindrical pores is completely liquid-filled, whereas larger pores are completely empty (Figure 2). This convenient idealization of soil pore space enables a linkage between the soil pore size distribution and the SWC based on the capillary-rise equation. (*See* **Capillarity**.) However, such representation imposes serious limitations on the practical application of the BCC model to natural porous media as discussed in the following section, including the unrealistic underlying geometry that precludes dual water–air occupancy within the same pores and lack of consideration of adsorbed water films.

Liquid Retention and Pore Shape

Liquid retention in the porous soil matrix is highly dependent on the shape and angularity of individual pores. Inspection of thin sections or soil micrographs (Figure 3) reveals that natural pore spaces do not resemble cylindrical capillaries, as often assumed for an idealized representation. Because natural porous media are formed by aggregation of primary particles and mineral surfaces, the resulting pore space is more realistically described by angular or slit-shaped pore cross-sections than by cylindrical capillaries. In addition to a more realistic representation of natural pore spaces, angular pores offer other advantages over cylindrical tubes in terms of liquid behavior. When angular pores are drained, a fraction of the wetting phase remains in the pore corners (Figure 3c). This aspect of 'dual occupancy' of the invaded portion of the tube, not possible in cylindrical tubes, more realistically represents liquid configurations and mechanisms for maintaining hydraulic continuity in porous media. Liquid-filled corners and crevices play an important role in displacement rates of oil and in other transport processes in partially saturated porous media.

The relationships between liquid retention and pore angularity are known. The water-filled cross-sectional area A_{wt} for all regular and irregular triangles and for regular, higher-order polygons is given by the following expression:

$$A_{wt} = r^2 \cdot F(\gamma) \qquad [2]$$

where r is the radius of curvature of the liquid–vapor interface that is dependent on chemical potential (μ) or capillary pressure (p_c) according to the Young–Laplace equation:

$$\mu = \frac{\sigma}{r\rho} \qquad \text{or} \qquad p_c = \frac{\sigma}{r} \qquad [3]$$

with σ as the surface tension of the liquid, ρ the density of the liquid, and $F(\gamma)$ a shape or angularity factor dependent on angularity of the pore cross-section only. Note that in this discussion we consider

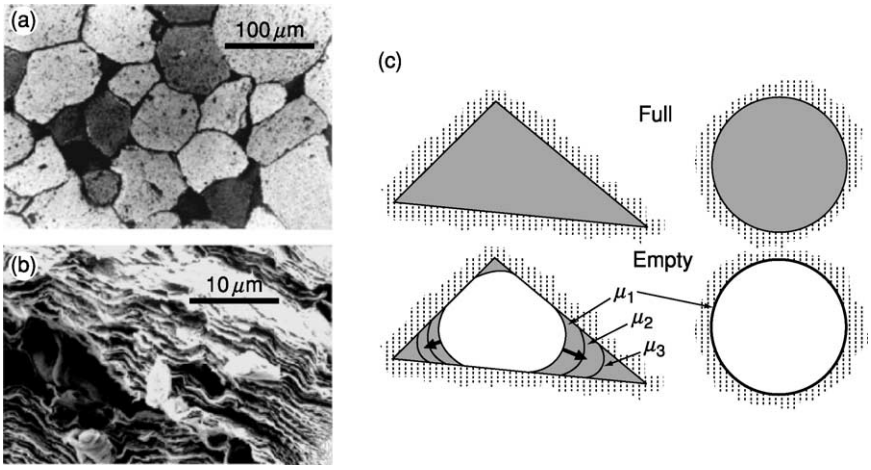


Figure 3 Angular pore space in natural porous media: (a) thin section of Devonian sandstone. (Reproduced with permission of the AAPG, whose permission is required for further use, from Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*; photograph by RF Sippel, AAPG © 1979. AAPG Mem. 28); (b) scanning electron micrograph (SEM) of calcium-saturated montmorillonite; (c) liquid retention in triangular and cylindrical pores. μ, chemical potential.

capillary behavior only, ignoring adsorbed liquid films that cover the flat pore surfaces.

In contrast to the piston-like filling or emptying of circular capillaries, angular pores undergo different filling stages and spontaneous displacement in the transition from dry to wet or vice versa. At relatively dry conditions (low chemical potentials), thin liquid films are adsorbed on all flat surfaces of the angular pore, and liquid accumulates in corners due to capillary forces. A further increase in chemical potential leads to further increase in film thickness and to an increase in the radius of capillary interface curvature in the corners, until the capillary corner menisci contact to form an inscribed circle. At this critical potential, liquid fills up the central pore spontaneously (snap-off). The radius of interface curvature at this critical point is equal to the radius of an inscribed circle in the pore cross-section:

$$r_{\text{imb}} = \frac{2 \cdot A}{P} = \frac{P}{4[F(\gamma) + \pi]} \qquad [4]$$

where r_{imb} is the radius of spontaneous imbibition, A is pore cross-sectional area, and P is the perimeter of the pore cross-section. For drainage, at a certain potential a liquid–vapor meniscus invades the pore and liquid is displaced from its center, leaving liquid in the corners. The radius of curvature at onset of drainage, r_{d}, is expressed in terms of perimeter (P) and the angularity factor $F(\gamma)$:

$$r_{\text{d}} = \frac{P}{2 \cdot [(F(\gamma) + \pi) + \sqrt{\pi \cdot (F(\gamma) + \pi)}]} \qquad [5]$$

An example of the effects of pore shape (and angularity) on imbibition and drainage processes is depicted in Figure 4, showing marked differences between cylindrical and angular pores having equal pore cross-sectional areas. Liquid displacement in cylindrical tubes during drainage is piston-like, leaving no liquid in the cross-section after the drainage threshold. Angular pores, on the other hand, show that liquid is displaced from the central region first (at a radius of curvature given by eqn [5], leaving some liquid in the corners. Subsequent decrease in chemical potential results in commensurate decreasing amounts of liquid in the corners ('lower' chemical potential or capillary pressure indicates more negative values, much like the terminology for the subzero temperature scale). The threshold chemical potential for different pore shapes with the same cross-sectional area increases (becomes less negative) with increasing angularity factor $F(\gamma)$. The same holds for the amount of liquid held in the corners.

The conditions during imbibition are slightly different. Liquid–vapor interfaces in corners of angular pores expand with increasing chemical potential to the point of snap-off (eqn [4]), where the pore completely fills up with liquid. The threshold chemical potential for snap-off increases with angularity $F(\gamma)$, and the amount of liquid held in the corners at a given chemical potential is directly proportional to angularity. Highly angular pore shapes such as triangular pores retain more liquid at the same potential than squares or hexagons (Figure 4). In the extreme case of cylindrical tubes, no liquid is held prior to spontaneous filling by snap-off (the 'empty–full' behavior).

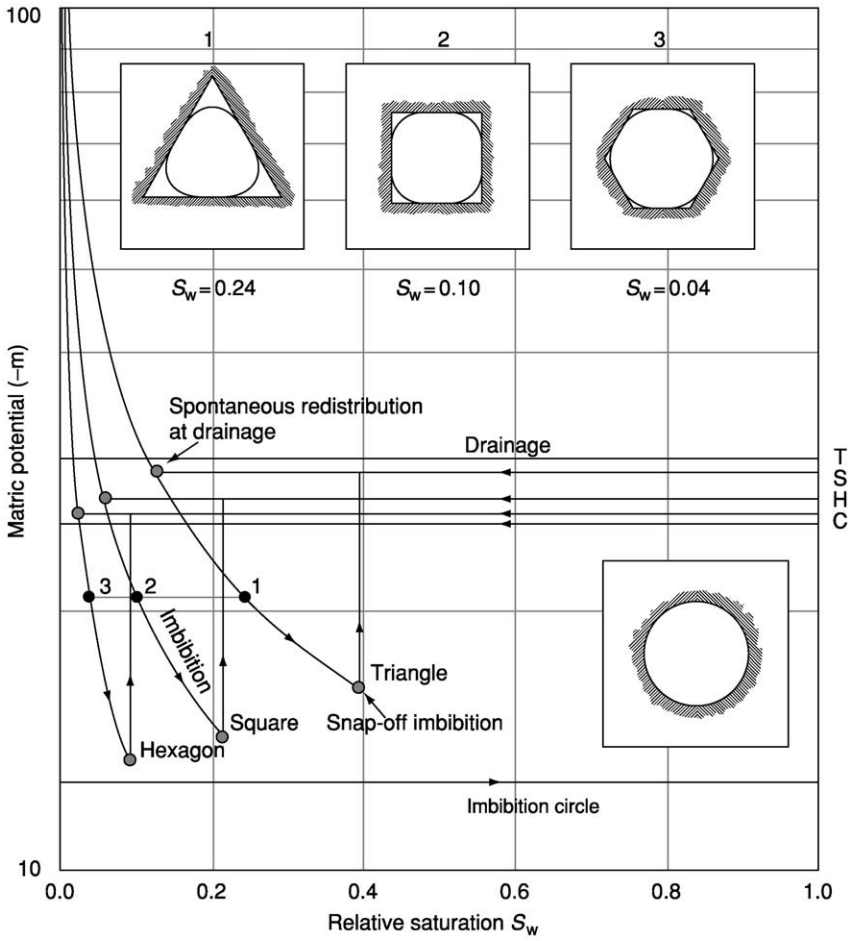


Figure 4 Capillary phenomena (imbibition and drainage) in pores with different cross-section shapes but identical pore cross-sectional area. T, triangle; S, square; H, hexagon; C, circle.

Modeling SWC

Measuring the SWC is laborious and time-consuming. Measured θ–ψ_m pairs are often fragmentary, and usually constitute relatively few measurements over the wetness range of interest. For modeling and analysis purposes, and for characterization and comparison between different soils and scenarios, it is therefore beneficial to represent the SWC in a mathematically continuous form. Several approaches, ranging from empirical parametric expressions to physically based models with parameters derived from measurable medium properties, to pore network or lattice Boltzmann simulations can be employed to represent a continuous SWC.

Empirical SWC Models

Key requirements for all parametric SWC expressions are parsimony (as few parameters as possible), to simplify parameter estimation, and accurate description of SWC behavior at the limits (wet and dry ends) while closely fitting the nonlinear shape of θ–ψ_m measurements.

An effective and commonly used parametric model for relating water content or effective saturation (Θ) to the matric potential was proposed by van Genuchten and is denoted as VG:

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left[\frac{1}{1 + (\alpha\,\psi_m)^n}\right]^m \qquad [6]$$

θ_r and θ_s are the residual and saturated water contents, respectively, ψ_m is matric potential, and α, n, and m are parameters directly dependent on the shape of the $\theta(\psi_m)$ curve. A common simplification is to assume that $m = 1 - 1/n$. Thus the parameters required for estimation of the model are θ_r, θ_s, α, and n. θ_s is sometimes known and is easy to measure, leaving only the three unknown parameters θ_r, α, and n to be estimated from the experimental data in many cases. Note that θ_r is sometimes taken as θ at -1.5 MPa, $\theta_{air\ dry}$, or a similar meaningful

value, though it is often advantageous to use it as a fitting parameter.

Another well-established parametric model is denoted by BC (Brooks–Corey):

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left(\frac{\psi_b}{\psi_m}\right)^{\lambda} \qquad \psi_m > \psi_b$$
$$\Theta = 1 \qquad \psi_m \leq \psi_b \qquad [7]$$

where ψ_b is a parameter related to the soil matric potential at air entry (b represents 'bubbling pressure'), and λ is related to the soil pore-size distribution. Matric potentials are expressed as positive quantities (i.e., in absolute values) in both VG and BC parametric expressions. Campbell uses the same power law function as BC to express degree of saturation ($S = \theta/\theta_s$) as a function of the air entry potential ψ_b and a factor b that can be related to soil texture:

$$\frac{\theta}{\theta_s} = \left(\frac{\psi_b}{\psi_m}\right)^{\frac{1}{b}} \qquad [8]$$

ψ_b and b are derived as functions of the geometric mean diameter and geometric standard deviation of the particle-size distribution. Equation [8] is often used in relation to the fractal idealization of the soil porous system as discussed below.

Estimation of VG or BC parameters from experimental data requires sufficient data points to characterize the shape of the SWC, and a program to perform nonlinear regression. Many computer spreadsheet software packages provide relatively simple and effective mechanisms to perform nonlinear regression.

Figure 5 depicts fitted parametric VG and BC models to silt loam $\theta(\psi_m)$ data. The resulting best-fit parameters for the VG model are: $\alpha = 0.417\,m^{-1}$; $n = 1.75$; $\theta_s = 0.513\,m^3\,m^{-3}$; and $\theta_r = 0.05\,m^3\,m^{-3}$ (with $r^2 = 0.99$). For the BC model, the best-fit parameters are: $\lambda = 0.54$; $\psi_b = 1.48\,m$; $\theta_s = 0.513\,m^3\,m^{-3}$; and $\theta_r = 0.03\,m^3\,m^{-3}$ (with $r^2 = 0.98$). Note that the most striking difference between the VG and the BC models is the discontinuity at $\psi = \psi_b$ in the BC model.

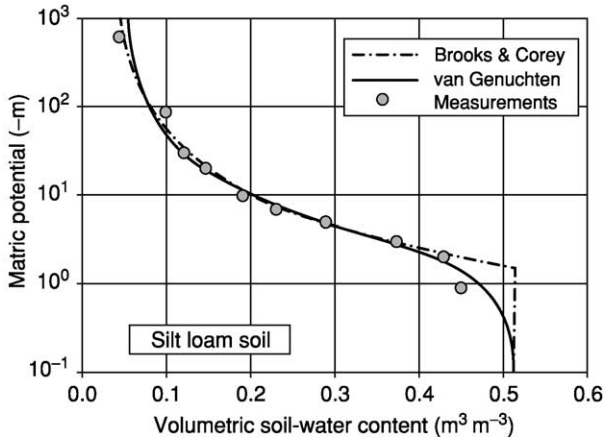


Figure 5 van Genuchten and Brooks and Corey parametric models fitted to measured data for silt loam soil.

Representative measured SWC information based on the VG and BC parametric models is available from a variety of sources. The UNSODA computer database compiled by the US Salinity Laboratory contains an exhaustive collection of soil-water retention and unsaturated hydraulic conductivity information for soils of different textures from around the world. While the authors have attempted to provide some indices concerning quality or reliability of the compiled data, users are advised (as always) to use their own experience and discretion in adapting others' data to their own applications. Regression studies provide a wealth of information on the BC parameter values for many soils. These include estimation of the hydraulic parameters based on other, often more easily available soil properties. These estimates may be sufficiently accurate for some applications and could be used to obtain first-order approximations. Table 1 contains listed values for the VG parameters α and n, and the residual and saturated water contents for various soil textural classes compiled from the UNSODA database. Note that substantial variation in SWC relationships for given soil textural classes is to be expected.

The models introduced so far can be categorized as empirical curve-fitting functions with free model parameters related to the specific shape of the employed

Table 1 Typical van Genuchten model parameters (α, n) including residual (θ_r) and saturated (θ_s) water contents compiled from the UNSODA database

Textural class	*N*	θ_r ($cm^3\,cm^{-3}$)	θ_s ($cm^3\,cm^{-3}$)	α (cm^{-1})	n
Sand	126	0.058	0.37	0.035	3.19
Loamy sand	51	0.074	0.39	0.035	2.39
Sandy loam	78	0.067	0.37	0.021	1.61
Loam	61	0.083	0.46	0.025	1.31
Silt	3	0.123	0.48	0.006	1.53
Silt loam	101	0.061	0.43	0.012	1.39
Sandy clay loam	37	0.086	0.40	0.033	1.49
Clay loam	23	0.129	0.47	0.030	1.37
Silty clay loam	20	0.098	0.55	0.027	1.41
Silty clay	12	0.163	0.47	0.023	1.39
Clay	25	0.102	0.51	0.021	1.20

N, the number of soils or samples of a given textural class from which the mean values are compiled.

Reproduced from Leij FJ, Alves WJ, van Genuchten MT, and Williams JR (1996) *The UNSODA Unsaturated Hydraulic Database*. EPA/600/ R-96/095. Cincinnati, OH: US Environmental Protection Agency.

mathematical function, rather than to physical properties of the porous medium. Relationships between the SWC and the pore-size distribution have been established, which can be described as a statistical log-normal distribution function. Brutsaert, and more recently Kosugi, established relationships between the SWC and the pore-size distribution, which can be described as a statistical log-normal distribution function. They idealized the soil porous system as a BCC with log-normal distributed radii, and apply the capillary-rise equation to establish a relationship between the matric potential and effective medium saturation. (*See* **Capillarity.**)

The applicability of the latter models is limited to coarse-textured soils and to conditions where capillary forces are the major contributor to the matric potential. They tend to fail in finer-textured soils with high specific surface area (e.g., clay soils) where adsorptive surface forces dominate, especially under drier conditions. (*See* **Water Potential.**)

Fractal Representation of the Soil Pore Space and the SWC

Fractals are hierarchical, often highly complex, spatial, or temporal systems that are generated with iterative algorithms obeying simple scaling rules. Patterns within such systems repeat themselves over a defined range of scales (self-similarity). This enables the reproduction of statistical properties of a particular pattern at other length or time scales. Fractal geometry can be applied to describe quantitatively irregularity and shape of natural objects by estimating their fractal dimension. Several theoretical models have been proposed to derive the SWC from the fractal representations of the soil porous system. There are two general approaches, based on either surface or mass fractals. Surface fractal models assume that water is only present in the form of adsorbed liquid films on pore surfaces, whereas mass fractal models assume that only capillary water obeying the capillary-rise equation is present within the fractal system. As with the BCC approach, fractal models for the SWC are based on derivation of the pore-size distribution from the fractal structure under consideration, neglecting pore connectivity and topology issues. Crawford presented the following relationship between the mass-fractal dimension (D_m) and the degree of saturation (S):

$$S = \left(\frac{\psi_m}{\psi_b}\right)^{(D_m - d_e)} \quad [9]$$

where ψ_m is the matric potential under consideration, ψ_b is the matric potential at the air entry point, and d_e is the embedding dimension. The embedding dimension equals 2 in two-dimensional systems and 3 in three-dimensional space. Note that the mass-fractal dimension D_m is always less than d_e. Due to the identical functional form, eqn [9] is commonly equated with Campbell's version of the B&C SWC function (eqn [8]) to derive Campbell's b-factor from the fractal dimension ($b = -1/(D_m - d_e)$).

Fractal approaches are limited due to the assumption that fractal scaling is valid over an infinite range of matric potentials. In reality natural porous media have lower and upper scaling limits related to their minimum and maximum pore sizes. Perfect proposed a mass-based model for the SWC that accounts for the finite range of the matric potential. Systems that are fractal over a finite range of scales are called prefractal.

Physically Based Models for the SWC

In contrast to the BCC and fractal approaches, angular pore-space models accommodate both capillary and adsorptive phenomena on internal surfaces. The basic unit pore element (**Figure 6**) is comprised of an angular central pore attached to slit-shaped spaces and is defined by the dimensionless slit-width and slit-length scaling parameters α and β, and the central pore length L. Variation of the cell parameters allows accommodation of a wide range of soil textural and structural classes. Sandy soils, for example, may be represented by a relatively large central pore length, and small values for β (i.e., small specific surface area). Fine-textured soils (e.g., clays) on the other

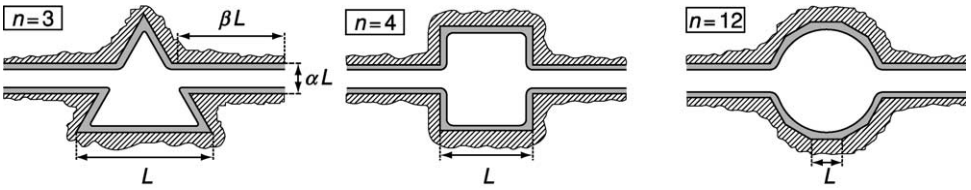


Figure 6 Basic unit elements of the angular pore space model of Tuller *et al.* (Reproduced from Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media-liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.)

hand can be represented by relatively small L and large β to account for the high internal surface areas often observed in such soils.

Despite the higher complexity of the angular pore-space model, it has distinct advantages over the more commonly applied BCC approach. First, angular pores more realistically represent natural porous medium pore space and allow dual occupancy of wetting and nonwetting phases (see the section Liquid Retention and Pore Shape, above). The potential of accommodating adsorptive surface forces leads to a more accurate derivation of the SWC for porous media with high specific surface areas as well as under dry conditions.

A modified form of the augmented Young–Laplace equation that considers capillary and adsorptive contributions to the matric potential is employed to calculate liquid–vapor interfaces within a cross-section of the angular pore model. In addition, dynamic liquid displacement mechanisms (see the section Liquid Retention and Pore Shape, above) derive SWC functions for imbibition or drainage at the pore scale. Figure 7 depicts a typical transition of a unit element from dry to wet (imbibition) that includes spontaneous liquid displacement (snap-off) in slits (Figure 7b) and in the central pore (Figure 7d).

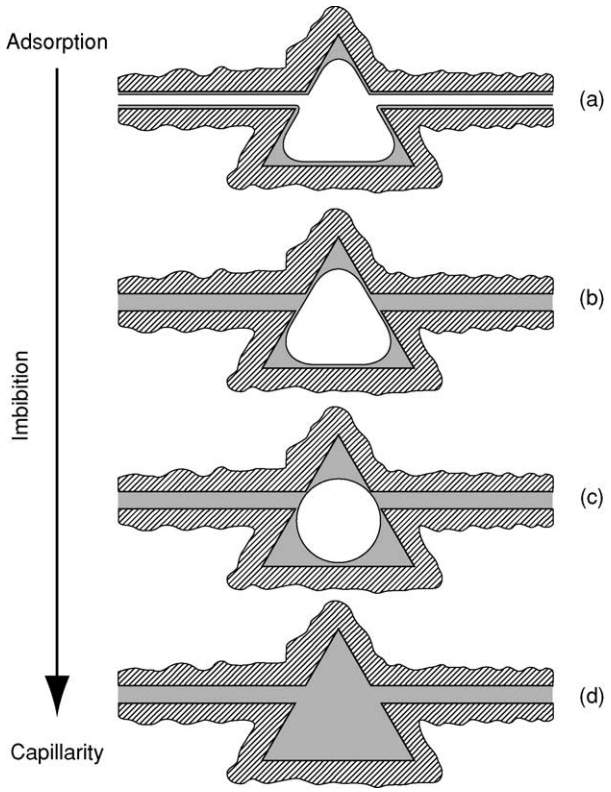


Figure 7 Unit element filling stages during imbibition. (a) Liquid films adsorbed on pore and slit walls and liquid held in corners due to capillary forces at low matric potentials; (b) spontaneous slit fill-up (capillary condensation); (c) pore snap-off; (d) full unit element.

A statistical upscaling scheme using a gamma density function for the central pore length L can be used to represent liquid behavior at the sample scale. The gamma distribution is given preference because it resembles the commonly observed positive skewness of soil pore-size distributions, and at the same time facilitates the derivation of analytical solutions for the SWC.

The upscaling procedure leads to a physically based expression for the SWC, where the degree of saturation (S) defined as a function of matric potential is expressed as the sum of functions corresponding to various pore-filling mechanisms and stages, as depicted in Figure 7. Figure 8a depicts a sample application for the Millville silt loam soil. Because the model calculates the configuration of liquid–vapor interfaces within pores and films, adsorptive and capillary contributions to the SWC can be separated (Figure 8a). Such separation is not possible with any of the empirical or semiempirical approaches presented above. Another important feature of this model is the ability to predict liquid–vapor interfacial area as shown in Figure 8b, comparing model predictions for sand with known surface area of 0.01–0.05 $m^2\,g^{-1}$. The magnitude and changes in liquid vapor interfacial area play an important role in multiphase flow processes, bioremediation of contaminated soils, gas-exchange phenomena, and virus and colloid transport.

Lattice Boltzmann Approach

Lattice Boltzmann models (LBM) are descendants of lattice gas cellular automata, which follow the motions of individual particles and were first presented as a viable means of solving Navier–Stokes equations. In the most simplistic sense, LBMs work with distributions of particles at each lattice point rather than with individual particles. LBMs simulate interactions of hypothetical particles confined to a regular lattice, which greatly enhances solvability of the Boltzmann equation. This has advantages for certain types of simulations in that averaging is not required to obtain smooth velocity fields.

The lattice Boltzmann method has emerged as a powerful tool for simulation of multiphase fluid systems, including water and water vapor. The LBM incorporates complex details of pore shape that characterize realistic porous media, fluid factors, and solid–fluid interactions in a physically sound way, leading to realistic simulation of interfaces between different fluids or between a liquid and its vapor and adsorption of wetting films. Interfaces arise, deform,

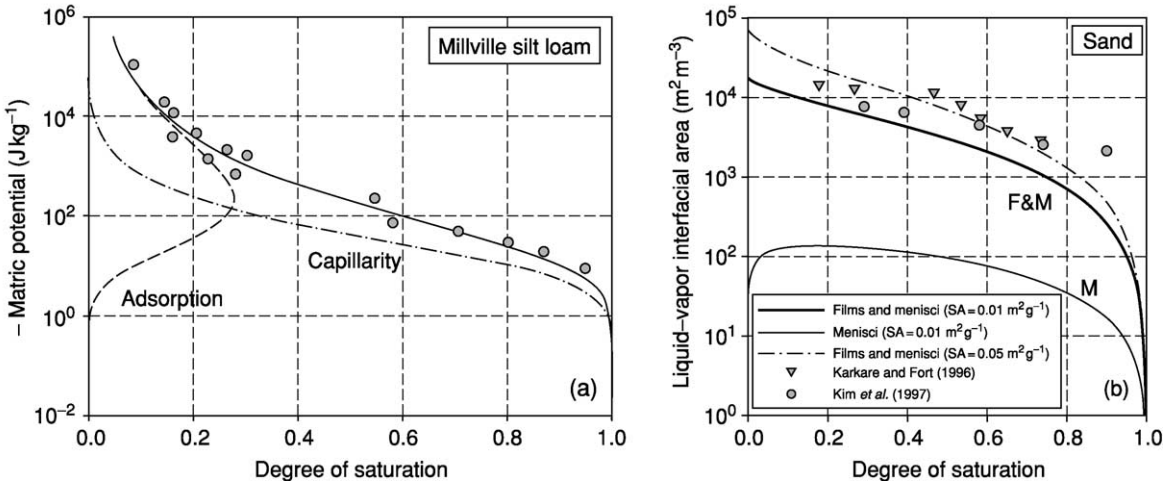


Figure 8 (a) Application of the angular pore-space model for Millville silt loam soil. Note the capillary and absorptive contributions to the soil-water characteristic (SWC); (b) comparison of model-calculated capillary (menisci) and adsorptive (films) contributions to liquid–vapor interfacial area as a function of saturation in an artificial sand mixture and interfacial area measurement results obtained with interfacial tracers. SA, surface area; F&M, films and menisci, Karkare & Fort (1996), Karkare MV and Fort J (1996) Determination of the air–water interfacial area in wet ''unsaturated'' porous media. *Langmuir* 12: 2041–2044; Kim *et al.* (1997), Kim H, Rao PSC, and Annable MD (1997) Determination of effective air–water interfacial area in partially saturated porous media using surfactant adsorption. *Water Resources Research* 33: 2705–2711. (Reproduced from Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.)

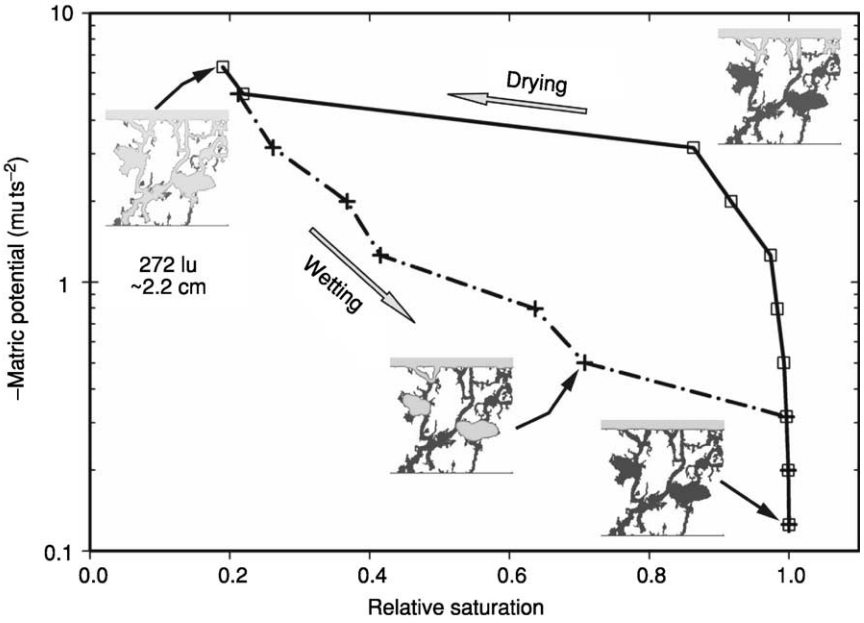


Figure 9 Lattice Boltzmann model simulation of liquid distribution in a complex pore system and computation of soil-water characteristic. (Reproduced from Sukop M and Or D (2003) Lattice Boltzmann method for modeling liquid–vapor interface configurations in porous media. *Water Resources Research* 40(1): W01509.)

and migrate in virtually any pore geometry rather naturally without the need for complex interface tracking algorithms.

An example of LBM application to compute water retention in complex porous media is shown as insets in Figure 9, which are based on two-dimensional (2-D) imagery of a real soil. Vapor and liquid boundaries of equal pressure were applied to the top and bottom of the domain, respectively, in steps corresponding to equal increments of log(−matric

potential). The matric potential is expressed in mass units per time step squared (equivalent to energy per unit volume in a 3-D system). Each potential step was terminated when the relative change in fluid mass in the domain was negligibly small. The simulated liquid behavior and SWC curve show appreciable differences in fluid configurations during wetting and drying that resulted from hysteresis.

Pore Network Models and the SWC

Network models were first developed by Fatt, based on the idea that pore space may be represented as an interconnected network of capillary tubes whose radii represent the dimensions of the pores within a porous medium. For a given matric potential, liquid–vapor interfacial configurations within the network can be determined exactly, based on pore-scale capillary and dynamic displacement considerations. The macroscopic SWC is then determined based on geometric volume averaging of the spatially distributed liquid within the network. A primary advantage of pore network models is explicit consideration of pore connectivity and topology in a simplified and mathematically tractable framework. Some of the limitations involve oversimplification of pore-scale physics (e.g., neglect of adsorptive pore-scale processes), incomplete understanding of interface migration and routing, inadequate technologies for inference of network parameters from real samples, and significant computational burden for detailed 3-D networks even at a core scale (greater than 100 mm).

Hysteresis of the SWC

Water content and the potential energy of soil water are not uniquely related, because the amount of water present at a given matric potential is dependent on the pore-size distribution and the properties of air–water–solid interfaces. An SWC relationship may be obtained by: (1) taking an initially saturated sample and applying suction or pressure to desaturate it (desorption), or (2) gradually wetting an initially dry soil (sorption). These two pathways produce curves that in most cases are not identical; the water content in the 'drying' curve is higher for a given matric potential than that in the 'wetting' branch (Figure 10a). This is called 'hysteresis,' defined as "the phenomenon exhibited by a system in which the reaction of the system to changes is dependent upon its past reactions to change."

The hysteresis in SWC can be related to several phenomena, including: (1) the 'ink bottle' effect, resulting from nonuniformity in shape and sizes of interconnected pores; drainage is governed by the smaller pore radius r, whereas wetting is dependent on the large radius R (Figure 10c); (2) different liquid–solid contact angles for advancing and receding water menisci (Figure 10b); (3) entrapped air in

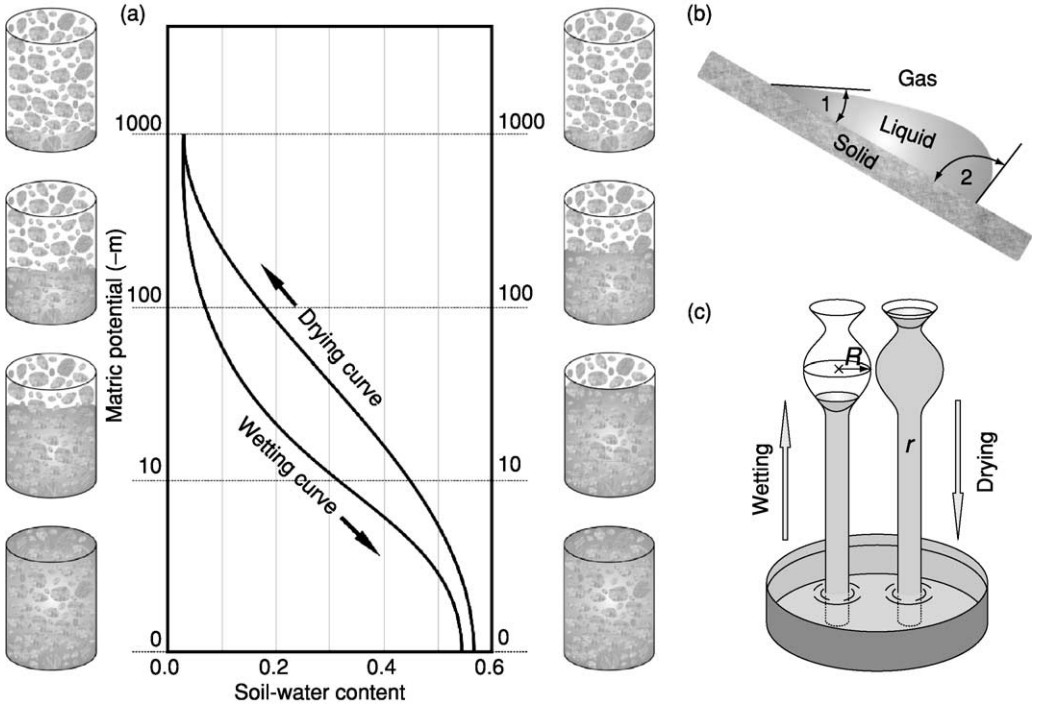


Figure 10 (a) Hysteresis of the soil-water characteristic. (b) The contact angle effect and (c) the ink bottle effect as potential mechanisms for hysteresis.

a newly wetted soil (e.g., pore doublet); and (4) swelling and shrinking of the soil under wetting and drying. The role of individual factors remains unclear and is subject to ongoing research.

Part of the hysteresis phenomena may be attributed to measurement artifacts, for example due to differences between tension- and pressure-induced desaturation. A potentially important aspect of desorption methods under tension is the possibility of liquid displacement (drainage) even in the absence of a continuous gaseous phase due to cavitation initiated by encapsulated gas bubbles or the liquid's own vapor pressure. Surface heterogeneity and impurities in soil and rock water are conducive to lowering the cavitation threshold.

Measurement of SWC Relationships

A variety of methods may be used to obtain requisite θ and ψ_m values to estimate the SWC. Potential experimental problems include: the limited functional range of the tensiometer, which is often used for *in situ* measurements; inaccurate θ measurements in some cases; the difficulty in obtaining undisturbed samples for laboratory determinations; and a slow rate of equilibrium under low matric potential (i.e., dry soils).

In situ methods are preferred in defining SWCs, as is measuring over a wide range of ψ_m and θ values. An effective method to obtain simultaneous measurements of ψ_m and θ_v is by installing time-domain

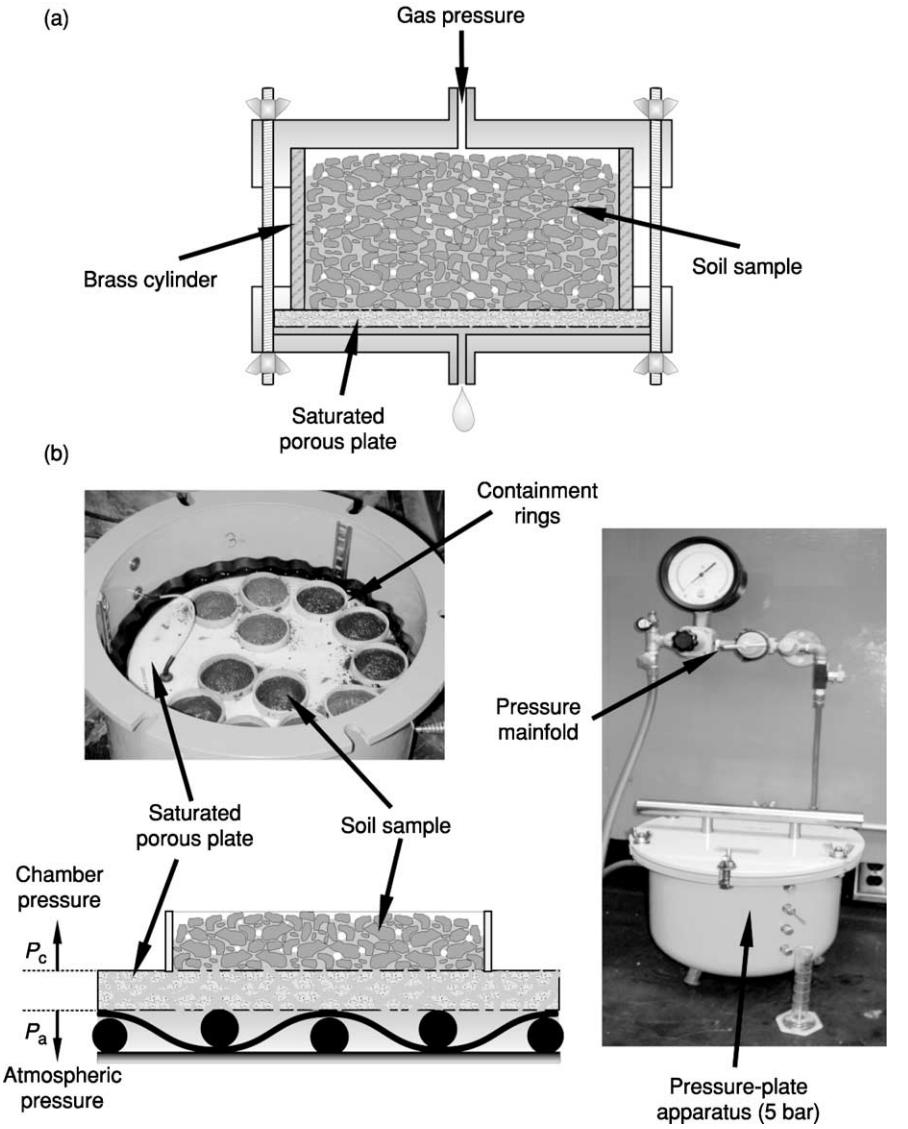


Figure 11 (a) Tempe pressure and flow cell; and (b) pressure-plate apparatus used to desaturate soil samples to specified matric potentials.

reflectometry (TDR) probes in close proximity to transducer tensiometers, with the attributes monitored during variable soil wetness. Large changes in ψ_m and θ_v are expected under highly evaporative conditions near the soil surface, or in the presence of active plant roots.

Pressure-Plate Apparatus and Pressure-Flow Cells

The pressure-plate apparatus comprises a pressure chamber enclosing a water-saturated porous plate, which allows water but not air to flow through its pores (Figure 11b). The porous plate is at atmospheric pressure at the bottom, while the top surface is at the applied pressure of the chamber. Soil samples, usually sieved to less than 2 mm, are placed in retaining rings in contact with the porous plate and allowed to saturate by immersion in water. The porous plate with saturated soil samples is then placed in the chamber and a known N_2 or air gas pressure is applied to force water out of the soil through the plate. Water flows out of the soil until equilibrium between the force exerted by the air pressure and the force by which soil water is being held by the soil (ψ_m) is attained.

Soil water retention at the wet end (less than 1 bar) is strongly influenced by the soil structure and its natural pore-size distribution. Hence, 'undisturbed' intact soil samples (cores) are preferred over repacked samples for this portion of the SWC. The pressure-flow cell (also known as Tempe cell) can hold intact soil samples encased in metal rings (Figure 11a). The operation of the Tempe cell follows that of the pressure plate, except its pressure range is usually limited to 0–1 bar or 0.1 MPa. The porous ceramic material used in pressure plates and flow cells must be completely water-saturated prior to use. Following equilibrium between soil matric potential and the applied air pressure, the soil samples are removed from the apparatus, weighed wet, then oven-dried to determine the mass water content gravimetrically. These may be converted to volume water contents through knowledge of the sample bulk densities. The water content of repacked soils at a given matric potential should not be used to infer θ of intact soils at the same ψ_m, due to modified pore sizes and pore geometry.

Multiple pressure steps may be applied to the same soil samples when using Tempe cells. The cells may be sequentially disconnected from the pressure source, weighed to determine the change in water content from the previous step, then reconnected and the next pressure step applied. Outflow of water from the cells may be collected to calculate or confirm changes in sample water contents.

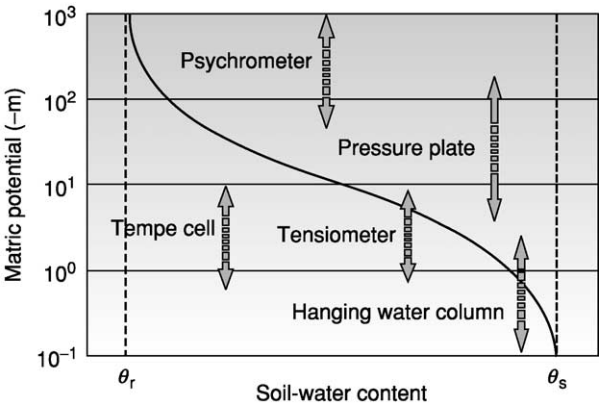


Figure 12 Some common soil-water characteristic measurement methods and their corresponding matric potential ranges.

Paired Sensors for Field SWC Measurement

In spite of the importance of measuring SWCs *in situ*, few suitable methods are available for field or *in situ* application. Paired sensors may be used to measure θ and ψ_m in the same or closely adjacent soil volumes, over a range of soil wetness. Examples include use of neutron moisture-meter access tubes or TDR waveguides, together with tensiometers. Application of paired sensors is often constrained by differences in soil volumes sampled by the respective sensors, different time constants required (e.g., many sensors obtain instantaneous measurements, while many ψ_m methods require equilibrium), and the fact that few matric-potential techniques function over the entire range of interest for wetness. This commonly provides limited overlap in soil water retention measured using different techniques, and measurements obtained using different methods may not be consistent in the ranges of overlap.

Figure 12 presents some common methods to measure or infer soil matric potential, including their respective ranges of application. Many of the available techniques have a limited range of overlap or do not overlap at all, and few of the methods are amenable to field measurements.

See also: **Capillarity**; **Hydrodynamics in Soils**; **Hysteresis**; **Porosity and Pore-Size Distribution**; **Water Content and Potential, Measurement**; **Water Potential**

Further Reading

Ahl C and Niemeyer J (1989) The fractal dimension of the pore volume inside soils. *Zeitschrift für Pflanzenernährung und Bodenkunde* 152: 457–458.

Berkowitz B and Ewing RP (1998) Percolation theory and network modeling applications in soil physics. *Surveys in Geophysics* 19: 23–72.

Bird NRA, Bartoli F, and Dexter AR (1996) Water retention models for fractal soil structures. *European Journal of Soil Science* 47: 1–6.

Brooks RH and Corey AT (1964) *Hydraulic Properties of Porous Media*. Hydrology Papers, No. 3. Fort Collins, CO: Colorado State University Press.

Celia MA, Reeves PC, and Ferrand LA (1995) Recent advances in pore scale models for multiphase flow in porous media. US National Report of the International Union of Geodesy and Geophysics 1991–1994. *Reviews of Geophysics* 33: 1049–1057.

Chen S and Doolen GD (1998) Lattice Boltzmann method for fluid flows. *Annual Review of Fluid Mechanics* 30: 329–364.

Crawford JW (1994) The relationship between structure and the hydraulic conductivity of soil. *European Journal of Soil Science* 45: 493–502.

Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure*, 2nd edn. San Diego, CA: Academic Press.

Giménez D, Perfect E, Rawls WJ, and Pachepsky YA (1997) Fractal models for predicting soil hydraulic properties: a review. *Engineering Geology* 48: 161–183.

Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Reiskirchen, Germany: Catena-Verlag.

Leij FJ, Alves WJ, van Genuchten MT, and Williams JR (1996) *The UNSODA Unsaturated Hydraulic Database*. EPA/600/R-96/095. Cincinnati, OH: US Environmental Protection Agency.

Li Y and Wardlaw NC (1986) Mechanisms of nonwetting phase trapping during imbibition at slow rates. *Journal of Colloid Interface Science* 109: 473–486.

Mandelbrot BB (1975) *Les Objets Fractals: Forme, Hasard et Dimension*. Paris, France: Flammarion.

Mandelbrot BB (1982) *The Fractal Geometry of Nature*. San Francisco, CA: Freeman.

Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12(3): 513–522.

Mualem Y (1984) A modified dependent domain theory of hysteresis. *Soil Science* 137: 283–291.

Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.

Or D and Tuller M (2002) Cavitation during desaturation of porous media under tension. *Water Resources Research* 38(5): 19.

Or D, Groeneveld DP, Loague K, and Rubin Y (1991) *Evaluation of Single and Multi-parameter Methods for Estimating Soil-Water Characteristic Curves*. Geotechnical Engineering Report No. UCB/GT/91-07. Berkeley, CA: University of California Press.

Pachepsky YA, Shcherbakov RA, and Korsunskaya LP (1995) Scaling of soil water retention using a fractal model. *Soil Science* 159: 99–104.

Perfect E (1999) Estimating soil mass fractal dimensions from water retention curves. *Geoderma* 88: 221–231.

Rawls WJ and Brakensiek DL (1989) Estimation of soil water retention and hydraulic properties. In: Morel-Seytoux HJ (ed.) *Unsaturated Flow in Hydraulic Modeling Theory and Practice*, pp. 275–300. NATO ASI Series. Series C: Mathematical and physical sciences, No. 275.

Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock*. Weinheim, Germany: VCH-Verlag.

Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*. American Association of Petroleum Geologists Memorandum 28. Julsa, OK: American Association of Petroleum Geologists.

Tuller M and Or D (2001) Hydraulic conductivity of variably saturated porous media – laminar film and corner flow in angular pore space. *Water Resources Research* 37(5): 1257–1276.

Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media-liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Tyler SWI and Wheatcraft SW (1990) Fractal processes in soil water retention. *Water Resources Research* 26: 1047–1054.

van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.

Water Table *See* **Groundwater and Aquifers**

WATER, PROPERTIES

D Hillel, Columbia University, New York, NY, USA

Introduction

Our planet is the planet of life primarily because it is blessed with the precise ranges of temperature and pressure that make possible the existence in a liquid state of a singular substance called water. So ubiquitous is water on our globe, covering nearly three-quarters of its surface, that the entire planet really should be called 'Water' rather than 'Earth.' However, as Coleridge's Ancient Mariner complained, most of the water everywhere is unfit to drink. Less than 1% of the water on earth is 'fresh' (i.e., non-saline) water, and that amount is unevenly distributed. Humid regions are endowed with an abundance of it, even with a surfeit, so that often the problem is how to dispose of excess water. Arid and semiarid regions, on the other hand, are afflicted with a chronic shortage.

Life as we know it began in an aquatic medium and water is still the principal constituent of living organisms. It is, literally, the essence of life. As Vladimir Vernadsky wrote a century ago: "Life is animated water." Though we appear to be solid, we are really liquid bodies, similar to gelatin, which also seems solid but is in fact largely water, made consistent by the presence of organic material. The analogous material in our bodies is protoplasm. The water content of a newborn infant is nearly 90% water by mass, and even in adults it is over 65%. Actively growing herbaceous plants typically contain over 90% water. Far from being a bland, inert liquid, water is a highly reactive substance, a solvent and a transporter of numerous substances.

The importance of water was recognized early in history, yet little was known about its real nature. In the Middle Ages, people believed that fresh water emanated magically from the bowels of the earth. They could not imagine that all the water flowing in innumerable springs and mighty rivers (such as the Nile, which appeared to the ancient Egyptians to come out of the driest desert!) could possibly result from so seemingly feeble a source as rain and snow. The first to conjecture this was Leonardo da Vinci, but only in the latter part of the seventeenth century did the English astronomer Edmond Halley and, separately, the Frenchman Claude Perrault prove the principle by calculation and measurement. Water was long thought to be a single element, until early in the eighteenth century, when it was found to consist of hydrogen and oxygen in combination.

Notwithstanding its ubiquity, water remains something of an enigma, possessing unusual and anomalous attributes. Perhaps the first anomaly is that, being a compound of two gases and having relatively low molecular weight, water is a liquid and not a gas at normal temperatures. (Its sister compound, hydrogen sulfide, H_2S, has a boiling-point temperature of $-60.7°C$.) Compared with other common liquids, water has unusually high melting and boiling points, heats of fusion and vaporization, specific heat, dielectric constant, viscosity, and surface tension.

Molecular Structure

One cubic meter of liquid water at 20°C contains about 3.4×10^{28} (34 billion billion billion) molecules, the diameter of which is about 3×10^{-10} meter ($3\times10^{-4}\,\mu$m, or about 3 Angstrom units). The chemical formula of water is H_2O, which signifies that each molecule consists of two atoms of hydrogen and one of oxygen. There are three isotopes of hydrogen (^{1}H, ^{2}H, ^{3}H) and three of oxygen (^{16}O, ^{17}O, ^{18}O), which can form 18 combinations. However, all isotopes but ^{1}H and ^{16}O are quite rare.

The hydrogen atom consists of a positively charged proton and a negatively charged electron. The oxygen atom consists of a nucleus having a positive charge of eight protons, surrounded by eight electrons, of which six are in the outer shell. Since the outer electron shell of hydrogen lacks one electron and that of oxygen lacks two electrons, one atom of oxygen can combine with two atoms of hydrogen in an electron-sharing molecule.

