THE ENVIRONMENT OF THE EARTH'S SURFACE

John Southard

Massachusetts Institute of Technology

LibreTexts™

The Environment of the Earth's Surface
(Southard)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the hundreds of other texts available within this powerful platform, it is freely available for reading, printing and "consuming." Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects.

Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts’ web based origins allow powerful integration of advanced features and new technologies to support learning.

The LibreTexts mission is to unite students, faculty and scholars in a cooperative effort to develop an easy-to-use online platform for the construction, customization, and dissemination of OER content to reduce the burdens of unreasonable textbook costs to our students and society. The LibreTexts project is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning by developing an Open Access Resource environment. The project currently consists of 14 independently operating and interconnected libraries that are constantly being optimized by students, faculty, and outside experts to supplant conventional paper-based books. These free textbook alternatives are organized within a central environment that is both vertically (from advance to basic level) and horizontally (across different fields) integrated.

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org. More information on our activities can be found via Facebook (https://facebook.com/Libretexts), Twitter (https://twitter.com/libretexts), or our blog (http://Blog.Libretexts.org).

This text was compiled on 11/19/2023

TABLE OF CONTENTS

11: Landscapes

12: The Hydrologic Cycle, the Sediment Cycle, and the Carbon Cycle

Index

Detailed Licensing

Introduction

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing****.*

CHAPTER OVERVIEW

1: Introduction

1.1: Some Basic Physics Relevant to The Earth's Surface

BACKGROUND: SOME BASIC (NON-MATHEMATICAL) PHYSICS RELEVANT TO THE EARTH'S SURFACE

1. Energy

First you need to feel comfortable about the nature of energy. We have to be concerned with three forms of energy: energy of motion (called ***kinetic energy***), energy of position (called ***potential energy***), and energy of thermal motions of atoms of molecules of a material (called ***thermal energy*** or ***heat energy***). Kinetic energy and potential energy together are often referred to as ***mechanical energy***.

If you are not already familiar with these concepts of energy, here's a little demonstration that should serve the purpose of giving you the basic ideas. You will need a lump of modeling clay or putty. Throw the lump straight up in the air. When it leaves your hand, it has a certain speed and therefore a certain kinetic energy. It's slowed by the downward force of gravity as it rises, so it loses kinetic energy, but that lost kinetic energy is stored as energy of position, by virtue of being located higher up in the gravity field of the Earth. When it reaches the top of its trajectory, all of its kinetic energy has been stored as potential energy. Then on the way down, the stored potential energy is converted back into kinetic energy. If you were doing this in a vacuum in some enormous evacuated bell jar or on the surface of the Moon, the putty would have the same speed when it lands back in your hand as when you threw it upward. The mechanical energy of the lump, kinetic plus potential, is the same all along the trajectory; the mechanical energy is said to be *conserved*.

On the real surface of the Earth, of course, the lump is bathed in air, and its motion is retarded by friction both on the way up and on the way down. That friction warms the lump (and the air it passes through), slightly, converting some of the mechanical energy into thermal energy.

Another thing you can do with the lump is drop it on the floor from some high place. It has