The strong intermolecular forces in liquid water are caused by the electrical polarity of the water molecule, which in turn is a consequence of the arrangement of electrons in its oxygen and hydrogen atoms (Figure 1). The oxygen atom shares a pair of electrons with each of the two hydrogen atoms, through overlap of the 1s orbitals of the hydrogen atoms with two hybridized sp^3 orbitals of the oxygen atom. The H–O–H bond in water is not linear but bent

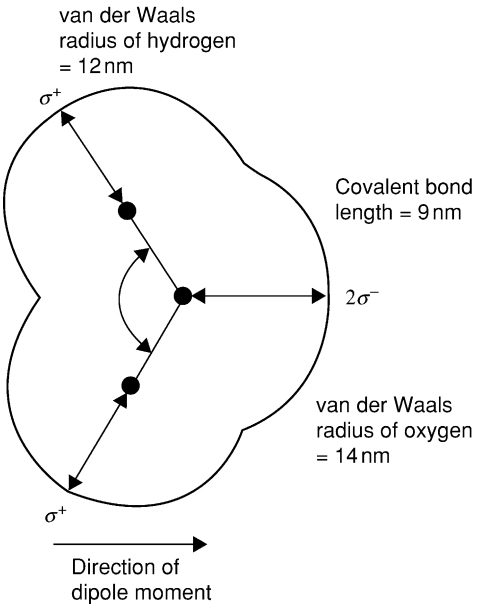


Figure 1 Model of a water molecule. The curved lines represent the borders at which van der Waals attractions are counterbalanced by repulsive forces. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

at an angle of 104.5°. That angle deviates slightly from a perfectly tetrahedral arrangement of the oxygen atom's four possible sp^3 orbitals, which would have an angle of 109.5°. The mean H-O interatomic distance is 9.65×10^{-5} μm. The arrangement of electrons in the molecule gives it electrical asymmetry. The electronegative oxygen atom tends to attract the single electrons of the hydrogen atoms, leaving the hydrogen nuclei bare. Hence, each of the two hydrogen atoms has a local partial positive charge. The oxygen atom, in turn, has a partial negative charge, located in the zone of the unshared orbitals. Thus, though the water molecule has no net charge, it forms an electrical dipole.

Hydrogen Bonding

Every hydrogen proton, while attached primarily to a particular molecule, is also attracted to the oxygen of the neighboring molecule, with which it forms a secondary link known as a hydrogen bond. Though the intermolecular link resulting from dipole attraction is not as strong as the primary link of the hydrogen to the oxygen of its own molecule, water can be regarded as a polymer of hydrogen-bonded molecules. This structure is most complete in ice crystals, in which each molecule is linked to four neighbors via four hydrogen bonds, thus forming a hexagonal lattice with a rather open structure (Figure 2). When the ice melts, this rigid structure collapses partially, so additional molecules can

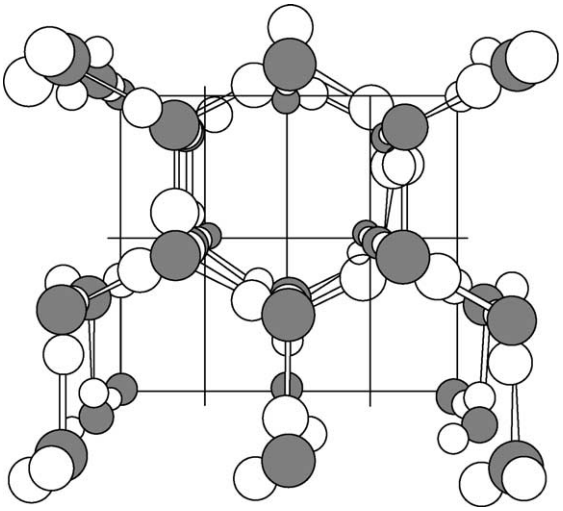


Figure 2 Schematic structure of an ice crystal. The oxygen atoms are shown in gray and the hydrogen atoms in white. The pegs linking adjacent molecules represent hydrogen bonds. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

enter the intermolecular spaces and each molecule can have more than four near neighbors. For this reason, liquid water can be more dense than ice at the same temperature, and thus lakes and ponds develop a surface ice sheet in winter rather than freeze solid from bottom to top as they would if ice were denser than liquid water.

States of Water

In the vapor or gaseous state, water molecules are largely independent of one another and occur mostly as monomers signified as $(H_2O)_1$. Occasionally, colliding molecules may fuse to form dimers $(H_2O)_2$ or even trimers, $(H_2O)_3$, but such combinations are rare. However, in the solid state a rigidly structured lattice forms with a tetrahedral configuration (Figure 2) that can be schematically depicted as sheets of puckered hexagonal rings (Figure 3). As many as nine alternative ice forms can occur when water freezes, depending on prevailing temperature and pressure conditions. Figure 3 pertains to ice 1, the familiar form, which occurs and is stable at ordinary atmospheric pressure.

The orderly structure of ice does not totally disappear in the liquid state. The polarity and hydrogen bonds continue to bind water molecules together, though the structural forms that develop in the liquid state are much more flexible and transient than in the rigidly structured solid state. Hydrogen bonds in liquid water form an extensive three-dimensional network, the detailed features of which appear to be short-lived. According to the 'flickering cluster' model discovered by Frank and Wen and modified

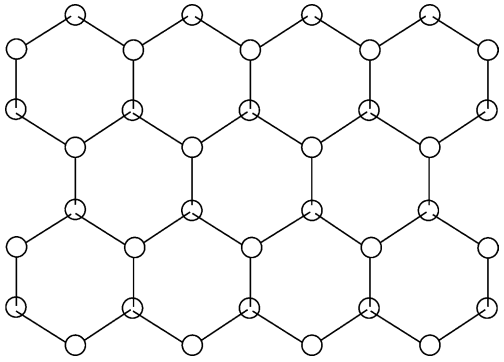


Figure 3 The crystalline structure of ice. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

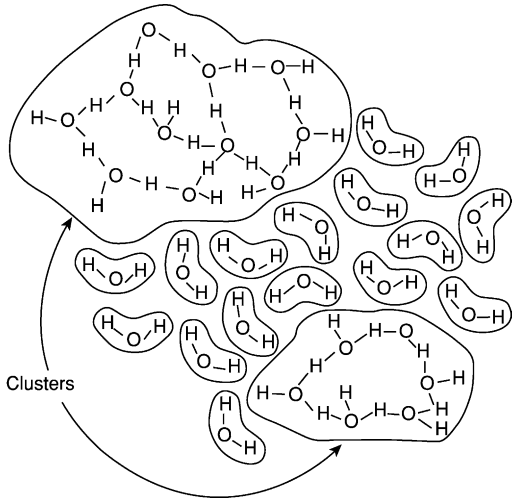


Figure 4 Schematic illustration of 'flickering clusters,' showing polymeric associations and monomeric molecules in liquid water. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

by Erland, the molecules of liquid water associate and dissociate repeatedly in transitory or flickering polymer groups, designated $(H_2O)_n$, having a quasi-crystalline internal structure. These microcrystals, as it were, form and melt so rapidly on the scale of picoseconds and randomly that, on a macroscopic scale, water appears to behave as a homogeneous liquid (Figure 4).

In transition from solid to liquid, and from liquid to gas, hydrogen bonds must be broken (while in freezing and condensation they are re-established). Hence relatively high temperatures and energies are required to achieve these transitions. To thaw 1 kg of ice, 3.35×10^5 J (80 cal g^{-1}) must be supplied. Conversely, the same energy (the latent heat of fusion) is released in freezing. At the boiling point (100°C at atmospheric pressure), water passes from the liquid to the gaseous state and in so doing it absorbs 2.26×10^6 J kg^{-1} (540 cal g^{-1}). This amount of heat is known as the latent heat of vaporization. Water can be vaporized at temperatures below 100°C, but such vaporization requires greater heat or lower atmospheric pressure. At 30°C, the latent heat is about 2.43×10^6 J kg^{-1} (580 cal g^{-1}). Sublimation is the direct transition from the solid state to vapor, and the heat absorbed by it is equal to the sum of the latent heats of fusion and of vaporization.

Ionization and pH

Because of its small mass and the tightness with which its single electron is bound to the oxygen atom, the nucleus of the hydrogen atom in the water molecule exhibits a finite tendency to dissociate from the oxygen with which it is covalently associated and to 'jump' to the adjacent water molecule, to which it is hydrogen-bonded. Such an event produces two ions: the hydronium ion (H_3O^+) and the hydroxyl ion (OH^-). The reaction described is reversible, and should be written as:

$$2H_2O \rightleftharpoons (H_3O)^+ + OH^-$$

However, by convention it is written simply as:

$$H_2O \rightleftharpoons H^+ + OH^-$$

and one speaks of 'hydrogen ions' rather than of 'hydronium ions.'

Although the self-ionization of water is small, its consequences are extremely important. The ionization is reversible, and it tends to an equilibrium state in which the rate of dissociation into ions equals the rate of ion reassociation to form molecules once again. For such a system in equilibrium (at which the concentration of each of the species H_2O, H^+, and OH^- remains constant), the law of mass action applies; i.e., the ratio of concentrations of the products and the reactants must be constant. Using brackets to denote concentration, we can write this in the following way:

$$K_{equil} = [H^+][OH^-]/[H_2O] \qquad [1]$$

Since the number of water molecules undergoing dissociation at any given time is very small relative to the total number of water molecules present, $[H_2O]$ can be considered constant. Assuming this concentration to be 55.5 mol l^{-1} (the number of grams per liter divided by the gram molecular weight: 1000/18 = 55.5 mol l^{-1}), we can simplify the equilibrium constant expression as follows:

$$55.5 \times K_{equil} = [H^+][OH^-], \text{or } K_w = [H^+][OH^-] \qquad [2]$$

in which K_w is a composite constant called the ion product of water. In fact, the concentrations of H^+ and OH^- ions in pure water at 25°C are 10^{-7} mol l^{-1}, an extremely small value when compared to the overall concentrations of (largely undissociated) water, namely 55.5 mol l^{-1}. Thus, K_w at 25°C is 10^{-14}. If the hydroxyl ion concentration $[OH^-]$ is changed, the hydrogen ion concentration $[H^+]$ changes automatically to maintain the constancy of the product, and vice versa. An excess concentration of hydrogen ions over the concentration of hydroxyl ions imparts to the aqueous medium the property of acidity, whereas a predominance of hydroxyl ions produces the opposite property of alkalinity or basicity. A condition in which the concentrations of H^+ and OH^- are equal is called neutrality.

The ion product of water, K_w, is the basis for the pH scale, a measure of the concentration of H^+ (and of OH^- as well) in any aqueous solution in the range of concentration between 1.0 mol l^{-1} H^+ and 1.0 mol l^{-1} OH^-. The pH is defined as

$$\mathrm{pH} = \log_{10} 1/[H^+] = -\log_{10}[H^+] \qquad [3]$$

As already stated, in a precisely neutral solution at 25°C,

$$[H^+] = [OH^-] = 1.0 \times 10^{-7} \mathrm{mol\,l^{-1}}$$

The pH of such a solution is

$$\mathrm{pH} = \log_{10}[1/(1 \times 10^{-7})] = 7.0$$

The pH value of 7.0 for a neutral solution is thus not arbitrary, but derives from the value of the ion product of water at 25°C. The pH scale is logarithmic, not arithmetic. If two solutions differ in pH by one unit, then one solution has 10 times the hydrogen ion concentration of the other. Thus, a pH of 6 implies a hydrogen ion concentration of 10^{-6} and a hydroxyl ion concentration of 10^{-8}. A pH of 5 indicates $[H^+] = 10^{-5}$ (acidity 10 times that of the above) and $[OH^+] = 10^{-9}$.

Solvent Properties of Water

Water dissolves or disperses many substances because of its polar nature. Hence it has been called the universal solvent. All chemical substances have finite solubilities in water, but these solubilities range widely. Many crystalline salts and other ionic compounds readily dissolve in water but are nearly insoluble in nonpolar liquids such as chloroform or benzene. Since the crystal lattice of salts, such as sodium chloride, is held together by very strong electrostatic attractions between alternating positive and negative ions, considerable energy is required to pull these ions away from one another. However, water dissolves sodium chloride because the strong electrostatic attraction between water dipoles and the Na^+ and Cl^- ions, forming stable hydrated Na^+ and Cl^- ions, exceeds the attraction of these ions to each other. In the case of Na^+, hydration is represented by the process:

$$Na^+ + (n+4)H_2O - Na(H_2O)_4(H_2O)_n^+$$

illustrated in Figure 5. In addition to hydration, there is also the hydrolysis of metal species, a reaction in which the metal ion displaces one of the protons (hydrogen) of water to form basic hydroxides.

Ion solvation is also aided by the tendency of the solvent to oppose the electrostatic attraction between ions of opposite charges. This is characterized by the dielectric constant D, which is defined by the relationship:

$$F = \kappa e_1 e_2 / \mathrm{D} r^2 \qquad [4]$$

Here F is the attractive force between two ions of opposite charge, κ is Coulomb's constant, e_1 and e_2 are charges on the ions, and r is the distance between them. Water has an extremely high dielectric constant, as shown in Table 1. For instance, the attractive force between Na^+ and Cl^- ions at a given distance in water is less than one-third that in ethanol and only

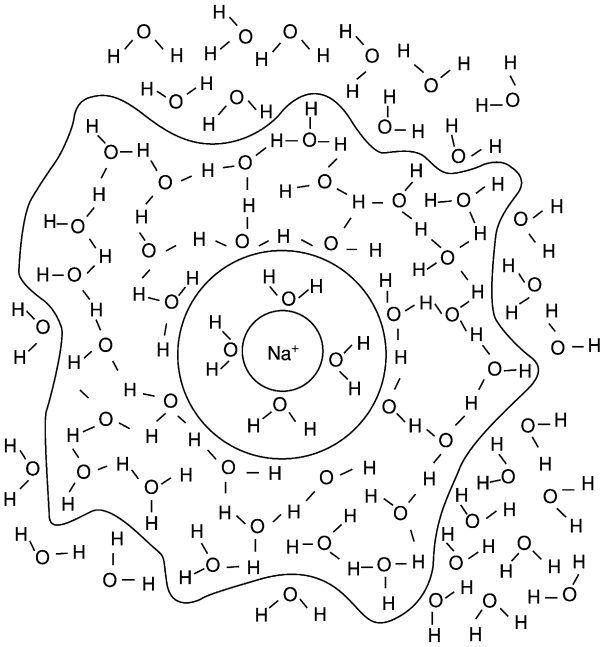


Figure 5 A model of the hydration 'atmosphere' of a sodium ion. An inner shell of more or less rigidly structured water is surrounded by a cluster of looser but still structure-enhanced water, the whole floating in a sea of 'free' water. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

Table 1 Dielectric constants of some liquids (20°C)

Water	80	Acetone	21.4
Methanol	33	Benzene	2.3
Ethanol	24	Hexane	1.9

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

one-40th that in benzene. This fact greatly facilitates hydration of ions and dissolution of the crystal lattices of salts in water.

The effect of a solute on the solvent is manifest in a set of properties, namely, the colligative properties of solutions, which depend on the number of solute particles per unit volume of solvent. Solutes produce such characteristic effects in the solvent as depression of the freezing point and of the vapor pressure, and elevation of the boiling point. They also endow a solution with the property of osmotic pressure. Theoretically, 1 mol of an ideal solute dissolved in 1 kg of water at a pressure of 760 mm of mercury (0.1 MPa, or 1 bar) depresses the freezing point by 1.86°C and elevates the boiling point by 0.543°C. Such a solution also yields an osmotic pressure of 2.24 MPa (22.4 atm) in an appropriate apparatus. However, aqueous solutions usually deviate considerably from ideal behavior, and the deviations are greater the higher the concentrations. The quantitative relationships given above are exact only at infinite dilution.

Osmotic Pressure

Owing to the constant thermal motion of all molecules in a fluid (above a temperature of absolute zero), solute species spread throughout the solution in a spontaneous tendency toward a state of equal concentration throughout. This migration of solutes in response to spatial differences in concentration is called diffusion.

If a physical barrier is interposed between two regions, across the path of diffusion, and if that barrier is permeable to molecules of the solvent but not to those of the solute, the former will diffuse through the barrier in a process called osmosis (from the Greek $\omega\sigma\mu o\sigma$, meaning 'push'). As in the case of unhindered diffusion, this process tends toward a state of uniform concentration even across the barrier. Barriers permeable to one substance in a solution but not to another are called selective or semipermeable membranes. Membranes surrounding cells in living organisms, for example, exhibit selective permeability to water while restricting the diffusion of solutes between the cells' interior and their exterior environment. Water molecules cross the membrane in both directions, but the net flow of water is from the more dilute solution to the more concentrated.

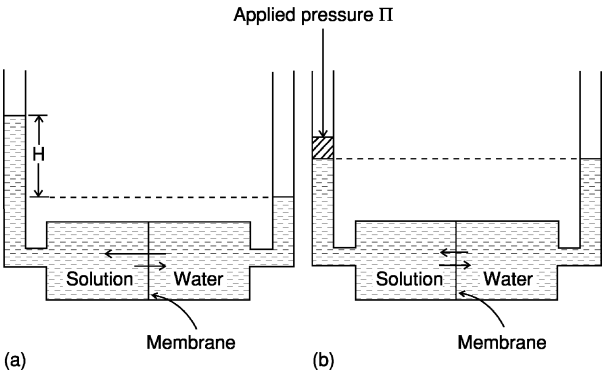


Figure 6 Osmosis and osmotic pressure: (a) in osmosis, the flow of water molecules through the membrane into the solutions is at first greater than the reverse flow from the solution into the water compartment. The hydrostatic pressure due to the column of expanded solution increases the rate of water flow from the solution to the water compartment until, at equilibrium, the opposite flows are equal. (b) The osmotic pressure of the solution is equal to the hydrostatic pressure Π, which must be applied to the solution to equalize the rate of flow to and from the solution and produce a net flow of zero. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

Figure 6 is a schematic representation of a pure solvent separated from a solution by a semipermeable membrane. Solvent will pass through the membrane and enter the solution compartment, driving the solution level up the left-hand tube until the hydrostatic pressure of the column of dilute solution on the left is sufficient to counter the diffusion pressure of the solvent molecules drawn into the solution through the membrane. The hydrostatic pressure at equilibrium, when solvent molecules are crossing the membranes in both directions at equal rates, is the osmotic pressure of the solution.

In dilute solutions, the osmotic pressure is proportional to the concentration of the solution and to its temperature according to the following equation:

$$\Pi = MRT \qquad [5]$$

Here Π is the osmotic pressure in atmospheres (to be multiplied by 0.101 to obtain megapascal units), M the total molar concentration of the solute (whether molecules or dissociated ions), T the temperature in degrees Kelvin, and R the gas constant (0.08205 liter atm deg^{-1} mole). The osmotic pressure increase with temperature is associated with the corresponding increase of the molecular diffusivity (self-diffusion coefficient) of water, D_w. According to the Einstein–Stokes equation,

$$D_w = kT/6\pi\eta$$

where $k = R/N$, the Boltzmann constant (1.38×10^{-23} JK^{-1}); r is the rotation radius of the molecule ($\sim 1.5 \times 10^{-4}$) and η is the viscosity.

Solubility of Gases

The concentration of dissolved gases in water in equilibrium with a gaseous phase generally increases with pressure and decreases with temperature. According to Henry's law, the mass concentration of a dissolved gas c_m is proportional to the partial pressure of the gas p_i:

$$c_m = s_o p_i / p_o \quad [6]$$

where s_o is the solubility coefficient of the particular gas in water and p_o is the total pressure of the atmosphere. The volume concentration is similarly proportional:

$$c_v = s_v p_i / p_o \quad [7]$$

where s_v is the solubility coefficient expressed in terms of volume ratios (i.e., c_v is the volume of dissolved gas relative to the volume of water). The values of s_c and s_v are determined experimentally. If the gas does not react chemically with the liquid, these properties should remain constant over a range of pressures, especially at low partial pressures of the dissolved gases. Solubility is, however, strongly influenced by temperature. Table 2 gives the s_v values of several atmospheric gases at various temperatures.

The solubilities of various gases (particularly oxygen) in varying conditions strongly influence such vital soil processes as oxidation and reduction, and respiration by roots and microorganisms.

Adsorption of Water on Solid Surfaces

Adsorption is an interfacial phenomenon resulting from the differential forces of attraction or repulsion occurring among molecules or ions of different phases at their exposed surfaces. As a result of cohesive and adhesive forces coming into play, the zones of contact among phases may exhibit a concentration or a density of material different from that inside the phases themselves. As different phases come in contact, various types of adsorption can occur: adsorption of gases on solids, of gases on liquid surfaces, and of liquids on solids.

Table 2 Solubility coefficients of gases in water

Temperature (°C)	*Nitrogen (N_2)*	*Oxygen (O_2)*	*Carbon dioxide (CO_2)*	*Air (without CO_2)*
0	0.0235	0.0489	1.713	0.0292
10	0.0186	0.0380	1.194	0.0228
20	0.0154	0.0310	0.878	0.0187
30	0.0134	0.0261	0.665	0.0156
40	0.0118	0.0231	0.530	—

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

The interfacial forces of attraction or repulsion may themselves be of different types, including electrostatic or ionic (Coulombic) forces, intermolecular forces such as van der Waals and London forces, and short-range repulsive (Born) forces. The adsorption of water upon solid surfaces is generally of an electrostatic nature. The polar water molecules attach to the charged faces of the solids and to the ions adsorbed on them. This adsorption of water is the mechanism causing the strong retention of water by clay at high suctions.

The interaction of the charges of the solid with the polar water molecules may impart to the adsorbed water a distinct structure in which the water dipoles assume an orientation dictated by the charge sites on the solids. This adsorbed 'phase' may have mechanical properties of strength and viscosity that differ from those of ordinary liquid water at the same temperature. The adsorption of water on clay surfaces is an exothermic process, resulting in the liberation of an amount of heat known as the heat of wetting.

Vapor Pressure

According to the kinetic theory, molecules in a liquid are in constant motion, which is an expression of their thermal energy. These molecules collide frequently, and occasionally one or another at the surface absorbs sufficient momentum to leap out of the liquid and into the atmosphere above it. Such a molecule, by virtue of its kinetic energy, thus changes from the liquid to the gaseous phase. This kinetic energy is then lost in overcoming the potential energy of intermolecular attraction while escaping from the liquid. At the same time, some of the randomly moving molecules in the gaseous phase may strike the surface of the liquid and be absorbed in it.

The relative rates of these two directions of movement depend upon the concentration of vapor in the atmosphere relative to its concentration at a state of equilibrium (i.e., when the movement in both directions is equal). An atmosphere that is at equilibrium with a body of pure water at standard atmospheric pressure is considered to be saturated with water vapor, and the partial pressure of the vapor in such an atmosphere is called the saturation (or equilibrium) vapor pressure. The vapor pressure at equilibrium with any body of water depends on the physical condition of the water (pressure and temperature) and its chemical condition (solutes) but is independent of the absolute or relative quantity of liquid or gas in the system.

The saturation vapor pressure rises with temperature. As the kinetic energy of the molecules in the liquid increases, so does the evaporation rate.

Consequently, a higher concentration of vapor in the atmosphere is required for the rate of return to the liquid to match the rate of escape from it. A liquid arrives at its boiling point when the vapor pressure becomes equal to the atmospheric pressure. If the temperature range is not too wide, the dependence of saturation vapor pressure on temperature is expressible by the equation (Table 3):

$$\ln p_o = a - b/T \qquad [8]$$

where $\ln p_o$ is the logarithm to the base e of the saturation vapor pressure p_o, T is the absolute temperature, and a and b are constants.

As mentioned, the vapor pressure also depends on the hydrostatic pressure of the liquid water. At equilibrium with drops of water (which have a hydrostatic pressure greater than atmospheric), the vapor pressure is greater than in a state of equilibrium with free water (which has a flat interface with the atmosphere). On the other hand, in equilibrium with adsorbed or capillary water under a hydrostatic pressure smaller than atmospheric, the vapor pressure is smaller than in equilibrium with free water. The curvature of drops is considered to be positive, as these drops are convex toward the atmosphere, whereas the curvature of capillary water menisci is considered negative, as they are concave toward the atmosphere.

For water in capillaries, in which the air–water interface is concave, the Kelvin equation applies:

$$-(\mu_1 - \mu_1^o) = RT \ln (p_1^o/p_1) = 2\gamma v_i \cos \alpha / r_c$$

in which $(\mu_1 - \mu_1^o)$ is the change in potential of the water due to the curvature of the air–water interface, γ is the surface tension of water, α is the contact angle, v_1 is the partial molar volume of water, and r_c is the radius of the capillary.

Water present in the soil invariably contains solutes, mainly electrolytic salts, in variable concentrations. Thus, soil water should properly be called the soil solution. The composition and concentration of the soil solution affect soil behavior. While in humid regions the soil solution may have a concentration of but a few parts per million, in arid regions the concentration may become as high as several percent. The ions commonly present are H^+, Ca^{2+}, Mg^{2+}, Na^+, K^+, NH_4^+, OH^-, Cl^-, HCO_3^-, NO_3^-, SO_4^-, and CO_3^-. The vapor pressure of electrolytic solutions is less than that of pure water. The equation is:

$$v_1 \Pi_o = RT \ln (p_1^o/p_1) = \mu_1 - \mu_1^o$$

wherein Π_o is the osmotic pressure of a nonvolatile solute, μ_1^o and p_1^o are the chemical potential and vapor pressure of the liquid in its pure state, and μ_1 and p_1 are the same for the solution. Thus the soil solution has a lower vapor pressure than pure water, even when the soil is saturated. In unsaturated soil the capillary and adsorptive effects further lower the potential and hence also the vapor pressure.

Surface Tension

Surface tension is a phenomenon occurring typically, but not exclusively, at the interface of a liquid and a gas. The liquid behaves as if it were covered by an elastic membrane in a constant state of tension that tends to cause the surface to contract. To be sure, no such membrane exists, yet the analogy is a useful one if not taken too literally. If we draw an arbitrary line of length L on a liquid surface, there will be a force F

Table 3 Physical properties of water vapor

	Saturation vapor pressure (torr)		*Vapor density in saturated air (kg m^{-3})*		
Temperature (°C)	*Over liquid*	*Over ice*	*Over liquid ($\times 10^{-3}$)*	*Over ice ($\times 10^{-3}$)*	*Diffusion coefficient (m^2 s^{-1}) ($\times 10^{-4}$)*
−10	2.15	1.95	2.36	2.14	0.211
−5	3.16	3.01	3.41	3.25	—
0	4.58	4.58	4.85	4.85	0.226
5	6.53	—	6.80	—	
10	9.20	—	9.40	—	0.241
15	12.78	—	12.85	—	
20	17.52	—	17.30	—	0.257
25	23.75	—	23.05	—	
30	31.82	—	30.38	—	0.273
35	42.20	—	39.63	—	
40	55.30	—	51.1	—	0.289
45	71.90	—	65.6	—	
50	92.50	—	83.2	—	

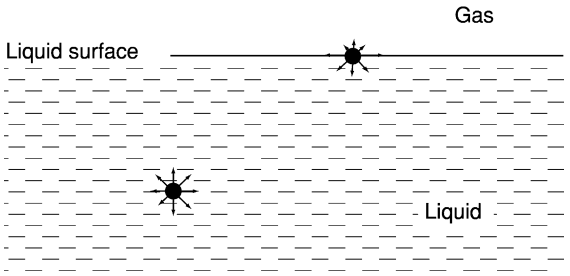


Figure 7 Cohesive forces acting on a molecule inside the liquid and at its surface.

pulling the surface to the right of the line and an equal force pulling the surface leftwards. The ratio F/L is the surface tension and its dimensions are those of force per unit length. The same phenomenon can also be described in terms of energy. Increasing the surface area of a liquid requires work, which remains stored as potential energy in the enlarged surface, just as energy can be stored in a stretched spring. That potential energy can perform work if the enlarged surface is allowed to contract again. Energy per unit area has the same dimensions as force per unit length.

An explanation for occurrence of surface tension is given in Figure 7. A molecule inside the liquid is attracted in all directions equally by the cohesive forces of neighboring molecules, while a molecule at the surface of the liquid is attracted into the relatively dense liquid phase by a net force greater than that attracting it toward the rarified gaseous phase. This unbalanced force draws the surface molecules inward into the liquid and results in the tendency for the surface to contract. This is why drops of a liquid in air as well as bubbles of air in a liquid assume the shape of a sphere, which is a body of minimal surface exposure relative to its volume.

Different liquids exhibit different surface tension values, as in the following list:

Water, 7.27×10^{-2} N m^{-1}(72.7 dyn cm^{-1}) at 20°C;
Ethyl ether, 1.7×10^{-2} N m^{-1} (17 dyn cm^{-1});
Ethyl alcohol, 2.2×10^{-2} N m^{-1}(22 dyn cm^{-1});
Benzene, 2.9×10^{-2} N m^{-1}(29 dyn cm^{-1});
Mercury, 0.43 N m^{-1} (430 dyn cm^{-1}).

Curvature of Water Surfaces and Hydrostatic Pressure

Wherever an interface between fluids (say, between water and air) is not planar but curved, the resolution of forces due to surface tension creates a pressure differential across that interface. For a spherical interface (as in the case of a bubble of air immersed in a body of water), the pressure difference is proportional

BOX 1 A Farewell to Teardrops

For centuries, conventional wisdom held that larger stones fall faster than small ones, that the Earth is flat yet the sun revolves around it, and that the sun and moon are of equal size. We would like to believe that in our time all baseless notions have been replaced by sound science. But have they?

Consider the shape of a raindrop. The conventional standard is a teardrop, rounded at the bottom and tapering to a point at the top. So prevalent is that image that it is printed in textbooks and used as a logo by irrigation companies and even by conferences sponsored by the United Nations. Alas, the vertically elongated, top-pointed raindrop is a physical impossibility.

A drop forming at the tip of a spout assumes that shape only at the very instant of detachment. Because of surface tension, any drop suspended in air 'balls up' spontaneously into a sphere. In a cloud, spherical droplets tend to grow by condensation or coalescence until reaching a critical weight, at which they begin to fall.

Air bypassing a falling drop acquires greater velocity around the curved 'waist' of the drop than near its top or bottom. By Bernoulli's law ($p + \rho v^2/2 = \text{constant}$), the pressure of the air alongside the drop is lowered relative to that of the air above and below the drop. Consequently the drop compresses vertically and comes to resemble an ellipsoid. Such is indeed the shape of small drops (< 2 mm).

In the case of larger drops, the laminar streams of air flowing past the drop may not converge smoothly above the top, but may leave a turbulent wake there. In such a wake, the pressure is lower than it would be in an ideally laminar flow regime. The reduced air pressure at the top causes the drop to bulge a bit there, thus acquiring the appetizing shape of a miniature hamburger bun (described by the appropriately named McDonald as early as 1954).

If a drop were to continue accelerating, it might eventually spread out to form a pancake. But Stokes' law intervenes, decreeing that the drop's acceleration be countered by the increasing viscous resistance of the air. When air resistance equals the gravitational force, acceleration ceases and the drop continues to fall at a constant 'terminal' velocity and with a more or less constant shape. It finally slaps the ground with its flattened face going 'plop'.

Some of us may think it unfair that such an exquisitely sculpted natural body should have no more glorious a fate than to splatter down on some dry bit of earth. But that is where – having at once lost its distinctive shape as it enters the labyrinthine interstices of the soil – our drop might well give life to a thirsty plant, perhaps even to a sunflower. Anyhow, we bid farewell to the sad countenance of the teardrop, a singularly inappropriate symbol for happy rain.

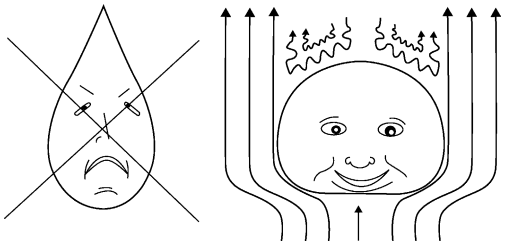


Figure B1 Conventional and real raindrops. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

to the surface tension and inversely proportional to the curvature:

$$\Delta P = 2\gamma/R \qquad [9]$$

Thus, the smaller the bubble is, the greater is its pressure.

If the bubble is not spherical, then instead of eqn [9] we obtain:

$$\Delta P = \gamma(1/R_1 + 1/R_2) \qquad [10]$$

where R_1 and R_2 are the principal radii of curvature for any given point on the interface. Eqn [10] reduces to [9] whenever $R_1 = R_2$.

Contact Angle of Water on Solid Surfaces

If we place a drop of liquid upon a dry solid surface, the liquid will usually displace the gas that covered the surface of the solid and it will spread over that surface to a certain extent. Where its spreading ceases and the edge of the drop comes to rest, it will form a typical angle with the surface of the solid. This angle, termed contact angle, is illustrated in **Figure 8**.

We now consider what factors determine the magnitude of the angle α. We can expect that angle to be acute if the adhesive affinity between the solid and liquid is strong relative to the cohesive forces inside the liquid itself and to the affinity between the gas and the solid. We can then say that the liquid 'wets' the solid. A contact angle of zero would mean the complete flattening of the drop and the perfect wetting of the solid surface by the liquid. On the other hand, a contact angle of 180° would imply a complete nonwetting or rejection of the liquid by the gas-covered solid. In that case the drop would retain its spherical shape without spreading over the surface at all (assuming no gravity effect). Surfaces on which water exhibits an obtuse contact angle are called water-repellent, or hydrophobic (Greek: 'water-hating').

The contact angle of a given liquid on a given solid is generally characteristic of its interaction under given physical conditions. This angle, however, may be different in the case of a liquid that is advancing over the solid surface than in the case of the same liquid receding over the surface. This phenomenon, where it occurs, is called contact angle hysteresis. The wetting angle of pure water upon clean and smooth mineral surfaces is generally zero, but where the surface is rough or coated with adsorbed surfactants of a hydrophobic nature, the contact angle, and especially the wetting angle, can be considerably greater than zero. This is illustrated in **Figure 9**.

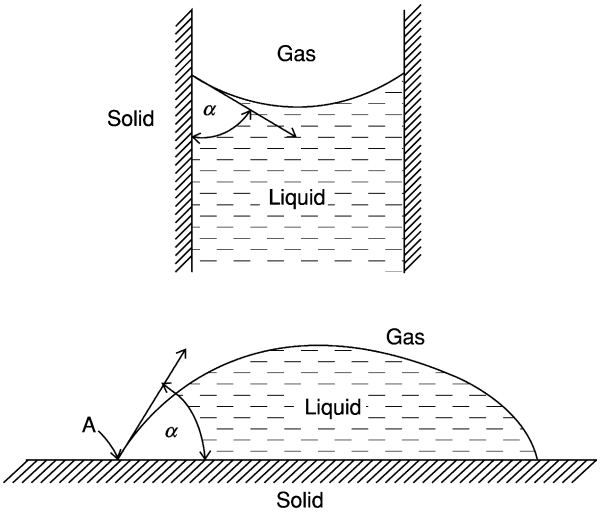


Figure 8 The contact angle of a meniscus in a capillary tube and a drop resting on the surface of a solid. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

The Phenomenon of Capillarity

A capillary tube dipped in a body of free water will form a meniscus as the result of the contact angle of water with the walls of the tube. The curvature of this meniscus will be greater (i.e., the radius of curvature smaller) the narrower the tube. The occurrence of curvature causes a pressure difference to develop across the liquid–gas interface. A liquid with an acute contact angle (e.g., water on glass) will form a concave meniscus, and therefore the liquid pressure under the meniscus (P_1) will be smaller than the atmospheric pressure (**Figure 10**). Hence, the water inside the tube will be driven up the tube from its initial location (shown as a dashed curve in **Figure 10**) by the greater pressure of the free water (i.e., water at atmospheric pressure, under a horizontal air–water interface) outside the tube at the same level. The

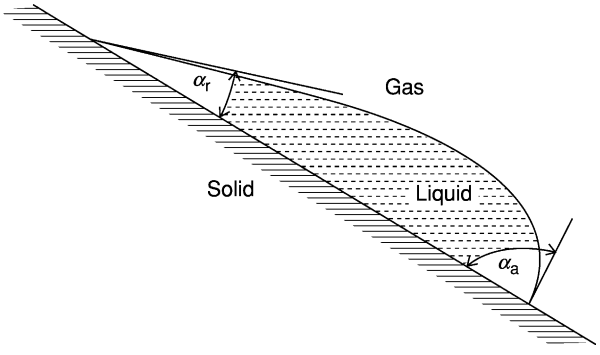


Figure 9 Hypothetical representation of a drop resting on an inclined surface. The contact angle α_a, at the advancing edge of the drop, is shown to be greater than the corresponding angle α_r at the receding edge. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

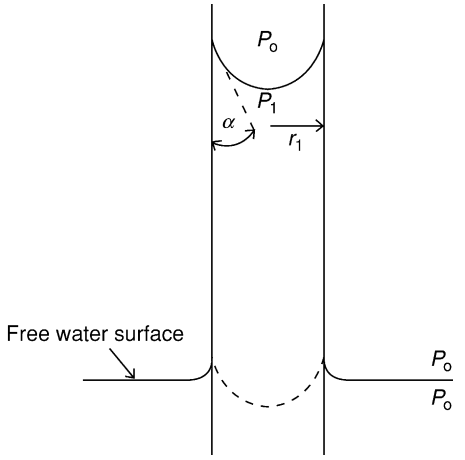


Figure 10 Capillary rise. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

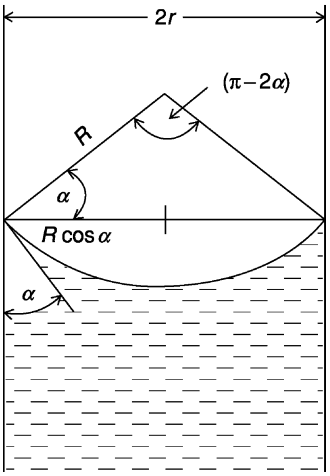


Figure 11 The geometric relationship of the radius of curvature R to the radius of the capillary r and the contact angle α. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

upward movement will stop when the pressure difference between the water inside the tube and the water under the flat surface outside the tube is countered by the hydrostatic pressure exerted by the water column in the capillary tube.

If the capillary tube is cylindrical and if the contact angle of the liquid on the walls of the tube is zero, the meniscus will be a hemisphere (and in a two-dimensional drawing can be represented as a semicircle) with its radius of curvature equal to the radius of the capillary tube. If, on the other hand, the liquid contacts the tube at an angle greater than zero but smaller than 90°, then the diameter of the tube ($2r$) is the length of a cord cutting a section of a circle with an angle of $\pi - 2\alpha$, as shown in Figure 11. Thus,

$$R = r/\cos\alpha \qquad [12]$$

where R is the radius of curvature of the meniscus, r the radius of the capillary, and α the contact angle.

The pressure difference ΔP between the capillary water (under the meniscus) and the atmosphere, therefore, is:

$$\Delta P = (2\gamma \cos\alpha)/r \qquad [13]$$

Recalling that hydrostatic pressure is proportional to the depth d below the free water surface (i.e., $P = \rho g d$, where ρ is liquid density and g is gravitational acceleration) we can infer that hydrostatic tension (negative pressure) in a capillary tube is proportional to the height h above the free water surface. Hence the height of capillary rise is:

$$h_c = (2\gamma \cos\alpha)/g(\rho_1 - \rho_o)r \qquad [14]$$

where ρ_g is the density of the gas (which is generally neglected), ρ_1 the density of the liquid, g the acceleration of gravity, r the capillary radius, α the contact angle, and γ the surface tension between the liquid and the air.

When the liquid surface is concave, the center of curvature lies outside the liquid and the curvature, by convention, is regarded as negative. Thus, for a concave meniscus such as that of water in a clean glass capillary, ΔP is negative with reference to the atmosphere, indicating a capillary pressure deficit, or subpressure, called tension. On the other hand, in a capillary tube that forms a convex meniscus (such as that of mercury in glass, or of water in an only or otherwise water-repellent tube), ΔP is positive and capillary depression, rather than capillary rise, will result.

Density and Compressibility

The open packing of water molecules in ice and liquid water accounts for their relatively low densities. Unlike most substances, water exhibits a point of maximum density (at 4°C), below which the substance expands due to the formation of a hexagonal lattice structure, and above which the expansion is due to the increasing thermal motion of the molecules. The coefficient of thermal expansion of water is rather low, and in the normal temperature range of, say, 4–50°C, the density diminishes only slightly, from 1000 to 988 kg m^{-3}.

The compressibility of water, C_w, can be defined as the relative change in density with change in pressure:

$$C_w = (1/\rho_w)(d\rho_w/dP) \qquad [15]$$

At 20°C and at atmospheric pressure, the compressibility of pure water is about 4.6×10^{-10} m^2 N^{-1}. In

Table 4 Physical properties of liquid water

Temperature (°C)	Density(kg m^{-3}) ($\times 10^3$)	Specific heat (J kg^{-1}deg) ($\times 10^3$)	Latent heat (vaporization) (J kg^{-1}) ($\times 10^6$)	Surface tension (kg s^{-2}) ($\times 10^{-2}$)	Thermal conductivity (J m^{-6} s deg)	Dynamic viscosity (kg m^{-1} s) ($\times 10^{-2}$)	Kinematic viscosity (m^2 s^{-1}) ($\times 10^{-6}$)
−10	0.99794	4.27	2.53	—	—	—	—
−5	0.99918	4.23	2.51	7.64	—	—	—
0	0.99987	4.22	2.50	7.56	0.561	0.1787	1.79
4	1.00000	4.21	2.49	7.5	0.570	0.1567	1.57
5	0.99999	4.207	2.49	7.48	0.574	0.1519	1.52
10	0.99973	4.194	2.48	7.42	0.587	0.1307	1.31
15	0.99913	4.19	2.47	7.34	0.595	0.1139	1.14
20	0.99823	4.186	2.46	7.27	0.603	0.1002	1.007
25	0.99708	4.18	2.44	7.19	0.612	0.0890	0.897
30	0.99568	4.18	2.43	7.11	0.620	0.0798	0.804
35	0.99406	4.18	2.42	7.03	0.629	0.0719	0.733
40	0.99225	4.18	2.41	6.95	0.633	0.0633	0.661
45	0.99024	4.18	2.40	6.87	0.641	0.0596	0.609
50	0.98807	4.186	2.38	6.79	0.645	0.0547	0.556

the normal situations encountered on the surface of the Earth, water can usually be considered incompressible. The compression of water cannot be ignored, however, in the case of deep confined aquifers, which may be subject to a pressure of, say, 10 MPa or more.

Dynamic and Kinematic Viscosity

When a fluid is moved in shear (i.e., when adjacent layers of fluid are made to slide over each other), the force required is proportional to the velocity of shear. The proportionality factor is called the viscosity. As such, it is the property of the fluid to resist the rate of shearing and this can be visualized as an internal friction. The coefficient of viscosity η is defined as the force per unit area necessary to maintain a velocity difference of 1 m s^{-1} between two parallel layers of fluid which are 1 m apart. The viscosity equation is:

$$\tau = F_s/A = \eta \, \mathrm{du/dx} \qquad [16]$$

wherein τ is the shearing stress, consisting of a force F_s acting on an area A; η (dimensions: mass/(length $\times$ time)) is the coefficient of dynamic viscosity; and du/dx is the velocity gradient perpendicular to the stressed area A.

The ratio of the dynamic viscosity of a fluid to its density is called the kinematic viscosity, designated ν. It expresses the shearing-rate resistance of a fluid independently of the density. Thus, while the dynamic viscosity of water exceeds that of air by a factor of about 50 (at room temperature), its kinematic viscosity is actually lower.

Fluids of lower viscosity flow more readily and are said to possess greater fluidity (which is the reciprocal of viscosity). As shown in Table 4, the viscosity of water diminishes by over 2% per 1°C rise in temperature, and thus decreases by more than half as the temperature increases from 5 to 35°C. The viscosity is also affected by the type and concentration of solutes present.

See also: **Capillarity**; **Hydrodynamics in Soils**; **Water Cycle**

Further Reading

Acheson DJ (1990) *Elementary Fluid Dynamics*. Oxford, UK: Clarendon.

Gleick PH (ed.) (1993) *Water in Crisis: A Guide to the World's Freshwater Resources*. New York: Oxford University Press.

Hillel D (1998) *Environmental Soil Physics*. Boston: Academic Press.

Kandel R (2003) *Water from Heaven*. New York: Columbia University Press.

Kramer PJ and Boyer JS (1995) *Water Relations of Plants and Soils*. San Diego: Academic Press.

WATER-REPELLENT SOILS

J Letey, University of California–Riverside, Riverside, CA, USA

Dry soil that does not spontaneously imbibe water, but rather has water remain on the surface, is referred to as a water-repellent or hydrophobic soil. Soils which spontaneously imbibe water are considered to be wettable but they can have different degrees of wettability.

The most common observations made on water-repellent soils are: (1) the water resists penetration and 'rolls' off the soil, and (2) very wet soil can be found next to very dry soil even though both receive equal water from rain or irrigation.

Water repellency has been reported in most continents of the world for varying land uses and climatic conditions and has been the topic of two international conferences. The first conference was held at Riverside, California, in 1968, and the second was convened 30 years later at Wageningen, the Netherlands.

The theory and description of soil-water processes described in soil physics textbooks are valid for wettable soils. That theory does not apply to water-repellent soils. Techniques that quantify the degree of soil water repellency are important for characterizing and mitigating such soil conditions.

Characterizing the Degree of Water Repellency

The extent of water repellency of a flat, smooth surface is commonly determined by measuring the water–solid contact angle when a drop of water is placed on the surface. Water will form a 'ball' on a very repellent surface. Soils do not have surfaces that conveniently allow the geometric measurement of a contact angle. Thus, an alternative to geometric measurement is required. Soils have pores and occasionally have been represented as being composed of a bundle of capillary tubes. The capillary tube is a vast oversimplification of the complex geometric arrangements of soil pores. Nevertheless, helpful insights can be achieved by assuming the capillary tube model for soils.

If the tip of a capillary tube is brought in contact with water, the water will be drawn into the tube. The height that the water will rise is given by eqn [1].

$$h = 2\gamma_w \cos\theta / r\rho g \qquad [1]$$

where h is the height of rise, γ_w the water–air surface tension, θ the water–solid contact angle, r the capillary radius, ρ the water density, and g the gravitational constant. According to eqn [1], water will not spontaneously enter the soil if cos θ is zero or a negative number ($\theta \geq 90^\circ$).

A soil is commonly classified as being water-repellent if a drop of water placed on the soil does not spontaneously enter the soil. By this convention, a water-repellent soil is one in which the water–solid contact angle is equal to or greater than 90°. The water-drop penetration time (WDPT) is an index commonly used to specify the degree of water repellency. This procedure involves placing a drop of water on the soil and measuring the time for it to penetrate. Since water only penetrates the soil if $\theta < 90^\circ$, WDPT is the measure of the time required for θ to change from its original value, which was more than 90°, to a value approaching 90°. Therefore, it is a measure of the stability of the repellency when the soil is brought in contact with water, and not necessarily an index of θ. The fact that the degree of repellency is not static but can change with time after contact with water complicates the analysis of temporal effects of water repellency on water flow.

Liquid surface tension that wets the soil material with a 90° contact angle is one index of water repellency. A series of aqueous ethanol solutions producing a range of surface tensions is prepared for this measurement. Drops of these solutions are placed on the soil. The higher-surface-tension solutions set on the surface and the lower-surface-tension solutions will spontaneously penetrate the soil. The 90° surface tension (γ_{nd}) is the surface tension of the solution where there is transition from sitting on the surface to penetration. Sometimes this procedure is followed but the molarity of the aqueous ethanol solution rather than the surface tension is reported as the index, referred to as the molarity of ethanol (MED) test. Others have reported the results of this test in terms of the volumetric ethanol percentage when penetration is initiated. The γ_{nd} has at times been referred to as the critical surface tension.

The solid–air surface tension, γ_s, is a fundamental physical–chemical property of a solid that affects its wetting properties. Therefore, characterizing the magnitude of water repellency by measuring the solid–air surface tension is valuable. Theoretical relationships between various surface tensions and contact angle have been combined to obtain the following two equations:

$$\gamma_s = \gamma_{nd}/4 \quad [2]$$

$$\cos\theta = [(\gamma_{nd}/\gamma_w)^{1/2} - 1] \quad [3]$$

Therefore the measurement of γ_{nd} can be used to calculate the value of γ_s and θ. The values of γ_s and θ determined by measuring γ_{nd} represent initial values before they have had an opportunity to change after contact with water. Both WDPT and γ_{nd} can be measured quickly and easily in either the laboratory or field. Measurement of γ_{nd} provides an index of the initial extent of water repellency, and WDPT provides an index of the stability of repellency after contact with water.

Water-Entry Pressure Head

When $\theta > 90°$, pressure must be applied to force the water into the soil. The water-entry pressure head, h_p, into a capillary tube is:

$$h_p = -2\gamma_w \cos\theta / r\rho g \quad [4]$$

For $\theta < 90°$, h_p is negative and water is drawn spontaneously into the tube. For $\theta > 90°$, h_p is positive and the water head equal to or greater than h_p must exist for water entry. Note that h_p depends on both the water-repellent index (θ) and the pore size (r). Measurement of h_p requires an apparatus that allows the depth of water on a soil column to increase gradually and some means of determining at what point the water penetrates the soil column.

Infiltration Rate

One of the most important effects of soil water repellency on the hydrologic cycle is its effect on infiltration rate. Infiltration will not occur until water has been in contact with the soil surface for a time equal to or greater than the WDPT unless the depth of ponding (h_o) is greater than the water-entry pressure (h_p). Even after conditions allowing water penetration, the infiltration and hydraulic property characteristics of water-repellent soils differ from wettable soils. Wettable soils typically have a high initial infiltration rate which decreases and then becomes relatively constant with increased time of water applied to the surface. In contrast, the infiltration rate into a water-repellent soil is slow during the initial phase of infiltration and increases with time. For soils with a finite WDPT, the degree of repellency changes with time after contact with water, so the increase in infiltration rate with time might be attributed to the changes in the degree of repellency.

In order to isolate and clearly identify the effects of water repellency that are not time-dependent on infiltration behavior, studies have been conducted on soils that have an infinite WDPT (water never penetrates the soil). Because natural soils do not typically have infinite WDPT values, soil materials artificially treated with octadecylamine provide a stable water repellency with an infinite WDPT. Different concentrations of octadecylamine create sands with different degrees of water repellency. Studies have been conducted using such treated sands.

Whereas the infiltration rate into wettable soils is only slightly affected by the depth of ponded water, infiltration into the water-repellent materials is drastically affected by the depth of ponded water (**Figure 1**). The h_p value for this material is 8.4 cm. As expected, no water infiltrates unless h_o is greater than h_p. At low values of h_o, the infiltration rate increases with time. However, for the higher values of h_o, infiltration rate decreases with increasing time, which is typical of wettable soils. Intermediate values of h_o ($h_o/h_p \approx 2.6$) produce a nearly constant infiltration rate as a function of time.

The hydraulic conductivity of the treated sand can be measured by ponding water on the surface of a column and measuring the steady-state water flow through the column. Increasing the depth of water ponding on the surface increases the hydraulic head gradient and therefore should induce an increased water flow. However, the hydraulic conductivity should be unaffected by the depth of water ponding, and this result is typically observed for wettable soils. In contrast, the hydraulic conductivity of the treated sands increases with depth of ponding until a critical depth of ponding is reached, after which the hydraulic conductivity becomes constant (**Figure 2**). The hydraulic conductivity in the water-repellent sand attains a value of the untreated sand when h_o is high enough. The ratio of h_o/h_p that results in maximum K equivalent to the untreated K is approximately 3.1 for each treatment.

Researchers have found that the average water content in the water-repellent sand decreases as the value of h_o decreases. Thus the decreasing K is caused by a decrease in water content, which is consistent with the well-recognized transport phenomenon that hydraulic conductivity is a function of soil water content. The water-entry pressure increases as the pore size decreases (eqn [4]); therefore the increase in water content could be the result of smaller pores being filled with water as h_o increases. Alternatively, finger flow through water-repellent sand could have caused only a portion of the sand to be wet. Whether the decrease in average water content is uniformly distributed in the sand or the result of finger flows

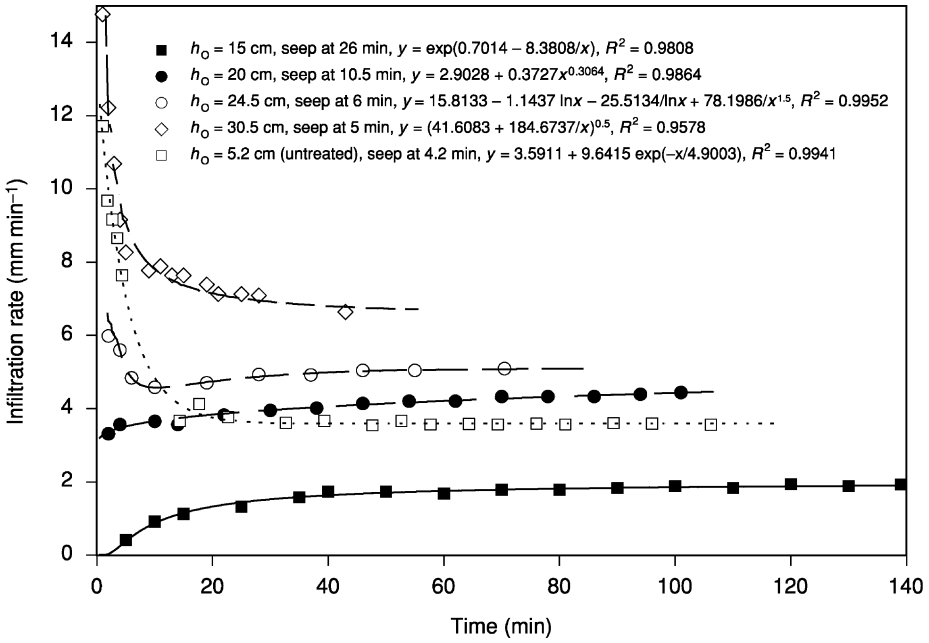


Figure 1 The infiltration rate as a function of time for sand ($h_p = 8.4$ cm) for different values of ponded-water depth (h_o). The equations are for the best-fit curve to the data points. (Reproduced with permission from Feng GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.)

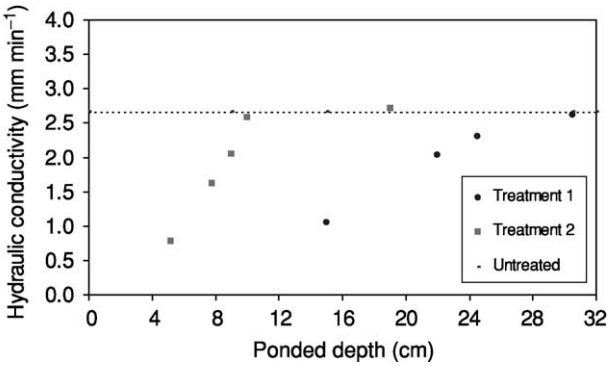


Figure 2 The relationship between hydraulic conductivity and ponded-water depth for sands receiving two treatments to make them water-repellent. The dotted line is the hydraulic conductivity of the untreated sand, which is independent of ponded depth. (Reproduced with permission from Feng GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.)

where only a fraction of the sand is wet has not been determined conclusively.

The h_o/h_p value at which there is a transition from infiltration rate increasing with time to decreasing with time (Figure 1) is close to the h_o/h_p value that produces a hydraulic conductivity close to that of untreated sand.

The effect on infiltration rate of placing a layer of wettable, untreated sand over the water-repellent sand has been studied. Infiltration rate is rapid into the overlying wettable sand, but stops when reaching the repellent layer until the hydraulic head at the interface exceeds the value of h_p. Thereafter, infiltration into the repellent layer is initiated. The consequence of having the untreated layer is approximately equivalent to increasing the ponded depth. In other words, ponding water 1 cm above a 5-cm layer of wettable sand is equivalent to ponding 6 cm of water on a soil that is water-repellent at the surface.

The hydraulic heads immediately behind the wetting front differ between wettable and water-repellent soils. In a wettable soil, the hydraulic head is negative immediately behind the wetting front and increases as

it moves upward from the wetting front. Conversely, the hydraulic head immediately behind the wetting front is positive in a water-repellent soil and decreases away from the wetting front. These are conditions that contribute to unstable or preferential flow in the soil.

The term 'preferential flow' has been used to describe nonuniform water flow or wetting-front advance through the soil. However, the term 'preferential flow' can have different connotations. Ambiguity arises because of different dimensional scales, ranging from pore size to several centimeters. Indeed, wet and dry zones in a water-repellent field may have dimensions in meters where the term 'preferential flow' might not be appropriate.

Instability of wetting-front advance has been observed in the laboratory on water-repellent soil materials. Instability of a dynamic wetting front is defined as the unconstrained growth of randomly occurring, small perturbations of the wetting front. The effect is to have a wetting-front advance rapidly in random zones, creating a 'finger' wetting pattern.

The depth of ponded water h_o has a profound effect on the extent of finger flow in a water-repellent soil. If the water-repellent layer is at a depth L below the soil surface, the effect of L is comparable to the effect of h_o on finger formation. Water penetrates the wettable surface layer but is constrained by the water-repellent sublayer, which allows pressure to build up in the water to a value comparable to the value that would exist without the wettable layer. One study has found that no water penetrates the water-repellent layer for values of $(h_o + L)/h_p < 1$; unstable flow develops for the values between 1 and 1.5; and a stable wetting front develops for values greater than 1.5.

The basic concepts of flux through a water-repellent layer as affected by the total head may explain the finger formation. Instability of water flow occurs if the flow through the system is less than the saturated hydraulic conductivity. The fact that finger flow and reduced hydraulic conductivity (**Figure 2**) are both associated with low values of h_o is consistent with the theory that instability of water flow occurs when the flow through the system is less than the saturated hydraulic conductivity.

Temporal effects of infiltration rate and preferential flow are influenced by the WDPT value. Infiltration will occur when a WDPT has been reached even without a ponding depth. However, one study has found that the wetting front is unconditionally unstable when h_o is less than h_p, resulting in finger flow. Conversely, the flow has been observed to be stable when h_o is greater than h_p if WDPT is not too large. Finger flow does not occur for a soil that has a WDPT of 1 min. When the WDPT is 10 min, fingers form; however, the fingers broaden and converge after continued flow and an almost uniform wetting front eventually develops.

Uneven wetting patterns, whereby zones of very dry soil can exist next to zones of very wet soils, are commonly observed in the field on water-repellent soils. Location of wet and dry zones usually cannot be determined by visual observations of the landscape. The phenomenon, however, can be triggered by any factor that initiates infiltration. Once infiltration occurs, the infiltration rate increases with time, drawing water away from adjacent zones where infiltration has not been initiated. The 'trigger' may be associated with special distribution of WDPT values, whereby the water starts to infiltrate in those zones with a lower WDPT than an adjacent zone with a higher WDPT. Also the value of h_p is associated not only with differences in repellency, but also with pore-size distribution. Possibly, slight differences in texture or packing that lead to different pore-size distributions could contribute to the variable wetting. Also slight differences in topography that allow slightly more depth of water ponding on one zone than the other could initiate the uneven infiltration.

Wildfires have been found either to create or enhance the severity of water repellency on watersheds. Increases in streamflow, quickflow volumes, and sediment in bedload yields have been observed on fire-induced water-repellent soils. Fire-induced water repellency produces localized runoff and sediment movement on hillsides, but does not appreciably affect the performance of a whole watershed. Because of the effects of cracks, animals burrows, root channels, etc., it is difficult to characterize the effects of water repellency at the watershed scale.

Mitigating Water Repellency

Surfactants can be used to mitigate the effects of soil water repellency. Adding a surfactant to water lowers the surface tension of the water. Reduction of the liquid surface tension also reduces the liquid–solid contact angle, which makes the soil behave as if it were wettable. However, if a soil is wettable, reducing the surface tension of water may actually decrease the infiltration rate. Surfactant molecules are adsorbed by the soil as the solution moves downward. As the surfactant is adsorbed from the solution, surface tension of the solution increases and its ability to wet diminishes. Therefore, surfactants are most effective in treating soil water repellency when it is associated with the surface layer of the soil.

Adsorption of the surfactant molecule converts a water-repellent soil to a wettable soil upon drying.

This phenomenon has been referred to as the 'rewet properties of a surfactant.'

Surfactants have been extensively used to overcome water repellency in turf. Water repellency is a common phenomenon in highly managed golf course soils. Water repellency has been managed through the application of surfactants to areas of turf exhibiting visual symptoms commonly referred to as localized dry spots.

Although there are theoretical relationships between liquid surface tension and soil wetting that should allow the computation of the optimal concentration of surfactant to use for treatment, the adsorption of surfactant by the soil (which alters the liquid surface tension) makes a theoretical basis for describing surfactant treatment virtually impossible. Therefore, developing the most effective surfactant treatment is most commonly done by an empirical approach of applying various treatments and observing the effects.

The application and incorporation of clay into the surface 10-cm layer of a water-repellent sand ameliorates water repellency. The clay, which is hydrophilic, converts the surface layer from a water-repellent to a wettable surface. A wettable layer overlying a water-repellent layer acts as if the depth of water ponding is increased equal to the depth of the wettable layer. Therefore if the water entry pressure is less than 10 cm, providing a 10-cm layer of wettable material by incorporating clay would allow the water to penetrate the underlying water-repellent soil.

One field study has compared soil-wetting patterns on two adjacent plots: one with a water-repellent top layer, and one treated with clay to remove the water repellency in the top 30-cm layer. Dye placed on the soil surface to identify wet and dry zones revealed a mosaic surface wetting of water-repellent soil and a uniform surface wetting of wettable soil. Although water repellency existed in the treated plot below the 30-cm depth, it was uniformly wet. Thus stable flow occurs if the depth to the water-repellent layer is sufficiently deep.

Summary

Naturally occurring water-repellent soils have been reported in many areas of the world. Wildfires either create or enhance the water repellency of watersheds. Processes within the hydrologic cycle such as infiltration and erosion are drastically affected by soil water repellency. Basic water flow principles that have been developed for wettable soils must be modified to be applicable to water-repellent soils. Except under unique conditions where treatment with surfactants is practical, large-scale mitigation of water-repellency is not feasible.

List of Technical Nomenclature

γ_{nd}	Surface tension of solution that wets soil at 90° contact angle
γ_s	Surface tension of solid
γ_w	Surface tension of water
θ	Contact angle (degrees)
h	Height of capillary rise (cm)
h_o	Depth of ponded water (cm)
h_p	Water entry pressure head (cm)
K	Hydraulic conductivity
WDPT	Water-drop penetration time

See also: **Capillarity**; **Infiltration**

Further Reading

Bauters TWJ, DiCarlo DA, Steenhuis TS, and Parlange J-Y (1998) Preferential flow in water-repellent sands. *Soil Science Society of America Journal* 62: 1185–1190.

Cann MA (2000) Clay spreading on water repellent sands in the south east of South Australia – promoting sustainable agriculture. *Journal of Hydrology* 231–232: 333–341.

Carrillo MLK, Letey J, and Yates SR (1999) Measurement of initial soil–water contact angle of water repellent soils. *Soil Science Society of America Journal* 63: 433–436.

Carrillo MLK, Letey J, and Yates SR (2000) Unstable water flow in a layered soil. I. Effects of a stable water-repellent layer. *Soil Science Society of America Journal* 674: 450–455.

DeBano LF (2000) Water repellency in soils: a historical overview. *Journal of Hydrology* 231–232: 4–32.

DeBano LF and Dekker LW (2000) Water repellency bibliography. *Journal of Hydrology* 231–232: 409–432.

Dekker LW and Ritsema CJ (1994) How water moves in a water repellent sandy soil. I. Potential and actual water repellency. *Water Resources Research* 30: 2507–2517.

Dekker LW and Ritsema CJ (2000) Wetting patterns and moisture variability in water repellent Dutch soils. *Journal of Hydrology* 231–232: 148–164.

Feng, GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.

Hendrickx JMH, Dekker LW, and Boersma OH (1993) Unstable wetting fronts in water repellent field soils. *Journal of Environmental Quality* 22: 109–118.

Kostka SJ (2000) Amelioration of water repellency in highly managed soils and the enhancement of turfgrass

performance through the systematic application of surfactants. *Journal of Hydrology* 231–232: 359–368.
Letey J (2001) Causes and consequences of fire-induced water repellency. *Hydrological Processes* 15: 2867–2875.
Ritsema CJ and Dekker LW (2000) Preferential flow in water repellent sandy soils: principles and modeling implications. *Journal of Hydrology* 231–232: 308–319.
Robichaud PR and Waldrop TA (1994) A comparison of surface runoff and sediment yields from low- and high-severity site preparation burns. *Water Resources Bulletin* 30: 27–34.
Scott DF and van Wyk DB (1990) The effects of wildfire on soil wettability and hydrological behavior of an afforested catchment. *Journal of Hydrology* 121: 239–256.
Shakesby RA, Doerr SH, and Walsh RPD (2000) The erosional impact of soil hydrophobicity: current problems and future research directions. *Journal of Hydrology* 231–232: 178–191.
Watson CL and Letey J (1970) Indices for characterizing soil-water repellency based on contact angle–surface tension relationships. *Soil Science Society of America Proceedings* 34: 841–844.

WATERSHED MANAGEMENT

M D Tomer, USDA Agricultural Research Service, Ames, IA, USA

Introduction

Watershed management may be defined as a set of resource-management practices that are planned and implemented to provide a sufficient source of quality water to sustain human society and natural ecosystems. The practice of watershed management is interdisciplinary, because it recognizes linkages between land and water resources, and because it seeks to balance the needs of society with the capacities of natural resources to meet them. Land-use practices have impacts on hydrologic processes and water quality. Identifying ways to manage these impacts across a mosaic of soils, terrain, and land-use settings is a key challenge in watershed management. Effective watershed management is an iterative process of assessment, planning, and implementation. It begins with an assessment of current land-use practices and their impacts on water resources. Opportunities to improve land management, considering roles of soil, vegetation, and terrain, are then identified and prioritized. Stakeholder groups should be included in planning, to inform citizens about water-resource management issues and provide feedback to ensure recommendations are realistic as well as effective. A range of computerized tools are available to assist with assessing watersheds and alternative management scenarios. Implementation should include a commitment to reassess water-resource management periodically and develop opportunities for further improvement.

An Interdisciplinary Task

Watershed management is aimed at land and water resources, and is applied to an area of land that drains to a defined location along a stream or river. Watershed management aims to care for natural resources in a way that supports human needs for water, food, fiber, energy, and habitation, while supporting other agreed attributes linked to recreation, esthetics, and/or ecologic function. Because of these multidisciplinary concerns, the development of watershed-management strategies can involve complex scientific and public policy issues. Each watershed is unique in physiography, ecology, climate, water quality, land use, and human culture. Therefore any generalized approach to watershed management must be customized to each setting when put into practice. Watershed management requires a long-term commitment that is adaptive to changes in population, climate, culture, and resource-use demands. These issues are unique to each watershed and each nation. Watershed-management experiences from around the globe have dealt with a wide range of issues.

Hydrology and Streamflow Variation

The need to develop a watershed management plan is often identified in response to impacts of floods or drought on society. Long-term monitoring of precipitation, snowpacks, streamflow, and groundwater levels can help society to predict and prepare for these events. Although rainfall (or lack thereof) is not subject to management control, management can reduce the frequency of adverse impacts caused

by extremely wet or dry conditions. Water conservation practices, water storage and control structures, vegetation management, and land-use planning are basic tools that can help manage effects of floods and drought.

Stormflow and Floods

Precipitation may be intercepted by vegetation, or reach the soil surface to infiltrate, pond, or flow over it as runoff. Water at (or near) the soil surface can evaporate, while water below the soil surface may be transpired by plants or may percolate downward. Percolating water can recharge groundwater or move laterally downslope (as interflow) to accumulate at lower parts of the landscape. Some of these low-lying areas have water tables near the surface and can become saturated to the soil surface during a rainfall event. These are called variable-source areas, because they vary in size during an event and become a source of stormflow discharge to streams (Figure 1). Variable-source areas are hydrologically sensitive, meaning that, in many watersheds, management of these areas can help attenuate floods and maintain water quality.

Changes in streamflow that follow precipitation are determined by many attributes that affect water storage on the landscape, the prominence and timing of different pathways of water movement, and the location and extent of variable-source areas. Even in small watersheds, water flow pathways and their timing are difficult to decipher, despite numerous research efforts to deconvolute these pathways using physical, chemical, and isotopic methods. Therefore a hydrograph (Figure 2), a plot showing the response of discharge to precipitation, remains a commonly used basis for watershed characterization. In a small, first-order watershed (the smallest land area generating perennial streamflow), this response will take place more quickly than in large river basins, where a range of soils, terrain, land uses, and travel distances act to desynchronize and lengthen the response.

Flooding occurs when stream discharge exceeds the channel's conveyance capacity and forces water over the stream banks. A flood's magnitude can be characterized by its average frequency of recurrence in years (Figure 2). Engineered hydrologic structures are designed to accommodate a maximum flow defined by a specific recurrence interval. Thereby the probability that the structure will fail can be defined for any given time period. Design criteria may allow a fairly large probability of failure if the cost of failure is small (e.g., culverts beneath low-use forest roads), but will require an infinitesimally small chance of failure if that failure were to be catastrophic (e.g., large dams). The discharge associated with a given recurrence interval must be accurately

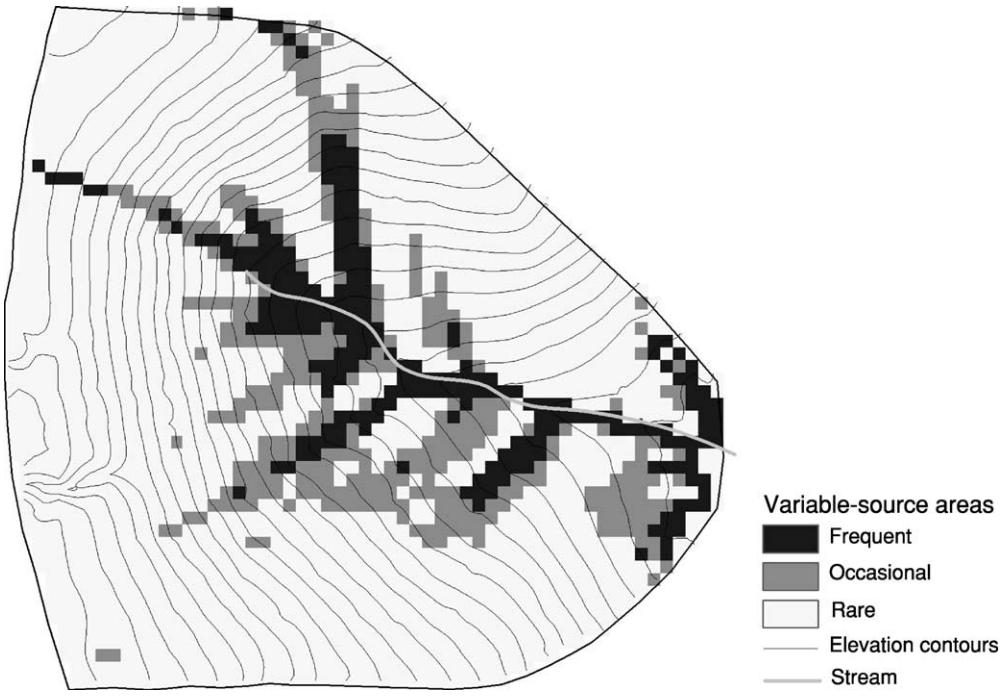


Figure 1 Following precipitation, water is discharged to streams from 'variable-source areas' (or 'partial-contributing areas') that are prone to saturation. The locations of these areas can be estimated based on topographic relationships. This conceptual map indicates the relative frequency of runoff contributions following precipitation.

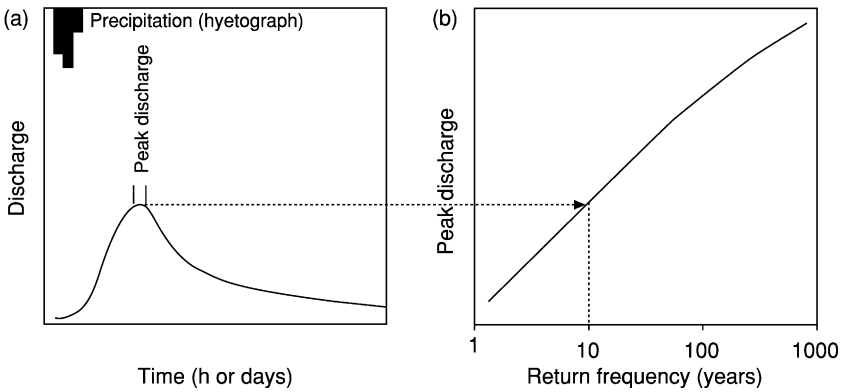


Figure 2 A hydrograph is the response in streamflow to a precipitation event (a). The peak discharge of a hydrograph will have an expected return period; the peak discharge of this event represents a flow that, on average, should only be exceeded once in a 10-year period (b).

known or conservatively estimated. A long-term monitoring record, spanning at least several decades, helps identify the frequency distribution of large flows. Methods of estimation appropriate for small, unmonitored watersheds are also available. These may be based on simple equations (e.g., the synthetic unit-hydrograph method, developed by the US Soil Conservation Service) or sophisticated simulation models.

Snow Hydrology

Seasonal timing of stream discharge is a critical issue, particularly for watersheds with mountainous headwaters where annual snowmelt is an important water source. In many of these watersheds, water yielded by snowmelt is stored in reservoirs for flood mitigation and for later use. Monitoring of annual snowpacks and predictions of water yields are important water management activities in these basins. Snowpacks vary each year, and advance knowledge of water yields helps to plan allocations and warn of potential shortages. Terrain (elevation, slope, and aspect) and wind patterns influence the spatial distribution of snow accumulation and the timing of melt. Vegetation can be managed to moderate these influences.

Drought and Low Streamflow

Prolonged drought can place tremendous pressure on land and water resources, and on society. Dry, hot weather increases demands for water by municipalities and irrigated agriculture when water supply is most limited, causing social conflict. Low streamflow and accompanying warm water temperatures also impact aquatic ecosystems. Drought may contribute to overgrazing and wildfire, which expose soils to erosion once rain returns. Management of groundwater can help protect water supplies and ensure adequate baseflow contribution to streams during drought.

Reservoir water storage, vegetation management, and water conservation measures are among the effective means to reduce impacts of drought. Water harvesting (*See* **Water Harvesting**) is also practiced in some areas. Water conservation can occur through improved methods and scheduling of irrigation, water recycling, and education, and/or regulatory programs to reduce urban water use. Improved irrigation methods may provide the greatest benefit, but return on the investment in conveyance and application systems may take years, in terms of the value of saved water.

Water Quality

Improvements in watershed management may become necessary if water quality becomes impaired. Water quality refers to the capacity of a water body to support certain uses or ecologic functions. It is determined by amounts of dissolved and suspended materials in water, presence or absence of certain microorganisms, and/or physical attributes such as temperature and clarity. Water-quality standards are threshold values (e.g., concentrations, loads) above (or below) which a specific use for water becomes impaired. Water quality can be degraded by natural events or processes, and by pollution resulting from human activities. Common sources of pollution are wastes from municipalities, livestock, and industry, and losses from applications of fertilizer and pesticides to land. Pollution can originate from point sources, where a pollutant enters a water body at a specific location. It can also originate from nonpoint sources where distributed land management activities promote movement of pollutants to ground or surface waters along natural flow pathways. Water quality

management is a process that identifies water uses that are (or may become) impaired, and the contaminants that impair (or threaten to impair) each use. The processes, sources, and pathways that contribute pollutants to a water body need to be understood, and then practices that can sustain water quality according to identified standards must be identified and implemented. This process can raise complex technical and sociopolitical issues. Table 1 lists water quality parameters most commonly of concern, typical sources of pollution for each, possible adverse effects, and some management practices that can control the pollutant. Several types of biologically active trace compounds are included that do not have water-quality standards developed, but that may be of greater concern in the future.

Land-Use Impacts

Soil and vegetation have a major impact on watershed hydrology because they affect partitioning of energy and water near the land surface. While basic soil properties such as texture and depth cannot be influenced through management, soil management can alter the effects of soil disturbance and influence a soil's capacity to support plant growth. Vegetation management can modify canopy interception, evapotranspiration, rooting depths, and seasonal patterns of water use, which affect key hydrologic processes (runoff, infiltration, evapotranspiration, percolation). Soil and vegetation management generally cannot mitigate extreme flood and drought events. However floods with frequent return periods (e.g., less than 10 years) may be mitigated, and low flows associated with annual to perhaps decadal dry periods may be increased. Logically, managing for increased infiltration can achieve both goals, because runoff is reduced, which mitigates flooding, and because groundwater recharge is increased, which contributes baseflow to streams during drought. However, the correspondence of these two benefits is moderated by influences of management on evapotranspiration.

Terrain is an important consideration in soil and vegetation management, and steep lands may require a particular focus, because water is routed from them quickly, contributing to flooding and sedimentation. Risks posed by instability of slopes may be important, especially in populated watersheds with steep terrain where landslides can cause loss of life.

Most watersheds are comprised of agricultural land, forestland, grazing land, urban areas, mine land, riparian zones, wetlands, and lakes in varying proportion and configuration. Each kind of land use poses different challenges and opportunities to improve watershed management. Watershed assessment aims to identify how each land use is influencing hydrology and water quality.

Agricultural Lands

Agricultural lands can contribute to nonpoint pollution, particularly by nutrients, sediment, and agricultural chemicals. It is important to identify erosion control, nutrient, and pest and/or weed management practices to minimize these pollutants effectively. In much of the developed world, crop and livestock production systems have become intensified and spatially aggregated, so that nutrients in animal feed are transferred over longer distances. Treatment, handling, and land application of livestock wastes is a key to effective nutrient management in areas that import feed. Generally, nutrient management aims to apply nutrients at the proper rates and timing to optimize efficiency of crop uptake. Erosion control practices range in scale and include reduced intensity of tillage (including no-tillage), contour farming, vegetative filter strips, and constructed terraces.

Cropping systems can be inefficient in their use of water under rain-fed and irrigated agriculture. Improvement in irrigation efficiencies helps reduce the frequency and severity of water shortages, and can minimize salinization problems in arid regions. Rain-fed agriculture often consists of annual crops that only transpire water during part of the growing season, in contrast to native plant communities. This seasonal restriction on plant water use can increase deep percolation and lead to excess nutrient leaching in humid areas, and salinization in semiarid areas. These problems can be addressed by increasing plant water use with crop rotations that, depending on the setting, may include trees, perennial forages, and/or cover crops.

Forest Lands

Trees have a large capacity to intercept precipitation within their canopy and deep roots that extract soil water from depth. These characteristics diminish the fraction of precipitation that can recharge groundwater and streamflow. Forest harvesting or conversion to other types of vegetation reduces evapotranspiration, which can increase baseflow contributions to streams and allow streamflow responses to small precipitation effects. Historically, observations of severe flooding after extensive forest harvesting brought about early research on the role of vegetation in watershed hydrology, beginning in the early twentieth century. This research quantified relationships between forest cover and streamflow for a number of

Table 1 A summary of common water quality problems associated with land management activities, their impacts, and practices that may address them

Contaminant	*Major causes/sources*	*Adverse impacts*	*Typical solutions*
Physical properties			
Sediment/ turbidity	Erosion, channelization, poor riparian zone management	Fisheries (spawning areas, gill function), aquatic ecosystems, aesthetics, reservoir water storage, water intake and supply systems	Streambank stabilization, erosion control, riparian vegetation management
Temperature	Removal of shading, streambed aggradation, some industry discharges	Fisheries (dissolved oxygen), aquatic ecosystems	Riparian zone management for shading
Inorganic constituents			
Nitrogen	Leaching from soils (NO_3), human and livestock waste (NH_4)	Drinkability (NO_3), toxicity to aquatic animals (NH_4)	Nutrient management, improved waste treatment
Phosphorus	Erosion, accumulation in manured soils	Fisheries, aesthetics, aquatic ecosystems (eutrophication), animal and human health (blue-green algae)	Erosion control, manure management
Dissolved oxygen	High temperature, eutrophication	Fisheries, aquatic ecosystems, aesthetics	Riparian zone management for shading, channel structures to create turbulence, nutrient management
Trace elements	Soil leaching, some pesticides, municipal/industrial wastes	Bioaccumulation, aquatic ecosystems	Increased plant water-use efficiency, industry-specific waste treatment technologies
Salinity	Soil leaching, evaporative discharge of groundwater	Irrigability, drinking, aquatic ecosystems	Management of irrigation water and/ or plant water use in recharge areas, water-table management
Organic materials			
Organic carbon (biological or chemical oxygen demand)	Organic wastes, erosion, eutrophication	Fisheries (reduces dissolved oxygen), aquatic ecosystems, turbidity	Waste treatment, erosion control, nutrient management
Pesticides (herbicides, insecticides, fungicides, nematocides)	Leaching or runoff	Drinkability, potential carcinogenic effects of some compounds and metabolites	Use and losses of nonpersistent compounds, management strategies for reduced pesticide use (e.g., crop selection, rotation, tillage)
Endocrine disruptors	Municipal and industrial wastewaters	Subject of debate, affects metabolic and reproductive function of some organisms	Not clear, but could include waste treatment technologies, alternative formulation of source products that enter the waste stream
Pharmaceuticals	Domestic and livestock wastes	Subject of debate, increased antibiotic resistance in microbial communities hypothesized	Not clear, but could include waste treatment technologies, alternative practices for reduced use in livestock feeding, veterinary, and medical practices
Biological organisms			
Pathogens	Domestic and livestock wastes, riparian grazing, wildlife	Drinkability	Waste treatment, riparian pasture management, improved methods for land application
Algae	Nutrients	Fisheries, aesthetics, dissolved oxygen	Nutrient management, riparian zone management

experimental watersheds. Increases in water yield after harvest may result from greater peak discharges or from greater baseflow contributions to streams, but not necessarily both. Specific factors of harvest disturbance, terrain, and geology determine specific responses in flow regime.

Water yield from forestland can be managed through rotational harvesting that restricts the extent

of harvest during a given period of time (e.g., 20% of the commercial forest under a 25-year rotation may be harvested during a 5-year period). This strategy can desynchronize flows in the watershed and decrease flood frequency. Forest management should consider effects of infrastructure, because forest roads, skid trails, and landings can be major sources of sediment. Harvesting can also increase the risk of mass movement (i.e., rockfalls, landslides) in steep terrain.

Grazing Lands

Rangelands occupy large areas with semiarid to arid climates where extensive crop production is not feasible. Overgrazing and subsequent erosion are prevailing concerns that affect hydrology and water quality. Proper stocking, herd management, and rotational grazing systems comprise effective management strategies, especially if designed to consider the frequency of drought. Managing the distribution of water to livestock is critical, because livestock will not wander far from reliable water sources. In humid areas, pasture management involves more intensive forage production and grazing. Limiting access of cattle to surface waters may address concerns for water quality along small streams.

Mined Land

Mining operations are often small, but the drastic nature of disturbance often requires that effects on water quality and supply be considered. Dewatering of aquifers may occur with subsurface or pit mining, and surface waters may be diverted off-site, or collected in holding ponds. Water quality impacts will vary depending on the method and extent of excavation, the type of rock and/or ore material being mined, the rock and/or ore processing occurring on-site, and reclamation and/or revegetation practices carried out as mining concludes or proceeds to new areas. Acidification, heavy metals, and sediment are common water-quality concerns, depending on the type of mine operation.

Urban Areas

Urban areas also occupy a small part of most basins, but are of predominant concern for the people and the economic activity within a watershed. Daily demands for drinking water, sanitation, and waste treatment are overriding issues due to obvious implications for health and quality of life. Demands for these services increase with development and population. If the growth in demand for water approaches the available supply, or if water quality becomes impaired, then stakeholders will demand improvements in water management. In watersheds with limited water supplies, planning should aim to provide a sustainable match between supply and demand in the longer term, including conservation measures to cope with effects of drought.

Management of storm runoff can be critical in urban areas. Urban development increases the extent of impervious surfaces (rooftops, roads, parking lots), and runoff from these areas must be accepted by a stormwater detention and conveyance system. Construction can cause soil compaction, increasing runoff, and diminishing infiltration and groundwater recharge. Urban development can change flood-frequency characteristics downstream, causing existing floodwater control and conveyance measures to become inadequate. As an example, a 0.1-ha residential lot may have pavement and roofing occupying 40% of that area. If a 10-mm rainfall generates 7 mm of runoff from the impervious surface, then 2.8 m^3 of water will be conveyed down-gradient. Prior to development, a 10-mm rainfall would probably generate little, if any, runoff. Alternative practices being used in some areas include stormwater detention basins with beneficial dry-weather uses (e.g., recreation), small detention basins ('rain gardens') for individual residential lots, 'green' (planted) roofs, and permeable pavements. Urban planning should consider future development and implications for stormwater hydrology.

Riparian Areas and Wetlands

In most watersheds, riparian zones and wetlands also occupy small areas, but they are present in every watershed and their management is frequently a focus of attention. Riparian zones contain variable-source areas that generate storm flow, influence the interaction between groundwaters and surface waters, and form the boundary between aquatic and terrestrial ecosystems. They support biological diversity and wildlife habitat, and provide aesthetic value. Biological processes (e.g., plant growth, microbial activity) are intensified in riparian areas because water and nutrients are usually abundant compared with upslope areas. Management of riparian zones offers the opportunity to improve streamflow regimen (timing) and water quality. Wetlands can detain storm flows and remove nutrients, particularly nitrate via denitrification. Riparian vegetation can take up nutrients and encourage trapping of sediments delivered from upslope areas. It is important to identify objectives for riparian management, locations where those objectives can be met most effectively, and vegetation management systems that can help achieve them. The potential benefits of improved riparian

management, however, can only be fully realized when uplands are also managed to achieve water resource goals.

Approaches, Challenges, and Tools for Watershed Management

The unique attributes of each watershed must be considered in developing and implementing a watershed management plan. While administrative approaches may vary, they usually involve four phases, including problem definition, assessment, selection of alternatives, and implementation and/or evaluation (Table 2). Ideally, watershed management is a continuous process that identifies and develops opportunities to improve environmental quality and resource sustainability.

Watershed planning must balance a number of legitimate, but often competing, resource-use requirements and concerns. Both environmental and social attributes of a watershed must be considered from the outset if planning is to result in successful implementation. Technical experts provide little long-term benefit unless local land managers and community interest groups (stakeholders) perceive their recommendations as balanced and realistic. Therefore stakeholder representation and involvement are necessary from the first stage of planning. An early consensus on resource issues to address and stakeholder groups to include (phase 1) benefits the community and its watershed resources.

Effective watershed planning depends on an objective assessment (inventory) of the current situation (phase 2), best carried out by technical experts from various resource-management and engineering disciplines. Stakeholder involvement can facilitate public education on the causes and impacts of water-resource problems, and can help establish trust between stakeholders and technical experts.

A number of computer tools can help to accomplish a watershed assessment. Use of a geographic information system (GIS) database provides ways to view and evaluate combined map data that include soil survey, land use and ownership, terrain analyses, and remote sensing. While GIS can provide information in compelling graphic formats, its output should be reviewed critically, considering the quality and scale of input data.

Major concerns of watershed management are related to streamflow regime, groundwater availability, and water quality. These all result from a mosaic of interactions among climate, land-use practices,

Table 2 A summary of the watershed management planning process, with a focus on stakeholder involvement

Phase	*Key questions*	*Sources of information*	*Analysis tools*	*Stakeholder role*
1. Problem definition	What are the key problems and who needs to be involved in solving them?	Public, public interest groups, focus group meetings, monitoring data	Statistical analysis of data, public opinion survey results	Source of information; help define formal and informal roles for stakeholders in project
2. Watershed assessment	What information can be gathered to facilitate an assessment of the watershed? Are new assessment/monitoring efforts needed and how can they be accomplished?	Mapped information: population census data, land use, soil survey, remote sensing, surveys of existing conservation systems; intensified and/or synoptic sampling of water resources	Geographic information systems, terrain- and image-analysis techniques, process models	Learning of assessment results, review, and feedback
3. Identify and select management alternatives	What kinds of changes can help solve the problem(s)? Are these changes feasible? Who would need to implement them and what incentives would be needed?	Design criteria and literature on management alternatives and their environmental impacts (e.g., best-management practices)	Process models, decision-support systems, economic analysis	Evaluate feasibility of alternatives; prioritize alternatives; facilitate education of wider community
4. Implement and evaluate	Are incentives effective in encouraging management changes? Are the changes effective in reaching water-resource management goals? What further improvements can be made?	Monitoring data, surveys of new conservation practices, stakeholder feedback	Process modeling, model validation, statistical analysis of monitoring data	Full participants in ongoing project review, help determine if/when new alternatives are required

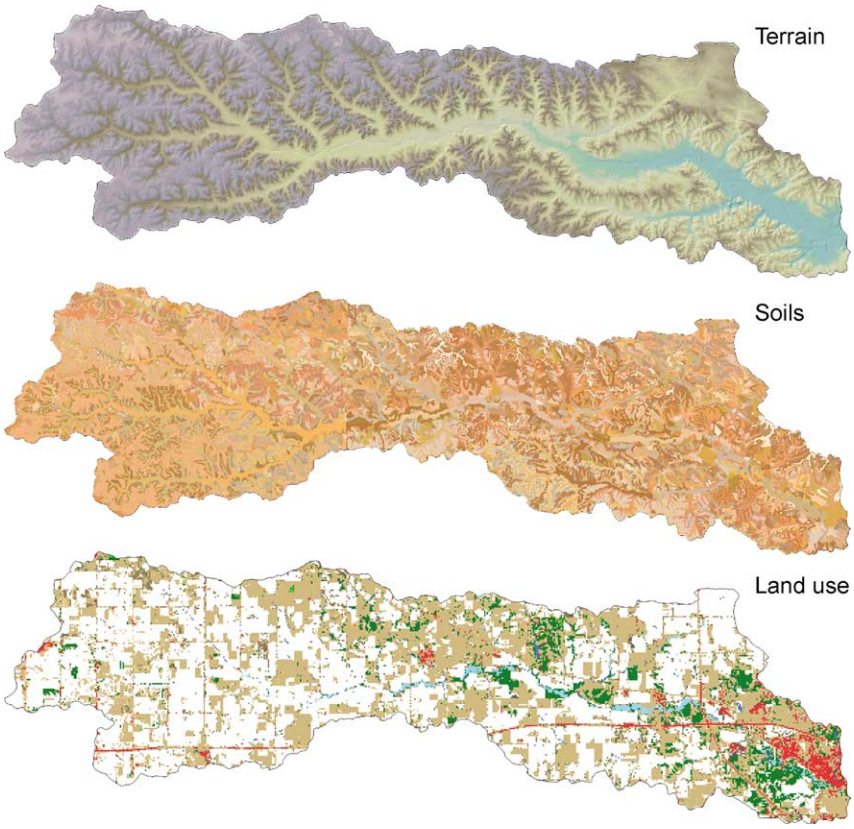


Figure 3 Watershed management and assessment must consider a spatial mosaic of land use, soil, and terrain.

soil, terrain, and geology (Figure 3). Existing hydrologic and water-quality conditions result from these interactions in space and their integration across time. Groundwater quality and stream baseflow are particularly influenced by past land-use practices, because of the slow rate of water movement through most aquifers. A key task for watershed assessment is to segregate this mosaic of spatial and temporal effects, and identify locations and activities that disproportionately influence water resources. Judgment is needed to do this. While there may be obvious areas of focus (e.g., shallow sandy soils, steep slopes, riparian areas), analytic capabilities in this area are not yet fully developed. Many of our working assumptions about effects of land use on hydrology and water quality have been gained through research carried out in small plots, hillslopes, or at small watershed scales. However, analyses conducted at a fine scale may not transfer directly to a large basin. The prominence of biophysical processes affecting hydrology and water quality can vary with the scale of observation. New techniques to analyze hydrologic processes across complex terrain, and analyze relationships between patterns and physical processes, are being developed that may provide new approaches to watershed assessment.

Trends in land use and population must also be considered during watershed assessment. The challenges of watershed management increase with population. The impact can be similar to drought, because pressures on water supplies and on land resources that yield water can increase concurrently. Urban expansion may be accompanied by increased disturbance from overgrazing and deforestation. Trends toward improved management practices should also be documented during an assessment, to help develop information about adoptability and effectiveness of new practices.

Monitoring of flows and water quality can also be important in assessment, because this helps to document water-resource problems. However an adequate set of monitoring data will not always be available, and usually the planning process cannot wait to acquire one. Nevertheless, it is important to continue monitoring, or initiate and/or expand it as soon as possible. Long-term data, gathered using consistent protocols for sampling and analysis, provide the best basis to assess the effects of new practices that are implemented.

Identifying management alternatives (phase 3) aims to propose and select realistic solutions to water-resource problems. Potential solutions include many

land-use practices, vegetation management changes, water-conservation measures, and engineered structures, across all types of land use throughout the watershed. These alternatives are not equal in terms of acceptability, cost, or effectiveness. Decision-support systems and economic analyses can be used to help rank social impacts (e.g., stakeholder acceptance, costs) of each alternative. The rankings also depend on the benefit of each alternative for water resources. Often, these benefits can be estimated using simulation models, which calculate how an alternative might affect biological and physical processes that determine a watershed's hydrology and/or water quality. A wide array of models are available, and the best model to use depends on the specific resource concern(s) and experience of the technical team. It may be possible to use monitoring data to validate model accuracy under current conditions, and this can improve the confidence in using the model to assess the effectiveness of alternatives. The final listing of ranked alternatives is likely to require several iterations of assessing social and environmental impacts.

Once alternatives are ranked, implementation of new practices is encouraged using some set of incentives, and then the management changes are evaluated (phase 4). This takes a sustained commitment. Monitoring data are needed to assess effects of new measures. Changes may not have immediate impact, particularly if they are aimed to improve groundwater conditions. New vegetation management strategies, particularly for forests, may take years to become fully effective. Seasonal trends and extreme events may make it difficult to confirm changes based on interpretation of monitoring data. External factors may also affect success; for example, an understanding of how a changing global climate is affecting water resources may become important. (*See* **Climate Change Impacts**.) These factors reinforce the need for a long-term commitment to watershed management, sustainable development, and periodic review of water-resource management goals and best ways to achieve them.

List of Technical Nomenclature

Baseflow — The portion of stream discharge that originates as groundwater, and that enters the stream channel from the saturated zone via lateral or upward-moving groundwater flow (volume per time)

Drought — A period of time with below-average precipitation that results in significantly diminished stream discharge and water-table elevations

Flood — An event during which a stream's discharge exceeds the capacity of the stream channel to convey water, so that water rises above the stream banks and covers some of the adjoining land surface

Hydrograph — A plot of stream discharge versus time that is often used to show the response of stream discharge to a precipitation event

Infiltration — Vertical entry of water into the soil at the land surface (length per time)

Interception — An amount or fraction of precipitation that wets vegetation and is eventually evaporated from vegetation so that it does not reach the soil surface (length)

Interflow — Water that moves laterally downslope beneath the land surface and at shallow depth above the saturated zone. This flow generally takes place at the scale of an individual hillslope (volume per time)

Perennial stream — A channel that conveys water throughout the year in most years. Usually, perennial flow only occurs where some of the stream discharge originates as baseflow

Physiography — Physical attributes of the terrain, including topography, soils, landform, and surficial geology

Recharge — Water that percolates down through the unsaturated zone and is added to groundwater storage (length)

Runoff — Water that flows across the land surface (volume per time)

Stakeholder — Individuals and/or groups that participate in assessing and solving resource management problems of direct importance to their community or representative constituency

Stream discharge — The quantity of water passing a location on a stream or river (volume per time)

Variable-source area — An area on the landscape that is prone to become saturated to the soil surface in response to precipitation or snowmelt, and that, when saturated, generates runoff that contributes to stream discharge. ('Partial contributing area' is an equivalent term)

Water-quality standard — A threshold value for a constituent concentration or other attribute of water, above or below which a specific use or ecologic attribute of the water becomes impaired

Watershed — A mapped area that contributes runoff or baseflow to a perennial stream or river at a defined location

Water yield This is equivalent to a unit-area stream discharge (volume per area per time)

See also: **Climate Change Impacts**; **Environmental Monitoring**; **Erosion:** Water-Induced; **Geographical Information Systems**; **Land-Use Classification**; **Overland Flow**; **Remote Sensing:** Soil Moisture; **Sustainable Soil and Land Management**; **Terraces and Terracing**; **Water Harvesting**

Further Reading

American Society of Civil Engineers (1996) *Hydrology Handbook*. ASCE Manual No. 28, Task Committee, Management Group D. New York: ASCE.

Brooks KN, Folliott PF, Gregersen HM, and DeBano LF (1997) *Hydrology and the Management of Watersheds*. Ames, IA: Iowa State University Press.

Downes BJ, Barmuta LA, Fairweather PG *et al.* (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge, UK: Cambridge University Press.

Folliott PF, Baker MB Jr, Edminster CB, Dillon MC, and Mora KL (2002) *Land Stewardship through Watershed Management: Perspectives for the 21st Century*. New York: Kluwer Academic/Plenum.

Grayson R and Bloschl G (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*. Cambridge, UK: Cambridge University Press.

Heathcote IW (1998) *Integrated Watershed Management: Principles and Practice*. New York: John Wiley.

Singh VP (ed.) (1995) *Computer Models of Watershed Hydrology*. Highlands Ranch, CO: Water Resources Publications.

Trudgill ST (ed.) (1995) *Solute Modelling in Catchment Systems*. Chichester, UK: John Wiley.

Vieux BE (2001) *Distributed Hydrologic Modeling using GIS*. Boston, MA: Kluwer Academic.

Wilson JP and Gallant JC (eds) (2000) *Terrain Analysis*. New York: John Wiley.

Younos T (ed.) (2001) *Advances in Water Monitoring Research*. Highlands Ranch, CO: Water Resources Publications.

WATER-USE EFFICIENCY

M B Kirkham, Kansas State University Manhattan, KS, USA

Introduction

For centuries, humans have been concerned with the efficient use of water in the production of crops. The ability to grow crops and manage their needs for water is necessary for civilization. Water-use efficiency is defined as the aboveground biomass production per unit area per unit water evapotranspired. The biomass is usually determined as dry weight rather than as fresh weight. Therefore, water-use efficiency is expressed in equation form as follows:

$$\text{Water-use efficiency (WUE)} = \frac{\text{Dry weight production (DW)}}{\text{Evapotranspiration (ET)}} \quad [1]$$

Water-use efficiency can be expressed on the basis of vegetative growth or reproductive (grain) growth; the basis must be stated. Different units can be used in the numerator and denominator of Eqn [1]. Old units were pounds or tons of dry weight produced per acre-inch of water evapotranspired. We now usually express it as a simple ratio, such as kilogram of dry weight per kilogram of water transpired. Water-use efficiency expressed in the latter dimensionless form is similar to the reciprocal of the old terms 'water requirement' and 'transpiration ratio.' New portable gas analyzers allow us to measure photosynthetic rate and transpiration rate on individual leaves or on parts of individual leaves as small as $6\,cm^2$. So we now can express water-use efficiency on a leaf basis, as follows:

$$\text{Leaf water-use efficiency} = \frac{\text{Photosynthetic rate}}{\text{Transpiration rate}} \quad [2]$$

The unit for photosynthetic rate is micromoles of CO_2 per square meter per second, and the unit for transpiration rate is micromoles of H_2O per square meter per second, so the unit for water-use efficiency on a leaf basis is micromoles of CO_2 per micromole of H_2O.

Water-use efficiency can be based either on evapotranspiration ('ET efficiency') or on crop transpiration ('T efficiency'). The difference is important, because suppression of soil-water evaporation and prevention of weed transpiration can improve ET efficiency. However, it need not improve T efficiency, which is a measure of crop performance. These two water-use efficiencies may also be based on either the

total dry-matter production or the marketable yield, and the basis should always be stated.

In dry regions where the proportion of water returned to the atmosphere by evaporation may be 90% or more, it is necessary to increase water-use efficiency. That is, we want to produce more plant material with smaller amounts of water. Many schemes have been proposed to obtain water for these dry regions: artificial rainmaking, sea-water distillation, and towing icebergs to regions where the melted water can be used for irrigation. However, increasing water-use efficiency may be the best way to provide enough water for crop production in areas where evapotranspiration exceeds rainfall.

History

At the start of the twentieth century, considerable work was in progress measuring water requirement, the inverse of T efficiency. Franklin Hiram King (1848–1911), at the University of Wisconsin (USA), was the first in the USA to research the water required to produce field crops. King was also the author of a classic work entitled *Farmers of Forty Centuries*, which describes his trip to China in the early 1900s. He made careful notes to help him understand how people could farm the same fields for 4000 years without destroying their productivity. King used small lysimeters in a greenhouse and in fields to determine the water requirement of crops. Widtsoe, at the University of Utah (USA), and Kiesselbach, at the University of Nebraska (USA), initiated research in 1902. Widtsoe is also well known for his studies on dry farming. The methods that he described for agricultural production in areas with low rainfall apply equally well today as they did at the beginning of the twentieth century.

Briggs and Shantz, based in Akron, Colorado, compiled much of the early water-requirement data from container experiments that began in 1910. The water-requirement work of Kiesselbach and of Briggs and Shantz was concerned with total dry matter and transpiration on a plant basis rather than on a land-area basis. Briggs, Shantz, and Kiesselbach reviewed the work that had been done up to that time, although they omitted water-culture experiments and those on seedlings. They found that 14 researchers had worked on water-use efficiency prior to 1900. Since that time, there has been an increasing amount of work on water use and water-use efficiency, all done with more elaborate equipment and experiments than those of the last half of the 1800s and early 1900s. However, these early experiments still remain the foundations of our modern understanding of water-use efficiency. One of the best summaries of the water requirement of many different plants is that compiled in 1927 by Shantz and Piemeisel.

Factors that Influence Water-Use Efficiency

Soil Factors

Soil-water content Briggs and Shantz felt that they could make no firm conclusion about the effect of soil-water content on the water requirement. In some experiments, however, the water requirement usually decreased a little when growth was limited by water deficits. The same result was obtained by Kiesselbach, who found that, with a deficient water regime, the dry matter of corn stalks, ears, and leaves were reduced 37%, 28%, and 10%, respectively. The water requirement based on ear weight decreased 4.3%, and the water requirement based on the total dry matter decreased 10%. Later Briggs and Shantz showed that the water requirement of wheat was essentially the same under deficient soil water (dry matter decreased 40%) as when water was adequate. King reached the same conclusion from experiments with corn, oats, and potato. Several decades later, de Wit analyzed experimental data that also showed similar T efficiencies regardless of water shortages.

Measurement of soil-water evaporation independently of plant transpiration is difficult, and models have been developed to separate the two. Evaporation can be measured directly with evaporimeters. Once a canopy covers the ground, most water is lost by transpiration. However, because water is lost by soil-water evaporation, investigators have studied evaporation-retardant chemicals, mulches, and plastic films. Results vary; for example, Letey and Peters have compared water-use efficiency of corn on plots with and without polyethylene film. The covered plots were initially wet and then covered to cut off summer rainfall and evaporation. Control plots received natural rainfall. Water-use efficiency for the covered plots was 345 kg per 100 m^3 of water and for the natural rainfall plots it was 148 kg per 100 m^3. However, in another year of the study, results showed no advantage from the plastic film. Plastic film is expensive and has been used only on high-value horticultural crops. But as water becomes more valuable, antitranspirants, mulches, and plastic films might be economically feasible for field crops, and water-use efficiencies with them need to be determined.

Method of irrigation Israel has dramatically increased its water-use efficiency through the use of new irrigation methods begun approximately in the 1960s (Figure 1). They are high-frequency,

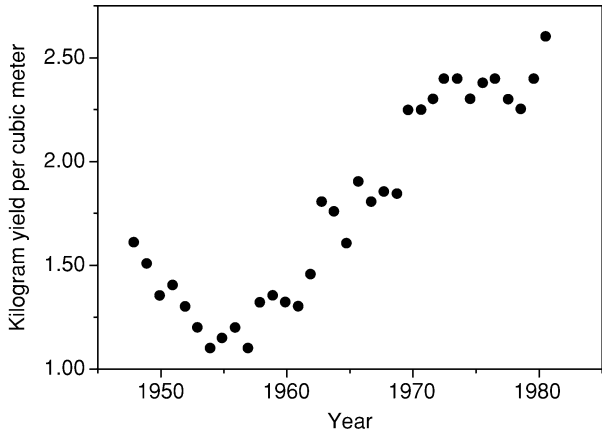


Figure 1 The improvement of crop water-use efficiency in Israel. (From *Rivers of Eden: The Struggle for Water and the Quest for Peace in the Middle East* by Daniel Hillel, © 1994 by Daniel Hillel. Oxford University Press, Inc.)

low-volume techniques of drip (trickle) and microsprayer irrigation. Before the new methods were developed, irrigations (flood or sprinkler) were usually infrequent and large, because it was assumed that soil moisture remained essentially equally available to crops until evaporation and extraction by roots depleted it to the permament wilting point. With the old methods, the topsoil was saturated after an irrigation, a condition that leached nutrients and deprived roots of aeration, followed by a period in which the crops were allowed to dry to a degree that desiccated the roots. Crops often show a pronounced increase in yield when irrigations are provided in sufficient quantity and frequency that water never becomes limiting. The newer irrigation methods became possible with the development of porous tubes for subirrigation and the advent of low-cost tubing that can be fitted with specially designed drip emitters. Drip and microsprayer irrigations can maintain the root zone in a moist yet unsaturated condition, so the roots do not lack water or oxygen. The irrigations are targeted precisely at the base of crop plants, thus wetting only a small fraction of the surface and not the interrow areas, where compaction by traffic, evaporation, and weed proliferation is reduced. Prevention of weed growth further conserves water.

Plant Factors

Species adaptation Because most species are adapted to grow with specific water requirements, plants must be matched with the rainfall. The record of human domestication of plants dates back at least 6000 years, and plant introductions (i.e., nonnative plants introduced into a region from another region) permitted our first great stride forward in efficient water use. In the USA, plant introductions were sponsored by the US Department of Agriculture. Immigrants also brought plants; for example, Turkey winter wheat was introduced into the Great Plains by Russian Mennonites, and winter wheat still is the most important crop in the region.

Plant breeding Briggs and Shantz have shown that plants differ greatly in water requirement. They carried out their research before the photosynthetic pathways had been elucidated. In all plants, photosynthesis involves the 'C_3' process that converts carbon dioxide into molecules with three carbons. In some species, called 'C_4' plants, carbon dioxide is first converted into molecules with four carbons, which are then transported within the leaf to sites (bundle-sheath cells) where carbon dioxide is released, providing a high concentration of carbon dioxide for the C_3 process. The same enzyme that catalyzes the first step in the C_3 process (ribulose 1, 5 bisphosphate carboxylase) can also catalyze an oxidation that leads to photorespiration, which is respiration that occurs in the light. Photorespiration consumes as much as a third of the carbon dioxide that the plant has absorbed in the light. It is slowed by high concentrations of carbon dioxide and, hence, occurs more rapidly in C_3 plants such as wheat than in C_4 plants such as maize. Plants with the C_4 type of photosynthesis have a lower water requirement than plants with the C_3 type of photosynthesis. The reasons for this difference are not fully understood, but it is probably due, in part, to the fact that, under the same environmental conditions, C_4 plants generally have a higher stomatal resistance than C_3 plants. The difference in water use between C_3 and C_4 plants was evident even in the work of Briggs and Shantz (Table 1). Plants which we know today have the C_4 type of photosynthesis (e.g., maize, millet, sorghum) have a lower water requirement than plants with the C_3 type of photosynthesis (e.g., alfalfa, barley, wheat). On average, plants with the C_4 type of photosynthesis have a water requirement of 250–350 g $H_2O\,g^{-1}$ dry weight, and plants with the C_3 type of photosynthesis have a water requirement of 450–950 g $H_2O\,g^{-1}$ dry weight. Alfalfa's water requirement can be greater than this; it has the highest water requirement of any crop species.

A third type of photosynthesis occurs, but the only commercially important plant that has this type of photosynthesis is pineapple. It is called the crassulacean acid metabolism (CAM) photosynthetic pathway. Unlike most plants that have stomata open in light, when carbon dioxide is taken up (and water is lost by transpiration), plants with CAM photosynthesis keep their stomata open in darkness and take

Table 1 Water requirement based on aboveground vegetative dry matter for various crops as determined by Briggs and Shantz before the knowledge of C_3 and C_4 photosynthetic pathways. Photosynthetic pathway of each crop is noted in parentheses

Crop	*Water requirement (g H_2O g^{-1} dry wt)*
Wheat (C_3)	507
Oats (C_3)	614
Barley (C_3)	539
Rye (C_3)	724
Corn (C_4)	369
Sorghum (C_4)	309
Millet (C_4)	275
Peas (C_3)	800
Sweet clover (C_3)	709
Alfalfa (C_3)	1068
Buckwheat (C_3)	578
Rape (C_3)	441
Potatoes (C_3)	448
Sugar beets (C_4)	377
Salsola (saltwort of the Chenopodiaceae family) (C_4)	366
Amaranthus (C_4)	303
Artemisia (C_3)	765

Source: Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. II. A Review of the Literature.* US Department of Agriculture, Bureau of Plant Industry, Bulletin No. 285, p. 90. Washington, DC: US Department of Agriculture.

up carbon dioxide at night. The photosynthetic enzymes that CAM plants use are the same ones as found in C_4 plants, but their activity depends upon whether it is light or dark. Because CAM plants keep their stomata open at night and not during the daylight hours when heat and light cause great amounts of water to be lost through the stomata, they have the lowest water requirement of any plants. On average, their water requirement is 50–55 g H_2O g^{-1} dry weight.

Not only do species differ in water requirement, but also cultivated varieties (cultivars) within a species vary. This was noted in 1913 by Briggs and Shantz, who suggested the possibility of developing varieties that are efficient in the use of water.

When water becomes limited during a plant's development, yield differences among species and varieties are due to a plant's water efficiency. Plant characteristics that contribute to such efficiency include maturity (including planting date); leaf area and angle of inclination; leaf rolling; number, distribution, and size of stomata; presence of waxy or corky epidermal cells; the ability to become temporarily dormant; and extensiveness of the root system. Distribution of roots is especially important. Even though roots may be at an optimum depth, the surface roots are often the ones that preferentially extract water after rewatering.

Cultural Factors

Planting patterns Planting patterns have a direct effect on yield, solar-energy capture, and soil-water evaporation and thus an indirect effect on water-use efficiency. Soil-water evaporation is diminished with denser plantings. In humid regions where rainfall exceeds evapotranspiration, plant populations can be increased with a concomitant increase in yield. However, in semiarid regions, when plant populations are increased, the plants most often need to be irrigated. For example, dense planting of sorghum has been advocated as a beneficial method for rain-fed land in Kansas, because it reduces soil-water evaporation, runoff, erosion, and weeds. But the super-thick plantings fail to produce grain if soil moisture is not adequate.

Seed quality A crop-management problem, sometimes overlooked, is seed quality. For most efficient use of water, a grower must start with high-quality seed. Poor seed may mean not only poor germination and weak plants that cannot take advantage of the available water, but also such seed may include weed seeds and thus provide competitors for the water present.

Weeds One of the primary management means of obtaining more efficient water use is the elimination of weeds in crops. Weeds compete with crops for soil nutrients, water, and light. Except in high-rainfall areas, the primary concern is the water factor. The water requirement of many weeds is greater than that of crop plants, because many weeds are C_3-type plants. For example, the average ragweed plant requires three times as much water per pound (0.45 kg) of dry matter produced as a maize plant, a C_4 plant.

Disease and insect pests Few data concerning effects of disease and insect pests on water-use efficiency are available. Rusts decrease transpiration in the early stage of infection, but, on sporulation, they cause epidermal leakage of water vapor and increased water loss in light and dark conditions. The powdery mildews cause a similar decrease in transpiration and then an increase in transpiration. These transpiration trends have been found to decrease the transpiration efficiency. For example, studies show that early rust infection halves the transpiration efficiency of wheat, when based on total dry matter, and, when based on grain yield, it is 25-fold smaller. Infection at later stages of development produces correspondingly smaller decreases. Water-use efficiency data for vascular wilts, root rots, and insect pests are lacking. Breeding for resistant varieties is probably the most economic and longest-lasting method to control

diseases and insects. Efficient crop and water management also includes the wise use of herbicides and insecticides, and their effect on water-use efficiency needs to be determined.

Tillage The surface characteristics of soil can have a profound effect on the water that infiltrates and runs off. Reduced or minimum tillage has been taken up by some farmers in the semiarid Great Plains. While yields are often reduced during the first few years after reduced tillage is initiated (due, for example, to difficulty in controlling diseases and insects), water is conserved. This results in less erosion in wet years and longer survival of crops in droughts, because the mulch that remains on the surface of the soil reduces evaporation. However, water-use efficiency may not be increased. Stubble-mulch tillage in the Great Plains does not necessarily alter rainfall-use efficiency.

Rotations While studies indicate that rotations can improve infiltration and water use, they do not always. For example, a grass–legume crop may have low infiltration rates because of heavy grazing. In Kansas, a deep-rooted alfalfa–fescue mixture was grown on a claypan to increase root penetration of maize, planted the following year. However, the deep-rooted plants removed moisture throughout the soil profile, resulting in low yields of maize. Continuous maize yielded more than maize grown in rotation. Sunflower, with its deep roots and a high water requirement, removes more water from a soil profile than sorghum. Under dry-land conditions, crops planted following sunflower will lack stored moisture. Thus, rotations may not generally improve available water and may result in less available moisture when the preceding crop has a high water requirement.

Fertilization Data have been reviewed from about 20 experiments showing the effect of varying fertility on water requirement and the conclusion is that, with poor soils, the water requirement may be reduced one-half to two-thirds by increasing fertility. In all but five experiments, evaporation from the soil was included in the water use and could bias the water requirement for smaller plants with less transpiration. In experiments where evaporation was prevented, the water requirement did not increase significantly as the fertility and yield decreased, until the dry weight of the plants had decreased approximately 50% because of malnutrition. It appears from these data and others that, unless malnutrition is severe, transpiration efficiency based on total dry matter is not greatly affected by poor fertility. But, when nutrient deficiency reduces yield to about half that on a well-fertilized soil, then transpiration efficiency is reduced markedly.

Climate

The seasonal changes in both transpiration and evapotranspiration efficiencies found with exposures to different radiation, temperature, and humidity regimes show that climate exerts a major influence on water-use efficiencies. Even the early researchers recognized that, with well-watered plants, variation of transpiration with climate causes changes in water-use efficiency as the season progresses, between years and between locations, and that free-water (pan) evaporation or saturation deficit could be used to normalize the transpiration component (Table 2).

The carbon dioxide concentration in the atmosphere is increasing due largely to the increased burning of fossil fuel by industry and automobiles. Since 1958, the carbon dioxide concentration has risen from 316 ppm ($316\,\mu l\,l^{-1}$) to approximately 370 ppm. The concentration increases at a rate of $1.5–2.0\,\mu l\,l^{-1}$ per year. When atmospheric carbon dioxide is relatively low, net photosynthesis is faster in C_4 plants than in C_3 plants, but, at higher levels of carbon dioxide, the change in photorespiration per change in carbon dioxide concentration leads to greater increases in net photosynthesis for C_3 plants than for C_4 plants. Therefore, increasing levels of carbon dioxide should benefit C_3 plants more than C_4 plants. The increasing levels of carbon dioxide do increase the water-use efficiency of C_3 plants. When winter wheat was grown for 3 years with four different levels of carbon dioxide (ambient or $340\,\mu l\,l^{-1}$

Table 2 Water requirement (WR) of an alfalfa cultivar (Grimm) and pan evaporation at four different stations in the Great Plains of the USA, as reported by Briggs and Shantz

Location	*Growth period (1912)*	*Water requirement (g $H_2O\,g^{-1}$ dry wt)*	*Mean pan evaporation (mm day^{-1})*	*WR/pan*
Williston, N. Dakota	29 Jul–24 Sept	518 ± 12	4.04	128
Newell, S. Dakota	9 Aug–6 Sept	630 ± 8	4.75	133
Akron, Colorado	26 Jul–6 Sept	853 ± 13	5.74	149
Dalhart, Texas	26 July–31 Aug	1005 ± 8	7.77	129

Reproduced with permission from Tanner CB and Sinclair TR (1983) Efficient water use in crop production: research or re-search? In: Taylor HM, Jordan WR, and Sinclair TR (eds) *Limitations to Efficient Water Use in Crop Production,* pp. 1–27. Madison, WI: American Society of Agronomy, Crop Science Society of America/Soil Science Society of America.

and 485, 660, and 825 $\mu l\,l^{-1}$), the water requirement decreased with increasing levels of carbon dioxide (Figure 2). During the 3 years of the study, water required to produce a gram of grain under the highest carbon dioxide concentration (825 $\mu l\,l^{-1}$) and dry conditions (half-field capacity) was less (547 $ml\,g^{-1}$) than that required to produce a gram of grain under well-watered conditions (field capacity) at the ambient carbon dioxide concentration (642 $ml\,g^{-1}$). Under well-watered conditions, the water requirement of the wheat grown with ambient and elevated carbon dioxide was 642 $ml\,g^{-1}$ and 458 $ml\,g^{-1}$, respectively (a reduction of water requirement, based on grain, of 29%).

Even though predictions indicate that C_3 plants will have a higher water-use efficiency with elevated carbon dioxide than C_4 plants, data also show that C_4 plants have increased water-use efficiency. When sorghum, kept under watered conditions, was grown under ambient (330 $\mu l\,l^{-1}$) and elevated (795 $\mu l\,l^{-1}$) levels of carbon dioxide, water requirement for grain production was 1090 $ml\,g^{-1}$ and 776 $ml\,g^{-1}$, respectively (a 29% reduction in water requirement with elevated carbon dioxide).

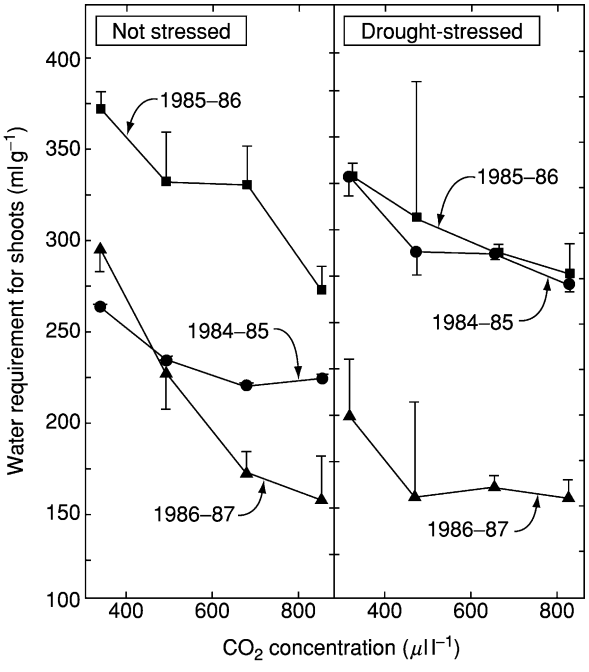


Figure 2 Water requirement for winter wheat grain grown under well-watered (field capacity) and drought (half-field capacity) conditions as affected by carbon dioxide concentration during a 3-year study. Vertical bars, ± standard deviation. Only half of each bar is drawn for clarity. (Adapted from Chaudhuri UN, Kirkham MB, and Kanemasu ET (1990) Carbon dioxide and water level effects on yield and water use of winter wheat. *Agronomy Journal* 82: 637–641, with permission.)

Measurement of Water-Use Efficiency

Water-use efficiency is usually measured by harvesting plants, determining dry weight of the vegetative portion or grain, and dividing that by the rainfall or irrigation plus rainfall. Weighing lysimeters have allowed more precise measurements of water used. With the development of portable photosynthetic systems, researchers can now measure in the greenhouse or field the water-use efficiency of individual leaves or parts of leaves. Figure 3 shows water-use efficiency determined on different locations of a leaf of puka (*Meryta sinclairii* Seemann), a native tree of New Zealand. The leaves are large and elliptical, up to 0.6 m long and 0.3 m wide. The center of the leaf (to the side of the main vein) and the middle edge have the highest photosynthetic rates and transpiration rates, but the tip of the leaf has the highest water-use efficiency. The base of the leaf has the lowest photosynthetic rate, the lowest transpiration rate,

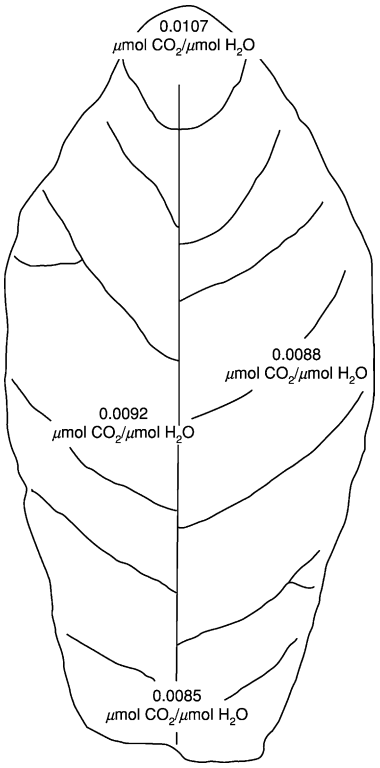


Figure 3 Water-use efficiency at four locations on a leaf of puka, a native New Zealand tree. The locations are tip, edge of middle, base, and center to the side of the main vein. To obtain these values, photosynthetic rates and transpiration rates were measured on puka leaves during a 1-month period in the summer of 1998 in Lincoln, New Zealand. From 14 to 34 values were measured. These values were divided to get the water-use efficiencies shown. The leaf traced for the figure is 25.7 cm long and 12.7 cm wide.

and the lowest water-use efficiency. These results show the power of modern measurement techniques. We can now define with fine spatial resolution the water-use efficiency of parts of individual leaves and try to understand the physiological reasons for the differences.

Another method of determining transpiration efficiency is carbon isotope discrimination. In this method, the discrimination (Δ) of the stable isotope of carbon, ^{13}C, relative to the more abundant ^{12}C, is determined. The discrimination occurs because of the different diffusivities of ^{13}C and ^{12}C across the stomata and the fractionation by the C_3 enzyme, ribulose bisphosphate carboxylase. The method is limited to laboratories that can afford an expensive mass spectrometer with the ability to measure the ratio of ^{13}C to ^{12}C with high precision. However, the method has shown that genetic variation in transpiration efficiency in wheat cultivars does exist, and the variation is approximately twofold. The genetic variation in transpiration efficiency is negatively correlated in C_3 species with the discrimination of ^{13}C relative to ^{12}C, as shown in Figure 4. In C_4 species, Δ does not correlate with transpiration efficiency.

Possibilities of Increasing Water-Use Efficiency

With careful isotopic measurements, C_3 plants that vary slightly in water-use efficiency can be selected. If one cultivar has only a small increase in water-use efficiency compared with another and if grown over many hectares, this small difference could have a significant effect on water conservation.

Genetic engineering for increased water-use efficiency might be a possibility in the far future; for example, incorporating the C_4 photosynthetic pathway into C_3 plants. In the meantime, the increasing carbon dioxide concentration in the atmosphere seems to offer the greatest potential for increasing water-use efficiency. Leaf photosynthetic rates and leaf transpiration rates of a C_3 plant (Kentucky bluegrass) and a C_4 plant (big bluestem) grown in the field under ambient and elevated (twice ambient) levels of carbon dioxide have been measured. The C_3 plant has a photosynthetic rate similar to the C_4 plant under elevated carbon dioxide (Figure 5). The transpiration rate of the C_3 plant under the low and high levels of carbon dioxide is similar. So the leaf water-use

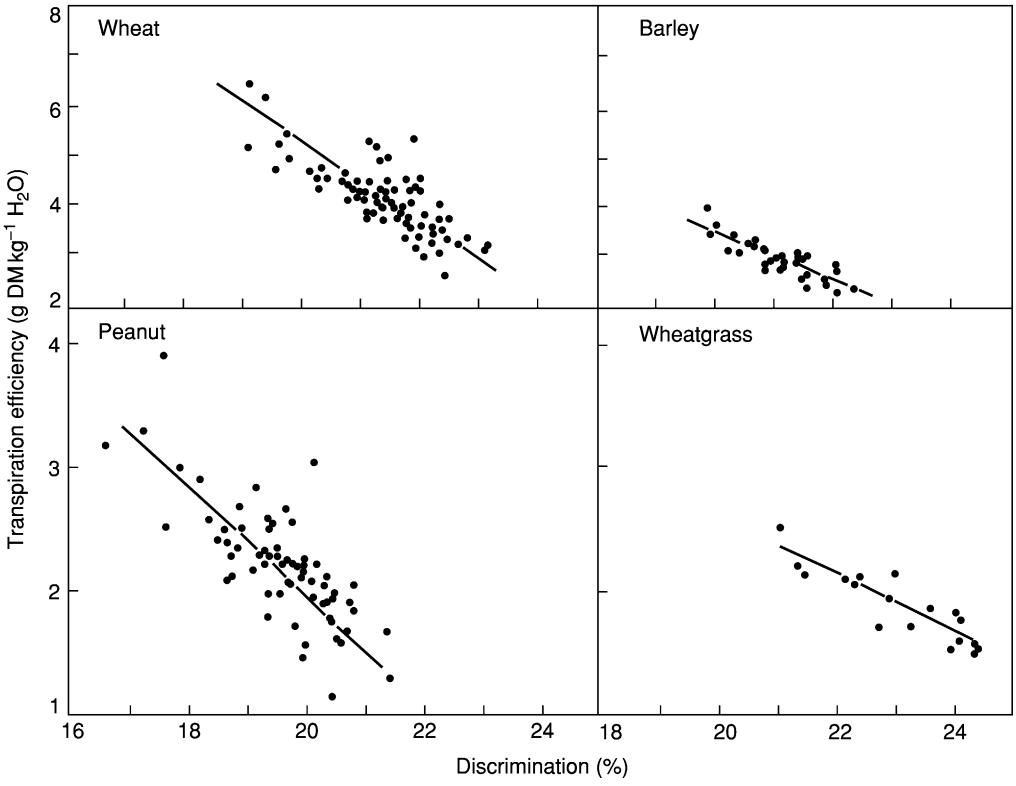


Figure 4 Relationship between transpiration efficiency and carbon isotope discrimination in genotypes of four C_3 plants: wheat, barley, peanut, and crested wheatgrass. DM, dry matter. (Reproduced with permission from Turner NC (1993) Water use efficiency of crop plants: potential for improvement. In: Buxton DR, Shibles R, Forsberg RA, *et al.* (eds) *International Crop Science I,* pp. 75–82. Madison, WI: Crop Science Society of America.)

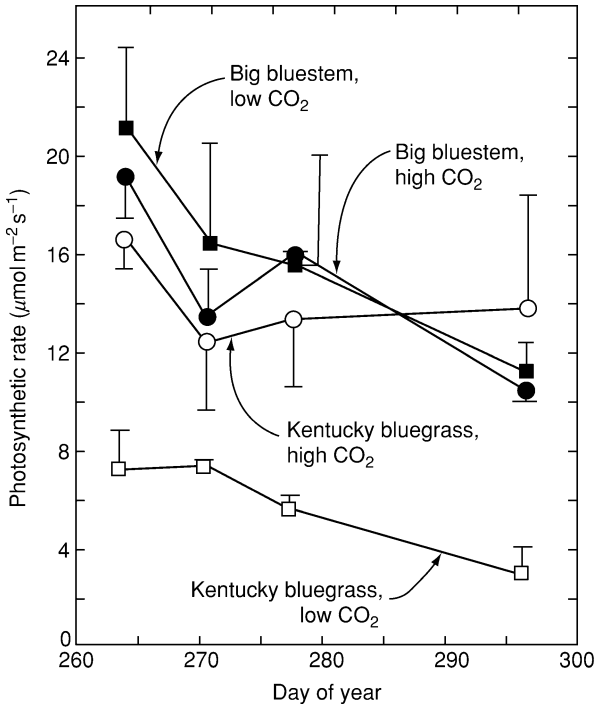


Figure 5 Photosynthetic rate of Kentucky bluegrass (open symbols) and big bluestem (closed symbols) with a high (twice ambient; circles) and a low (ambient; squares) atmospheric carbon dioxide concentration. Vertical bars, ±standard deviation. (Reproduced with permission from He H, Kirkham MB, Lawlor DJ, and Kanemasu ET (1992) Photosynthesis and water relations of big bluestem (C_4) and Kentucky bluegrass (C_3) under high concentration carbon dioxide. *Transactions of the Kansas Academy of Science* 95: 139–152.)

efficiency of the C_3 plant is increased three times due to the doubling of the ambient carbon dioxide–just about the same amount as the photosynthetic rate is increased, if the values are averaged throughout the experiment. Even though increasing carbon dioxide concentrations may have negative consequences (e.g., favoring of weed growth over crop growth), it appears that this does have the positive effect of increasing water-use efficiency.

See also: **Crop Water Requirements**; **Plant–Soil–Water Relations**; **Soil–Plant–Atmosphere Continuum**

Further Reading

Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. Irrigation and Drainage Paper No. 56. Rome, Italy: Food and Agricultural Organization of the United Nations.

Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. I. Investigations in the Great Plains in 1910 and 1911*. US Department of Agriculture, Bureau of Plant Industry Bulletin No. 284. Washington, DC: US Department of Agriculture.

Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. II. A Review of the Literature*. US Department of Agriculture, Bureau of Plant Industry Bulletin No. 285. Washington, DC: US Department of Agriculture.

de Wit CT (1958) *Transpiration and Crop Yields*. Verslag Landbouwkundig Onderzoek (Agricultural Research Report) No. 64.6. Instituut voor Biologisch en Scheikundig Onderzoek van Landbouwgewassen. Wageningen, The Netherlands: Institute of Biology and Chemical Research on Field Crops and Herbage.

Doorenbos J and Pruitt WO (1975) *Guidelines for Predicting Crop Water Requirements*. Irrigation and Drainage Paper No. 24. Food and Agricultural Organization of the United Nations. Rome, Italy: FAO.

Heschel MH, Donohue K, Hausmann N, and Schmitt J (2002) Population differentiation and natural selection for water-use efficiency in *Impatiens capensis* (Balsaminacea). *International Journal of Plant Science* 163: 907–912.

Hillel D (1994) *Rivers of Eden. The Struggle for Water and the Quest for Peace in the Middle East*. New York: Oxford University Press.

Hillel D (1998) *Environmental Soil Physics*, pp. 630–634. San Diego, CA: Academic Press.

King FH (1901) *A Text Book of the Physics of Agriculture. Strive to Know Why, for this Teaches How and When*, 2nd edn. Madison, WI: F.H. King.

Kirkham MB (ed.) (1999) *Water Use in Crop Production*. New York: The Haworth Press.

Letey J and Peters DB (1957) Influence of soil moisture levels and seasonal weather on efficiency of water use by corn. *Agronomy Journal* 49: 362–365.

Pendleton JW (1966) Increasing water use efficiency by crop management. In: Pierre WH, Kirkham D, Pesek J, and Shaw R (eds) *Plant Environment and Efficient Water Use*, pp. 236–258. Madison, WI: American Society of Agronomy/Soil Science Society of America.

Rosenberg NJ (1974) *Microclimate: The Biological Environment*. New York: John Wiley.

Salisbury FB and Ross CW (1978) *Plant Physiology*, 2nd edn. Belmont, CA: Wadsworth.

Shantz HL and Piemeisel LN (1927) The water requirement of plants at Akron, Colo. *Journal of Agricultural Research* 34: 1093–1190.

Tanner CB and Sinclair TR (1983) Efficient water use in crop production: research or re-search? In: Taylor HM, Jordan WR, and Sinclair TR (eds) *Limitations to Efficient Water Use in Crop Production*, pp. 1–27. Madison, WI: American Society of Agronomy, Crop Science Society of America/Soil Science Society of America.

Widtsoe JA (1911) *Dry-farming. A System of Agriculture for Countries Under a Low Rainfall*. New York: Macmillan.

WEED MANAGEMENT

D D Buhler, Michigan State University, East Lansing, MI, USA

Introduction

Soil tillage has been a major element of crop production for centuries. However, tillage has negative side-effects, including increased soil erosion and high labor and fuel requirements. Methods for reducing tillage were proposed as early as the 1920s. For row-crop production, no-tillage methods became much more feasible beginning in the late 1950s and 1960s as herbicides became available to replace tillage for control of unwanted vegetation.

Weed control refers to actions used to eliminate an existing weed population. Weed management is more than control of existing weed problems and places greater emphasis on preventing weed reproduction, reducing weed emergence after crop planting, and minimizing weed competition with the crop. Integrated weed management emphasizes combinations of techniques and knowledge that consider the causes of weed problems rather than react to problems after they occur. The goal of weed management is to optimize crop production and grower profit through the concerted use of preventive tactics, scientific knowledge, management skills, monitoring procedures, and efficient use of control practices.

One of the first steps in effectively managing weeds is understanding weed populations and their responses to tillage systems. As the impacts of tillage systems on weed population dynamics are examined, we must apply basic ecologic principles. Weeds are successful because of their genetic diversity and their ability to adapt to and take advantage of conditions created by crop-production systems. Therefore, modifying tillage practices results in an altered competitive environment in which the morphological and physiological traits that confer success are altered. Understanding weed population shifts leads to the identification of vulnerable stages in weed life cycles that can be exploited in management systems. Understanding population shifts also helps to identify species that are favored as tillage practices are changed, allowing for the development of weed-management systems that target the appropriate weed species.

Impacts of Tillage on Weed Management

General Effects

Crop-production practices exert selection pressure on weed communities and create niches that favor or disfavor various species. Since tillage has been an integral part of many cropping systems for centuries, it has played a major role in shaping the nature of weed communities in most agricultural lands.

Tillage buries crop residue and alters the characteristics of the surface soil, regulating the germination environment of seeds by reducing soil surface cover, altering soil temperature and moisture patterns, changing weed seed distribution in the soil, and disrupting the growth of winter annual, biennial, and perennial species. The term 'safe site' has been used to describe the complex conditions that are required for successful seed germination and seedling establishment. Several studies have pointed out that safe-site conditions vary greatly among plant species. The changes in safe-site characteristics, along with the physical disruption caused by tillage, combine to play a major role in weed-population dynamics in crop-production systems.

Control Options and Efficacy

Tillage systems impact weed-control options available to producers. Certain herbicides may not be used in conservation-tillage systems because of the need for mechanical incorporation into the soil after application. With less tillage and more plant residue on the soil surface, mechanical weed-control operations may become less effective. Rotary hoeing is especially difficult in untilled soil covered with residue of the previous crop. Interrow cultivation can be a component of some conservation-tillage systems and is an integral component of ridge-tillage. Combining interrow cultivation with reduced herbicide use has provided weed control similar to full-rate herbicide treatments in conservation-tillage systems. However, the effects of interrow cultivation on the long-term impacts of no-tillage to soil properties and soil-erosion reduction have not been documented.

Plant residue on the soil surface can alter the behavior of soil-applied herbicides. Depending on percentage surface cover, residue type, and herbicide formulation, up to 60% of the herbicide applied may be intercepted by residue. However, much of this herbicide is often washed off by subsequent

Table 1 Influence of tillage systems on control of selected summer annual weed species with soil-applied herbicides in *Zea mays* and *Glycine max* production systems

	Change relative to moldboard-plow systems[a]		
Weed species	*No-tillage*	*Chisel plow*	*Ridge-tillage*
Abutilon theophrasti	+	+, 0	+
Amaranthus retroflexus	+, 0,−	0,−	+, 0
Chenopodium album	+, 0,−	0,−	+, 0
Setaria faberi	−	−	0

[a]+, increased control relative to moldboard plow; 0, no change relative to moldboard plow;−, decreased control relative to moldboard plow.

rainfall and irrigation. The ultimate effect on weed control is determined by the timing of rainfall or irrigation relative to herbicide application, the level of residue cover, and the nature of the weed population.

Setaria faberi (giant foxtail) is more difficult to control with soil-applied herbicides in conservation- versus conventional-tillage systems (Table 1). Conversely, *Abutilon theophrasti* (velvetleaf) control increases as tillage is reduced. The effects of tillage systems on the control of small-seeded dicotyledon weed species vary by species and location. Herbicides applied after weed emergence are influenced less by tillage systems because they are applied directly to the target weeds and do not interact with residue on the soil surface.

The effect of tillage systems on herbicide efficacy is due to the combination of surface-residue interception of herbicides and changes in weed-population dynamics. For example, control of annual grass species is often reduced as tillage is reduced, while control of some annual dicot weeds increases. These variable responses indicate that choices of appropriate control tactics and application timing are critical in conservation-tillage systems. Because of the reliance on herbicides for weed control in most conservation-tillage systems, knowledge of the interactions of weed biology and ecology with herbicide activity is essential to managing weeds in conservation tillage without increasing herbicide use to environmentally and economically unacceptable levels.

Biology and Ecology of Weed Responses to Tillage Systems

Classification of Weed Species

Weeds can be classified in many ways, and understanding weed life cycles is important in predicting population shifts and designing management strategies. Weed species with different life cycles have different requirements for seed germination, seedling establishment, growth, and reproduction. Because tillage greatly alters the environment where weed species survive, altering tillage systems favors certain life cycles over others. In no-tillage systems, there has often been a marked drop in the variety of annual species. Concomitantly, there has been an increase in winter annual, biennial, and perennial weed species.

Weeds of agronomic crops are a unique group of plant species because of their ability to infest and often thrive in intensively disturbed habitats. In addition to classification by life cycle, a classification appropriate to understanding weed-population dynamics under different tillage systems follows:

1. Arable-response weed species that require periodic, regular disturbance of the soil for survival;
2. Inverse-response weed species that require undisturbed soil for survival;
3. Intermediate-response weed species that survive in both disturbed and undisturbed soil.

Combining characteristics of life cycles and tillage response of weed species is useful in predicting weed population shifts. For example, weed populations in reduced-tillage systems where summer annual crops are grown are dominated by arable- and intermediate-response summer annual species, because the soil is disturbed annually and management practices favor summer annual species. In no-tillage systems, weed populations shift toward inverse- and intermediate-response species. Winter annual, biennial, and perennial species are well adapted to no-tillage, because there is little or no tillage to interrupt the life cycle. In addition, summer annual species that do not require burial for establishment are well-adapted to no-tillage systems. However, since soil disturbance is required to plant crop seeds and incorporate fertilizers in annual cropping systems, arable-response species may continue to exist in no-tillage fields. Seed buried by previous tillage may also persist for several years and continue to be a source of arable-response species in no-tillage systems.

Summer Annual Species

Summer annual species are major weed problems in production systems dominated by summer annual crops, because they are well-adapted to the environment created by the production systems (i.e., similar life cycle, high fertility, annual disturbance, production in rows, and herbicide use). Changes that occur in the dynamics of annual weed communities as a function of tillage practices are regulated by species present in the field, biology of the species, and efficacy of weed-management practices.

The most commonly reported change in annual weed populations with decreased tillage is an increase in summer annual grass species, including *S. faberi*, *S. viridis* (green foxtail), *S. glauca* (yellow foxtail), *Panicum dichotomiflorum* (fall panicum), and *Cenchrus incertus* (field sandbur) (Table 2). The magnitude of changes in summer annual grass densities has varied among experiments, but the potential for rapid changes in density and emergence patterns of these weed species clearly exists.

Changes in the population dynamics of summer annual dicotyledon species with changing tillage practices have been less consistent. The most common response among annual dicotyledon species has been reduced density of large-seeded species such as *Ab. theophrasti*, *Xanthium strumarium* (common cocklebur), and *Cassia obtusifolia* (sicklepod) as tillage is reduced (Table 2). The response of small-seeded summer annual dicotyledon weed species to changes in tillage practices has been variable among weed species and experiments. This variable response is typified by *Chenopodium album*. In different studies at different locations, *Ch. album* densities increased, stayed the same, or decreased in response to reduced tillage. Species that germinate under cool soil conditions, such as *Ch. album*, may be prevalent in no-tillage fields prior to crop planting. Early emergence may allow such species to suppress later emerging weeds or result in a significant portion of the population being destroyed by tillage or herbicide prior to crop planting. *Amaranthus retroflexus* (redroot pigweed) is another small-seeded summer annual dicotyledon species that has responded inconsistently to changes in tillage systems. *Am. retroflexus* may be best-suited to systems with moderate tillage, because its seeds are well-adapted to shallow burial.

Table 2 Influence of tillage systems on densities of selected summer annual weed species in *Zea mays* and *Glycine max* production systems

Weed species	*Tillage response*[a]
Abutilon theophrasti	−
Amaranthus retroflexus	+, 0,−
Cassia obtusifolia	−
Cenchrus incertus	+
Chenopodium album	+, 0,−
Echinochloa crus-galli	+
Panicum dichotomiflorum	+
Setaria complex	+
Xanthium strumarium	−

[a]+, increased density with reduced tillage; 0, no change with reduced tillage;−, decreased density with reduced tillage.

Weed Seed Bank and Seedling Biology

It is often difficult to determine whether changes in population dynamics are due to biological properties of the weed species or differences in control efficacy. To address this issue, research has been conducted to separate various aspects of tillage–weed species–weed-control efficacy interactions to elucidate the mechanisms of weed population shifts.

The primary source of arable-response, summer annual weed species is the weed seed bank in the soil, with the largest source of seed being that produced in previous years in the same field. Changes in the weed seed bank occur with time and weed-management programs. Tillage is the primary cause of vertical seed movement in arable soils. Since tillage systems affect weed management, weed seed production, and soil disturbance, changing tillage systems changes the distribution and density of weed seed in agricultural soils.

Moldboard-plow plots have fewer weed seeds in the upper 20 cm of soil than chisel-plow or no-tillage plots after 5 years. Moldboard plowing results in the most uniform distribution of seed over soil depths. In a no-tillage system, more than 60% of all weed seeds are found in the upper 1 cm of soil and few seeds are found at more than 10 cm. The concentration of weed seed in no-tillage decreases logarithmically with increasing depth. In the chisel-plow system, more than 30% of the weed seeds are in the upper 1 cm of soil and seed concentration decreases linearly with depth.

Differences in emergence depths in different field tillage systems reflect differences in weed seed distribution in the soil. Mean seedling emergence depths in no-tillage are the least, followed by chisel-plow and moldboard-plow systems in two soil types (Table 3). At least 40% of the *S. faberi* and *S. viridis* plants emerge from the upper 1 cm of soil in no-tillage compared with approximately 25% in chisel-plow and less than 15% in moldboard-plow plots. Changes in seed depth in the soil and the corresponding differences in emergence depth may contribute to shifts among weed species. In greenhouse research, *Ab. theophrasti* establishment from seed germinating on the soil surface is much less than for seed planted 6 cm deep, and seedlings from surface-placed seed are less vigorous. *S. faberi* seed germinating on the soil surface has an establishment percentage similar to seed planted 1–4 cm deep, but establishment is reduced by 50% when the seeds are planted 6 cm deep.

Surface topography influences the emergence location of *Ab. theophrasti* seedlings when a field has not been tilled the previous two growing seasons. No *Ab. theophrasti* seedlings emerge from bare soil, and the majority of the *Ab. theophrasti* seedlings emerge

Table 3 Effect of tillage systems on depth of emergence of *Setaria* species under field conditions

	Depth of emergence (cm below soil surface)	
Tillage system	*Range*	*Mean*
Moldboard plow	0–10	3.1
Chisel plow	0–9	2.4
No-tillage	0–10	1.4

Source: Buhler DD (1998) Tillage systems and weed population dynamics and management. In: Hatfield JL, Buhler DD, and Stewart BA (eds) *Integrated Weed and Soil Management*, pp. 223–246. Chelsea, MI: Ann Arbor Press.

from cracks in the soil. Another significant portion of the *Ab. theophrasti* emerge from under surface residue. The presence of surface residue increases soil moisture and reduces soil crusting compared with bare soil, thus providing adequate moisture conditions for germination for a longer period of time. These data suggest that the safe-site concept should be extended below the soil surface to the depth where soil surface conditions influence germination conditions and that soil cracks may be a factor in weed emergence patterns.

The effect of tillage practices on the population dynamics of summer annual weed species is complex and involves several factors. However, seed depth in the soil appears to be the most important factor. Weed species that have the ability to germinate and become established when the seeds are at or near the soil surface have the greatest potential to increase under conservation-tillage systems. These species tend to be small-seeded and are represented by small-seeded, annual dicotyledon and most annual grass species. Deep burial of seed of small-seeded species by moldboard plowing reduces germination and emergence. Conversely, seeds of large-seeded species remain near the soil surface in conservation-tillage systems, inhibiting establishment.

Winter Annual and Biennial Species

Weed species not previously observed in fields planted to summer annual crops have rapidly appeared following elimination of preplant tillage. Species most rapidly and commonly observed are winter annual and biennial species that are characterized by an inverse tillage response. These weed species are unable to complete their life cycles in association with summer annual crops if the soil is disturbed after harvest or prior to crop planting the following spring.

Changes in soil surface characteristics and disturbance patterns in no-tillage systems create an environment where winter annual and biennial species may become established and complete their life cycle. Winter annual species germinate in the late summer and autumn, survive the winter, and flower the following spring. Because of the growth habit and ability to adjust photosynthetic responses to temperature, many winter annual species can grow even during the winter. Established winter annual plants preempt space, suppress summer annual species, and can be very competitive with summer annual crops. Rapid growth early in the spring can also make winter annual species difficult to control with herbicides if allowed to grow until or after the time of crop planting.

When the soil is tilled in the spring, rosettes and seedlings of winter annual and biennial species recruited during the previous growing season are destroyed. However, new seedlings of winter annual and biennial species may emerge under the canopy from seeds in the seed bank or that move in from adjacent areas. If tillage is removed from the system, these seedlings survive and grow during the autumn and winter and are established at the start of the second year. Thus, these species can rapidly infest fields (often during the first year) when tillage is eliminated. The rapid appearance of winter annual species does not appear to result from differences in innate competitive ability relative to summer annual species, but rather from the timing of disturbance, which changes competitive hierarchies.

Conyza canadensis (horseweed) is one of the most common and troublesome winter annual weeds in no-tillage systems in the central and eastern USA and an excellent example of an inverse-tillage response winter annual species. *Co. canadensis* can infest no-tillage fields very rapidly because it is commonly found in field edges and roadsides, produces many wind-disseminated seeds, is tolerant to many commonly used herbicides, and can become established under a wide-range of soil and climatic conditions, making it well-adapted to the no-tillage soil environment. Other winter annual species such as *Bromus tectorum* (downy brome) and *Capsella bursa-pastoris* (shepherd's purse) have also been observed to infest no-tillage fields.

Perennial Species

Perennial weed populations increase as tillage is reduced, because the underground system is not disturbed and most herbicides used to control annual weeds are not fully effective on established perennial plants. A preponderance of the literature indicates that weed flora associated with reduced-tillage systems favor perennial monocotyledon and dicotyledon species. Perennial monocotyledons might be

the greatest threat to the adoption of reduced-tillage systems because of the importance of monocotyledon crops in many regions of the world.

In most studies, more diverse and dense populations of perennial weeds develop in conservation-tillage systems than where moldboard plowing is conducted. Increases in perennial weed populations vary among sites because tillage interacts with weed-management practices, environment, and initial perennial weed populations. Therefore, the general statement that perennial weeds increase as tillage decreases must be made with caution. Management practices such as selective application of glyphosate to tall-growing perennial weeds, the use of crop cultivars resistant to herbicides such as glyphosate, and interrow cultivation help reduce perennial-species densities regardless of tillage system. Conservation-tillage systems have the potential to increase densities and species diversity of perennial weeds, but management practices such as crop rotation, selective application of glyphosate or other herbicides, and interrow cultivation can prevent perennial weed species from increasing to levels that reduce crop yields or increase production costs.

Summary

Although general trends in weed population dynamics in conservation tillage have been observed, it is important to note that location, environment, the type of conservation-tillage system used, producer management skills, and weed-management inputs regulate responses in individual fields. In addition, individual species respond differently among sites or within a site over time. This indicates that weed-management systems and cultural practices interact with tillage to regulate weed populations.

Effective and economical weed management in conservation-tillage systems, especially no-tillage, is a challenge. Due to the changes in biological and management systems, it is important to take a systematic approach to planning and executing weed management in conservation-tillage systems. Assessing the strengths and weaknesses of the available control options relative to expected weed populations is essential when designing weed-management systems.

The intensive tillage traditionally conducted in row-crop production has had a profound effect on weeds and weed management. Effective, economical, and environmentally sound weed management in conservation-tillage systems over the long term will require integration of new information with established principles of weed management. Crop rotation, tillage rotation, judicious use of tillage, herbicide-resistant crop cultivars, and innovative application methods for herbicides will aid weed management in conservation-tillage systems. New technologies are also needed to deal with the altered agroecosystems created by conservation-tillage production systems. New control options such as biological agents and allelopathic cover crops can provide useful new management tools. Many weed species and weed-control tactics behave differently as tillage is reduced or eliminated. These changes must be taken into consideration to develop economically and environmentally sound weed-management systems. It is essential that crop producers and pest managers consider the idiosyncrasies of the biological interactions among weeds, crops, residues, and environment as they develop weed-management plans for conservation-tillage systems.

List of Technical Nomenclature

Abutilon theophrasti Medikus	velvetleaf
Amaranthus retroflexus L.	redroot pigweed
Bromus tectorum L.	downy brome
Capsella bursa-pastoris (L.) Medikus	shepherd's purse
Cassia obtusifolia L.	sicklepod
Cenchrus incertus M.A. Curtis	field sandbur
Chenopodium album L.	common lambsquarters
Conyza canadensis (L.) Cronq.	horseweed
Glycine max (L.) Merr.	soybean
Glyphosate	*N*-(phosphonomethyl)glycine
Panicum dichotomiflorum Michx.	fall panicum
Setaria faberi Herrm.	giant foxtail
Setaria glauca (L.) Beauv.	yellow foxtail
Setaria viridis (L.) Beauv.	green foxtail

Xanthium strumarium L.	common cocklebur
Zea mays L.	corn

See also: **Conservation Tillage**; **Cover Crops**; **Crop Rotations**; **Crop-Residue Management**; **Cultivation and Tillage**; **Pesticides**

Further Reading

Bazzaz FA (1990) Plant–plant interactions in successional environments. In: Grace JB and Tilman D (eds) *Perspectives on Plant Competition*, pp. 239–263. San Diego, CA: Academic Press.

Buhler DD (1998) Tillage systems and weed population dynamics and management. In: Hatfield JL, Buhler DD, and Stewart BA (eds) *Integrated Weed and Soil Management*, pp. 223–246. Chelsea, MI: Ann Arbor Press.

Buhler DD (ed.) (1999) *Expanding the Context of Weed Management*. New York: Food Production Press.

Buhler DD, Liebman M, and Obrycki JJ (2000) Theoretical and practical challenges to an IPM approach to weed management. *Weed Science* 48: 274–280.

Derkson DA, Lafond GP, Thomas AG, Loeppky HA, and Swanton CJ (1993) Impact of agronomic practices on weed communities: tillage systems. *Weed Science* 41: 409–417.

Forcella F, Buhler DD, and McGiffen (1994) Pest management and crop residues. In: Hatfield JL and Stewart BA (eds) *Crop Residue Management*, pp. 173–189. Boca Raton, FL: Lewis Publishers.

Froud-Williams RJ, Chancellor RJ, and Drennen DSH (1981) Potential changes in weed flora associated with reduced-cultivation systems in cereal production in temperate regions. *Weed Research* 21: 99–109.

Harper JL (1977) *Population Biology of Plants*. New York: Academic Press.

Liebman M, Mohler CL, and Staver CP (2001) *Ecological Management of Agricultural Weeds*. New York: Cambridge University Press.

Sims GK, Buhler DD, and Turco RF (1994) Residue management impact on the environment. In: Unger PW (ed.) *Managing Agricultural Residues*, pp. 77–98. Boca Raton, FL: Lewis Publishers.

Sprague MA and Triplett GB (1986) *No-Tillage and Surface Tillage Agriculture*. New York: John Wiley.

WETLANDS, NATURALLY OCCURRING

E K Hartig, New York, NY, USA

Introduction

Wetlands are areas that are located where land and water meet, where the habitat is neither aquatic nor fully terrestrial, the soil being for the most part too wet and oxygen-depleted for any but specially adapted plants to survive. Wetlands occur along lakes, creeks, rivers, bays, and deltas; they are found on every continent except Antarctica, in the low-lying peatlands of Indonesia and Siberia as well as in high-altitude montane depressions. Worldwide estimates vary somewhat; e.g., estimated freshwater coverage of 5.3 million km^2, based on soil maps prepared by the Food and Agricultural Organization of the United Nations (FAO); and the commonly used estimate of 8.6 million km^2. Some examples of huge contiguous wetlands are the Pantanal in parts of five countries in South America, the Sudd Marshes of the Sudan along the White Nile River, and the permafrosted peatlands in the northern high latitudes of Canada, Alaska, and Siberia. Examples of very small but highly productive wetlands can be found in the Prairie Pothole region in North Dakota, Minnesota, and parts of Canada. These comprise a vast complex of tens of thousands of individual shallow wetlands that vary in size, some being as small as 0.1 ha.

Historically wetlands have been treated as wastelands to be drained and farmed, or filled and built upon. Traditional attitudes are exemplified by such expressions as 'bogged down,' 'swamped,' 'muddied,' 'mired,' 'murky,' 'morass,' 'malarial,' and 'quagmire.' Great drainage schemes have long dominated approaches to wetlands worldwide. An example of a government works program was the mosquito ditching conducted in the USA during the 1930s by the Work Projects Administration. Most US East Coast marshes bear the scars of these labor-intensive ditching projects. More than 50% of the wetlands in the USA have been drained since colonial times. In the UK more than 90% of present-day agricultural lands were formerly wetlands. Many wetlands remain under threat either from direct human interference or from indirect causes such as sea-level rise inundating coastal areas, or changes in hydrological regimes due to upstream alterations for dams, dikes, or levees. The swamps of southern Sudan are among the largest of the world's wetlands and home to one of the greatest

concentrations of large mammals (such as the buffalo, elephants, hippopotami, gazelles, waterbucks, and zebra). These wetlands are threatened by the planned diversion of the White Nile.

In recent decades, the perception of wetlands has changed, and there is increasing recognition of the many vital functions provided by wetlands. These include providing wildlife habitat, flood and erosion control, aesthetics and recreation, carbon sequestration, and high biological productivity.

Wetland Definitions

Wetland definitions used by scientists and by administrative agencies commonly refer to vegetation types, soils, and/or hydrology. Administrative agencies use variations of the three-parameter approach (vegetation, soils, hydrology). The definition developed by the US Fish and Wildlife Service (USF-WS) and used by the US Army Corps of Engineers (Corps) for delineating wetland boundaries under Corps jurisdiction states:

> Wetlands are areas that are inundated or saturated by surface or groundwater at a frequency and duration sufficient to support, and that under normal circumstances do support, a prevalence of vegetation typically adapted for life in saturated soil conditions. Wetlands generally include swamps, marshes, bogs, and similar areas. (Definition of Waters of the United States; 33CFR328.3(b); 1984)

Wetland Classification

Wetland classifications abound, but usually they are divided according to their relation to water bodies, vegetation types, and substrates. The widely used Cowardin classification takes into account wetlands that can be found in five environments. Of these, Marine, Riverine (along freshwater nontidal rivers and streams), and Lacustrine (along lakes) wetlands are mainly aquatic and have little or no soil development. Estuarine (tidal embayments and tidal portions of rivers) and Palustrine (inland freshwater forested and nonforested) wetlands contain hydric soils.

Estuarine Wetlands

This category includes intertidal salt and brackish low and high marsh zones, nonvegetated tidal flats, brackish waters of coastal rivers and embayments, and mangrove swamps:

1. Estuarine emergent wetlands: These are wetlands dominated by herbaceous grasses and are referred to as salt marsh, low marsh (intertidal marsh), high marsh, and brackish tidal marsh. They are found in protected embayments or behind barrier beaches where the water has slowed. Many, such as the *Spartina alterniflora* intertidal marshes of the northeastern USA, formed within the last 4000–7000 years as the postglacial rise in sea level slowed. Typically there is strong zonation of species in the salt marsh, depending on the frequency and duration of tidal ebb and flow (Figure 1);
2. Estuarine scrub–shrub wetlands: These wetlands contain woody vegetation such as marsh elder or high-tide bush. With exposure to salt spray and infrequent flooding they are adapted to a high-salt environment. This category also includes mangrove swamps.

Palustrine Wetlands

This category includes the majority of freshwater wetlands. It encompasses marshes, bogs, swamps (forested), and ponds. Shallow-water bodies such as

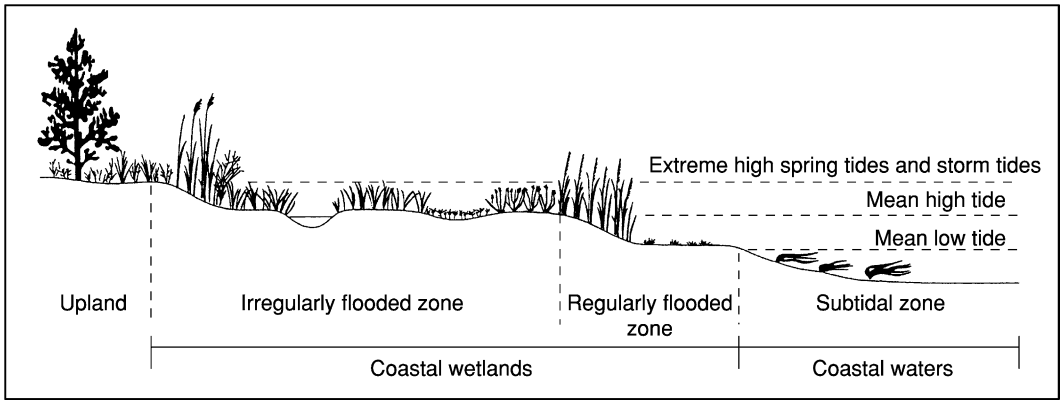


Figure 1 Hydrology of a tidal wetland indicating different zones of flooding. The regularly flooded zone is subject to at least once-daily tidal inundation, while the irregularly flooded zone is inundated less frequently. (Reproduced from Tiner RW and Burke DG (1995) *Wetlands of Maryland*. USFWS, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.)

vernal ponds and playa lakes (less than 2 m deep) are also in this category:

1. Palustrine emergent wetlands: These wetlands are dominated by herbaceous vegetation with many species of grasses, rushes, and sedges. They are referred to as freshwater marshes, wet meadows, and fens. The marshes may be irregularly to permanently flooded. In the US Prairie Pothole region, glacial depressions flood, especially during the spring season following snowmelt (and offer a protein-rich diet to migrating waterfowl). Palustrine emergent wetlands occur in a variety of habitats, including peatlands and freshwater marshes;
2. Palustrine scrub–shrub wetlands: These are characterized by woody vegetation less than 6 m tall. North American bogs with vegetation such as bog laurel, Labrador tea, cranberry, and leatherleaf, and with stunted trees such as black spruce, larch, and balsam fir are examples of scrub–shrub wetlands;
3. Palustrine forested wetlands: These are wetlands with trees greater than 6 m in height. Tree species found in northeastern US wetlands include red maple, pin oak, sweet gum, black spruce, and larch.

Hydric Soils

The definition of a hydric soil is as follows:

> A hydric soil is a soil that is saturated, flooded, or ponded long enough during the growing season to develop anaerobic conditions in the upper part of the soil.

In organic wetland soils, a thick organic layer develops over time where microbial decomposition is severely slowed or remains incomplete under waterlogged conditions. Organic wetland soils develop in depressions or in coastal areas where anaerobic conditions prevail (such as in bogs or intertidal marshes). Included are mucks (saprists), peats (fibrists), and hemists – mucky peat or peaty muck. The accumulated peat or muck layers are 0.6 to more than 9 m in depth. While organic matter is present in all soils, it constitutes greater than 20% (based on dry weight) in organic soils. These soils also have a lower bulk density and higher water-holding capacity than mineral soils.

Mineral wetland soils form either in moving or periodically standing water and can be found in low-lying depressions or slopes (Figure 2). They may

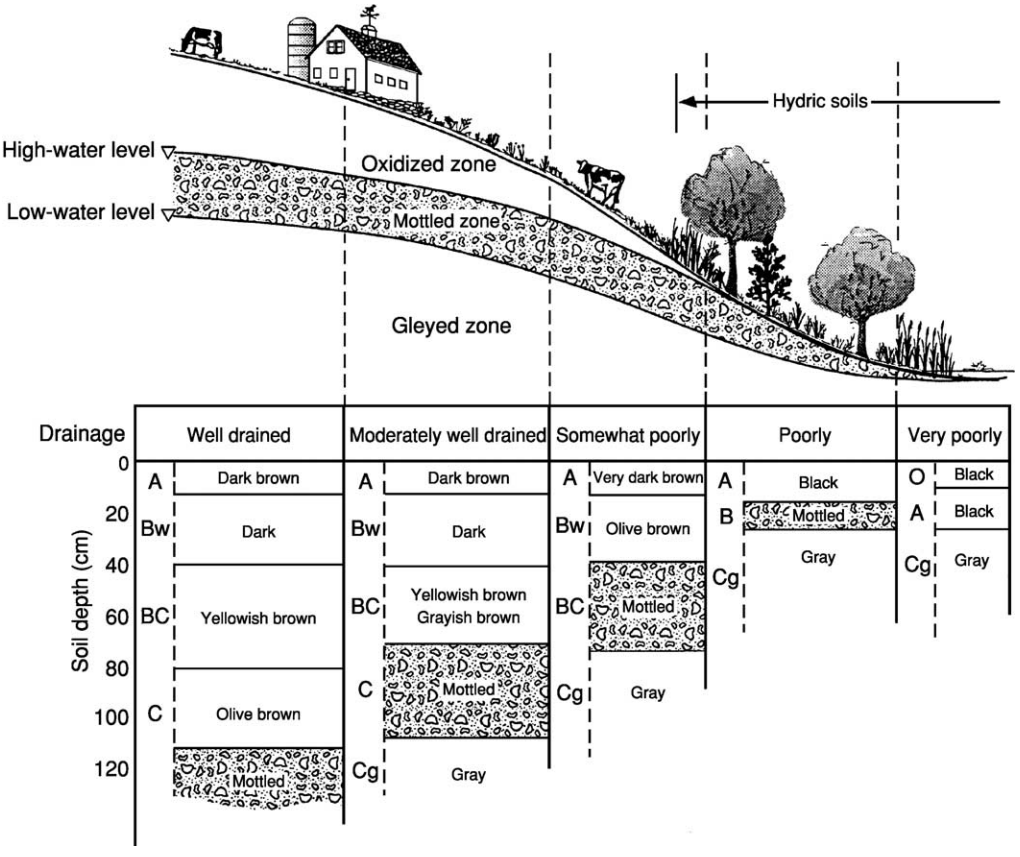


Figure 2 Soil properties are shown along a wetland-to-upland landscape gradient. Note that the seasonal high-water table, indicated by the mottled zone, reaches the surface only at the lower elevations. Reproduced from Tiner RW and Burke DG (1995) *Wetlands of Maryland*. USFWS, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.

receive surface water runoff or groundwater seepage, and there may be the presence of an only slowly permeable soil layer (e.g., a fragipan, clay, or hardpan layer, or confining bedrock) that slows the movement of water. Criteria used to determine whether a soil is hydric are listed in Table 1. Mineral wetland soils have a dry weight consisting of less than 20–35% organic material.

Observable indications (redoximorphic features) of mineral wetland soil include its dominant 'matrix' color chroma that can vary from very dark to washed-out blue-gray or green-gray colors (with often brighter colors found as minor components, as 'mottles'). Mottles are yellow-orange to brown accumulations that form where oxygen-rich zones are encountered, and iron and manganese oxidize and precipitate from solution. These can coat soil peds, walls of voids, or the root channels (rhizospheres), where they can be observed readily. The extent of gleying and mottle abundance, size, and color depends on the prevailing hydrology (e.g., fluctuating water table) and the extent of microbial activity: if it is too cold, then there is little activity and wetland indicators such as gleying and mottling will not be present.

One way to determine whether a soil is hydric is by using a standard Munsell color chart to identify its color strength. Low chromas, generally a 2 or less, indicate hydric conditions. The dark matrix color is the result of prolonged saturation that converts iron from its rust-colored, oxidized (ferric) form to its reduced (ferrous) form (see Iron and Manganese, below). Exceptions occur where hydric color indicators are masked, such as with soils that remain redder, in spite of hydric conditions, having formed from red glacial till material. In these cases other indicators such as iron or manganese concretions and the oxidized rhizospheres can be observed.

Redoximorphic features such as mottling, gleying, and oxidized rhizospheres that develop in hydric soils generally remain visible even after wetlands have been disturbed and thus remain useful in determining former wetland extent.

Table 1 Criteria for hydric soils

Description
1 All Histels except Folistels and all Histosols except Folists
2 Soils in aquic suborders, great groups, or subgroups, Albolls suborder, Historthels great group, Histoturbels great group, Pachic subgroups, or Cumulic subgroups that are: Somewhat poorly drained with a water table equal to 0.0 m from the surface during the growing season or Poorly drained or very poorly drained and have either: Water table equal to 0.0 m during the growing season if textures are coarse sand, sand, or fine sand in all layers within 50 cm, or, for other soils, Water table at less than or equal to 15 cm from the surface during the growing season if permeability is equal to or greater than 15 cm h^{-1} in all layers within 50 cm, or Water table at less than or equal to 30 cm from the surface during the growing season if permeability is less than 15 cm h^{-1} in any layer within 50 cm, or
3 Soils that are frequently ponded for long duration or very long duration during the growing season, or
4 Soils that are frequently flooded for long duration or very long duration during the growing season

Source: USDA/NRCS Website http://soils.usda.gov

Soil Chemistry

At the soil surface, a thin layer of oxidation is crucial in maintaining wetland function. Chemical transformations and nutrient cycling depend on the oxidized ions found in this layer. These oxidized ions are Fe^{3+}, Mn^{4+}, NO_3^-, SO_4^{2-}. Below the oxidized layer, and below the water table where anaerobic conditions dominate, the reduced forms of compounds are found such as Fe^{2+}, Mn^{2+}, NH_4^+, S^{-2}, and CH_4. Table 2 lists the oxidized and reduced forms of these elements.

Nitrogen

As in many ecologic communities, nitrogen is one of the most limiting of nutrients in many wetlands. Much of the nitrogen in soil is derived from dead organic material such as proteins, amino acids, and nucleic acids. These nitrogen-rich compounds decompose into simple compounds by soil-dwelling saprobic bacteria and fungi. These microbes then release the excess nitrogen in the form of ammonium ions (NH_4^+). Figure 3a shows the nitrogen cycle: after ammonium diffuses to the aerobic soil layer, nitrification occurs, converting it to nitrate. The nitrate may

Table 2 Oxidized and reduced forms of several elements in order of their redox potential

Element	*Oxidized form*	*Reduced form*
Nitrogen	NO_3^- Nitrate	N_2O nitrous oxide, N_2 Nitrogen gas, NH_4 Ammonium
Manganese	Mn^{4+} Manganic	Mn^{2+} Manganous
Iron	Fe^{3+} Ferric	Fe^{2+} Ferrous
Sulfur	SO_4^{2-} Sulfate	S^{2+} Sulfide
Carbon	CO_2 Carbon dioxide	CH_4 Methane

Adapted from Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley, with permission.

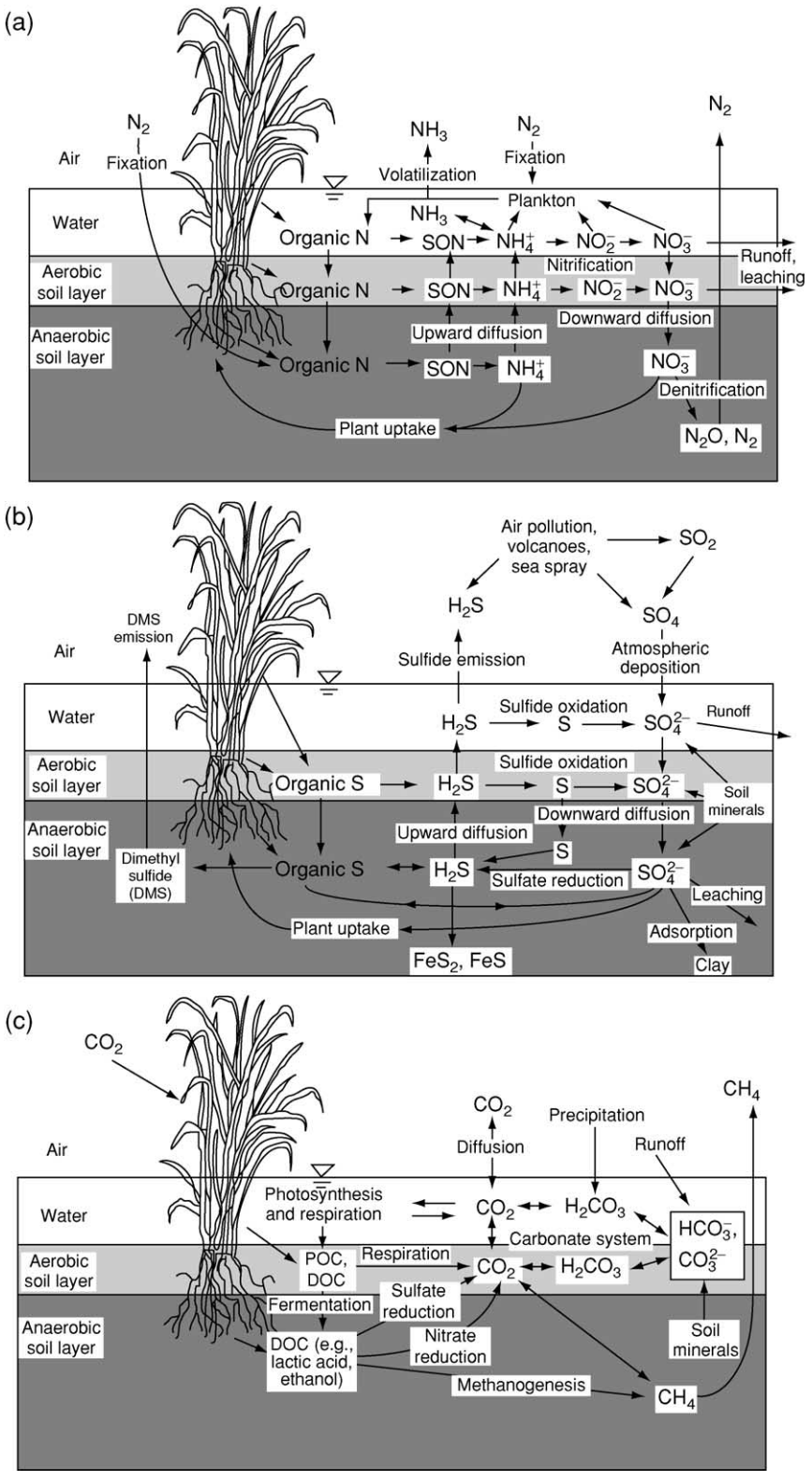


Figure 3 Chemical transformations in wetlands: (a) nitrogen transformations in wetlands. SON, soluble organic nitrogen; (b) sulfur transformations in wetlands; (c) carbon transformations in wetlands. POC, particulate organic carbon; DOC, dissolved organic carbon. Reproduced with permission from Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley.

then diffuse back into the anaerobic layer where denitrification transforms it into nitrogen gas. Ammonium nitrogen is oxidized to nitrite ions (NO_2^-) through nitrification:

$$2NH_4^+ + 3O_2 \rightarrow 2NO_2^- + 2H_2O + 4H^+ + \text{energy}$$

Nitrite is toxic to higher plants but it does not have a chance to accumulate in the soil owing to the presence of *Nitrobacter*, a genus of bacteria that oxidizes the nitrite to form nitrate ions (NO_3^-), again with a release of energy:

$$2NO_2^- + O_2 \rightarrow 2NO_3^- + \text{energy}$$

Nitrification in wetland soils also takes place within the oxidized rhizospheres of plant roots where sufficient oxygen may facilitate the conversion of NH_4^+ to NO_3^-.

Nitrate can also undergo reduction reactions. Nitrate being more mobile, if not rapidly assimilated by plants or microbes, or lost through groundwater flow, will probably be reduced to ammonium, nitrous oxide (N_2O), or molecular nitrogen (N_2) under anaerobic conditions. Denitrification typically occurs under waterlogged conditions in swamps and marshes. The availability of fresh organic matter together with denitrifying bacteria promotes dentrification:

$$C_6H_{12}O_6 + 4NO_3^- \rightarrow 6CO_2 + 6H_2O + 2N_2$$

In the highly acidic peat soils of northern bogs, nitrifying bacteria are lacking due to the low pH. In such environments carnivorous plant species have evolved that are able to use animal proteins directly as their nitrogen source. The sundew (*Drosera intermedia*) and the Venus flytrap (*Dionaea muscipula*) both attract insects and obtain minerals, including fixed nitrogen from animal prey via sticky hairs that contain digestive enzymes.

Iron and Manganese

Both iron and manganese occur in two oxidation states. Under the conditions that prevail in wetlands, they exist primarily in reduced form. Iron is transformed from ferric (Fe^{3+}) iron compounds to ferrous (Fe^{2+}) iron:

$$Fe_2O_3 + 4e^- + 6H^+ \rightarrow 2Fe^{2+} + 3H_2O$$

Manganese is transformed from manganic to manganous compounds:

$$MnO_2 + 2e^- + 4H^+ \rightarrow Mn^{2+} + 2H_2O$$

Ferrous iron causes gleying of mineral soils, the gray colors observable in highly reduced soils. In the presence of oxygen, when, for example, the water table drops, the reaction can be reversed and insoluble ferric oxides precipitate out. These reactions can be mediated through bacterial activity, forming 'bog iron' ore. Bog iron deposits found in anaerobic groundwater in northern peatlands have been used in iron and steel industries.

Sulfur

Sulfur has several oxidation states and is transformed through microbial activity. Sulfur is usually readily available and so not limiting to organisms. Elevated concentrations of hydrogen sulfide in saltwater wetlands may be toxic to plants and animals. When wetland sediments are disturbed, the smell of rotten eggs is emitted, a smell that is familiar to those working in wetlands. Sulfur-reducing bacteria in anaerobic soils reduce sulfates to sulfides:

$$4H_2 + SO_4^{2-} \rightarrow H_2S + 2H_2O + 2OH^-$$

As shown in Figure 3b, sulfur can be released to the atmosphere as methyl and dimethyl sulfide, organic sulfur compounds, and hydrogen sulfide. Among wetlands, salt marshes have the highest rate of emission of hydrogen sulfide per unit area. The dark color of many wetland peats comes from ferrous sulfide. Sulfur found in peat bogs and coal deposits usually occurs as the mineral pyrite (FeS_2).

A number of sulfur reactions occur in the upper aerobic portion of the wetland soil. *Thiobacillus* obtains energy from the oxidation of hydrogen sulfide to sulfur. Other species further oxidize sulfur to sulfate:

$$2H_2S + O_2 \rightarrow 2S + 2H_2O + \text{energy}$$
$$2S + 3O_2 + 2H_2O \rightarrow 2H_2SO_4 + \text{energy}$$

Purple sulfur bacteria found in salt marshes and mudflats produce organic matter in the presence of light:

$$CO_2 + H_2S + \text{light} \rightarrow CH_2O + S$$

Hydrogen sulfide is used as an electron donor instead of water. This reaction requires anaerobic conditions, ample hydrogen sulfide, and sufficient sunlight near the surface.

Carbon

Major processes for carbon transformations are given in Figure 3c for both aerobic and anaerobic conditions. Through fermentation, sugars are transformed to lactic acid, ethanol, and CO_2. Methane production (methanogenesis) is the result of extremely reduced conditions whereby bacteria (methanogens) reduce CO_2 to CH_4, as in:

$$CO_2 + 8e^- + 8H^+ \rightarrow CH_4 + 2H_2O$$

The resulting methane is often called 'swamp gas' or 'marsh gas.' It is known for being highly combustible especially in peatland environments. Methane production is higher in freshwater than in marine wetlands but has a broad productivity range in each. It is season-dependent, taking place only when warm temperatures prevail. Rice paddies are a major source of methane to the atmosphere. Peatlands are usually a net source of methane; however, under drought conditions, when oxidation occurs below the surface layers, the peat can become a source of CO_2 emissions.

Table 3 US (48 states) wetlands and former wetlands (thousands of hectares) where land use has been determined (nonwetland hydric soils indicate former wetland areas)

1992 land use	*Wetland soils (ha)*	*Nonwetland hydric soils (ha)*	*Total (ha)*
Cropland	4270	22 430	26 699
Pastureland	3232	2611	5843
Rangeland	3142	2426	5568
Forest land	24 732	3829	28 561
Misc.	9753	2501	12 254
Total	45 129	33 797	78 925

Adapted from Economic Research Service/USDA compilation of 1992 National Resources Inventory, Wetlands and Agriculture: Private Interests and Public Benefits/AER-765. http://www.ers.usda.gov with permission.

Policies, Regulations, and Protection

There are several international treaties that protect wetlands. The first was the Ramsar Convention. The role of wetlands to support migratory waterfowl that cross national boundaries was recognized when the Convention on Wetlands (originally entitled the Convention on Wetlands of International Importance Especially as Waterfowl Habitat) was signed in Ramsar, Iran, in 1971. It is an "intergovernmental treaty which provides the framework for national action and international cooperation for the conservation and wise use of wetlands and their resources." There are 136 signatory nations, and, as of May, 2003, 1283 wetlands sites are listed, comprising 108.7 million hectares. Often the wetlands encompass tracts with multiple uses, including water supply, fisheries, agricultural use, and more. The number of sites and their sizes vary from country to country. The UK has more than 150 sites, whereas the USA has 17 sites. Australia lists sites from 1 ha (Christmas Island Territory) to almost 2 Mha in size in the southern Coongie Lakes. Many countries use the Ramsar Convention as their main instrument for wetland protection.

The USA relies mainly on separate federal, state, and local laws and regulations to limit development in wetlands. In the 1970s the Federal Water Pollution Act was followed by the Clean Water Act of 1977, Section 404, that considered wetlands as waters of the USA covered in part under the Rivers and Harbors Act of 1899. Hence jurisdiction of private property could be given to public agencies as waters are considered a public good. Under Section 404 a permit is needed to conduct dredge-and-fill activities in jurisdictional wetlands even when on private property. The 1985 Food Security Act's 'Swampbuster' provisions denied government farm subsidies to farmers who drained wetlands after December 23, 1985.

In the USA, wetland inventories are conducted every 10 years and their status and trends are compiled. Wetlands comprise approximately 91 million hectares in these 48 contiguous states. Of these, land-use categories of 79 million hectares can be documented through knowledge of soil types (Table 3). Table 3 indicates that about 34 million hectares are no longer wetlands but retain hydric soil characteristics.

Where once the US government actively encouraged drainage, with cost-sharing and expertise, the paradigm has shifted. One of the more successful federal programs, the Wetland Reserve Program (WRP), promotes wetland restoration and offers cost-sharing and expertise to break drainage tiles and return areas to wetlands. Every 5 years Congress sets policies, guidelines, and funding for such agriculture programs. The WRP offers financial incentives and expertise from the USFWS and Natural Resources Conservation Service (NRCS) to take marginal land out of production.

Conserving wetlands is increasingly recognized as beneficial for flood control in a range of rural to urban areas, and as essential habitats for many fish, bird, and other species. In spite of conflicts that occur with pressure from expanding populations, local, state, and national regulations, as well as international agreements, have slowed the drainage of wetland areas.

See also: **Anaerobic Soils**; **Carbon Emissions and Sequestration**; **Hydric Soils**; **Land-Use Classification**; **Paddy Soils**; **Sulfur in Soils:** Biological Transformations

Further Reading

Cowardin LM, Carter V, Golet FC, and LaRoe ET (1979) *Classification of Wetlands and Deepwater Habitats of the United States*. Fish and Wildlife Service, FWS/OBS-79/31. Washington, DC: US Government Printing Office.

Crum HA (1988) *A Focus on Peatlands and Peat Mosses*. Ann Arbor, MI: University of Michigan Press.

Dahl TE (1990) *Wetlands Losses in the United States, 1780s to 1980s*. US Department of the Interior, Fish and Wildlife Service. Washington, DC: US Government Printing Office.

Diers R and Anderson JL (1984) Development of soil mottling. *Soil Survey Horizons* winter: 9–12.

Gambrell RP and Patrick WH Jr (1978) Chemical and microbiological properties of anaerobic soils and sediments. In: Hook KK and Crawford M (eds) *Plant Life in Anaerobic Environments*, pp. 375–423. Ann Arbor, MI: Ann Arbor Scientific Publishing.

Heimlich RE, Wiebe KD, Claassen R, Gadsby D, and House RM (1998) *Wetlands and Agriculture: Private Interests and Public Benefits*. Agriculture Economics report no. 765. Available online at: www.ers.usda.gov/publications/aer765/ (see chapter entitled Wetland Status and Trends, Settlement to 1992, 117 Kb).

Hillel D (1991) Endangered wetlands. In: *Out of the Earth: Civilization and the Life of the Soil*, pp. 215–224. Berkeley, CA: University of California Press.

Hillel D (1994) *Rivers of Eden*. New York: Oxford University Press.

Hurt GW, Whited PM, and Pringle RF (eds) (2003) *Field Indicators of Hydric Soils in the United States,* Version 5.01. USDA–NRCS National Technical Committee for Hydric Soils. Fort Worth, TX: US Government Printing Office.

Matthews E and Fung I (1987) Methane emission from natural wetlands: global distribution, area, and environmental characteristics of sources. *Global Biogeochemical Cycles* 1: 61–86.

Mitsch WJ and Wu X (1995) Wetlands and Global Change. In: Lal R, Kimble J, Levine E, and Stewart BA (eds) *Soil Management and Greenhouse Effect*, pp. 205–230. Boca Raton, FL: CRC Press.

Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley.

Redfield AC (1972) Development of a New England salt marsh. *Ecological Monographs* 42: 201–237.

Richardson JL and Vepraskas MJ (2000) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: CRC Press.

Thomas E and Varekamp JC (1991) Paleo-environmental analyses of marsh sequences (Clinton, Connecticut): evidence for punctuated rise in relative sea-level during the latest Holocene. *Journal of Coastal Research* special issue 11: 125–158.

Tiner RW (1984) *Wetlands of the United States: Status and Trends*. National Wetland Inventory, Fish and Wildlife Service, US Department of Interior. Washington, DC: US Government Printing Office.

Tiner RW and Burke DG (1995) *Wetlands of Maryland*. US Fish and Wildlife Service, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.

US Army Corps of Engineers (1987) *Corps of Engineers Wetlands Delineation Manual*. Technical Report Y-87-1. Washington, DC: US Government Printing Office.

WIDTSOE, JOHN A. AND GARDNER, WILLARD

G S Campbell, Decagon Devices, Inc., Pullman, WA, USA
W H Gardner, Washington State University, Pullman, WA, USA

Soil physics can be defined as the study of the state and movement of energy and mass in the continuum that includes the soil, plants, and surface boundary layer of the atmosphere. Since soil physics is both a theoretical and an experimental science, it includes equations from physics, and knowledge and measurement tools built up through years of laboratory and field experience. Understanding the state and movement of water in soil is an important part of the overall body of soil physics knowledge.

The relationship between water flux and driving force:

$$J_w = -k\nabla\psi$$

called Darcy's law, was already well established for saturated flow of water in porous materials before the beginning of the twentieth century. Darcy recognized that the driving force for flow was a gradient in what we now call the pressure and gravitational potentials. These ideas were extended to unsaturated flow in the early part of the twentieth century by Edgar Buckingham. He recognized that water in unsaturated soil is under negative pressure, i.e., the matric or capillary potential. Buckingham also recognized that gradients in matric potential, as well as in the other potentials, result in water flow. He generalized Darcy's equation by allowing k to vary as a function of the matric potential. While the second laws of both Fick and Fourier (combining the flux equation with the continuity equation to obtain the second-order partial-differential equation) were well known at the beginning of the twentieth century, they had not been extended to water flow. By the end of the twentieth century, the second-order equation for water was well known, with both analytical and numerical solutions available for a wide range of theoretical and practical problems.

Buckingham had no direct way to measure matric potential. He attempted to establish equilibrium in vertical soil columns and then related the water content of the soil in the columns to the height above a free water surface at the bottom of the column. Water content was measured by weighing and oven-drying the soil. By the end of the twentieth century, one could measure soil-water content using neutron scattering, neutron and gamma attenuation, time-domain reflectometry, capacitance, and microwave-drying. Matric potential could be measured using tensiometers, gypsum blocks, heat-dissipation matric probes, thermocouple psychrometers, filter paper, and freezing point. The relationship between water content and water potential is now routinely determined using a pressure plate.

Many of the important soil physics developments of the twentieth century, both in theory and in measurement methods, have direct or indirect ties to the laboratory of Willard Gardner, a professor of physics at Utah State Agricultural College from 1918 to 1948. While Willard Gardner made important contributions to soil physics in his own right, his influence has been amplified many-fold through the disciples he converted and trained. It is an interesting story of science on the frontier, but the story starts much earlier, in the mid nineteenth century.

On July 21, 1847, Orson Pratt and Erastus Snow entered the valley of the Great Salt Lake, the first of the vanguard group of Mormon (Church of Jesus Christ of Latter-Day Saints) pioneers fleeing west to escape religious persecution. Their only hope was to plant crops, even though it was late in the season. However, it was impossible to plow the hard-baked valley soil, so they channeled water from the mountain streams on to the thirsty ground. This was the beginning of modern irrigation in the US Mountain West. Since irrigation was necessary to produce almost all crops in the arid mountain valleys, the practice became widespread. Pratt was a physicist and mathematician who lectured at the University of Deseret (later the University of Utah), but he apparently took little interest in the physics of water flow, preferring astronomy and mathematics instead.

The practical aspects of designing irrigation systems and managing water flow in soil were a part of ordinary life in the frontier towns of the Great Basin, but the scientific study of irrigation in Utah did not begin until 50 years later. It was started by John A. Widtsoe soon after he became Director of the Utah Agricultural Experiment Station in 1900 (Figure 1).

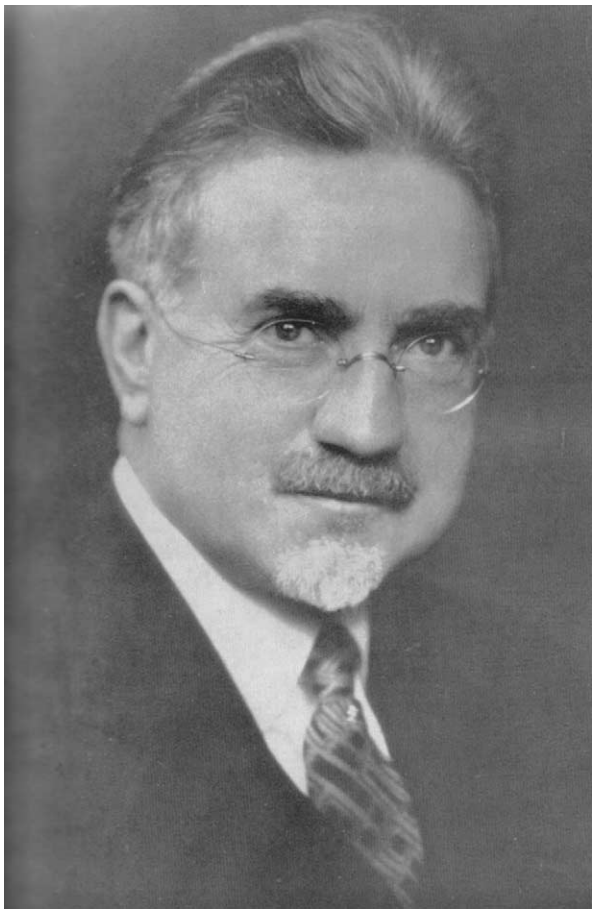


Figure 1 John A. Widtsoe (1872–1953).

Widtsoe was born in Daloe, Norway, in 1872, the oldest son of John Andersen Widtsoe, a schoolmaster, and Anna Karine Gaarden. When John was 6 years old, his father died, leaving his mother, himself, and his 2-month old brother, Osborne. When Widtsoe was 11 years old, he moved with his mother and brother to Utah, and settled in the town of Logan. He worked in Logan for several years to help support the family, but finally graduated from Brigham Young College (a junior college) in 1891. He was then selected as one of six talented Utah students to attend Harvard University, in Cambridge, Massachusetts, where he majored in chemistry with minors in physics and metallurgy, graduating in 1894. On his return to Utah, he joined the faculty of the new Utah State Agricultural College in Logan, where he began studying ways to improve agricultural production. In 1898, he and his new wife, Leah Eudora Dunford, went to Germany for his graduate study. In less than two years, he earned his doctoral degree in chemistry at the University of Goettingen and completed additional studies at Berlin and Zurich.

While still in Europe, Widtsoe was offered the position of Director of the Experiment Station at the Utah State Agricultural College. He served as Director from 1900 to 1905. The following excerpt from

Widtsoe's autobiography indicates the kind of research he and his colleagues undertook:

> By the planting season of 1901, a part of the College farm was supplied with troughs or flumes, by which measured quantities of water could be applied at will to a series of plots planted to various crops. Later, another and more suitable farm was purchased and equipped more extensively for the controlled application of irrigation water to the experimental plots. A vegetation house was also built, housing a large number of tanks, which could be wheeled in and out of the house. These tanks, protected from natural rainfall, filled with various soils, received carefully measured quantities of water. The farm and the vegetation house were undoubtedly the first experimental plant of its kind in the world dedicated to the scientific study of the use of water in irrigation. The results proved that they were very effective tools.
>
> The investigations that followed answered questions concerning the movement of water in irrigated soils; the control of loss of soil moisture by seepage and evaporation; the relation between the water lost by evaporation from soils, and by transpiration from plants; the relation between soil fertility and plant transpiration; the actual quantities of water required in crop production; the yields of crops and their chemical composition under varying quantities and times of application of irrigation water; and many other irrigation problems.

They also studied dryland farming, and the results of both irrigation and dryland research were shared throughout Utah by Dr. Widtsoe and his co-workers through extension activities. Results of the research were published in experimentation bulletins, and in two books: *Dry Farming* (published in 1911) and *Principles of Irrigation Practice* (published in 1914). Reports on the findings of plant–water relations and evapotranspiration research a century later differ little from the reports Widtsoe gave of these few years at the Experiment Station.

Politics were an important part of university life, then as now, and this was certainly true for Dr. Widtsoe. In spite of his amazing accomplishments, he was fired in 1905. He worked briefly at Brigham Young University in Provo, UT, and was then hired back as president of Utah State Agricultural College. This time the politics were in his favor. He served as president from 1907 to 1916, when he was made president of the University of Utah in Salt Lake City, UT. In 1921 he was called to serve as an apostle in the Church of Jesus Christ of Latter-Day Saints (LDS), a position he held until his death in 1953.

Widtsoe's early research on plant–water relations and irrigation practice was, without question, pioneering work, well ahead of its time. It is difficult to trace the influence of that work directly to modern plant–water relations theory. It is not difficult, however, to trace his influence on people who have had a great impact on modern soil physics. For example, in 1920, during the time when he was president of the University of Utah, Dr. Widtsoe joined Willard Gardner in writing a general treatise with the title, *The Movement of Soil Moisture*. A footnote to this article reads:

> The formal development of the mathematical material of this paper was made by the senior author as an outgrowth of the earlier work of the junior author, who has assisted in clarifying this material for publication.

Such participation with, and encouragement of, Utah soil scientists has had enormous impact on the careers of many soil physicists and on soil physics in general.

By the end of the nineteenth century, physics of soil processes had received considerable attention in many places in the USA, in Europe, and in other parts of the world. The names 'soil physics' and 'soil physicist' began to be used to identify subject matter and scientists on the faculties of colleges and universities and in governmental programs. In Utah, this occurred when Willard Gardner become a member of the faculty of the physics department and physicist for the Utah Agricultural Experiment Station in 1918 (**Figure 2**).

Figure 2 Willard Gardner (1883–1964).

Willard Gardner was born in 1883 in the little logging community of Pine Valley, Utah. His parents, Robert Gardner and Cynthia Lovina Berry, encouraged their children to educate themselves. Because they did so, several generations of Gardners have made important scientific contributions throughout the twentieth century. In 1903, Willard completed high school and traveled to Salt Lake City, where he attended business college, learning to type and take shorthand dictation. He became highly proficient at both, typing at faster than 90 words per minute and taking dictation at more than 150 words a minute. He then went to Logan, Utah, where he was employed by the Cache Valley Banking Company as stenographer. He did this work so well that he was often loaned to the local Justice Court to take dictation. In 1910, when the Agricultural Experiment Station at the college offered him a position as a clerk in their campus office, he accepted the position against the advice of his banker employers. This made it possible for him to pursue college classes on the side, and he was able to complete a bachelor's degree in physics. In 1912, he obtained a departmental assistantship in Physics at the University of California–Berkeley, graduating in 1916. His was the eighth PhD degree awarded by the University of California–Berkeley Physics Department.

After graduating, Gardner became principal of the Murdock Academy in Beaver, Utah, where his soil physicist son, Walter, was born in 1917. The following year he taught physics at the Brigham Young Academy in Logan, and in 1918 was invited to join the faculty in physics at Utah State Agricultural College (USAC). He held that position until his retirement in 1949. He became Professor Emeritus and continued to be active in university work until 1954. He died in 1964.

Dr. Gardner retained his stenographic skills and used them to good advantage throughout his career. For him, note-taking was easy, and he always typed his own letters rapidly, using a typewriter with an extra dozen Greek letter keys needed in writing mathematical equations. Not only did this save the college the cost of a secretary, but it was also more economical than to train secretaries to deal with mathematics.

Gardner's PhD thesis was titled *The Photo-Electric Current as a Function of the Angle of Emission and the Thickness of the Emitting Film*. He evidently found the physics of water flow and retention to be an important area to pursue in an agricultural experiment station. O.W. Israelsen, a friend and colleague from graduate school, and professor of irrigation engineering at USAC, encouraged Gardner to undertake irrigation research. Several other faculty members working in soil science had reason to seek his help in dealing with problems involving physics or mathematics.

Widtsoe's early work, Buckingham's 1907 monograph on water in soil, and Dr. Gardner's physics training formed the foundation for the soil physics work carried out by Gardner and his colleagues at USAC. A few of the titles of early papers show that the research direction was set from the beginning. Some are "The movement of moisture in soil by capillarity," "Capillary moisture-holding capacity," and "A capillary transmission constant and methods of determining it experimentally." These papers and others show consistent skill in using physics and mathematics to analyze soil-water flow problems. Research topics included saturated and unsaturated water flow, erosion, drainage, and measurement. The tensiometer, which is still a mainstay of soil physics research, was a product of Gardner's early research efforts, as was the tension or disk infiltrometer. It is no wonder that Dr. Gardner is referred to as the father of modern soil physics.

USAC did not begin offering the PhD degree until after World War II, so Willard Gardner had no official PhD students. In 1950, his son Walter received the first PhD given at USAC, with Wynn Thorne as major professor and his father as research advisor. However, numerous students and colleagues became interested in soil physics as a consequence of studying or working with Willard Gardner. The USAC library and Agricultural Experiment Station records are incomplete, but, from publications and personal files, it is possible to identify most of the students who studied with Gardner. Seventeen MS theses on topics closely related to soil physics are listed in the USAC Graduate School, and 14 published scientific papers with acknowledgements to Willard Gardner are in the library files.

There is little doubt that a number of students who have pursued careers in soil physics developed that interest while attending USAC and working with Gardner. L.A. Richards, one of his earliest students, obtained his masters degree at USAC and then went to Cornell University for his PhD degree. (*See* **Richards, Lorenzo A.**)

Another colleague who was strongly influenced by Gardner was Don Kirkham (*See* **Kirkham, Don**). Kirkham joined the physics faculty at USAC in 1937, having obtained a PhD in physics from Columbia University. Like Gardner, he came with no background in soil physics, but became interested through working with Gardner. After World War II, Kirkham was hired as a soil physicist by Iowa State University and became one of the most influential soil physicists of the twentieth century, both through his own research and through his large number of well-trained PhD students. Kirkham furnishes another tie between the two main characters of this brief history. Before

attending graduate school, Kirkham was a missionary for the LDS Church in Germany. John A. Widtsoe was in charge of the European missions at that time. Widtsoe and the young missionary found themselves together on a train at one point during Kirkham's mission, talking about careers. Widtsoe advised Kirkham to complete his education (i.e., attain the PhD) before he sought employment. Later, as Kirkham was in graduate school, he was offered a good job that would require that he drop out of school. This was tempting, since it was the depth of the depression, and jobs were hard to find. He remembered Widtsoe's advice, however, and completed his studies. Thus an important chapter in the lives of many soil physicists (Kirkham's students), as well as in the history of soil physics itself, would never have been written without the influence of both Widtsoe and Gardner in the life of Don Kirkham. Many similar stories could be told of the influence these men had on students and colleagues.

L.A. Richards, late in his professional career, recorded a revealing story about Willard Gardner's propensity to utilize mathematics in his scientific papers. The story is as follows:

> When I came on the scene as an undergraduate at Utah State College at Logan in 1922, Dr. Willard Gardner was pretty much alone in the subject matter field of soil physics and was making important applications of the Buckingham capillary potentials to the mathematical formulation of water flow in soil. This approach, though basic to progress, limited his current audience to people who understood his papers. I can well remember a conversation that took place between Sir Bernard Keen, then head of the Rothamsted Experimental Station [in the UK], and Dr. Gardner along about 1925. I was in the study room adjoining Dr. Gardner's office and with the door ajar. I, of course, was all ears concerning the conversation that was in progress. At a certain stage Dr. Keen said, 'Dr. Gardner I find your papers hard to read. Why don't you try to express things more simply?' Dr. Gardner was a man of some piety and had a quiet manner of speaking, but he had a prominent lower jaw. I could feel the air crackle in the next room as he paused for reply, which was, 'Sir Bernard, God made physics hard, not Willard Gardner.'

Hard or not, there have been many willing to follow in Willard Gardner's footsteps and learn his approaches to this new science. The result has been a century of solid progress. Like Widtsoe, he carried out pioneering work in many of the areas of modern soil physics. Unlike Widtsoe, he published many papers from which key ideas in modern soil physics can be traced.

See also: **Kirkham, Don**; **Richards, Lorenzo A.**

Further Reading

Buckingham E (1907) Studies of the movement of soil moisture. *Soils Bulletin* 38.

Gardner W (1919) The movement of moisture in soil by capillarity. *Soil Science* 7: 314–315.

Gardner W (1919) Capillary moisture-holding capacity. *Soil Science* 7: 319.

Gardner W (1920) A capillary transmission constant and methods of determining it experimentally. *Soil Science* 10: 106–107.

Gardner W and Widtsoe JA (1921) The movement of soil moisture. *Soil Science* 11: 215–232.

Taylor SA (1965) Willard Gardner. *Soil Science* 100: 79–82.

Widtsoe JA (1952) *In a Sunlit Land, the Autobiography of John A. Widtsoe*. Salt Lake City, UT: Deseret News Press.

Wind Erosion *See* **Erosion:** Wind-Induced

WINDBREAKS AND SHELTERBELTS

E S Takle, Iowa State University, Ames, IA, USA

Introduction

Recent interest by agricultural and environmental scientists in finding more sustainable ways of managing landscapes has highlighted the need for a better understanding of plant–atmosphere–soil interactions. Windbreaks and shelterbelts are used to modify meteorologic conditions at the soil surface to meet specific management objectives. Both artificial barriers (fences) and living barriers (trees and shrubs) are used to modify flow near the ground. The terms 'windbreak' and 'shelterbelt' are often used interchangeably. However, a fence is usually considered a windbreak, whereas a living barrier, although also a windbreak, offers shelter not only from wind but also from intense solar radiation, driving rain, and erosive movement of water across the soil surface. The emergence of the fields of agroecosystems and agroforestry has brought further attention to the use of living barriers to modify plant and soil microclimates through changes in micrometeorologic conditions.

Several centuries of practical application, followed recently by a few decades of research, have provided a wide range of reasons for using windbreaks and shelterbelts. The most frequent uses relate to reducing wind speed for purposes such as reducing soil erosion or drifting soil and sand, capturing snow for soil-moisture recharge, or protecting highways, rail lines, farmsteads, and animal confinement areas. Recent research has focused on other benefits such as changing microclimates for crops, preserving soil moisture, improving landscape and biological diversity, and creating recreational opportunities or aesthetically desirable landscapes. Recent advances in numerical modeling of flow fields and microclimate factors around shelterbelts have increased understanding of the complexities of the soil–atmosphere–plant interactions.

Use of Windbreaks and Shelterbelts in Soil Management

The diversity of vegetation offered by shelterbelts in regions of monoculture-managed landscapes promotes biological diversity both above and below the surface. By providing shade, detritus layers, wind speed reduction, soil ventilation, and changes in temperature, humidity, and soil moisture, perennial living barriers offer richer spatial variation in microclimates for plants and belowground ecosystems. In intensively managed agricultural environments, tree shelterbelts provide islands having reduced concentrations of agricultural chemicals and increased biodiversity, with soil and aboveground ecosystems that include earthworms, small mammals, birds, perennial grasses, and woody plants and that deliver a range of beneficial ecosystem services.

Shelterbelts will become increasingly important as the regional impacts of global warming become more clearly identified, both for sequestering carbon and to suppress the negative agricultural impacts relating to reduced soil moisture and increased likelihood of erosion. Simulations of future-scenario climates show higher likelihood of more extreme events (both droughts and floods), and shelterbelts offer protection of crops under such impending changes. Prevention of soil loss due to high wind is a historic benefit of shelters, but they also suppress soil loss due to floods or intense rains on sloped surfaces. Soil erosion decreases soil productivity locally owing to loss of fine soil particles containing organic matter and nutrients and causes off-site damage to structures and unwanted deposition of soil particles. Irreversible damage to soil ecosystems due to extended drought is suppressed by shelters. Perennial living barriers in agricultural fields sequester carbon aboveground by creating biomass and belowground by deep roots, litter production, and providing regions of undisturbed soils that reduce microbial decomposition rates. Shelterbelts restore soil organic matter lost through agricultural practices.

Influence of Shelters on Aerodynamics and Microclimate

Wind Speed Reduction

The main impact of a windbreak or shelterbelt is to reduce mean wind speed near the surface. Acting as a porous barrier to the flow, the shelter creates a small region of wind reduction on the windward side and a low-speed, turbulent wake zone followed by a region of gradual wind speed recovery in the lee, as shown in Figure 1.

Shelterbelts typically reduce wind speed substantially on the leeward side and to a lesser extent on the windward side (Figure 2). The observed wind speed at 2.88 m above the ground reveals the region of wind speed reduction from the undisturbed value of 5.6 m s^{-1} to a sheltered value of 1.5 m s^{-1} in the

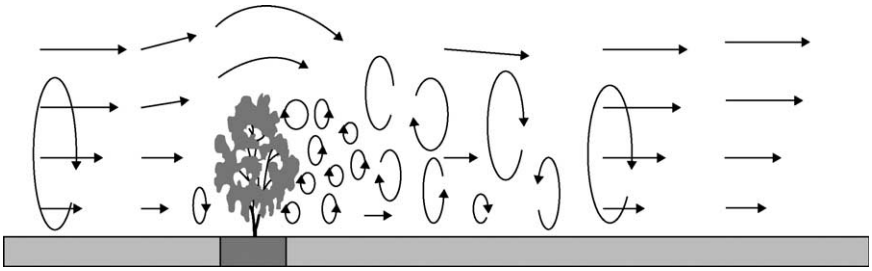


Figure 1 Mean and turbulent flow around a living shelter.

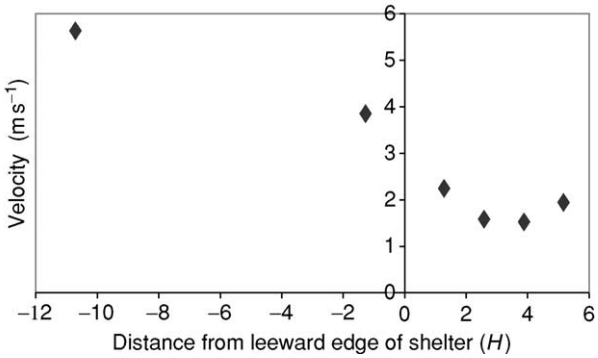


Figure 2 Observed horizontal wind speed at 2.88 m above the ground at various distances upwind and downwind of a two-row shelterbelt with height (H) of 14 m. Undisturbed wind speed at height 2.88 m is 5.6 m s^{-1}. Downwind edge of the shelter is taken as $x = 0$. (Adapted from Zhou XH, University of Nebraska.)

near lee. The barrier shelters strongly in a region approximately 10H leeward (H is height of the shelter), with modest sheltering extending leeward from the shelter approximately 20H.

The length of the sheltered region is reduced for more dense shelters and is larger but with less wind-speed reduction for highly porous shelters. A very dense shelter may even create a recirculation zone behind the shelter. When the wind direction is not perpendicular to the shelter, the sheltered region is reduced in size for successively more oblique angles of attack.

Shelterbelts have been used for many decades to suppress negative effects on soils due to high winds, namely to stop or slow drifting soil and sand. Suppression of drifting snow, on the other hand, makes important contributions to growing-season soil moisture in dry regions. Shelterbelts have been used widely to protect crops whose foliage or fruit are prone to damage from abrasion due to wind-blown material.

Turbulence Fields

The atmosphere contains turbulent eddies that connect the upper part of the atmospheric boundary layer to the surface (e.g., left-most eddy in Figure 1). These large eddies regulate surface exchange of heat, moisture, and trace gases between the atmosphere and the soil. Shelterbelts break up the large and efficient eddies that ventilate the surface into smaller eddies (e.g., see small eddies just downwind of the shelter in Figure 1) that are less efficient, thereby suppressing the transport of heat, moisture, and trace gases away from the surface. These small eddies have short lifetimes and dissipate within a few H of the shelter. Reduced turbulent fluxes of heat and moisture lead to higher near-surface air temperatures and soil temperatures within approximately 8H of the shelter compared with unsheltered regions during both day and at night.

Further downwind, where the large eddies reattach to the surface and where mean wind speed is reestablished, enhanced ventilation leads to slightly lower temperatures than in unsheltered areas. At night under low wind speeds and laminar flow (very small eddies) the shelter may increase the eddy size and transport heat downward. But if winds are already near calm, the barrier may reduce both mean and turbulent flow and allow the sheltered region to be cooler than unsheltered areas.

Pressure Fields

A shelterbelt, as a barrier to the flow, increases the static pressure upwind and reduces pressure in the lee of the shelter, as revealed by measurements shown in Figure 3. Since it is not a direct microclimate factor affecting crops and soil, pressure is often overlooked as one of the key influences of a shelter. However, fluctuations in the stationary pressure field created by the shelter can have substantial influence on movement of trace gases in soil or mulch layers.

Surface Fluxes of Heat and Moisture

The breakdown of large, efficient eddies in the lee of the shelter reduces vertical transport of both heat and moisture away from the near-surface air, which increases humidity near the surface and reduces surface evaporation or evapotranspiration. Lower evaporation allows more incoming radiant energy to be

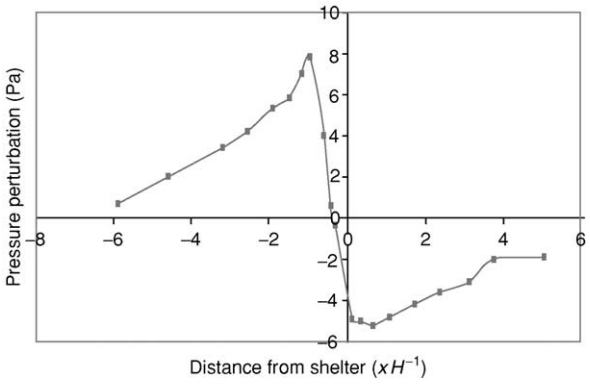


Figure 3 Perturbation of static pressure at ground level at various distances upwind and downwind of a two-row shelterbelt with height (H) of 14 m. Undisturbed wind speed at height 2.88 m was observed to be 5.6 m s^{-1}. Downwind edge of the shelter is taken as $x = 0$. (Data provided by Zhou XH, University of Nebraska.)

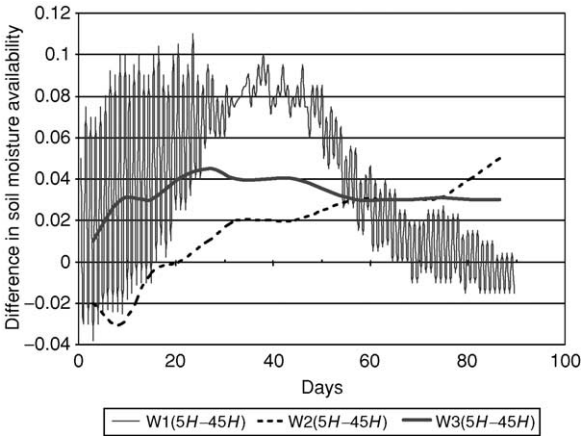


Figure 4 Differences in simulated soil moisture availability (dimensionless) between a sheltered region (5H downwind) and unsheltered region (45H downwind) at depths of 2.0 cm (W1), 1.0 m (W2), and 2.0 m (W3). The surface is assumed to be bare soil. (Adapted from Wang H, Shen J, and Takle ES (1997) Influences of agroforestry ecosystem on evapotranspiration and soil moisture. *Proceedings, 13th Conference on Hydrology*. Boston, MA: American Meteorological Society.)

partitioned to sensible heat, which thereby raises the surface temperature of the sheltered area. Energy loss by longwave radiation, which is proportional to the fourth power of temperature, keeps the surface temperature from getting too far above the temperature of the overlying atmosphere.

These modifications to the surface, and hence the near subsurface, of the soil lead to increased soil temperature and soil moisture compared with unsheltered regions, which in turn cause many soil chemical and biological processes to proceed at different rates in sheltered areas compared with open fields.

Living barriers consisting of woody plants also create a layer of litter or detritus mulch in the near vicinity to the stem or trunk. This layer does more than simply reduce extreme temperatures at the soil surface and preserve soil moisture; it also provides an additional zone for chemical and biological processes, such as methanogenesis. Being highly porous, such layers are subject to enhanced ventilation by mechanisms described below.

Soil Moisture

Most agricultural regions have deficient soil moisture at some point in the growing season. Observations suggest that shelterbelts raise both the daytime temperature and humidity in the sheltered region under most conditions, but exceptions do occur, in some cases due to soil-moisture conditions. Changes in soil moisture occur much slower than changes in air properties, so heat and humidity supplied due to the presence of soil moisture can modify micrometeorologic conditions in ways that may sometimes be counterintuitive.

Soil-moisture preservation in the lee of a shelter after a rain event that saturated the soil at all levels is shown in Figure 4. In this simulation, the soil surface was covered by a transpiring crop and was protected by a medium-density shelterbelt of height H. The time series of this plot gave the differences in soil moisture availability between sheltered and unsheltered locations at three depths for 90 days after a simulated precipitation event.

During the first 15 days after the onset of saturated conditions, daytime soil moisture at a depth of 2 cm (W1 in Figure 4) is higher in the sheltered region during the day (positive difference), but is lower at night (negative difference), presumably due to higher near-surface nighttime temperatures in the sheltered region. After 15 days, soil in the sheltered region has higher moisture both day and night. The relative benefit of shelter to soil-moisture preservation increases with time to a maximum near 40 days after the saturating rain event. At this time the soil-moisture availability near the shelter is approximately 25% higher than in unsheltered regions. Beyond this time the sheltered region loses moisture faster than unsheltered regions out to approximately 80 days, by which time the soil-moisture availability is everywhere uniform at approximately 0.14.

Deeper in the soil (1.0 m, plotted as W2, and 2.0 m, plotted as W3 in Figure 4), soil moisture is preserved in the sheltered region throughout the period except for the accelerated loss in the first 15 days from the 1-m layer due to higher temperatures. The region 4–8H downwind from the shelter is the region of highest protection, although evapotranspiration is

reduced by 10% or more compared with unsheltered areas out to approximately 13*H*.

Meteorologic Mechanisms that Move Gases in Soil

Most descriptions of gas movement in soils refer to the process as diffusional; however, ambient meteorologic conditions frequently create processes that move gases much more effectively than pure molecular diffusion. Numerous external (to the soil) factors are responsible for the movement of gases in soils. Most notably, pressure changes in the atmosphere are transmitted through airways within the soil and force movement of gases by nondiffusive means.

Static air pressure at a point (in either the atmosphere or soil above the saturated zone) is the horizontally averaged accumulated weight per unit area of atmospheric (gaseous) mass and the mass of any solid or liquid suspended by the atmosphere (birds, airplanes, liquid or solid H_2O, particulate matter, etc.) above that point. And changes in static pressure are caused by changes in the accumulated mass above the point. A point in the soil (we exclude soil locations not having a gaseous connection to the free atmosphere) may therefore experience pressure changes due to changes in the mass accumulation in the atmosphere above the point or due to phase changes of liquids or solids to vapors in the soil or on its surface. The region below the surface has zero mean flow in the absence of pressure gradients and volume expansion resulting from liquid-to-gas phase change. Any nonzero flow below the possible thin, turbulent layer in the soil is related to a pressure change or phase change, i.e., some change in the overlying mass accumulation.

Several mechanisms can lead to movement of gases within the soil and to effluxes from the soil. Many of these are related to changes in atmospheric pressure on various time scales or to liquid-to-gas phase changes. Because shelterbelts and windbreaks create spatial and temporal variations in surface pressure and temperature, they indirectly and importantly contribute to gas movement in the below-surface environment. Figure 5 provides a stylized view of pathways linking atmospheric gases with soil gases to assist in visualizing various mechanisms for gas movement.

Type 1: Large-Scale Pressure Changes

Changes in the static pressure of the atmosphere (and hence soil) due to movement of large-scale weather systems represent lowest-frequency pressure changes that might lead to externally driven transport of gases

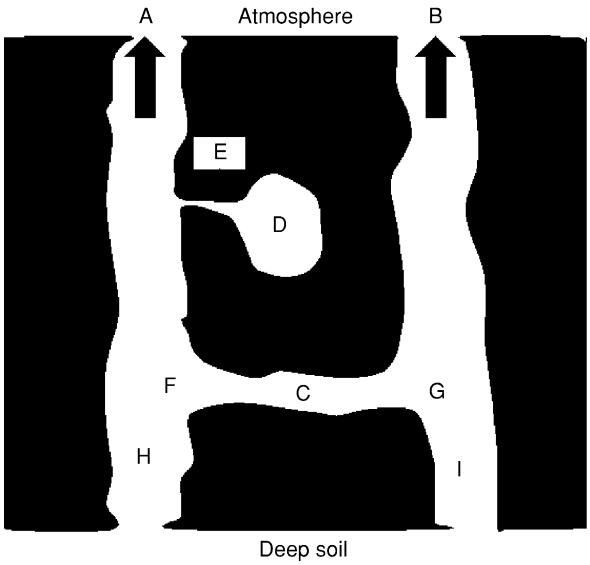


Figure 5 Stylized view of soil reservoirs and pathways for movement of trace gases.

in soils near the surface. Barometric pressure changes with periods of hours to days due to mesoscale and synoptic-scale weather systems cause changes in static pressure on the order of ±2% of total pressure over large areas, leading to uniform flow in or out at the surface.

Type 2: Traveling Atmospheric Pressure Waves

Traveling atmospheric pressure waves (of frequency assumed to be high enough to exclude the previous type of flow) transported over a flat, featureless surface can generate pressure fluctuations. For instance, observations of pressure, temperature, and wind fluctuations at several levels within tree canopies have revealed 'coherent structures' consisting of episodic bursts that penetrate and flood the canopy with overlying air. Pressure changes of this type at the soil surface cause oscillatory flow at A and B in Figure 5, with possibly different amplitudes and phases depending on the relative spatial separation as compared to the wavelengths of turbulent eddies. Point C experiences transient flow depending on conditions at A and B, and flow through E is oscillatory with zero mean. This type of flow offers a mechanism for drawing air with trace gases (water vapor, methane, nitrous oxide, carbon dioxide, oxygen) of high concentration out of D and replacing it with air of low trace-gas concentration.

Other atmospheric disturbances leading to traveling pressure waves include passage of a sea-breeze front, mountain lee waves, and outflow boundaries from thunderstorms and mesoscale convective complexes.

Type 3: Static Horizontal Pressure Gradient in Soil (Standing Wave)

Flow can be generated if points A and B, for illustrative purposes, are on opposite sides of an obstacle (e.g., tree, fence, rock outcropping, topographic feature) that creates a static horizontal pressure difference, as shown in Figure 3, due to a uniform (nonturbulent) wind blowing from left to right. Point A experiences high pressure and B low pressure that moves gas at a uniform rate from A to B through channel C owing to a pressure difference between F and G. As in type 2, passage E experiences only weak or zero flow. This process leads to horizontal as well as vertical motion in soils.

Type 4: Fluctuations on the Standing Wave

Wind that is not uniform in speed but has frequency ω (also assumed to be high enough to exclude the first type of flow but probably is lower than turbulence frequencies) may encounter an obstacle between A and B, thereby giving a fluctuating pressure of:

$$p = p_o e^{\omega t - k \cdot r} \quad [1]$$

at A, where k is the horizontal wave number and r is the horizontal distance vector. At B the pressure will be out of phase by π and possibly have different magnitude. Note that this forcing arises from changes in the static pressure due to dynamic interaction of the wind with the obstacle and is different from flow related to turbulence in the free atmosphere (see Type 5: Turbulence). This fourth kind of pressure change causes flow through C to be unidirectional and oscillatory, and flow through E to be oscillatory with zero mean. The venting of trace gases from cavity D, as described in type 2 flow, more effectively contributes to surface fluxes under this regime due to the unidirectional flow between A and C. If the obstacle is not between A and B, the pressure fluctuation is not out of phase, but the pressure gradient across C still leads to unidirectional flow.

Type 5: Turbulence

A turbulent atmosphere has stochastic fluctuations of pressure as well as wind speed, and larger turbulent pressure fluctuations generally accompany higher wind speeds. Despite the fact that the magnitude of these pressure fluctuations increases exponentially with mean wind speed, the resulting pumping action at the soil surface may displace an air parcel only a few micrometers at the surface. However, these pressure fluctuations may penetrate with minor loss of amplitude to the depth of pore D in Figure 5, and contribute to a substantial venting of the isolated pore. The resulting pumping action can move trace gases out of D at a much higher rate than is possible by molecular diffusion. Although this mechanism itself does not contribute strongly to surface effluxes, it may lead to higher availability of trapped gases for movement by other mechanisms.

Type 6: Venturi Effects

Differences in sizes of the soil channels through which air is ventilated lead to different flow speeds and therefore different rates of vaporization and gas transfer from solid or liquid surfaces lining the channel. Hollow stems of plants also have Venturi-like constrictions that cause irregular flow speeds.

Type 7: Rainfall

Rainwater ventilates soil gases by displacing air from pores and also enriches the soil air with dissolved O_2. Differential penetration of rainwater into the soil can lead to heterogeneous flushing of soil air. As rainwater percolates downward in the soil, the upper liquid surface also moves downward, drawing atmospheric air into the soil.

Type 8: Phase Change of Liquid to Vapor

When a liquid changes phase to a vapor, the accompanying volume expansion creates gaseous flow. For instance, evaporation of 1 cm of liquid water from a horizontally homogeneous soil and resulting expansion by a factor of 1000 leads to a net vertical discharge of 10 m ($10\,m^3\,m^{-2}$) of water vapor from the soil. This vertical efflux of water vapor carries other trace gases also. Horizontal differences in soil temperature lead to different evaporation rates and therefore different rates of soil-gas venting.

Spatial and temporal variability in soil pressures and temperatures created by shelterbelts, windbreaks, and isolated trees indirectly influence soil-gas movement, particularly by mechanisms 3, 4, 5, and 6 listed above. The presence of windbreaks or shelterbelts can therefore be expected to influence movements and distributions of trace gases in soil in complex ways.

Aesthetic and Recreational Value

Shelterbelts add diversity to agricultural landscapes, which, in turn promotes other biological diversity both above and below the surface and helps maintain delivery of a range of ecosystem services. Shelterbelts prevent loss of soil productivity owing to soil erosion. Finally, the addition of shelterbelts to the landscape, with their attendant diversity, enhances the aesthetic

and recreational value of rural environments, thereby promoting a sense of connectedness to the land that may be compromised in an increasingly urbanized society.

List of Technical Nomenclature

Angle of attack	Angle between the wind direction and a line perpendicular to a linear shelterbelt
Anthropogenic	Of human origin
Eddy	A transient circulation in a fluid that interacts with the mean flow to redistribute momentum, energy, and mass
Mesoscale convective complexes	Massive and long-lived storms that are relatively circular in shape, persist for 36 or more hours and produce large amounts of rain and possibly severe weather

See also: **Agroforestry**; **Biodiversity**; **Energy Balance**; **Erosion:** Wind-Induced; **Evapotranspiration**; **Forest Soils**; **Mulches**; **Soil–Plant–Atmosphere Continuum**

Further Reading

Brandle JR, Hinz DL, and Sturrock JW (1988) *Windbreak Technology*. Amsterdam, the Netherlands: Elsevier Science.

Brandle JR, Hodges L, and Wight B (2000) Windbreak practices. In: Garrett HE, Rietveld WJ, and Fisher RF (eds) *North American Agroforestry: An Integrated Science and Practice*, pp. 79–118. Madison, WI: American Society of Agronomy.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Jury W, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

McNaughton KG (1988) 1. Effects of windbreaks on turbulent transport and microclimate. *Agriculture, Ecosystems, and Environment* 22/23: 17–39.

Miller JM, Bohm M, and Cleugh HA (1995) *Direct Mechanical Effects of Wind on Selected Crops: A Review*. Technical Report No. 67. Canberra, Australia: Center for Environmental Mechanics ACT.

Rosenberg NJ, Blad BL, and Verma SB (1983) *Microclimate: The Biological Environment*, 2nd edn. New York: John Wiley.

Sumner ME (ed.) (2000) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Wang H (2001) *High-Performance Cluster Computing, Algorithms, Implementations and Performance Evaluation for Computation-Intensive Applications to Promote Complex Scientific Research on Turbulent Flows*. PhD Dissertation. Ames, IA: Iowa State University Press.

Wang H, Takle ES, and Shen J (2001) Shelterbelts and windbreaks: mathematical modeling and computer simulation of turbulent flows. *Annual Reviews of Fluid Mechanics* 33: 549–586.

WOMEN IN SOIL SCIENCE (USA)

M J Levin, USDA Natural Resources Conservation Service, Washington, DC, USA

Published by Elsevier Ltd.

Introduction

Women have been involved in soil science and soil survey since interest in this most basic of natural resources emerged. Women employed by the early US Soil Survey were largely restricted to office activities, but soon they ventured into fieldwork. Albeit brief, the first appointment of a woman to a field party came about in 1901, only 2 years after the soil survey began in earnest. It would be another 45 years, however, before the Soil Conservation Service (SCS) would appoint a woman as a field soil scientist.

The delay in women joining the ranks of field scientists has significance for telling the story of women's place in the soil survey. Soil science is first and foremost a field-based science. Soil mapping begins by studying the landscape and building a conceptual model of how the topography, geology, plants, climate, water, and animals interact to determine soil characteristics. Mapping then continues in the field where the soil scientists validate their determinations of soil types by digging, describing, and sampling the soils and vegetation. Benchmark samples are analyzed in the laboratory to validate conceptual models. Finally, the information is consolidated on to aerial photo maps, associated computer databases, and manuscripts for publication. Until women gained an equal place as scientists in the field, they remained in supporting roles in the soil survey.

The Pioneers (1895–1965)

Women's roles in the earliest years of the soil survey appear to have been limited to clerical work, copy-editing of manuscripts, and cartographic drafting of maps, although women with appropriate academic

training soon began to work in the laboratories. Janette Steuart and Sorena Haygood, who maintained the laboratory and field records for the US Department of Agriculture's (USDA) Division of Agricultural Soils, were among the first women to work for the soil survey. According to Macy H. Lapham's account, Steuart was hired January 4, 1895, making her the first woman appointed to the Division of Soils, then part of the US Weather Bureau, located in the USDA. Both Steuart and Haygood were career employees; Steuart retired in the late 1920s with nearly 30 years of service and Haygood retired some time later.

The soil survey remained resistant to women in field parties. Julia R. Pearce, one of the earliest female pioneers in soil survey, joined the soil survey as a member of the field party at Hanford, California, in June 1901, but never had the opportunity to join her party in the field. Pearce had been one of only two 1901 graduates in agriculture from the University of California at Berkeley (UCB). In his commencement address, Secretary of Agriculture 'Tama Jim' Wilson lamented the small size of the graduating class and emphasized that the Department of Agriculture needed candidates trained for technical positions. After the speech, Pearce sought Wilson out and told him 'she was ready and willing to come to the relief of the Departments.' Pearce found herself almost immediately appointed as an assistant to Macy Lapham's all-male field party at Hanford.

While Secretary Wilson was sympathetic to the idea of women in the workplace, Lapham was uncomfortable with having women in the field with an all-male crew. On the day that Pearce arrived in Hanford, it was said that Lapham sent a telegram that stated, 'Miss Pearce is here, what in hell shall I do with her?' In the end, he put her to work copying maps. In 1903, she was transferred to the Bureau of Soils in Washington as an assistant in soil survey and later transferred to the Bureau of Plant Industry as a laboratory assistant.

From Lapham's memoirs, it appears fieldwork was out of the question for women before the 1940s. Women found places in the field in unofficial capacities, however. Mary Baldwin, the wife of soil inspector Mark Baldwin (employed by the Soil Survey 1912–1944), mapped with her husband in northern Wisconsin and the Boundary Waters of Minnesota during the early 1920s. They worked during the summer months, camping and using a small boat to go from island to island. Mary would drop Mark off on one side of the island, he would map, and she would wait for him with the boat on the other side of the island. While she waited, she might search for survey markers or make observations on her own of the general area. Mary recalled that there were times when she wished that she could have tried mapping on her own. As it was, she accompanied her husband everywhere during his remote mapping experiences, transcribing or taking field notes for him and assisting with the sampling.

Mary Baldwin might still have had difficulty finding employment with the USDA, even if she had the background and training to map soils. At the time, married couples were discouraged from working for the same federal agency, and the USDA already employed Mary's husband. A clause in a 1932 appropriations law even stated that married persons with a federally employed spouse would be dismissed first in the case of government reductions in force and the preference should be given to others for new appointments. Although this legislation was repealed in 1937, it limited married women's employment by the federal government at that time and perhaps set a precedent for the future.

During the decades of the 1930s, 1940s, and 1950s, women contributed to the soil survey through editing, writing erosion history, and conducting laboratory work. Lillian H. Weiland – the first female employee of the newly established Soil Erosion Service and secretary to Hugh Hammond Bennett – put together a 'Bibliography on Soil Conservation Compiled in the Office of the Chief of SCS' in 1935. The bibliography consolidated ideas for soil erosion control technology for the new agency. In 1937, Lois Olson and Dr Arthur Hall spoke on studies in erosion history as part of a series of research seminars for SCS staff. Some of today's thinking on interpretations of the soil survey and field practices to control erosion can be attributed to this series of lectures. Olson, a geographer by training, headed the SCS's Erosion History Section.

Charlotte Whitford (Coulton), a graduate of Ohio State University with a master's degree in botany, joined the SCS as a secretary with a field soils staff in Zanesville, Ohio, in the mid-1930s. J. Gordon Steele, an old classmate, soon recruited Whitford to work as an assistant soil technologist in Washington, DC, on a series of reports on soil erosion. She later worked as an editor on soil surveys and eventually became head of the SCS publication staff. She retired in the 1980s with almost 50 years of service.

Dorothy Nickerson, a soil color technologist for the USDA from the late 1920s through the 1940s, was instrumental in developing the soil color standards for soil survey. Nickerson had been an assistant manager of the Munsell Color Company before joining the USDA in 1927. She made extensive colorimetric tests in the laboratory and worked with soil scientists in the field to match soils to the Munsell colors and to create a new set of color names, first introduced in preliminary form in 1941. She then

worked with Thomas D. Rice, Kenneth Kelly, and Albert H. Munsell to adapt the Munsell color chart system for describing soil color in the laboratory and the field. The US Soil Survey adopted the Munsell color charts and new color names in 1949.

The first woman soil scientist officially assigned in the field for the SCS was Mary C. Baltz (Tyler). Mary Baltz graduated from Cornell University and joined the soil survey as a 'junior soil surveyor' in 1946. Labor shortages during World War II provided the opportunity for her to work in a job that, up to that time, appeared to be reserved for men. By 1951, Mary was responsible for mapping in Madison and Oneida Counties in New York, and later she was assigned the task of map measurement for the entire state. In contrast with today's electronic techniques, the work was done by cutting out the soil map delineations on copies of the field sheets. Areas with the same label were weighed together and a factor converted the weight to acres. She hired a team of women to do this conversion job in the winter months.

Erwin Rice, a retired soil scientist in New York, mapped under Mary Baltz's direction. He remembered Mary as a confident, petite woman who enjoyed mapping in the field, was comfortable with the all-male crews, and had a good sense of humor. He called her a 'splitter,' a soil scientist who tends to separate out concepts for new soils as opposed to lumping them together under general categories of old soil names. Mary Baltz worked for the SCS until about 1965.

Ester Perry was a major figure in the California soil survey effort. Her 1939 PhD in soil science from UCB was the first received by a woman in the USA. For her doctoral research, 'Profile Studies of the More Extensive Primary Soils Derived from Granitic Rocks in California,' she may have been one of the first students to use X-ray diffraction to look at clay mineralogy structure in soils. She studied under Charles Shaw at Berkeley, and Kelly, Doer, and Brown were mentors and coworkers with her in Riverside, California, where she worked at the Subtropical Horticulture Research Center during her graduate studies. From 1928 to 1939 she worked as an associate soil technologist for the California Agricultural Experiment Station.

From 1939 until she retired in 1965, Perry essentially ran the USDA soil survey laboratory in Room 33 in the basement of Hilgard Hall at UCB. She moved up through the ranks during that time, from 'junior soil technologist' (1939–1954) to 'associate specialist soils' (1954–1960) to 'specialist' (1960–1965). In a 1952 presentation to the Western Soil Science Society in Corvallis, Oregon, Perry praised the benefits of close collaboration between a soils laboratory and a field soil scientist for quick turnaround of information in support of mapping. After the establishment of the Beltsville Agricultural Research Center in 1952, the Berkeley Soils Laboratory was slowly phased out.

In soil science, or more specifically, in pedology (the study of soil genesis), as with all the earth sciences, there were very few women working in the field before the 1970s. Gary Sposito, who was a student in Ester Perry's laboratory for a year, recalls that she was well aware of being a pioneer in her profession. As one of 'Ester's boys' (as students who worked part-time in her laboratory were known), he thought she effectively mentored many young men and women into a soil science career. Dr Perry maintained an all-business approach in the laboratory, but also remembered to bring birthday cakes for her students and provided a bed in the laboratory for those who might work through the night on important projects. Despite her accomplishments, Ester Perry was never promoted to associate professor or put on the tenure track. This was not unusual, however, since many women and men worked as researchers or technicians for their entire careers at the agricultural experiment station without receiving academic status. She also was not acknowledged in USDA records as an official soil survey collaborator.

Foundations: Building on the Pioneers (1959–1975)

During the 1950s and 1960s few women ventured directly into the field of soil science. Some arrived through other disciplines, such as geology, microbiology, or one of the plant sciences. They found through their graduate studies that soil science was a key element in their research and then continued on to further studies in soil science. Some were mentored and encouraged by major professors (as Ester Perry had been by Shaw and Kelly) to pursue soil science and stick with it. As scientists and teachers, the women soil scientists who started in the 1950s and 1960s spent a good deal of their careers as mentors themselves, and many placed a high value on that aspect of their work.

Ester Perry, who bridged the gap from the pioneer era to this period of building foundations, was, herself, one such mentor, not only through her work with students in the Berkeley Soils Laboratory, but also particularly in her effort to bring equal access to soil survey field training to women in the mid-1950s. In that decade, a summer field course – Soils 105 field trip – was a requirement for a soil science major from UCB and the University of California at Davis (UCD). The course was offered each summer, and since the 1930s had convinced many a prospective soil science student that soil survey could be a lifelong interest and career path. In 1953, Eva Esterman, a soil science honors student, was the

first woman to request to take the course. The UCB Soils Department offered Eva an option to graduate without participating in the field course, but she wanted to take the trip just like all of the other students. For 6 weeks in the field, the university arranged for Eva to have separate sleeping facilities and comfort stops. Dr Earl Story's wife served as chaperone, accompanying the students in a separate car and with some discomfort. The academic dean at the time, Dr Frank Haridine, considered the experiment a complete disaster and swore publicly that no women would ever again take the field course.

The event triggered Ester Perry to step in and offer a Soils 105F course ('F' for female), which she planned and made available from 1956 to 1959. The trip was soil survey-oriented but because Ester had different professional contacts, the course had a somewhat different approach. Three women took the course in 1959, the last year the specialized course was offered. In 1965, the 'Soils 105' course officially became co-ed and included two women students. Perry accompanied the group as a chaperone. Seven years later, the class was almost 50% women, and there was no women's chaperone.

In addition to mentoring students of soil science, many women professors and researchers made substantial contributions to our understanding of soils and soil science during the 1960s and 1970s. In the USA, Cornelia Cameron, Jane Forsyth, Jaya Iyer, Eva Esterman, Nellie Stark, and Elizabeth Klepper have been among the most prominent, although there are no doubt many others. Scientific publications usually list the first names of researchers only by their initials, which makes identifying the authors by gender difficult at best; the individuals included here were identified by their students and colleagues.

Dr Cornelia C. Cameron completed her PhD in geology (with an emphasis in geomorphology) at the University of Iowa in 1940. After 11 years teaching earth sciences, Dr Cameron joined the US Geological Survey (USGS) in 1951 and spent the next 43 years in the field. Her field career began in military geology, with terrain analysis of military sites in over 30 countries on five continents, many of them dangerous militarized zones at the time. Dr Cameron's colleagues remember her as quite a character in the field.

In part, Dr Cameron's reputation for eccentricity had to do with her mother. One of the USA's first female PhDs in botany, Dr Cameron's mother had a strong interest in her daughter's work and accompanied Cornelia on field expeditions until she was 103 years old. The younger Dr Cameron joked that when her mother got so old that her eyesight had deteriorated, she put a cow bell on her so she could find her if she wandered off. In a story about daughter and mother's adventurous military terrain investigations in the Caribbean area in early 1961 before the Bay of Pigs invasion, Dr Cameron recounted that "Mother and I were a perfect pair. We told everyone that we were Canadian tourists. One time, as I was doing traverses along the slopes of one of the islands, Mother stayed in the car. I was upslope from her when I saw a truck full of guerrillas pull up. Mother simply charmed them and they drove off."

Dr Cameron was an internationally recognized authority on peat soils and their use as a soil additive and source of energy, and on the impact of peat removal on swamp and bog environments; she wrote prolifically – 110 publications – on the subject. Both the USGS and the Department of the Interior recognized her accomplishments in research and public service. Dr Cameron received the USGS's Meritorious Service Award in 1977 and its Distinguished Service Award in 1986, and received the Department of the Interior's Public Service Recognition Award in 1990.

Dr Jane L. Forsyth, a professor of geology at Bowling Green State University in Ohio, has contributed much to our understanding of the age relationship of soils and till to northern Ohio glacial geology. Among her peers she has been affectionately dubbed the 'Queen of the Pleistocenes,' according to her colleague Peter Birkeland, retired from the faculty at University of Colorado, Boulder. Dr Forsyth earned her PhD from Ohio State University in 1956 and taught at University of Cincinnati, Miami University, UCB, and Ohio State University, before joining the faculty at Bowling Green State University in 1965.

Dr Jaya Iyer, professor of soil science at the University of Wisconsin, Madison, has focused her research on relating soil properties and tree growth. Dr Iyer earned her PhD in botany from the University of Bombay, India, in 1959. The external referee for her PhD research, soil scientist Dr Sergei Wilde of the University of Wisconsin and a member of the Wisconsin Forestry Hall of Fame, encouraged her to take an interest in soils, and she completed a second PhD in Soil Science at the University of Wisconsin, Madison in 1969. Dr Iyer has had a highly successful career as a national expert in soils for tree nurseries, specializing in urban, Christmas tree, and forestry production.

Dr Eva Esterman, who received her PhD in soil science from UCB in 1958, nearly 20 years after Ester Perry, was only the second woman to earn a soil science PhD from Berkeley. She went on to become a professor at San Francisco State University in 1960, where she taught botany and later added soil science to the curriculum. She focused her soils research on soil microbiology and biochemistry before she retired in 1982 and began raising sheep.

Dr Nellie Stark, tenured professor and forest soil ecologist at the University of Montana, Missoula, from 1970 to 1992, gave us the theory of the 'biological life of a soil,' which describes how soils and plants interact during development and decline phases of soil genesis. The theory explains the variation in nutrient uptake by plants based on the stage of soil genesis. Indirect nutrient cycling, which involves uptake of ions by the roots from the soil, predominates when a soil is young; direct nutrient cycling, which involves uptake of ions by the roots directly from the litter, by-passing the soil, occurs as the soil becomes older and depleted by weathering.

Stark earned a PhD in botany (ecology) from Duke University in 1962, with a minor in soils based on credits she collected at Oregon State University in 1961. Dr Stark's research with the Desert Research Institute, Reno, Nevada, focused on soils and nutrient cycling of litter in the tropical ecosystems of Brazil and Peru. The soil chemistry laboratory for forestry that Dr Stark operated at the University of Montana was well-known and received and processed samples from all over the world.

Finally Dr Elizabeth L. Klepper, a research leader and plant physiologist at the Columbia Plateau Conservation Research Center, Pendleton, Oregon, concentrates her research on root growth and functioning under field conditions and plant–soil water relations. Dr Klepper holds degrees from Vanderbilt University and Duke University. She has been recognized for her accomplishments by all three professional agronomic research societies of the USA: the American Society of Agronomy (ASA), the Crop Society of America (CSA), and the Soil Science Society of America (SSSA). She was the first woman ever to receive the Fellow award from the SSSA (American Society of Agronomy).

These are but a few of the women who dedicated their lives and research to soil science and soil survey during this period. It is inspiring to consider their achievements. Over the course of about 50 years, women in soil science and soil survey moved out from under their restriction to clerical support to become influential researchers and field investigators. Both Dr Stark's theory of the biological life of a soil and Dr Cameron's important and risky fieldwork would have been virtually unimaginable when the soil survey began in 1899.

In the Classroom, in the Field, and in the Laboratory (1970–1990)

Despite great strides in the field of soil science, women were still not actively recruited into the USDA's Soil Survey Division in the 1960s. For example, in a 1962 recruitment speech to the Agronomic Education Division of the American Society of Agronomy in Ithaca, New York, Assistant Soil Survey Administrator Charles Kellogg expressed his agencies' concern about recruiting good candidates, "especially of well-trained, broadly educated young men who can develop rapidly." His comments were not surprising perhaps, since the professional workforce of the country at that time was still predominantly male.

The transition during the 1960s was profound, however. By the 1970s, career counseling documents were beginning to discuss ways to channel girls into nontraditional careers, and encouraging young women to enter nontraditional occupations continued as a theme into the 1980s. Corresponding changes occurred in the classrooms; as more and more young women began to enter previously male science and employment territories, materials and approaches to education changed to meet the needs of this more diverse student population.

In the soil survey, as well as in some of the other earth sciences professions, a woman still needed to be persistent in the 1970s to obtain a field appointment. In the SCS of the 1970s, there were fewer than 15 women in the federal employment series soil scientist (470 series) at anyone time nationally, despite an acceleration in soil survey mapping and a general increase in field crews.

Most of these women soil scientists thought they were the only female soil scientist in the agency. In addition, there were no formal professional organizations for women field soil scientists; the Association of Women Soil Scientists (AWSS), organized by a group of women soil scientists in the US Forest Service, was not formed until the early 1980s. But slowly more career opportunities began to emerge for women. Title VII of the Civil Rights Act of 1964 and the Civil Service Reform Act of 1978 helped increase opportunity by prohibiting sex discrimination in employment and requiring diversity in the workforce, and the Women in Science and Technology Equal Opportunity Act of 1980 opened up more opportunities for women to receive support in university settings.

In the 1970S and early 1980s, the SCS soil survey staff in California included five women field soil scientists (Arlene Tugel, Nancy Severy, Chris Bartlett, Lisa Holkolt, and Maxine Levin), a crowd compared to other states. Many states had only one woman working as a soil scientist in the field – Carole Jett in Nevada; Carol Wettstein in Florida; Sue Southard in Utah; Margaret Rice in Mississippi; Caryl Radatz in Minnesota; Mary Collins in Iowa; and Debbie Brasfield in Tennessee, for example. Some states had two or more women in the field, and there may have

been more. Records of employees for those years are spotty and have not been saved comprehensively. But, in any event, women still comprised a small percentage of the total field soil scientist staff. Nevertheless, their contribution to soil survey was sizable, with millions of acres mapped, at times with some physical hardship.

Many of the women who worked in field parties during the 1970s went on to achieve greater responsibilities and position. By the late 1980s, the SCS had appointed the first woman state soil scientist, followed by others in the early 1990s, and women grew more prominent as soil survey party leaders. Carol Wettstein became the first woman state soil scientist, serving in Maryland from 1988 to 1989 and as state soil scientist in Colorado from 1990 to 1995. Carole Jett served as state soil scientist in California in 1991, and Carol Franks was state soil scientist in Arizona in 1994. In 2000, Maxine Levin was appointed national program manager of the Soil Survey Division. In the 1980s, there were at least three published soil surveys for which women were the party leaders or the principal field investigators: Sacramento County, California (Arlene Tugel); City of Baltimore, Maryland (Maxine Levin); and Indian River County, Florida (Carol Wettstein).

We can anticipate seeing more women listed in soil surveys of the 1990s, as the number of women party leaders increased significantly during the decade. An all-female crew of soil scientists led by Deborah Prevost mapped the Hualapai-Havasupai Indian Reservation, Arizona, in the late 1980s and early 1990s. This soil survey was published in 1999 – exactly 100 years after the establishment of the USDA Soil Survey.

Women have also contributed to soil surveys in many other ways that are not reflected in publications. Soil correlators and data management specialists make significant contributions to soil survey data and manuscripts, and a number of women have held these positions: Sue Southard (California), Renee Gross (Nebraska), Carmen Santiago (Puerto Rico), Panola Rivers (Pennsylvania), Kathy Swain (New Hampshire), Laurie Kiniry (Texas), Diane Shields (Delaware), Susan Davis (Maryland), Marjorie Faber (Connecticut), Tammy Cheever (Nebraska), and Deborah Anderson (North Carolina). In the last few years, women soil scientists have been instrumental in the effort to digitize the soils information that is used in the publications, including Vivian Owen (Texas), Jennifer Brookover (Sweet) (Texas), Darlene Monds (Massachusetts), Barbara Alexander (Connecticut), Caroline Alves (Vermont), Lindsay Hodgman (Maine), Caryl Radatz (Missouri), Adrian Smith (Nebraska), Amanda Moore (Oregon), Sharon Schneider (Oregon), Marcella Callahan (Arkansas), Brandi Baird (Oregon), and Jackie Pashnik (Rhode Island).

In the National Cooperative Soil Survey there are also a number of women field soil scientists who work mostly with soil survey interpretations and education, including Sue Southard (California – volcanic soils and vertisols), Lenore M. Vasilas (Maryland – hydric soils), Sheryl Kunickis (Washington, DC – landscape analysis), Susan Ploetz (Minnesota – resource inventory), Susan Casby-Horton (Texas – soil geomorphology), Christine Clarke (Maryland – geographic information systems), Jeannine Freyman (Virginia), Karen Kotlar (New York), Lisa Krall (Connecticut), Gay Lynn Kinter (Michigan), Donna Hinz (Nebraska), Patricia Wright-Koll (Minnesota), Jeanette Bradley (Arkansas), and Deborah Prevost (Nevada). Like agricultural extension specialists, these soil scientists act as a bridge between university research, soil survey mapping, and the public, interpreting soil surveys for practical use by both agencies and individuals, including providing on-site field investigations.

Women in the SCS, now the Natural Resources Conservation Service (NRCS), have made significant contributions to soil science in the National Soil Survey Laboratory (NSSL) and the National Soil Survey Center (NSSC), and as researchers in the Soil Quality and Watershed Sciences Institutes. Carolyn G. Olson has been a lead research scientist at the NSSC, located in Lincoln, Nebraska, since 1989. Dr Olson's research focuses on soil geomorphology, quaternary geology, and clay mineralogy. Olson received the honor of being made a Fellow of the Soil Science Society of America in 1996. Other women soil scientists at NSSL and NSSC include Rebecca Burt (soil chemical properties), Joyce Scheyer (urban soil properties), Susan Samson-Liebig (soil quality), Deborah Harms (soil physical properties), Sharon Waltman (national soil survey databases and GIS interpretations), and Carol Franks (soil biology). In the Institutes, Arlene Tugel (New Mexico), Betty McQuaid (North Carolina), and Cathy Seybold (Oregon) have been working with soil quality and watershed health indicators.

Other federal agencies, such as the USGS, also provide opportunities for women in soil science research. Jennifer W. Harden, with USGS in Menlo Park, California, built on her PhD research using soil chronosequencing to develop the Harden Index, which used soil horizons and carbon-dating to measure time in the alluvium sequencing. Since then, she has worked on the effect of climate on soil, particularly as it relates to groundwater recharge and wetland assessment. She has been a front-runner in research on global change issues of soil carbon, carbon dioxide emissions, and soil carbon sequestration. Marith Reheis, with the USGS in Denver, has

also done significant research in using soil properties as a paleoclimatic record for chronosequence mapping in Rocky Mountain glacial outwash. Originally a geologist by training, she received her PhD in soil science under Pete Birkeland at the University of Colorado, Boulder, in 1984.

At NASA's Goddard Space Flight Center, Elissa Levine has been working with soils in forested ecosystems since 1987. She models soil physics and soil chemistry to assess watershed leaching, soil carbon ecosystem effects, and the effects of acid precipitation on soils and groundwater. She was recently appointed lead scientist for the NASA Global Change Master Directory. She was also selected as a Fellow of the Brandwein Institute for Science Education, an award based on her work as the principal scientist in the GLOBE program's soil investigations for teaching soil science worldwide to K-12 students.

Since the 1970s, women scientists with the US Forest Service have been involved with the National Cooperative Soil Survey ecological unit inventories, as well as with technical soil interpretations in the specialties of forest soil productivity, soil erodibility, fire ecology, and forest ecosystem health. Some of the prominent women involved in this effort include Gretta Boley (Washington, DC), Clare Johnson (Six Rivers National Forest, California), Carol Smith (SCS-USDA and Tahoe National Forest, California), Barbara Leuelling (Superior National Forest, Minnesota), Connie Carpenter (White Mountain National Forest) and Mary Beth Adams (Northeast Forest Experiment Station, West Virginia).

In 2000, three women held positions as pedology (soil genesis) professors in US universities: Janice L. Boettinger, Utah State University, Logan; Christine Evans, University of New Hampshire, Durham; and Mary Collins, University of Florida, Gainesville. Dr Boettinger is working on an extensive review of worldwide zeolite mineral occurrences in soils and the use of zeolite and clinoptilolite for waste disposal systems of animal production operations. She is also working on characterizing selected soil resources of Utah, which includes research on saline, wet soils and irrigation-induced hydric soil characteristics. Dr Evans is focusing her research in the field of describing anthropogenic (human-influenced) soils and developing terminology to describe soil properties derived from human activity.

Dr Collins' research at the University of Florida focuses on the genesis, morphology, and classification of soils; identifying and delineating hydric soils; using ground-penetrating radar to study subsurface properties; and pedoarcheology. She is best known for her dedication to soil survey fieldwork and reaching out to other countries to spread soils technology. As part of the People to People Program, she first opened the door to doing ground-penetrating radar soil investigations in China and Portugal. Dr Collins was made a Fellow of the American Society of Agronomy in 1996 and a Fellow of the Soil Science Society of America in 1997.

A number of women who have been made Fellows of the Soil Science Society of America have provided outstanding contributions to soil science. Mary Beth Kirkham, made a Fellow in 1987, is a professor at the Kansas State University Evapotranspiration Laboratory. Her work has focused on heavy-metal uptake by plants and soil–plant–water relations for over 20 years. Mary K. Firestone, who became a Fellow in 1995 and also received the Emil Truog Soil Science Award, is a professor of soil microbiology at UCB. Her research focuses on the microbial population-basis of carbon and nitrogen processing in ecosystems. Jean L. Steiner, made a Fellow in 1996, is director of the USDA Agricultural Research Service (ARS) Southern Piedmont Conservation Research Laboratory in Watkinsville, Georgia. Her research is on humid region water balance studies in complex topographies. Finally, Diane E. Stott, a 1997 American Society of Agronomy Fellow, is a soil microbiologist with ARS in West Lafayette, Indiana. Her work has focused on the effects of organic matter dynamics on soil structure and erodibility and modeling the effects of plant residue decay on erodibility.

The women described above are only a sample of the many women who have contributed research, mapping, applications, and education to the field of soil science in the last 25 years. Others, to name only a few more, include Nancy Cavallaro (University of Puerto Rico – soil chemistry and tropical soil fertility), Laurie Drinkwater (Rodale Institute, Pennsylvania – sustainable agriculture), Kate Scow (University of California, Davis – soil microbiology), Laurie Osher (University of Maine, Bangor – soil science), Samantha Langley (University of Southern Maine, Gorham – soil science education), Kate Showers (Boston University – soil conservation), Jeri Berc (NRCS-USDA – soil conservation), and Katherine Newkirk (Woodshole Marine Biological Laboratory – global warming).

1990 and Ahead

Clearly, women have made significant and numerous contributions to the field of soil science and soil survey through research, mapping, applications, and education. As the numbers of women have increased in the classroom, laboratory, and field, changes have also taken place within the soil science discipline. The women who are graduating in soil science in the 1990s are confident and intellectually engaged and

are quickly gaining recognition for their work. As an example, Eva M. Muller of Spokane, Washington, a soil survey project leader with only 7 years of experience, was awarded the National Cooperative Soil Survey Soil Scientist of the Year Award in 2001. Lenore M. Vasilas, who finished her MS in soils in 1997, remarked about any remaining stereotypes in her work, "Oh we don't think about it... We just go ahead and do it!"

There is a world of difference between Julia Pearce's experience in the early 1900s and Lenore Vasilas's reality in 1998. In the educational realm, between 1987 and 1996, soil science, along with education, communication, and social science, experienced the largest percentage growth of female participation. While overall enrollment of students (BS, MS, and PhD) in the soil sciences held relatively steady between 1987 and 1996, fluctuating between 1200 and 1500 students, enrollment of women in the soil sciences rose from 16.2% in 1987 to 32% in 1996. In 1996 there were 228 female BS graduates in soil science, almost double that of 10 years before. PhD and MS candidates in the soil sciences in 1996 were also about one-third female, once again double from 10 years before.

Progress in employment numbers in the SCS/NRCS appears a little less dramatic. In 1985, the SCS employed 85 women soil scientists at the federal level; in May 1998, there were 94 women soil scientists in various level positions. The progress is significant, however, given the overall reduction in the total number of soil scientists in the agency. Currently, about half of the NRCS's new soil scientists are women. We can anticipate a twenty-first century that witnesses a continuing trend of more women working in the field in soil survey and private consulting firms and as teachers and researchers in the field of soil science in university and laboratory settings across the USA and the world.

Acknowledgment

This article is adapted, with permission of Iowa State University Press from Levin MJ (2002) Opening opportunities: women in soil science and the soil survey. In: Helms D, Effland ABW, and Durana PD (eds) *Profiles in the History of the US Soil Survey*, pp. 149–168. Ames: Iowa State University Press.

Further Reading

Baker L (1976) Women in the US Department of Agriculture. *Agricultural History* 50(1): 190–201.

Bureau of the Census (1911) *Official Register: Persons in the Civil, Military, and Naval Service of the United States, and List of Vessels*, vol. 1. Washington, DC: Bureau of the Census, Department of Commerce and Labor.

Cattell J (ed.) (1944) *American Men of Science: A Biographical Directory*. Lancaster, Penn.: Science Press.

Cattell J, Cattell G, and Hancock D (eds) (1961) *American Men of Science: A Biographical Directory, the Physical and Biological Sciences*, 10th edn. Tempe, AZ: Jaques Cattell Press.

Helms D (1992) Women in the Soil Conservation Service. *Women in Natural Resources* 14(1): 88–93.

Jaques Cattell Press (ed.) (1972) *American Men and Women of Science, Formerly American Men of Science, the Physical and Biological Sciences*, vol. 4, 12th edn. New York: Jaques Cattell Press/R. R. Bowker Company.

Kellogg C (1963) Opportunities for soil scientists and agronomists in the Soil Conservation Service. *Agronomy Journal* 55: 575–576.

Lapham MH (1945) The soil survey from the horse-and-buggy days to the modern age of the flying machine. *Soil Science Society of America Proceedings* 10: 344.

Lapham MH (1949) *Crisscross Trails: Narrative of a Soil Surveyor*. Berkeley, CA: Willis E. Berg.

Prevost D and Linsay BA (1999) *Soil Survey of Hualapai-Havasu-pai Area, Arizona, Parts of Coconino, Mohave, and Yavapai Counties*. Washington, DC: Natural Resources Conservation Service, US Department of Agriculture.

WORLD SOIL MAP

H Eswaran and P F Reich, USDA Natural Resources Conservation Service, Washington, DC, USA

Introduction

Soil is commonly described as the mantle or the skin covering the land mass of planet Earth. In some places, bare rock is exposed and the skin is absent or may be just a few centimeters thick. In other locations, the mantle is several meters thick with distinct layers that reveal the formation or depositional history of the mantle. For most people, soil is the natural medium for the growth of nonaquatic plants and has a thickness that is determined by the rooting depth of plants. It is composed of solids (mineral and organic), liquids, and gases. Soil scientists differentiate between the initial material (if it was a sediment) and the soil derived through soil-forming processes. The upper

limit of the soil is the contact between soil and air or shallow water (less than 2.5 m deep). The soil on a landscape is a continuum, and across the landscape it grades into nonsoil materials such as rock, ice, or deep water.

The study of soil is 'pedology' and the scientists who study it are referred to as pedologists or soil scientists. The soil that is studied is the zone that extends to the nonsoil material (if the latter occurs at less than 200 cm) or to approximately 200 cm if the nonsoil material occurs deeper, as this is the zone where much of the biological or pedological processes operate. This is a convention, though; for some studies, depths of 20 m or more are considered. The different climates that prevail in the world and their interactions with the rocks or sediments produce a range of soils that vary in their physical, chemical, biological, and mineralogical properties. The kind of soil is also determined by the geomorphology of the land, the amount and rate of water moving in and on it, and the kind of disturbances it has been subject to, including human influence.

It is important to understand and know the different kinds of soils and how they form, as this helps in sustainable land management. Soil classifications, created to help organize knowledge about soils, have been structured to enable this. An important difference between classification of soils and classification of plants and animals is that, unlike the latter, soil is not discrete but a continuum, and as a consequence the boundaries between named soils are artificial. Depending on the objectives for which the classification was created, the class definitions and the architecture of the classification system may vary. Many countries have national soil classification systems but one that is widely used and employed is Soil Taxonomy. Soil surveys enable the depiction of soils across a landscape and soil maps are made to show the patterns of soils that exist and provide information on the management properties of the soils. Specialized maps with technical soil classifications are produced for specific objectives. The soil maps are produced at different scales: detailed maps (1:10 000 or larger scale) to show soils on farms, and general maps (1:100 000 or smaller scale) to depict soils over large areas such as countries, continents, or the world. Soil classification has the additional function of linking the information of these different maps.

Soil Classification

To appreciate the variety of soils that occur on this planet, it is necessary to understand the classification system used to demonstrate this. Two classification systems are widely used: the Food and Agriculture Organization of the United Nations (FAO) initiated the World Reference Base for soil classification; while Soil Taxonomy is a system developed in the USA with the collaboration of international soil scientists. Soil Taxonomy has six categories, as listed in Table 1. Each category serves specific functions determined by its information content. The soil series is the lowest category and is defined with the maximum amount of information. Information is generalized in the definition of each higher category. Consequently, each category can only be used for specific purposes. When employed to designate map units, the scale of the maps limits the category that is represented, as shown in Table 1. The highest category, the order, has only 12 classes. There are 64 classes in the suborder, 300 classes in the great 10 470 group, and 2400 classes in the subgroup. In the USA soil families and 21 000 soil series are recognized. No global estimates for the number of classes in these two categories exist.

Soils are grouped into 12 orders in Soil Taxonomy. Figure 1 shows the global distribution of the soil orders and Figure 2 shows representative examples. Soil-forming processes (Table 1) and the attributes they provide for the soil are used to define the soil. This is expressed in another manner in Table 2. As soils are composed of mineral or organic materials or mixtures, when these primary components dominate the system, the resulting properties define the soil: short-range minerals in Andisols, smectitic clays in Vertisols, organic matter in Histosols, and low-activity clays in Oxisols. In the second group of soils, processes listed in Table 1 result in horizons of accumulation or depletion. The next group of soils (Inceptisols and Gelisols) also results from soil-forming processes but the manifestations are slight or mainly physical. Finally, Entisols are the soils where the original material is least altered or not modified by soil-forming processes. As climate or hydrology controls many of the processes, each of these major kinds of soils has a specific place on the landscape. However, each soil may have properties transitional to other soils, and such secondary features are used to define the intergrades or the lower categories. Each order of soils also has specific subordinate properties that are employed to define the lower categories.

Unlike other soil classification systems, Soil Taxonomy incorporates soil climate by using the soil moisture regime (SMR) and the soil temperature regime (STR) to define lower categories. Soils with an aridic SMR (deserts) occupy approximately 36% of the total land area, making such land only available for agriculture if there is a source of irrigation. The soils with a xeric SMR (Mediterranean) occur in areas with winter rains and, like those with an ustic

Table 1 Defining characteristics of the categories in soil taxonomy

Category	*Definition*	*Functions*	*Potential uses*
Order	Soils having properties (marks) or conditions, resulting from major soil-forming processes that are sufficiently stable pedologically and that help to delineate broad zonal groups of soils	Depict zones where similar soil conditions have occurred for general understanding of global patterns of soil resources Establish global geographic areas within which more specific factors and processes result in the diversity of soils	General global, continental, or regional assessment Global climate-change studies AMS $<$1:10 000 000 MSD $>$40 000
Suborder	Soils within an order having additional properties or conditions that are major controls or reflect such controls on the current set of soil-forming processes and delineate broad ecosystem regions	Demarcate broad areas where dominant soil-moisture conditions generally result from global atmospheric conditions Delineate contiguous areas with similar natural resource endowments Ecosystems with distinct vegetation affinities usually determined by limiting factors of soil moisture or conditions	Demarcate areas in regions or large countries for assessment and implementation of economic development Analysis of international production and trade patterns Priority setting for multipurpose uses of land resources AMS 1:1 000 000–1:10 000 000 MSD 4000–40 000
Great group	Soils within a suborder having additional properties that constitute subordinate or additional controls or reflect such controls on the current set of soil-forming processes, including landscape-forming processes	To demarcate contiguous areas with similar production systems or performance potentials	Development of strategic plans for regional development Basis for coordinating national resource assessment and monitoring programs Infrastructure development to assure equity in development AMS 1:250 000–1:1 000 000 MSD 250–4000
Subgroup	Soils within a great group having additional properties resulting from a blending or overlapping of sets of processes in space or time that cause one kind of soil to develop from, or toward, another kind of soil: Intergrades show the linkage to the great group, suborder, or order level; Extragrades have sets of processes or conditions that have not been recognized as criteria for any class at a higher level, including nonsoil features The soil is considered as the 'typic' member of the class if the set of properties does not define intergrades or extragrades	To demarcate production land units with similar land use and management requirements	Targeting research and development for specific land use or cropping systems Implementing conservation practices and ecosystem-based assistance Community development projects and monitoring sustainability Basis for diversification of agriculture and uses of land resources Basis for implementing environmental management programs and modeling ecosystem performance AMS 1:100 000–1:250 000 MSD 40–250
Family	Soils within a subgroup having additional properties that characterize the parent material and ambient conditions The most important properties are particle size, mineralogy, and soil temperature regime	To demarcate resource-management domains characterized by similar management technology and production capabilities	Basic units for extension and/or technology transfer Modeling cropping systems' performance Addressing socioeconomic concerns AMS 1:25 000–1:100 000 MSD 2.5–40
Series	Soils within a family having additional properties that reflect relatively narrow ranges of soil-forming factors and processes, determined by small variations in local physiographic conditions, that transform parent material into soil	To delineate land units for site-specific management of farms	Implementing soil-specific farming Designing farm-level conservation practices AMS $>$1:25 000 MSD $<$2.5

AMS, appropriate map scale; MSD, minimum size delineation: smallest area that can be delineated on a map with a legible identification, in hectares.

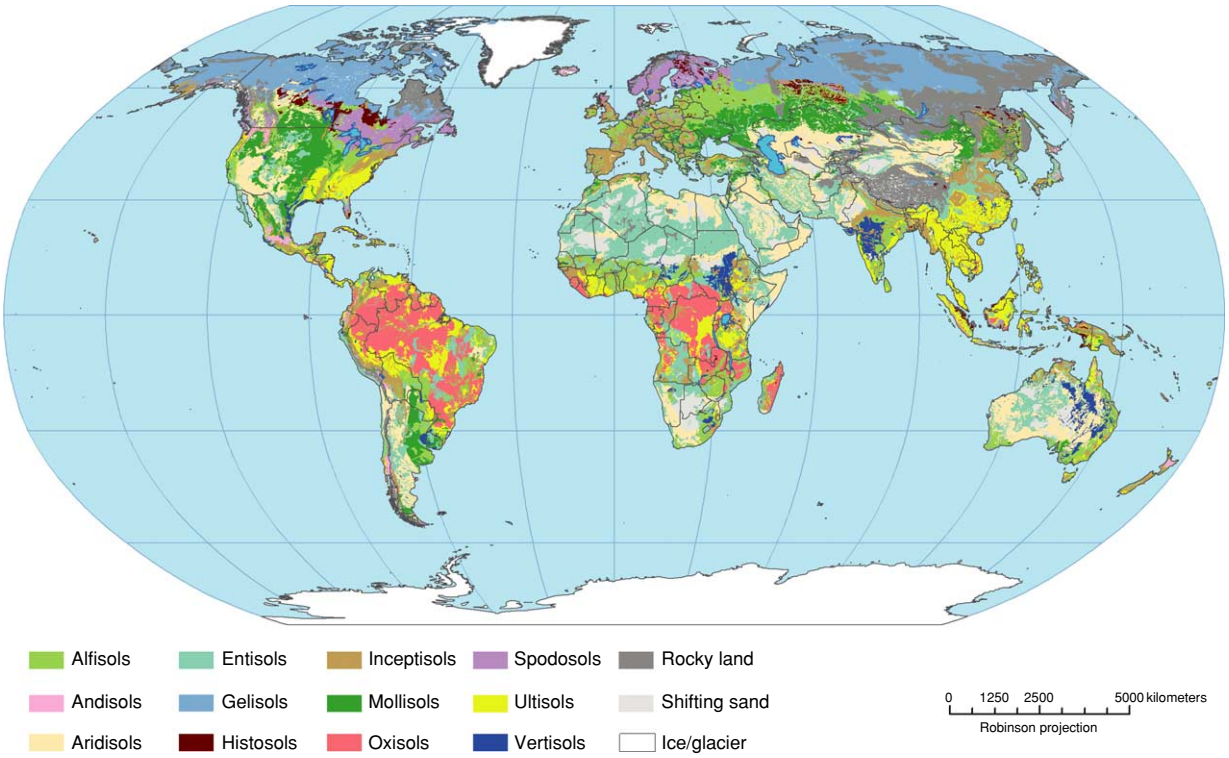


Figure 1 A generalized soil map of the world.

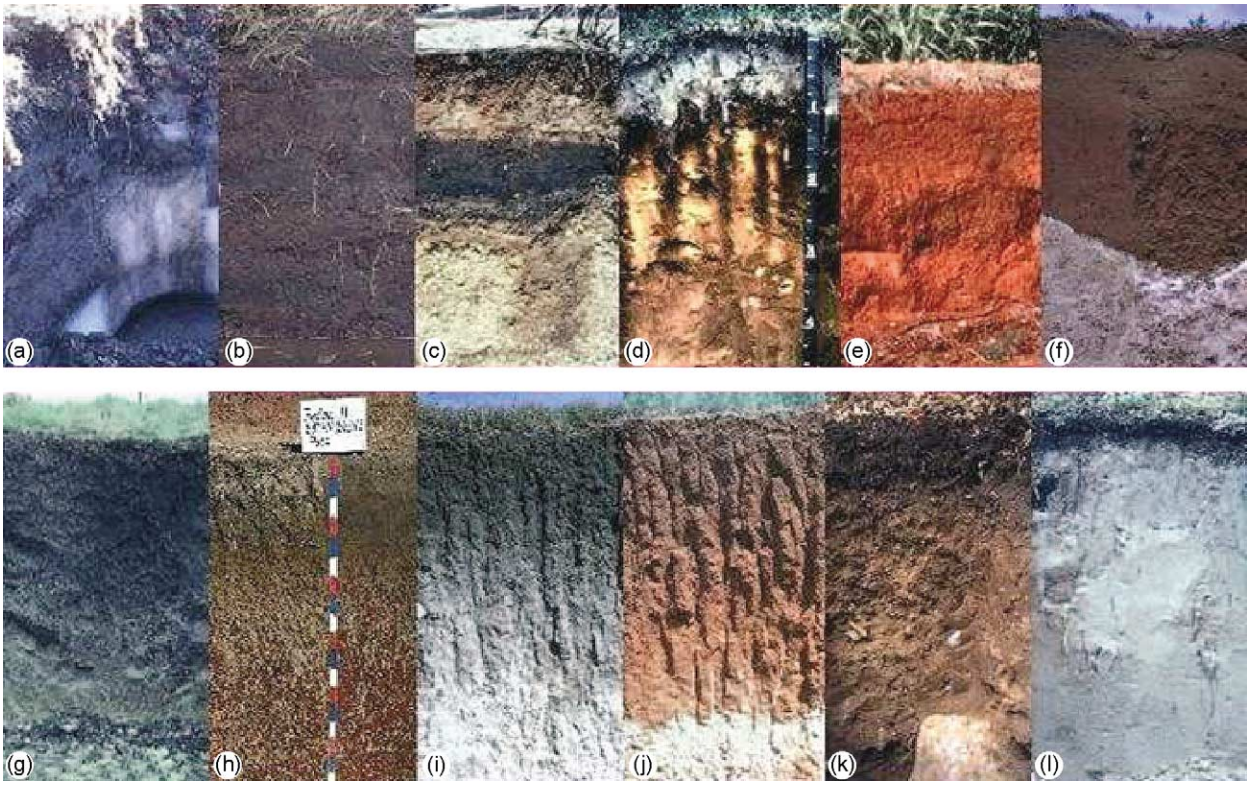


Figure 2 Examples of major soil orders: (a) Gelisol; (b) Histosol; (c) Andisol; (d) Spodosol; (e) Oxisol; (f) Aridisol; (g) Vertisol; (h) Ultisol; (i) Mollisol; (j) Alfisol; (k) Inceptisol; and (l) Entisol.

Table 2 Grouping of soils based on major differentiating criteria

Properties determined by the soil material	*Horizon results from an accumulation or depletion*	*Horizon formed by alteration of primary minerals or of aberrant properties*	*Features of sediment (stratification) retained*[a]
Histosols	Alfisols	Inceptisols	Entisols
Andisols	Ultisols	Gelisols	
Oxisols	Mollisols		
Vertisols	Spodosols		
	Aridisols		

[a]Only a small accumulation of organic matter.

SMR (semiarid), have moisture stress for prolonged periods during the year. These two groups of soils occupy approximately 21.5% of the land and form the semiarid zones, where a second crop requires supplemental irrigation. The soils with an aquic SMR represent the wetlands, which generally are ecologically fragile systems. Thus only approximately 33% of the total land is relatively free from moisture stress, although low temperatures may also stress part of this. The very cold pergelic soils, which occupy approximately 11% of the land, are generally under sparse forest vegetation or are free of vegetation. The cryic and frigid (boreal) soils are generally under forest but are cultivated in some areas with an annual grain crop of short duration. The temperate climates are reflected by soils with mesic, thermic, isomesic, and isothermic STRs, and in general if the SMR is conducive, they are the soils with the highest agricultural potential. The humid tropical soils are represented by the combination of an isohyperthermic STR and udic or perudic SMRs. They occupy approximately 9% of the land surface and have not only some of the poorest soils, but also a heavy incidence of pests and diseases. Each combination of SMR and STR represents a unique pedoenvironment that also contains a myriad of soils.

The distribution of the different kinds of soils is presented in Table 3. The Aridisols are the largest order of soils, occupying approximately 23% of the land. The Histosols, Andisols, and Vertisols individually occupy less than 3% of the area. These are soils formed on special materials or under special conditions and so occupy specific geographic areas. The Gelisols are confined to the very cold regions of the world. Due to intense weathering and leaching by the tropical environment, most of the Oxisols and Ultisols are located in the tropical zones. The unique soil-formation conditions of the tropics are also reflected in the distribution of the Alfisols and Mollisols, which are more extensive in the temperate regions. Inceptisols and Entisols are present under a range of climatic environments. Estimates of the distribution of the suborders of soils are also given in Table 3.

Global Distribution

Gelisols

In areas where the mean annual soil temperature is less than 0°C, the soils are frozen for long periods of the year and thaw out during the short warmer spells. The freezing and thawing processes promote physical changes in the soil. If there is sufficient water and the warm period is long enough, vegetation is established and organic matter accumulates on the soil. Organic-rich soils or peat develop. Due to low temperatures, these Arctic soils have unique features such as ice-lenses, or a layer of ice may underlie the soil. In situations where free water is limited, permafrost layers are present. Depth to the permafrost layer is determined by several factors, including the protective organic-rich layers on the soil surface. In arid areas, the soil particles are held together by dry permafrost. A thin film of frozen water forms a binding film around the particles. The three suborders in Gelisols are:

1. Histels: characterized by a histic epipedon with ice lenses; may be underlain by ice or permafrost, or rock;
2. Turbels: characterized by disrupted horizons or ice wedges penetrating upward into horizons;
3. Orthels: characterized by permafrost layers at depth; but the surface horizons do not show disruptions or distortions due to freezing and thawing. Frequent in the transition to warmer soils.

Histosols

Most soil classifications, including Soil Taxonomy, separate mineral soils from organic soils. Histosols are soils that consist of dominantly organic soil materials. Histosols comprise only a small portion of the world's land area (1.2%), but are widely distributed irrespective of climate. They develop where the rates of organic matter accumulation exceed decomposition and removal. Most of these soils have formed

under saturated conditions where the soil is saturated or nearly saturated with water most of the year. These soils are referred to as bogs, moors, peat, or mucks.

To be farmed, most Histosols must be drained. Management of the water table depth is critical to their use. When drained, Histosols oxidize and subside, and require further drainage. The rate of subsidence can be several centimeters per year and is affected most by depth of drainage. Experience has shown that Histosols in cool temperate areas have the lowest levels of subsidence under tillage. Climate is the main limitation for crops grown on Histosols. Many are used for vegetable crops. Fire and wind-erosion hazards increase after Histosols have been drained.

Histosols are and have been mined as a source of fuel and as a source of organic matter for soil amendment. Because of their extreme instability, Histosols present special problems when roads or other structures are built on them. They are removed if possible during road construction, and buildings constructed on them are normally placed on pilings driven into the mineral soil below them.

Vegetation found on Histosols ranges from grasses, sedges, and rushes to shrubs and trees, all of which are usually water-loving or water-tolerant. Most recently Histosols have been recognized for their importance worldwide as wetland wildlife habitats. Those associated with open bodies of water are especially important habitats. Histosols are recognized as equally important filters of pollutants such as phosphates, nitrates, and other agricultural and industrial contaminants that migrate in water.

The suborders of Histosols are separated by the degree of decomposition of the organic matter, and climate:

1. Folists consist of leaves, twigs, and branches resting on rocks, stones, or coarse fragments in which the interstices are partly filled with organic materials. They are not saturated with water for more than a few days per year if rock is present at shallow depths; they occupy a small area (0.14 million km^2), mainly in the northern latitudes;
2. Fibrists consist largely of plant remains so slightly decomposed that rubbing does not destroy them and their botanical origin can be determined easily. They occupy large areas (1.1 million km^2) in the northern latitudes;
3. Hemists consist of organic materials that are so decomposed that the botanic origin of as much as two-thirds of the materials cannot be readily discerned or the fibers can be largely destroyed by rubbing between the fingers. They occupy small areas in the northern latitudes and may also be found in the tropics;
4. Saprists consist of almost completely decomposed plant remains and their botanic origin cannot be determined. Only small areas of these soils are found in the cool and cold regions of the world; most of them are confined to the tropics.

Andisols

Soils formed on volcanic ash and cinders and having andic properties are distributed along the circum-Pacific belts and occur sporadically elsewhere (1.8% of the land mass). The mineralogical composition is a function of the age of the deposit and the climatic conditions it has been exposed to. The Andisols have mineralogical composition ranging from volcanic glass, short-range-order minerals such as allophane and immogolite, and variable amounts of halloysite. This mineralogical association gives unique properties to such soils, including a high phosphate-fixing capacity, low cation retention, and a high water-holding capacity. Many of these soils are found on volcanic slopes or are developed through the weathering of plateau basalts. These soils support a high human population density owing to their general ease of cultivation and also because of the cool environment of the volcanic mountains, which is generally free of pests and diseases.

The success of human habitation on Andisols is largely due to the soil and climatic environment. Though phosphorus availability is a problem, there is usually a sufficient supply to meet most needs. For intensive cultivation, fertilizers are needed. The soils are friable and well drained so tillage is not a problem. As the soils have been used for generations, conservation practices have been installed in most situations.

Seven suborders are recognized in the Andisols. These are:

1. Aquands: these soils are wet and experience extensive water saturation for prolonged periods during the year. Aquands are local and do not occupy enough land to be shown on small-scale maps;
2. Cryands: these soils have a cryic or pergelic STR. Cryands are found in small areas of Canada, the Kamchatka peninsula of Russia, and high elevations in the tropics;
3. Torrands: these soils have an aridic SMR. Torrands have been reported in the volcanic areas of Mexico and Syria; many have not been studied in detail and so classification is still in doubt;
4. Xerands: these soils have a xeric SMR. Xerands occur sporadically in the Mediterranean areas of

Table 3 Estimates of areas occupied by major soils of the world

Soil		*Ice-free land*		*Tropical*		*Temperate*		*Boreal*		*Tundra*		*Aridic*		*Xeric*		*Ustic*		*Udic*	
Order	*Suborder*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*
	Histels	1 013 358	0.77	0	0	0	0	0	0	1 011 295	0.77	0	0	0	0	0	0	0	0
	Turbels	6 332 748	4.84	0	0	0	0	0	0	6 316 202	4.83	0	0	0	0	0	0	0	0
	Orthels	3 914 016	2.99	0	0	0	0	0	0	3 903 320	2.98	0	0	0	0	0	0	0	0
Gelisols		11 260 122	8.61	0	0	0	0	0	0	11 230 817	8.59	0	0	0	0	0	0	0	0
	Folists	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Fibrists	197 387	0.15	0	0	0	0	194 056	0.15	0	0	11 674	0.01	0	0	42 689	0.03	139 694	0.11
	Hemists	988 264	0.76	0	0	99 508	0.08	884 897	0.68	0	0	103 452	0.08	8 551	0.01	268 205	0.21	604 197	0.46
	Saprists	340 781	0.26	317 753	0.24	22 201	0.02	0	0	0	0	0	0	0	0	40 540	0.03	299 415	0.23
Histosols		1 526 432	1.17	317 753	0.24	121 709	0.09	1 078 953	0.82	0	0	115 126	0.09	8 551	0.01	351 434	0.27	1 043 306	0.8
	Aquods	169 059	0.13	13 364	0.01	55 748	0.04	99 566	0.08	0	0	0	0	0	0	0	0	168 678	0.13
	Cryods	2 459 814	1.88	0	0	0	0	2 455 873	1.88	0	0	11 871	0.01	6 278	0	179 455	0.14	2 258 269	1.73
	Humods	57 870	0.04	29 242	0.02	28 476	0.02	78	0	0	0	78	0	6 546	0.01	10 081	0.01	41 091	0.03
	Orthods	666 784	0.51	18 103	0.01	508 527	0.39	138 221	0.11	0	0	0	0	32 739	0.03	95 150	0.07	536 962	0.41
Spodosols		3 353 527	2.56	60 709	0.05	592 751	0.45	2 693 738	2.06	0	0	11 949	0.01	45 563	0.03	284 686	0.22	3 005 000	2.3
	Cryands	255 195	0.2	0	0	0	0	254 426	0.19	0	0	13 990	0.01	0	0	47 144	0.04	193 292	0.15
	Torrands	1 598	0	1 598	0	0	0	0	0	0	0	1 598	0	0	0	0	0	0	0
	Xerands	32 128	0.02	0	0	32 118	0.02	0	0	0	0	0	0	32 118	0.02	0	0	0	0
	Vitrands	281 070	0.21	202 457	0.15	77 443	0.06	605	0	0	0	26 158	0.02	0	0	161 689	0.12	92 658	0.07
	Ustands	62 822	0.05	58 857	0.04	3 690	0	0	0	0	0	0	0	0	0	62 547	0.05	0	0
	Udands	279 427	0.21	185 821	0.14	89 490	0.07	1 765	0	0	0	0	0	0	0	0	0	277 076	0.21
Andisols		912 240	0.7	448 733	0.34	202 741	0.16	256 796	0.2	0	0	41 746	0.03	32 118	0.02	271 380	0.21	563 026	0.43
	Aquox	320 065	0.24	320 065	0.24	0	0	0	0	0	0	0	0	0	0	300 272	0.23	19 792	0.02
	Torrox	31 233	0.02	27 118	0.02	4 115	0	0	0	0	0	31 233	0.02	0	0	0	0	0	0
	Ustox	3 096 466	2.37	3 086 719	2.36	9 465	0.01	0	0	0	0	0	0	0	0	3 096 185	2.37	0	0
	Perox	1 161 980	0.89	1 010 135	0.77	151 490	0.12	0	0	0	0	0	0	0	0	0	0	1 161 626	0.89
	Udox	5 201 102	3.98	5 166 551	3.95	32 506	0.02	0	0	0	0	0	0	0	0	0	0	5 199 065	3.97
Oxisols		9 810 846	7.5	9 610 588	7.35	197 576	0.15	0	0	0	0	31 233	0.02	0	0	3 396 457	2.6	6 380 483	4.88
	Aquerts	5 484	0	763	0	4 721	0	0	0	0	0	0	0	0	0	0	0	5 484	0
	Cryerts	14 925	0.01	0	0	0	0	14 911	0.01	0	0	34	0	113	0	0	0	14 764	0.01
	Xererts	98 718	0.08	0	0	98 577	0.08	0	0	0	0	0	0	98 577	0.08	0	0	0	0
	Torrerts	889 353	0.68	238 410	0.18	647 662	0.5	0	0	0	0	886 072	0.68	0	0	0	0	0	0
	Usterts	1 767 647	1.35	1 169 403	0.89	594 367	0.45	2 288	0	0	0	0	0	0	0	1 766 059	1.35	0	0
	Uderts	384 358	0.29	86 105	0.07	297 273	0.23	0	0	0	0	0	0	0	0	0	0	383 378	0.29
Vertisols		3 160 485	2.42	1 494 681	1.14	1 642 600	1.26	17 199	0.01	0	0	886 106	0.68	98 690	0.08	1 766 059	1.35	403 626	0.31
	Cryids	943 285	0.72	0	0	0	0	940 532	0.72	0	0	795 230	0.61	20 339	0.02	55 348	0.04	69 615	0.05
	Salids	890 118	0.68	52 910	0.04	632 946	0.48	195 536	0.15	691	0	761 691	0.58	17 320	0.01	95 103	0.07	7 279	0.01
	Gypsids	682 963	0.52	228 484	0.17	429 405	0.33	24 126	0.02	0	0	601 964	0.46	22 692	0.02	57 359	0.04	0	0
	Argids	5 407 965	4.13	573 248	0.44	4 035 105	3.09	782 223	0.6	0	0	5 015 755	3.83	93 618	0.07	268 535	0.21	12 656	0.01
	Calcids	4 872 554	3.73	451 161	0.34	4 400 123	3.36	13 823	0.01	0	0	4 728 720	3.62	71 574	0.05	61 180	0.05	3 637	0
	Cambids	2 931 387	2.24	561 394	0.43	2 063 362	1.58	302 236	0.23	0	0	2 926 976	2.24	0	0	0	0	0	0

Aridisols		15 728 272	12.02	1 867 197	1.43	11 560 941	8.84	2 258 476	1.73	691	0	14 830 336	11.34	225 543	0.17	537 525	0.41	93 187	0.07
	Aquults	1 280 989	0.98	1 042 999	0.8	235 985	0.18	58	0	0	0	0	0	1 337	0	729 665	0.56	548 038	0.42
	Humults	343 518	0.26	277 802	0.21	61 013	0.05	4 691	0	0	0	0	0	18 159	0.01	38 007	0.03	287 339	0.22
	Udults	5 539 906	4.24	2 654 476	2.03	2 872 711	2.2	9 523	0.01	0	0	0	0	0	0	0	0	5 536 706	4.23
	Ustults	3 869 722	2.96	3 630 467	2.78	234 877	0.18	1 471	0	0	0	0	0	0	0	3 866 828	2.96	0	0
	Xerults	18 815	0.01	0	0	940	0	17 875	0.01	0	0	0	0	18 815	0.01	0	0	0	0
Ultisols		11 052 950	8.45	7 605 744	5.81	3 405 526	2.6	33 618	0.03	0	0	0	0	38 311	0.03	4 634 500	3.54	6 372 083	4.87
	Albolls	27 656	0.02	0	0	1 372	0	26 266	0.02	0	0	20 965	0.02	0	0	18	0	6 656	0.01
	Aquolls	118 072	0.09	1 156	0	84 787	0.06	31 974	0.02	0	0	0	0	9	0	18 191	0.01	99 717	0.08
	Rendolls	265 827	0.2	120 959	0.09	103 513	0.08	40 986	0.03	0	0	0	0	408	0	15 371	0.01	249 680	0.19
	Xerolls	924 394	0.71	0	0	873 511	0.67	50 365	0.04	0	0	0	0	923 876	0.71	0	0	0	0
	Cryolls	1 163 797	0.89	0	0	0	0	1 160 462	0.89	0	0	0	0	271 415	0.21	588 171	0.45	300 877	0.23
	Ustolls	5 244 636	4.01	184 731	0.14	2 370 624	1.81	2 682 438	2.05	0	0	3 387 540	2.59	0	0	1 850 262	1.41	0	0
	Udolls	1 261 051	0.96	54 220	0.04	1 058 004	0.81	146 619	0.11	0	0	0	0	0	0	0	0	1 258 841	0.96
Mollisols		9 005 433	6.89	361 066	0.28	4 491 811	3.43	4 139 110	3.16	0	0	3 408 505	2.61	1 195 708	0.91	2 472 013	1.89	1 915 771	1.46
	Aqualfs	836 077	0.64	407 123	0.31	373 655	0.29	54 760	0.04	0	0	0	0	9 925	0.01	548 919	0.42	276 695	0.21
	Cryalfs	2 517 693	1.92	0	0	0	0	2 509 517	1.92	0	0	0	0	77 883	0.06	814 915	0.62	1 616 711	1.24
	Ustalfs	5 663 916	4.33	3 773 322	2.88	1 719 484	1.31	165 070	0.13	0	0	0	0	0	0	5 657 884	4.33	0	0
	Xeralfs	896 915	0.69	0	0	848 514	0.65	46 146	0.04	0	0	0	0	894 661	0.68	0	0	0	0
	Udalfs	2 706 299	2.07	616 696	0.47	1 926 532	1.47	158 113	0.12	0	0	0	0	0	0	0	0	2 701 333	2.07
Alfisols		12 620 900	9.65	4 797 141	3.67	4 868 185	3.72	2 933 606	2.24	0	0	0	0	982 469	0.75	7 021 718	5.37	4 594 739	3.51
	Aquepts	3 199 286	2.45	1 498 377	1.15	1 183 134	0.9	502 454	0.38	0	0	0	0	7 500	0.01	1 050 621	0.8	2 125 835	1.63
	Anthrepts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cryepts	456 920	0.35	0	0	0	0	456 109	0.35	0	0	0	0	79 516	0.06	87 881	0.07	288 711	0.22
	Ustepts	4 241 101	3.24	2 804 601	2.14	1 372 212	1.05	60 352	0.05	0	0	0	0	0	0	4 237 166	3.24	0	0
	Xerepts	685 336	0.52	0	0	674 694	0.52	9 760	0.01	0	0	0	0	684 454	0.52	0	0	0	0
	Udepts	4 247 035	3.25	1 755 023	1.34	2 153 364	1.65	333 545	0.26	0	0	0	0	0	0	0	0	4 241 911	3.24
Inceptisols		12 829 678	9.81	6 058 001	4.63	5 383 404	4.12	1 362 220	1.04	0	0	0	0	771 470	0.59	5 375 668	4.11	6 656 457	5.09
	Aquents	116 087	0.09	105 614	0.08	10 222	0.01	0	0	0	0	3 589	0	0	0	83 087	0.06	29 160	0.02
	Psamments	4 428 052	3.39	2 799 768	2.14	1 625 304	1.24	1 333	0	0	0	1 466 816	1.12	56 315	0.04	2 324 022	1.78	579 248	0.44
	Fluvents	2 860 240	2.19	1 017 875	0.78	1 455 341	1.11	368 407	0.28	0	0	830 887	0.64	174 240	0.13	967 198	0.74	869 295	0.66
	Orthents	13 733 032	10.5	2 095 961	1.6	11 264 330	8.61	363 206	0.28	0	0	10 432 873	7.98	620 154	0.47	1 138 660	0.87	1 531 763	1.17
Entisols		21 137 411	16.16	6 019 218	4.6	14 355 197	10.98	732 946	0.56	0	0	12 734 165	9.74	850 709	0.65	4 512 967	3.45	3 009 466	2.3
	Shifting sands	5 321 875	4.07	545 835	0.42	4 680 128	3.58	92 912	0.07	0	0	5 279 510	4.04	7 255	0.01	27 137	0.02	4 972	0
	Rock	13 076 333	10	7 055	0.01	363 324	0.28	3 727 903	2.85	8 956 897	6.85	1 139 000	0.87	151 764	0.12	507 688	0.39	2 299 876	1.76
	Ice	14 640 098																	
Miscellaneous		397 692 802																	
Total		130 796 504	100	39 193 721	29.97	51 865 893	39.65	19 327 477	14.78	20 188 405	15.43	38 477 676	29.42	4 408 151	3.37	31 159 23	23.82	36 341 992	27.79

the Near East, and not much information on them exists;

5. Vitrands: these soils are recent ash or cinder deposits where some weathering and soil formation has taken place. Vitrands are normally very recent deposits such as those around Mount St. Helens in the USA or around Mount Pinatubo in the Philippines. Mostly they occur on steep upper slopes, and no reliable estimates on them are available;

6. Ustands: these soils have an ustic SMR. Ustands are widespread on the leeward side of tropical volcanoes and are generally used intensively for agriculture;

7. Udands: these soils have an udic SMR. Udands are most widely used for agriculture and generally support the highest population density. Large areas of South-East Asia, Central Africa, Mount Cameroon in West Africa, and the Andean range in South America, as well as many of the volcanic islands of the Pacific, including Japan, are the main locations of Udands.

Spodosols

A black, reddish-brown to dark-brown subsoil (spodic) horizon is the primary identifying characteristic of a Spodosol. It is often overlain by a gray to light-gray eluvial horizon. These distinctive and contrasting colors make Spodosols easily identifiable, although there are always exceptions. The simple explanation for this horizon sequence holds that, under cool, humid, or perhumid climates, organic acids from a litter layer leach amorphous mixtures of organic matter and aluminum with or without iron from the eluvial horizon and deposit them in the illuvial spodic horizon. Most Spodosols have formed under such conditions and thus are common in the northern latitudes where most of these soils are to be found. However, Spodosols vary widely depending on climate and other soil-forming factors. In some, the gray (albic horizon) may be absent; in others, it may be more than 2 m thick over a spodic horizon; furthermore, in some the spodic horizon may be cemented and is then called 'ortstein.'

Most Spodosols have few silicate clays. The particle-size class is mostly sandy, sandy-skeletal, coarse-loamy, loamy-skeletal, or coarse-silty. In hot, humid intertropical areas and other warm humid areas, they have for the most part formed in quartz-rich sands that have a fluctuating water table. Spodosols may form rather quickly (several hundred to several thousand years), again depending on climate and other soil-formation factors.

Many Spodosols are forested. They are generally used for forestry, cultivated crops, and pasture. Spodosols are naturally infertile, but with fertilization, commonly additions of large quantities of lime, nitrogen, and phosphorus, they are quite productive. They tie up considerable amounts of phosphorus, at times returning only approximately 0.45 kg of phosphorus for every 2.7 kg applied. In some Spodosols there may be a deficiency in heavy metals such as selenium or cobalt in forage used for ruminants. There are four suborders of Spodosols:

1. Aquods have an aquic SMR or the climate is extremely humid. They have a spodic horizon, with very high organic matter content, and are generally found in depressions or where the soils have an impermeable subsoil;

2. Humods have a spodic horizon enriched with organic matter. They are well drained and are generally found in cool climates, including high elevations in tropical mountains. On sandy coastal plains of the tropical littoral, such soils frequently occur with Histosols, the latter in depressions;

3. Cryods have a cryic STR;

4. Orthods have a spodic horizon enriched with both iron and organic matter. These are the most extensive Spodosols (2.7 million km^2) and are dominant in cool temperate areas of the northern latitudes.

Oxisols

Oxisols are reddish, yellowish, or grayish soils. They are most common on the gently undulating surfaces of geologically old surfaces in tropical and subtropical regions where they occupy approximately 11.8 million km^2. Oxisol profiles are distinctive because of the lack of obvious horizons. Their surface horizons are usually somewhat darker in color than the subsoil, but the transition to subsoil features is gradual.

Oxisols consist mainly of quartz, kaolinite, oxides, and organic matter. Both the structure and 'feel' of Oxisols are deceptive. Upon first examination they appear structureless and feel like a loamy particle-size class. While some may be loamy or even coarser, many are extremely clayey, but that clay is aggregated in a strong grade of fine and very fine granular structure. To obtain a true feel of the fine texture, a wet sample must be worked for several minutes in the hand to break down the sandy-textured, granular structure. The strong granular structure apparently causes most Oxisols to have a much more rapid permeability than would be predicted by the particle-size distribution class. Although compaction and reduction in permeability can be caused by cultivation, they are extremely resistant to compaction and so free drainage can take place soon after rain without puddling.

Oxisols are present in every soil-moisture regime from aridic to perudic and aquic. Natural vegetation ranges from tropical rainforests to savannas. Although many Oxisols are extremely infertile, there are some that have small but adequate supplies of nutrients and are immediately productive when cultivated. The reserves of plant nutrients even in the most fertile Oxisols are not great, and to sustain yields, fertilizers and lime are needed after only a few years of cultivation. In most of the Oxisols, fertilizers are needed for the first crop unless enough fertility for one or two crops is available from the ash derived from the burning of natural vegetation. Phosphorus is generally the most restrictive plant nutrient, mainly because of the tendency for the sesquioxide-rich clays to fix large amounts of fertilizer phosphorus. However, once this capacity to fix phosphorus is satisfied by initial applications, phosphorus fixation is no longer a problem.

Road-building and other engineering practices are relatively easy on most Oxisols because of the physical stability of the material. Soil organic carbon is generally much higher than indicated by the color, but, unlike most other soils, much of the carbon is inert and does not contribute to the nutrient-holding capacity or to the physical properties.

The most extensive areas of Oxisols are on the interior plateaus of South America, the lower portion of the Amazon basin, significant portions of the Central African basin, and important areas in Asia, northern Australia, and several tropical islands of the Pacific.

The suborders of the Oxisols are based on the soil moisture regimes:

1. Aquox have an aquic SMR and so are the very wet Oxisols. They occur in depressions associated with the better-drained upland Oxisols;
2. Torrox have an aridic SMR. Such soils are reported in the literature but no contiguous areas have been mapped;
3. Ustox have an ustic SMR. They are the second most extensive kind of Oxisols and occupy large areas in the Brazilian Shield;
4. Perox have a perudic SMR. They represent the most typical Oxisols of the very wet zones of Amazon, Central Zaire, and South-East Asia;
5. Udox have an udic SMR. They are the most extensive of the Oxisols, with large areas in the Amazon area, Central Africa, Borneo, and some of the Pacific Islands.

Vertisols

Vertisols are clayey soils, which have deep, wide cracks on some occasions during the year and slickensides within 100 cm of the soil surface. They shrink when dry and swell when moistened. Vertisols make up a relatively homogenous order of soils because of the amount and kind of clay that is common to them.

In many countries where Vertisols are common, they are known by their local names. For example, Gilgai soils (Australia), Adobe (Philippines), Sha Chiang (China), Black Cotton Soils (India), Smolnitza (Bulgaria), Tirs (Morocco), Makande (Malawi), Vleigrond (South Africa), and Sonsosuite (Nicaragua).

Vertisols generally have gentle slopes, although a few slope strongly. They develop commonly in large pedons and polypedons. The natural vegetation is a function of the soil climate. Most Vertisols are well suited for mechanized farming if there is plenty of rainfall or irrigation water and if suitable management practices are followed. Large areas of Vertisols in the world, however, are not farmed, because their cultivation would require too much energy, especially where traditional, low-input methods are used. This is one of the basic limitations of using Vertisols. Irrigation also presents special problems due to their low saturated hydraulic conductivity. Bypass flow in open cracks is the common situation. Because of their low permeability, irrigation of these soils may result in waterlogging and a buildup of salinity unless adequate artificial drainage is provided. A drainage system designed for Alfisols or Ultisols may be totally inadequate for Vertisols.

Six suborders of Vertisols are recognized based on the SMR or STR:

1. Aquerts: wet Vertisols with an aquic SMR, important locally;
2. Cryerts: cold Vertisols with a cryic or pergelic STR; little information is available;
3. Torrerts: Vertisols with an aridic SMR, present in areas where the SMR is transitional between aridic and ustic, although in most years it is aridic. Large areas are found in Sudan, the Near East, and Central Asia;
4. Xererts: Vertisols with a xeric SMR; the typical reddish-brown Vertisols of the Mediterranean areas, found in Jordan, Turkey, and Tunisia;
5. Usterts: Vertisols with an ustic SMR, geographically the most widespread, occupying approximately 1.8 million km^2; the dominant Vertisols in Africa, India, and Australia;
6. Uderts: Vertisols with an udic SMR, large areas being found in Bengal, some of the Caribbean Islands, Eastern Europe, and Argentina.

Aridisols

Aridisols, as their name implies, are soils that do not have water available to mesophytic plants for long

periods. During most of the time when the soil is warm enough for plants to grow, soil water is held at potentials less than permanent wilting point or if it is salty, or both. There is no period of 90 consecutive days when moisture is continuously available for plant growth.

The concept of Aridisols is based on the low availability of soil moisture for sustained plant performance. In areas bordering deserts, the absolute precipitation may be high but due to runoff or a very low storage capacity of the soil, or both, the actual soil-moisture regime is aridic. In these areas, tree vegetation may exist. Deep-rooted acacias are frequently the dominant vegetation but, if the trees are removed, the soil cannot support general farm crops. In general there is a 70% probability (7 of 10 years) that there will be a crop failure.

Many Aridisols, due to an extreme imbalance between evapotranspiration and precipitation, are similar to incipient evaporites. The dominant process is one of accumulation and concentration of salts. The high salt concentration is an adverse attribute in these soils and is the second most important constraint to the use of the soil. Many soluble salts may be eliminated or changed in concentration through irrigation. In Aridisols, however, the availability of good-quality irrigation water is a fundamental problem; secondly, together with irrigation, a mechanism for evacuation of salts must be provided or a rapid buildup of salinity and/or alkalinity will occur. Thirdly, irrigation and drainage systems must be well maintained to prevent the soils from reverting to their original state.

The classification of Aridisols includes these constraints or performance-restrictive qualities, at a high categoric level. Some Aridisols are also situated on geologic evaporites. It is often difficult to enter these substratum conditions into a classification system, but care must be taken to evaluate these deep-seated salt accumulations, particularly in irrigation projects. Some Aridisols also present inherited features, which may be attributed to earlier wetter or drier paleoclimatic conditions. These attributes, and specifically an argillic horizon, are also considered, as they are important in the use and management of soils.

The suborders reflect the results of dominant soil-forming processes. Unlike many other soils, the redistribution of soluble materials and their accumulation in some layers in the soil is a dominant process in Aridisols. The products of this process not only give special attributes that distinguish the soils, but also present constraints to the use of the soil. Four of the six suborders are defined on the composition and accumulation of the soluble fraction. Weathering and clay translocation also take place in Aridisols, although the expression of the products is not as vivid. The fourth and sixth suborders reflect these processes. The six suborders are:

1. Salids: characterized by accumulation of salts more soluble than gypsum; the typical soils of the playas or desert depressions, or closed basins;
2. Durids: characterized by accumulations of silica, but infrequently found; associated with soils formed from volcanic materials. They have been reported in the western USA but not in other parts of the world;
3. Gypsids: characterized by an accumulation of gypsum; extensive in the Near East, especially in Syria, Iraq, Saudi Arabia, and Iran. They have a total area of approximately 1.1 million km^2;
4. Argids: characterized by an accumulation of clay by translocation, the Argids are present in areas adjoining soils with an ustic SMR. They have not been mapped in large areas of North Africa or the Near East and are currently reported only in the USA. Correct classification of Argids is a problem;
5. Calcids: characterized by an accumulation of carbonates, the soils have a calcic, petrocalcic, or a hypercalcic horizon. There are extensive areas of these soils in the major deserts of the world, the total area exceeding 10.2 million km^2;
6. Cambids: characterized by a transformation of material, these soils are the most extensive of the Aridisols and occupy approximately 13.3 million km^2.
7. A special suborder, cryids, is provided for those aridisols with low soil temperature.

Ultisols

Ultisols are similar to Alfisols in having a subhorizon of clay accumulation but have few bases, especially at depth. Most Ultisols are acid, although some may have a high pH in the surface horizons owing to aerosolic additions of carbonate dust. The ideal Ultisol has a subsurface horizon of clay enrichment due to clay translocation from the surface horizons. However, in areas bordering the deserts, wind-blown sand may bury a former Oxisol, such as in western Zambia and Zimbabwe or in Mali and Niger. These soils have highly weathered subsoil with very low clay activity but, owing to the wind-blown material on the surface, show an increase in clay content with depth. There is little evidence of clay translocation in these soils. Similar situations prevail in the soils of old geomorphic surfaces where the subsoil is heavier-textured and with low clay activity but with little evidence of clay translocation. If the surface horizons have more than 40% clay, for practical purposes, these soils that change in texture with depth are considered as Ultisols. If there is less than 40% clay, they are classified as Oxisols.

The subsurface horizon of clay accumulation is an important feature in the semiarid environments. In general, the loamy soils have low water-holding capacity and so any horizon with more clay enhances the water-holding capacity. Moisture management is a most important feature of the semiarid Ultisols and even in some Ultisols in the humid areas. The additional acid nature of the soils exaggerates the moisture stress in them. Deep liming is essential to enable a proliferation of roots so that a larger volume of soil is exploited for moisture and roots.

The suborders of the Ultisols are based on the SMRs except for the Humults, which are characterized by an accumulation of organic matter:

1. Aquults are wet Ultisols, with an aquic SMR;
2. Humults have accumulations of organic matter and usually occur in areas with an udic SMR and cool temperatures;
3. Udults have an udic SMR;
4. Ustults have an ustic SMR;
5. Xerults have a xeric SMR.

Mollisols

Generally in Soil Taxonomy, it is the presence or absence of subsurface horizons and their characteristics that are used to separate the orders. In Mollisols, however, it is the presence of a thick, dark, humus-rich surface horizon (mollic epipedon) that is the key to placement. In the development of Soil Taxonomy, this surface horizon is the only common characteristic that can be found to tie together the grassland soils of North America, Europe, Asia, and South America. It is an important separation, because these soils are most easily cultivated, generally without irrigation. Thus most Mollisols have had grass vegetation at some time, and melanization – the process of darkening of the soil by organic matter additions – is probably the most important process in the formation of a Mollisol. It is the addition, decomposition, and accumulation of relatively large amounts of organic matter in the soil profile, with the presence of calcium, that forms the central concept of Mollisols. Mollisols have a variety of subsurface horizons and/or diagnostic characteristics, or horizons may be entirely absent.

To a large extent, Mollisols are the breadbasket of the world – the prairies in the USA, the steppes of Russia, and the pampas of Argentina. Most Mollisols are cultivated; in fact there are only limited areas in the world where they have not been cultivated. Mollisols may initially be farmed with no additions of fertilizers. However, to sustain the high yields of corn, soybeans, sorghum, and small grains of today, fertilizers must be used. Soil temperature and moisture are principally used to separate all but two (Albolls and Rendolls) of the seven suborders of Mollisols:

1. Albolls have a bleached subsurface horizon called the albic horizon and are usually wet;
2. Aquolls are wet Mollisols and may have a histic epipedon;
3. Rendolls are shallow and stony soils formed on limestone and have an udic SMR;
4. Xerolls have a xeric SMR;
5. Cryolls have a frigid, cryic, or pergelic STR;
6. Ustolls have an ustic or an aridic SMR;
7. Udolls have an udic SMR.

Alfisols

Most Alfisols were or are forested, with moderate to high base saturation; most formed under deciduous forest. Typically they have a light-colored surface layer over a horizon of silicate clay accumulation (argillic). The cooler Alfisols tend to form a belt between the grassland Mollisols and the Spodosols of the more humid climates. Where temperatures are warmer, they form a belt between the Aridisols and the older Ultisols and Oxisols. Along with Mollisols, Alfisols account for a major portion of soils that are used to grow crops in the world. They are found generally in climates favorable to crop production; in warm moist climates, they are used to grow many crops. They generally contain adequate plant nutrients, but like Mollisols they must be fertilized to obtain high yields. Crop production may become more difficult when these soils are eroded down to the argillic horizon. The higher clay content of the argillic horizon may impede root, water, and air movement.

All five suborders of Alfisols are defined by the SMR:

1. Aqualfs are wet Alfisols;
2. Cryalfs have a cryic STR, or, if the SMR is not xeric, have a frigid STR;
3. Ustalfs have an ustic SMR;
4. Xeralfs have a xeric SMR;
5. Udalfs have an udic SMR.

Inceptisols

The Latin word ‘inceptum’ means beginning and the central concept of Inceptisols is that of soils in the early stages of soil formation. The initial stage of soil formation is exemplified by several attributes, which are the result of the presence or absence of certain processes.

Soil formation on rocks consists of weathering of the rock which is essentially a geochemical process

accompanied by soil-forming processes acting on the weathered products. In cool, humid climates, the soil-forming process may be the accumulation of organic matter to give rise to a mollic or umbric epipedon. In warmer climates, cambic horizon formation takes place; this is expressed by clay formation or release of iron to form a 'color B' horizon. On steep slopes, even if moisture and temperature conditions are conducive to form an argillic horizon, the slow soil loss through erosion retards any of the other horizons from being expressed. Such soils may have just an ochric epipedon and a weak cambic horizon.

Inceptisols are more prevalent on sediments, and a major morphological change is the removal of the original stratification, which takes place through bio- and pedoturbation. Structure and color development are common marks of cambic horizons. If a permanent or a fluctuating water table is present, oxidation–reduction processes leave their marks and are frequently sufficient evidence for a cambic horizon. In extremely wet environments (perudic soil moisture regime) and even on steep slopes, or in soils with a fluctuating water table, weathering of primary minerals may release sufficient iron, which percolates through the soil and accumulates in a subsurface layer to form a thin iron pan or placic horizon. The placic horizon is sufficient evidence to consider the soils Inceptisols.

Some coastal sediments contain pyrite which, when exposed to oxidizing conditions, oxidizes to jarosite, releasing sulfuric acid. The presence of the straw-yellow jarosite aggregates and extreme acidic conditions of the soil is sufficient to place the soil in the order of Inceptisols.

The Inceptisols consequently comprise a wide array of soils, which range not only in properties but also in behavior. The lower categories in the Inceptisols attempt to cluster more homogeneous soils.

Five suborders of Inceptisols are recognized:

1. Aquepts are wet Inceptisols;
2. Plaggepts have a man-made surface horizon thicker than 50 cm called the plaggen epipedon (these soils are not extensive enough to be shown on the map);
3. Ustepts have an ustic SMR;
4. Xerepts have a xeric SMR;
5. Udepts have an udic SMR.

Entisols

The Entisols show little or no evidence of soil formation. They are most extensive on recent alluvial plains and valleys or on steep slopes where erosion is rapid. The rate of soil formation is reduced for several reasons. Generally time has not elapsed since deposition of the material for soil-forming processes to act. In some of these soils, peraquic conditions prevail where the soil is saturated with water during the whole year. The soil is permanently reduced, preventing cambic horizon formation. On steep slopes, rapid erosion results in shallow soils where weathered parent materials rest on hard rock. Rates of soil loss are much greater than soil formation and so cambic horizons do not form. Entisols may also occur on older deposits where, for example, the material is formed from quartzitic sands. There are no primary minerals in the deposit to weather and form clay or liberate iron. In fluvial deposits, Entisols show marked stratification. This is frequently evidence for recent deposits.

Five suborders are recognized in the Entisols:

1. Aquents are wet Entisols;
2. Arents have been subject to deep plowing;
3. Psamments are sandy Entisols;
4. Fluvents are recent alluvial soils showing stratification;
5. Orthents are shallow soils on steep slopes.

In many countries Entisols, on flat alluvial plains or riverine terraces, are widely used for annual crops. Terrain conditions are frequently suitable for low-input agriculture. The major civilizations of the world developed on Entisols and Inceptisols of the larger river terraces owing to good farming conditions and easy navigation.

Summary

A wide diversity of soils exists and each has its own potential and limitations. The performance-related attributes result from the properties and the prevailing environmental conditions. There are large land areas where it is too cold, or too dry, or the soils are on steep slopes and the potential for human use is limited. For most agricultural uses, only about 9% of the global land surface is relatively constraint-free. Understanding the soil and its relation to the rest of the environment is important for its sustainable use. Detailed soil surveys provide such information, but most countries of the world lack access to it. Less information is available on the state of soil resources. Assessment and monitoring of land conditions are essential to the judicious use of soil.

Further Reading

Ahrens RJ and Arnold RW (2000) Soil taxonomy. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 117–135. Boca Raton, FL: CRC Press.

Arnold RW (1990) Soil taxonomy, a tool of soil survey. In: Rozanov BG (ed.) *Soil Classification*, pp. 94–111. USSR State Committee for Environmental Protection. Moscow, Russia: Publishing Centre for International Projects.

Arnold RW, Ahrens RJ, and Engel RJ (1997) Trends in soil taxonomy – a shared heritage. *Communications of the Austrian Soil Science Society* 55: 167–170.

Bartelli LJ (1978) Technical classification system for soil survey interpretation. In: Brady NC (ed.) *Advances in Agronomy* 30: 247–289.

Buol SW, Hole FD, and McCracken RJ (1980) *Soil Genesis and Classification*. Ames, IA: Iowa State University Press.

Deckers JA, Nachtergaele FO, and Spaargaren OC (eds) (1998) *World Reference Base for Soil Resources: Introduction*. International Society of Soil Science (ISSS), International Soil Reference and Information Centre (ISRIC) and Food and Agriculture Organization of the United Nations (FAO). Leuven, Belgium: FAO.

Orvedal AC and Edwards MJ (1941) General principles of technical grouping of soils. *Soil Science Society of America Proceedings* 6: 386–391.

Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. US Department of Agriculture Handbook 436. Washington, DC: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. US Department of Agriculture Handbook 436. Washington, DC: US Government Printing Office.

Sumner ME (ed.) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Wilding LP, Smeck NE, and Hall GF (eds) (1983) *Pedogenesis and Soil Taxonomy*. I. *Concepts and Interactions*. Amsterdam, the Netherlands: Elsevier.

Wilding LP, Smeck NE, and Hall GF (eds) (1983) *Pedogenesis and Soil Taxonomy*. II. *The Soil Orders*. Amsterdam, the Netherlands: Elsevier.

Z

ZERO-CHARGE POINTS

J Chorover, University of Arizona, Tucson, AZ, USA

Introduction

Many soil chemical reactions occur at the interface between solid particles and the aqueous phase. Fundamental to the reactivity of this interface is the surface charge on soil particles, which affects a range of chemical phenomena including adsorption of charged solutes, the nature and extent of the electrical double layer on soil particles, kinetics of colloid aggregation, and mineral dissolution rates. Cation and anion exchange capacities, which influence the retention of plant-available nutrients, are directly affected by the nature of soil surface charge and its response to changes in solution of chemical parameters such as pH, ionic strength, and ion composition.

Charge properties have traditionally been used to characterize soils as 'permanent' or 'variable' charge media. These characterizations reflect the relative predominance of one or more soil solid-phase constituents. Whereas the surface chemistry of permanent charge soils is governed by isomorphically substituted mineral particles (i.e., 2:1 layer-type clays such as smectite and vermiculite), variable-charge soils are dominated by amphoteric hydroxylated surfaces (i.e., metal oxides, hydroxides or oxyhydroxides, and organic matter). As a result of the ubiquity of hydroxylated solids in soil environments, all soils comprise some portion of their total surface charge that is conditional upon aqueous chemical conditions.

The law of surface-charge balance states that, irrespective of the source of particle surface charge, it must be balanced by net accumulation of ions in the surrounding diffuse swarm. The balance of surface charge and relationships among components may be used to constrain the application of molecular models to the particle–water interface and to verify internal consistency of adsorption measurements. Particularly useful in this respect are points of zero charge, which define the physicochemical conditions wherein one or several components of surface charge sum to zero.

Components of Surface Charge

Electric charge develops on soil particle surfaces as a result of: (1) structural disorder or isomorphic substitutions among ions of differing valence within soil minerals, and (2) reactions of surface functional groups with ionic species in aqueous solution. These mechanisms of charge development are incorporated into four 'components of surface charge' that together contribute to the *total particle surface-charge density*, σ_P. Surface-charge density is conventionally expressed in coulombs per square meter, consistent with the International System of Units (SI). For complex solids, including soils, charge is likely to be distributed unequally among the many phases present, owing to differences in structural charge, surface site density, and specific surface area. Furthermore, measurement of total surface area alone is nontrivial and method-dependent. For these reasons, soil charge is often measured on a mass basis and expressed as moles of charge per kilogram. Values presented in each of these sets of units differ by the factor F/a_s, where F is the Faraday constant and a_s is specific surface area.

The *net structural surface-charge density*, σ_0, is created by isomorphic substitutions in soil minerals. These substitutions occur in both primary and secondary minerals, but they produce significant surface charge only in the 2:1 layer-type aluminosilicates (e.g., Figure 1). The *net proton surface-charge density*, σ_H, results from the difference between the moles of protons and the moles of hydroxide ions complexed by surface functional groups (Figure 2). The *net inner-sphere complex surface-charge density*, σ_{IS}, results from the net total charge of ions, other than H^+ and OH^-, which are bound into inner-sphere surface coordination. Inner-sphere complexation involves the formation of one or more direct bonds between the adsorbate molecule and adsorbent surface functional group, with no water molecules

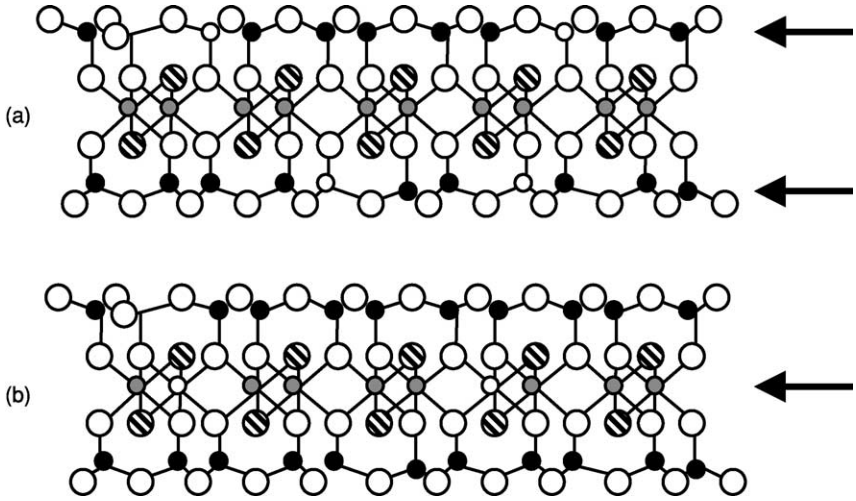


Figure 1 Isomorphic substitution of ions of differing valence gives rise to structural charge density (σ_0) that is independent of solution chemistry. Model structures of vermiculite (a) and smectite (b) show that substitutions (small white spheres) can give rise to charge in the (a) tetrahedral (Al^{3+} for Si^{4+} substitution) or (b) octahedral (e.g., Mg^{2+} for Al^{3+} substitution) sheets of 2:1 layer-type silicates. The location of substitution, particularly in regard to proximity to the surface, affects site reactivity. Large white spheres, O; black spheres, Si; gray spheres, Al; hatched spheres, OH.

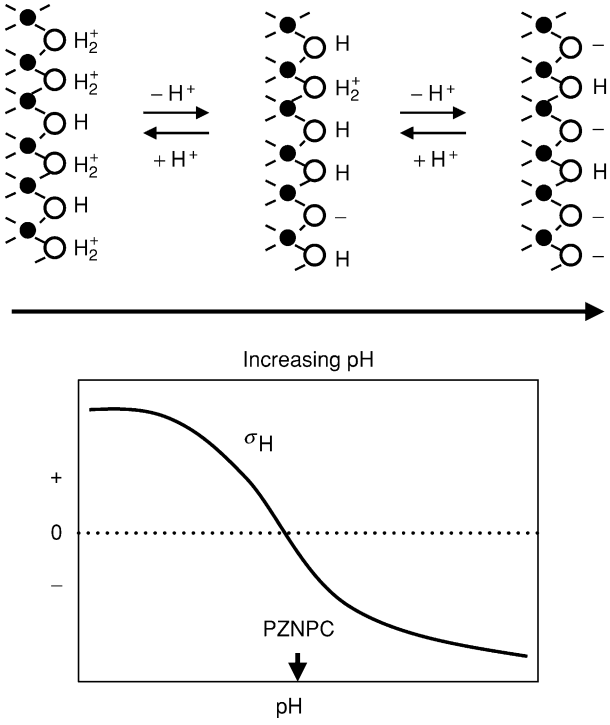


Figure 2 Development of proton surface-charge density (σ_H) derives from the net adsorption of protons on particle surfaces and reflects the Brönsted acid–base chemistry of surface functional groups. PZNPC, point of zero net proton charge.

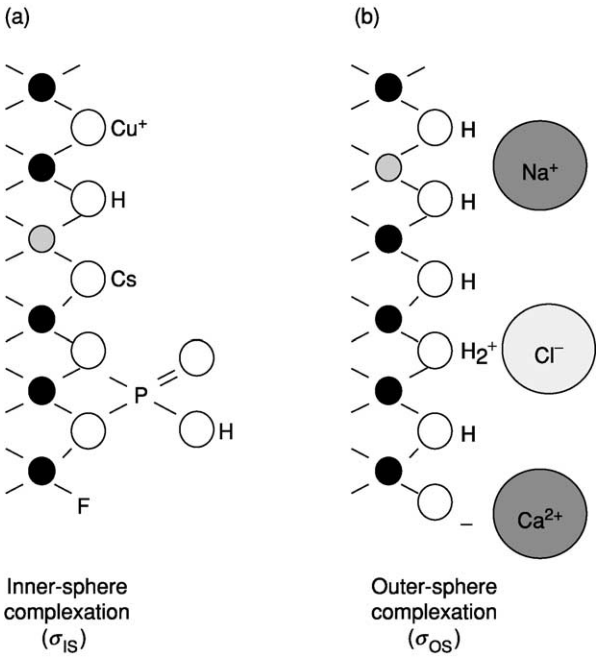


Figure 3 Surface coordination of ions in stable complexes results in the development of (a) inner-sphere complex surface-charge density (σ_{IS}) and (b) outer-sphere complex surface charge density (σ_{OS}), both of which contribute to total particle surface charge. Isomorphic substitutions, are shown schematically as gray spheres.

interposed (Figure 3a). The *net outer-sphere complex surface-charge density*, σ_{OS}, results from the net total charge of ions, other than H^+ and OH^-, which are bound into outer-sphere surface coordination. Outer-sphere complexes involve hydration water interposed between the adsorbate and the adsorbent surface (Figure 3b). The net total particle surface charge is the sum of these four components:

$$\sigma_P = \sigma_0 + \sigma_H + \sigma_{IS} + \sigma_{OS} \quad [1]$$

The right side of eqn [1] indicates contributions to particle surface charge from both permanent (σ_0) and conditional (σ_H, σ_{IS}, σ_{OS}) components; the value of σ_0 is dependent only upon the composition and structure of the solid phase, while the values of the remaining components depend upon the nature of both the solid and the solution phases. The last two terms on the right side of eqn [1] together define adsorbed ions that are immobilized in surface complexes, also known as the 'Stern-layer charge', σ_S.

Surface Charge Balance

Aqueous particle suspensions are electrically neutral. Therefore, if σ_P is nonzero, it must be balanced by a diffuse swarm of ion charge equal in magnitude but opposite in sign to σ_P. These ions differ from those bound into the Stern layer because their much greater mobility relative to the adsorbent surface is analogous to diffusive motions of free ions in aqueous solution. The net charge of such ions defines the *diffuse layer surface-charge density*, σ_d. The charge balance equation for soil particles may then be written:

$$\sigma_P = -\sigma_d \qquad [2]$$

and substituting the left side of eqn [1] into eqn [2] gives:

$$\sigma_0 + \sigma_H + \sigma_{IS} + \sigma_{OS} + \sigma_d = 0 \qquad [3]$$

The terms σ_{IS}, σ_{OS}, and σ_d correspond to different molecular mechanisms of ion adsorption that are difficult to distinguish on the basis of quantitative macroscopic measurements. These three terms in eqn [3] together give the *adsorbed ion charge density*, Δq, which is the net charge of ions, other than H^+ and OH^-, adsorbed at the soil–water interface (**Figure 4**):

$$\Delta q = \sigma_{IS} + \sigma_{OS} + \sigma_d = q_+ - q_- \qquad [4]$$

The value of Δq in eqn [4] can be determined by measuring the difference between *surface excess* of cation charge (q_+ corresponding to all cations other than H^+) and *surface excess* of anion charge (q_- corresponding to all anions other than OH^-). The charge balance may then be rewritten as:

$$\sigma_0 + \sigma_H + \Delta_q = 0 \qquad [5]$$

Points of Zero Charge

Points of zero charge are defined most commonly as pH values for which one or more of the surface-charge components is equal to zero at a given temperature, pressure, and aqueous solution composition. The

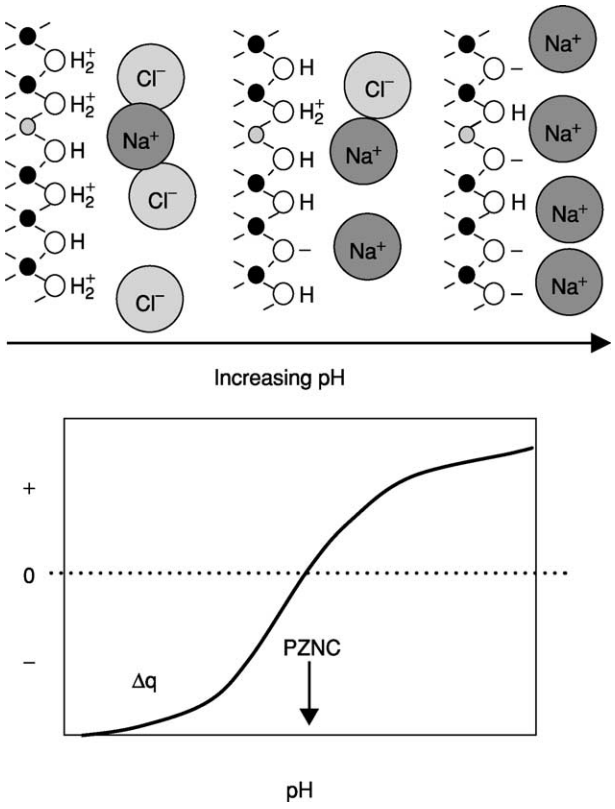


Figure 4 The net charge deriving from adsorption of ions (other than H^+ and OH^-) in surface complexes ($\sigma_{IS} + \sigma_{OS}$) plus the diffuse ion swarm (σ_d) constitute the net adsorbed ion charge density (Δq). PZNC, point of zero net charge.

emphasis on pH derives from the wide and dynamic range of proton concentrations in soil solutions, and the fact that complexation of H^+ and OH^- at soil particle surfaces strongly affects the adsorption of other cations and anions (e.g., see **Figure 4** and eqn [5]). Although pH will be used to illustrate points of zero charge here, it is important to note that one can also define them in terms of aqueous concentrations of other surface complexing ions, such as $H_2PO_4^-$ or Pb^{2+} (**Figure 3a**), which can affect the value of σ_P directly. Three principal points of zero charge are named conventionally as: (1) the point of zero charge (PZC), (2) the point of zero net proton charge (PZNPC), and (3) the point of zero net charge (PZNC).

Point of Zero Charge

The PZC ($\sigma_P = 0$) is defined as the pH value at which the net total particle surface charge, σ_P, is equal to zero. According to eqn [2], σ_d also vanishes at the PZC, so that all adsorbed ions (other than H^+ and OH^-) must be bound into surface complexes. Traditionally, the PZC has been assessed by measuring

the pH value at which soil particles exhibit zero electrophoretic mobility (i.e., the isoelectric point or IEP), assuming that a negligible portion of the diffuse ion swarm is advected during the electrophoresis measurement. However, this assumption has never been verified. Furthermore, model calculations suggest that surface charge heterogeneity or patchiness can also result in a nonzero electrophoretic mobility for a particle whose overall charge is equal to zero. This raises additional questions regarding the validity of equating PZC and IEP values. Values of PZC for soil minerals derived dominantly from electrophoresis measurements are presented in Table 1.

According to the DLVO (Derjuagin–Landau–Verwey–Overbeek) theory, a suspension of colloidal particles should undergo rapid flocculation when net charge in the diffuse ion swarm (σ_d) is reduced to zero. Therefore, the PZC may also be determined by measuring the pH where a colloidal suspension of particles undergoes rapid coagulation (flocculation). However, the influence of surface-charge heterogeneity probably affects collision efficiency as well. New

Table 1 Points of zero charge for selected soil mineral constituents

Solid	*Chemical formula*	*PZNPC*	*PZNC*	*PZC*
Albite	$NaAlSi_3O_8$	3.9	–	–
Allophane	$SiAl_2O_5 \cdot 0.5H_2O$	8.2	7.9	10
Alumina	γ-Al_2O_3	7.0	6.7	–
Birnessite	$(Na,Ca)Mn_7O_{14} \cdot 2.8H_2O$	2.2	1.9	1.7
Boehmite	γ-AlOOH	–	–	8.6
Calcite	$CaCO_3$	–	–	8.5
Corundum	α-Al_2O_3	–	–	9.1
Dolomite	$CaMg(CO_3)_2$	–	–	8.0
Forsterite	Mg_2SiO_4	–	–	4.5
Gibbsite	γ-$Al(OH)_3$	8.5	8.5	–
Goethite	α-FeOOH	7.3	7.5	–
Hematite	α-Fe_2O_3	8.2	8.3	8.2
Illite	$K_{1.75}Si_7Al(Al_3Fe^{(III)}_{0.25}Mg_{0.75})O_{20}(OH)_4$	2.7	–	–
Imogolite	$SiAl_2O_3(OH)_4$	6.8	8.4	11.5
Kaolinite	$Si_4Al_4O_{10}(OH)_8$	5.0	3.6	4.7
Magnetite	Fe_3O_4	6.8	–	–
Quartz	SiO_2	–	2.5	2.0
Rutile	TiO_2	6.0	–	5.9
Silica	SiO_2	2.0	2.4	2.4
Smectite	$Na_2Si_{7.5}Al_{0.5}(Al_{3.5}Mg_{0.5})O_{20}(OH)_4$	8.5	–	$<$2.3

PZNPC, point of zero net proton charge; PZNC, point of zero net charge; PZC, point of zero charge.

Sources of data: Albite: Mukhopadhyay B and Walther JV (2001) Acid–base chemistry of albite surfaces in aqueous solutions at standard temperature and pressure. *Chemical Geology* 174: 415–443. Allophane: Su C, Harsh JB, and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes. *Clays and Clay Minerals* 40: 280–286; Su C and Harsh JB (1993) The electrophoretic mobility of imogolite and allophane in the presence of inorganic anions and citrate. *Clays and Clay Minerals* 41: 461–471. Alumina: Goyne KW, Zimmerman A, Newalkar BL, Komarneni S, Brantley SL, and Chorover J (2002) Surface charge of variable porosity Al_2O_3 (s) and SiO_2 (s) adsorbents. *Journal of Porous Materials* 9: 243–256. Birnessite: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Boehmite: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348. Calcite: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Corundum: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Dolomite: Pokrovski OS, Schott J, and Thomas F (1999) Dolomite surface speciation and reactivity in aquatic systems. *Geochimica Cosmochimica Acta* 63: 3133–3143. Forsterite: Pokrovsky OS and Schott J (2000) Forsterite surface composition in aqueous solutions: a combined potentiometric, electrokinetic, and spectroscopic approach. *Geochimica Cosmochimica Acta* 64: 3299–3312. Gibbsite: Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press. Goethite: Sigg L and Stumm W (1981) The interaction of anions and weak acids with the hydrous goethite (α-FeOOH) surface. *Colloids and Surfaces* 2: 101–117; Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press. Hematite: Chibowski S and Janusz W (2002) Specific adsorption of Zn(II) and Cd(II) ions at the α-Fe_2O_3/electrolyte interface structure of the electrical double layer. *Applied Surface Science* 7853: 1–13. Illite: Sinisyn VA, Aja SU, Kulik DA, and Wood SA (2000) Acid–base surface chemistry and sorption of some lanthanides on K^+-saturated Marblehead illite: 1. Results of an experimental investigation. *Geochimica Cosmochimica Acta* 64: 185–194. Imogolite: Su C, Harsh JB, and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes. *Clays and Clay Minerals* 40: 280–286; Su C and Harsh JB (1993) The electrophoretic mobility of imogolite and allophane in the presence of inorganic anions and citrate. *Clays and Clay Minerals* 41: 461–471. Kaolinite (KGa-1): Schroth BK and Sposito G (1997) Surface charge properties of kaolinite. *Clays and Clay Minerals* 45: 85–91; Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Magnetite: Regazzoni AE, Blesa MA, and Maroto AJG (1983) Interfacial properties of zirconium dioxide and magnetite in water. *Journal of Colloid and Interface Science* 91: 560–570. Quartz: Eggleston CM and Jordan G (1998) A new approach to pH of point of zero charge measurement: Crystal-face specificity by scanning force microscopy (SFM). *Geochimica Cosmochimica Acta* 62: 1919–1923. Rutile: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348. Silica: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348; Goyne KW, Zimmerman A, Newalkar BL, Komarneni S, Brantley SL, and Chorover J (2002) Surface charge of variable porosity Al_2O_3 (s) and SiO_2 (s) adsorbents. *Journal of Porous Materials* 9: 243–256. Smectite: Avena MJ and de Pauli CP (1998) Proton adsorption and electrokinetics of an Argentinian montmorillonite. *Journal of Colloid and Interface Science* 202: 195–204.

methods, involving the use of scanning force microscopy, appear promising in regard to elucidating patchwise heterogeneity of surface charge and providing direct force measurements of the PZC (e.g., as used to obtain the value of PZC for quartz in Table 1).

Point of Zero Net Proton Charge

The pH value at which the proton surface-charge density, σ_H, is zero defines the PZNPC. Proton surface-charge density is given by:

$$\sigma_H = q_H - q_{OH} \quad [6]$$

where q_H and q_{OH} are the surface excess (e.g., in moles of charge per kilogram) of H^+ and OH^-, respectively. Therefore, at the PZNPC the moles of surface-adsorbed H^+ and OH^- are balanced. Proton and hydroxide adsorption reactions can be represented by the mass action expressions:

$$SOH(s) + H^+(aq) = SOH_2^+(s) \quad [7a]$$

$$SOH(s) + OH^-(aq) = SO^-(s) + H_2O(l) \quad [7b]$$

where SOH is a neutral surface hydroxyl group on a soil particle, and *s*, *aq*, and *l* represent the solid, aqueous, and liquid phases, respectively. If proton adsorption–desorption is assumed to occur dominantly at hydroxylated sites (e.g., as in the case of pure metal oxides), then the PZNPC defines the condition where the density or concentration of protonated sites is equal to that of proton-dissociated sites [i.e., $(SOH_2^+) = (SO^-)$]. Thus, both protonated and proton-dissociated sites can coexist at the PZNPC.

In 'indifferent' electrolyte solutions (i.e., those comprising monovalent cations and anions that do not form stable complexes at particle surfaces), proton adsorption behavior, as depicted in Figure 2, is a function of the Brönsted acidity of surface functional groups. Soil particles exhibit a wide range in surface acidity, as indicated by the PZNPC values shown in Table 1. Low PZNPC values are typical of more strongly acidic soil constituents such as silica, quartz, Mn oxides, and natural organic matter, whereas higher values are observed for (hydr)oxides of Fe and Al. Table 1 shows PZNPC values of soil minerals suspended in indifferent electrolyte solutions. However, the presence of cations (anions) that form inner- or outer-sphere surface complexes (Figure 3a) will tend to decrease (increase) the PZNPC because of adsorptive competition with H^+ (OH^-).

Proton- and hydroxide-promoted dissolution of soil minerals is found to increase at pH values both below and above the PZNPC. Correlations are observed between dissolution rate constants and the absolute magnitude of σ_H, suggesting that metal–oxygen bonds at the particle surface are weakened as reactions in eqns [7a] and [7b] proceed to the right.

Point of Zero Net Charge

The first two terms in eqn [5] are often combined to provide a measure of *intrinsic surface-charge density*, σ_{in}, which derives from the crystal and surface structure of the adsorbent:

$$\sigma_{in} = \sigma_0 + \sigma_H \quad [8]$$

The PZNC is the pH value at which σ_{in} is equal to zero. The intrinsic surface charge is balanced by the adsorption of ions. Since σ_{in} results from both proton adsorption–desorption reactions (pH-dependent) and isomorphic substitutions, it comprises conditional and permanent components (Figure 4). Since $\sigma_{in} = -\Delta q$, the PZNC may be determined: (1) by measuring the pH where surface excess values of cation and anion charge are equal (i.e., $\Delta q = 0$), or (2) by measuring the pH where $\sigma_0 = -\sigma_H$.

Charge balance dictates that, as σ_H decreases with increasing pH (affected by mass action as indicated in eqn [7]), ion adsorption–desorption reactions must balance the change in intrinsic charge density. The dependence of σ_H on pH is shown schematically in Figure 2, and corresponding effects on Δq are illustrated in Figure 4. For a soil adsorbent that is devoid of structural charge (i.e., $\sigma_0 \approx 0$), such as a highly weathered Oxisol that contains a negligible quantity of 2:1 layer-type silicates, $\Delta q = -\sigma_H$, and the PZNPC equals the PZNC (Figure 5, top). This is also the case for pure oxides and hydroxides that are free of structural defects. The presence of negative structural charge ($\sigma_0 < 0$), as occurs because of isomorphic substitutions in many temperate-zone soils, results in PZNC being less than PZNPC, and the value of Δq (or σ_H) at the PZNPC (PZNC) is equal to $-\sigma_0$ (Figure 5, middle). Whereas positive structural charge ($\sigma_0 > 0$) is rare in soils, it is detectable from PZNPC being less than PZNC, and its magnitude is given by the value of Δq (or σ_H) at the PZNPC (PZNC) (Figure 5, bottom).

In addition to the effects of σ_0, the nature of variable-charge behavior, such as that depicted in Figure 5, is highly dependent on the Brönsted acidity of the soil particles. For example, the magnitude and pH range of greatest local slope (i.e., $\delta\sigma_H/\delta$pH or $\delta\Delta q/\delta$pH) is dependent on the identity and quantity of surface sites undergoing proton adsorption–desorption reactions. Whereas oxides and hydroxides of Fe and Al are weakly acidic, with PZNPC values near neutrality, more strongly acidic groups such as those residing on natural organic matter functional

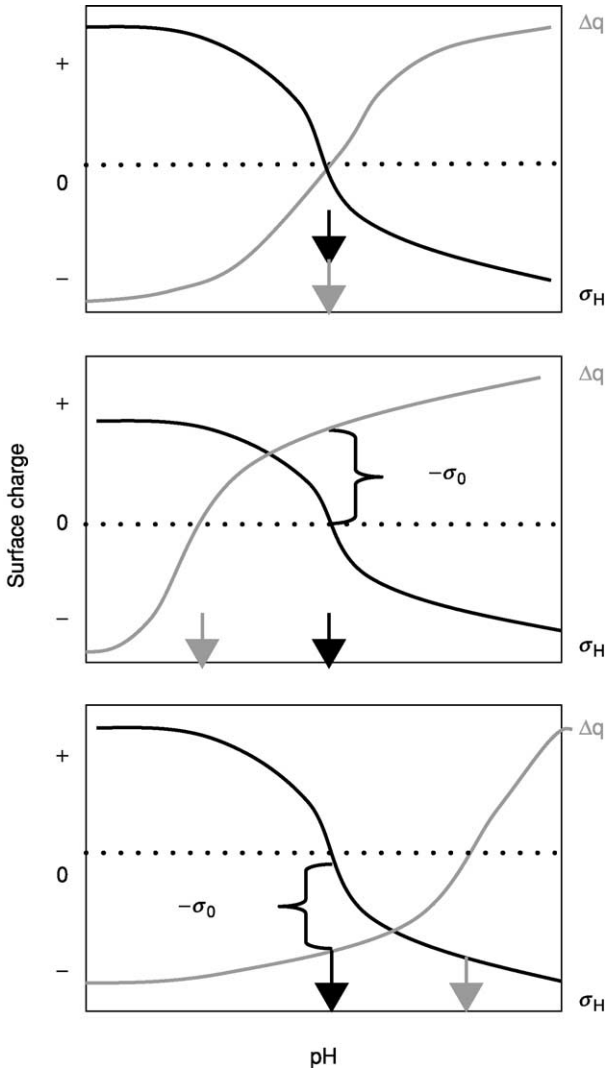


Figure 5 The inverse relation between σ_H and Δq is depicted schematically for (a) $\sigma_0 = 0$ (PZNC = PZNPC), (b) $\sigma_0 < 0$ (PZNC < PZNPC), and (c) $\sigma_0 > 0$ (PZNC > PZNPC). The PZNC is indicated by gray arrows and the PZNPC is indicated by black arrows. PZNC, point of zero net charge; PZNPC, point of zero net proton charge.

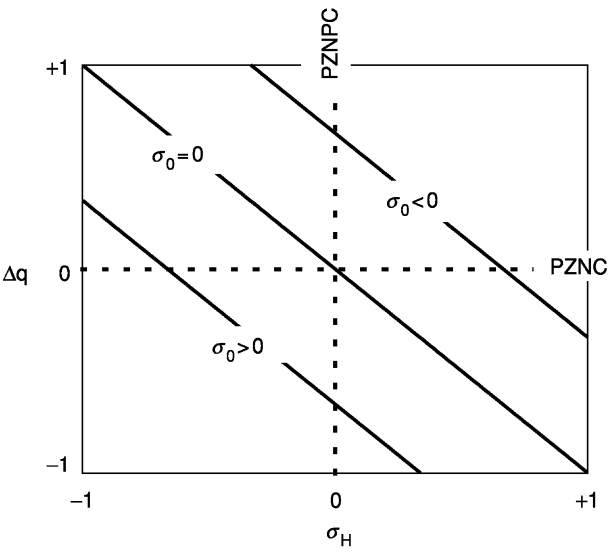


Figure 6 Surface-charge balance dictates that a plot of Δq versus σ_H must be linear, with a slope equal to -1 and x and y intercepts equal to $-\sigma_0$. PZNPC, point of zero net proton charge; PZNC, point of zero net charge.

groups (e.g., carboxylic acids), silica or quartz surfaces, and Mn oxide surfaces all tend to decrease PZNPC values for whole soils.

The surface-charge balance of eqn [5] is summarized in Figure 6; a plot of Δq versus σ_H should be linear with a slope equal to -1, and x and y intercepts equal to $-\sigma_0$. The x and y scales in Figure 6 are arbitrary and actual values depend on surface chemistry of the adsorbent and the units selected for plotting surface-charge data (e.g., moles of charge per kilogram or coulombs per square meter). Whereas curves such as those depicted in Figure 5 (with pH as independent variable) are strongly affected by ionic strength and electrolyte composition for a given adsorbent, the relation depicted in Figure 6 must be reproducible for a given soil, irrespective of changes in solution chemistry.

List of Technical Nomenclature

Adsorbate	Matter accumulating at the interface between adsorbent and the aqueous phase
Adsorbed ion charge density (Δq)	The net charge density created by adsorption of ions, other than H^+ and OH^-, into inner-sphere complexes, outer-sphere complexes, and the diffuse ion swarm
Adsorbent	The solid surface on which matter accumulates
Adsorption	The net accumulation of matter at the adsorbent surface
Diffuse layer surface-charge density (σ_d)	The net surface charge density created by adsorbed ions in the diffuse layer
Inner-sphere complex surface-charge density (σ_{IS})	The net charge density created by ions adsorbed via direct coordination to surface functional groups (i.e., with no water molecules interposed between adsorbate and surface)
Intrinsic surface-charge density (σ_{in})	The sum of structural and proton surface charge densities

Outer-sphere complex surface-charge density (σ_{OS})	The net charge density created by adsorption of hydrated ions that form stable complexes with adsorbent surface functional groups
Point of zero charge	The pH value at which particle surface charge is equal to zero
Point of zero net charge	The pH value at which adsorbed ion charge density is equal to zero
Point of zero net proton charge	The pH value at which proton surface charge density is equal to zero
Proton surface-charge density (σ_H)	Surface-charge density created by the formation of adsorption complexes involving proton or hydroxide ions as adsorbate; the difference between surface excess of protons and hydroxide ions ($\sigma_H = q_H - q_{OH}$)
Structural surface-charge density (σ_0)	Surface-charge density created by isomorphic substitutions among ions of differing valence in the crystal structure of an adsorbent

See also: **Cation Exchange**; **Dissolution Processes, Kinetics**; **pH**; **Sorption:** Metals

Further Reading

Everett DH (1972) Definitions, terminology and symbols in colloid and surface chemistry. *Pure and Applied Chemistry. Chimie pure et appliquée* 31: 578–638.

Sposito G (1998) On points of zero charge. *Environmental Science and Technology* 32: 2815–2819.

Stumm W and Wollast R (1990) Coordination chemistry of weathering: kinetics of the surface controlled dissolution of oxide minerals. *Reviews of Geophysics* 28: 53–69.

ZONE TILLAGE

J L Hatfield and A T Jeffries, National Soil Tilth Laboratory, Ames, IA, USA

Tillage is the act of disturbing the soil through some type of mechanical means. It has been practiced since the development of crude instruments such as a pointed stick to create a hole into which a seed was dropped. The largest advance in the progress of tillage implements has been the invention of the moldboard plow, which inverts the soil and allows rapid cultivation. Since that invention, there have been continual advances in the development of tillage equipment for the purposes of: creating a seed bed that is favorable for rapid germination; removing weeds; destroying crusts that impede the emergence of seedlings, or prevent water or gases from moving into or out of the soil; or removing compacted layers from the upper soil profile. There is a large array of tillage tools available to producers that reshape the soil into a more compliant medium for producing plants.

Tillage can be considered as a time-and-space operation within a given field. The time component is introduced because different tillage implements are used throughout a production sequence to achieve different goals in manipulating the soil profile. The space component is introduced because different tillage implements affect the soil in a variety of ways across a field. An example of this is that the preplant tillage operation often disturbs the entire soil surface, while cultivation only disturbs the area between the rows of plants. An entire production sequence can be considered as a series of tillage operations that changes the soil surface in different ways.

Zone tillage is a unique subset of tillage operations in which only a portion of the soil surface is disturbed. In a very broad sense, zone tillage can be considered to represent cropping in which strips of the field are planted to different crops and managed differently or a strip is left in a fallow area to conserve precipitation for the subsequent crop. In the Great Plains of the USA and the Prairie provinces of Canada, the practice of strip farming with cultivated and fallow strips could be considered as a large-scale form of zone tillage. In this case, the zone is created by the temporal distribution of tillage operations in which strips of the field are cultivated one year and the next year the fallow strip is cultivated.

Zone tillage is more often referred to as tilling only a narrow zone of the seedbed and is practiced on row crops more than on small-grain or forage crops. Zone tillage can take on many different forms, uses an array of equipment, and is often the primary tillage operation following crop harvest. Zone-tillage units provide the following attributes in soil disturbance: (1) arrangement of shanks or rippers is placed in combination with coulters and disks to till only a select portion of the soil in a regular pattern; (2) crop

residue is removed or incorporated in the tilled area, leaving the soil surface with little residue in this zone; and (3) fertilizers may or may not be incorporated into the tilled area. These operations occur either to till over the row of the previous crop or between the rows. The depth of tilling depends upon the desired effect either to remove a compaction layer beneath the soil surface or to place fertilizer into the soil. In either case, the disturbed area serves as the planting zone for the subsequent crop.

Zone tillage is shown in Figure 1, in which the disturbed area is relatively small compared with the undisturbed area and the volume of the tilled area is distributed more vertically than horizontally. Zone tillage is often referred to as vertical rather than horizontal tillage, since the goal is to till more deeply into the soil profile than to till extensively over the soil surface. There are advantages to this type of system: the primary reasons are protection of the soil surface from erosion due to water or wind coupled with placement of fertilizer into an area in which the plant roots can easily extend.

In zone-tilled systems, crops tend to proliferate their roots into the disturbed area first and then through the side walls of the tilled area. One can image that the root system in Figure 1 is confined to the tilled area, and early in the growing season this area supplies the nutrients and water required for proper plant growth. Root systems in zone-tilled systems have more of their root mass concentrated in the tilled area. They also increase their rooting depth earlier in the growing season, which leads to greater support for the plant and less lodging or breakage in high-wind conditions. Altering the root system also has an effect on the water and nutrient uptake patterns. Water and nutrients are removed from deeper in the soil profile, because more roots are concentrated there, as depicted in Figure 2. Nutrient placement in zone tillage becomes a critical management factor in crop production.

Changes in soil with zone tillage are isolated to the tilled area. Bulk density of the soil is reduced in this area because of the tillage practice and the incorporation of crop residue. In comparison, the area that is not tilled may have an increased bulk density and a surface crust that forms if heavy wheel-traffic is used when the soil is wet. The increased bulk density between the crop rows may reduce both water infiltration and exchange of gases, with a shift to greater exchange rates in the tilled area. In contrast, if zone tillage is practiced in fields that have a slope and the tilled areas are positioned parallel with the slope, then the tilled zone can act as a conduit for water, and erosion rates can be increased in this zone because of the loose nature of the soil surface.

The effects of zone tillage on soil properties have not been extensively documented. Observations suggest that the positive benefits of zone tillage on crops are the increased rooting depth and consequently more-rapid exploration of the soil volume by the plant roots. Increased resistance to lodging and plant breakage occurs because of the greater rooting depth and concentration of root mass below the stem. Higher water and nutrient uptake from the lower portion of the soil profile leads to better crop performance in periods of drought. The soil changes

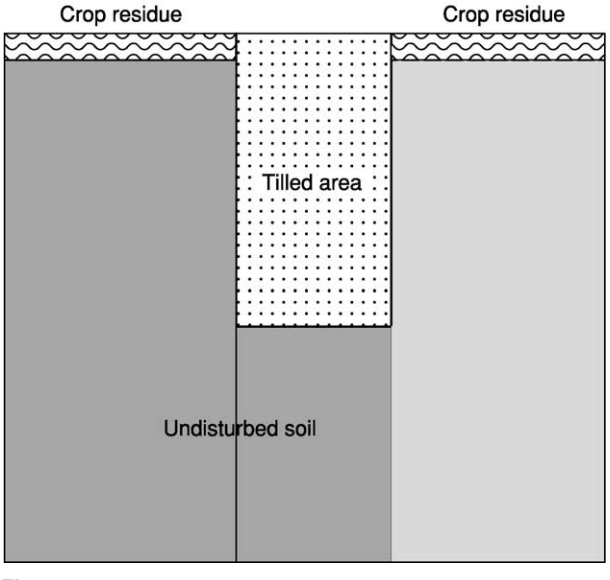


Figure 1 Zone-tilled soil in the central Corn Belt region of the USA.

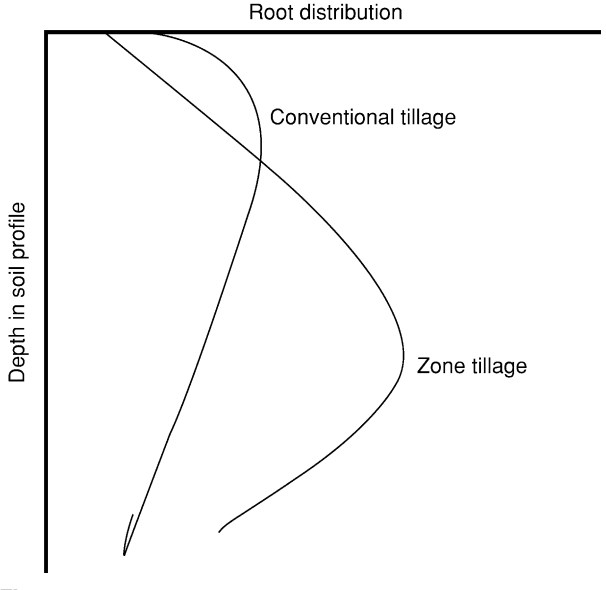


Figure 2 Root distribution patterns as a result of broad conventional tillage compared with zone tillage.

that occur in zone tillage are a decrease in the bulk density in the upper portion of the soil profile and removal of compaction layers from the soil, because the tillage operations are deeper than typical tillage systems. The zone-tillage system creates a more favorable seed zone for planting, and, e.g., in the US Midwest, plants emerge more quickly in zone-tillage systems because of better seed–soil contact, less crusting, and slightly warmer temperatures in the seed zone. Crop emergence is more uniform and early-season plant vigor appears to be enhanced through a combination of soil microclimate and seed-zone conditions.

Zone tillage is a practice that is beginning to be adopted in areas that need protection from soil erosion. Also it is being used where reduced tillage results in maintenance of crop residue on the soil surface. The types of implements include those employed exclusively after harvest in the autumn to incorporate nutrients (P and K) in the seed zone. These implements prepare an area into which the next crop is planted directly the following spring. There are single-operation planters that give a shallow form of zone tillage and only disturb a small portion of the soil surface into which the seed is placed, and nutrients are banded alongside the seed furrow. Emergence of tillage systems that are directed toward soil management and creation of a seed zone that increases crop production efficiency and reduces environmental problems, e.g, erosion or nutrient loss, will continue.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Conservation Tillage**; **Crusts:** Structural; **Erosion:** Water-Induced; Wind-Induced; **Evapotranspiration**; **Mulches**

Further Reading

Carter MR (ed.) (1994) *Conservation Tillage in Temperate Ecosystems*. Boca Raton, FL: Lewis Publishers/CRC Press.

Hatfield JL and Stewart BA (1994) *Crops Residue Management*. Boca Raton, FL: Lewis.

Hatfield JL, Buhler DD, and Stewart BA (1998) *Integrated Weed and Soil Management*. Boca Raton, FL: Lewis.

Michalson E, Papendick RI, and Carlson J (1998) *Conservation Farming in the United States: Methods and Accomplishment of the STEEP Program*. Boca Raton, FL: CRC Press.

Midwest Plan Service (1992) *Conservation Tillage Systems and Management*. MWPS-45. Ames, IA: Iowa State University Press.

Sprague MA and Triplett GB (1986) *No-Tillage and Surface-Tillage Agriculture*. New York: Wiley Interscience.

Unger PW (1994) *Managing Agricultural Residues*. Boca Raton, FL: CRC Press.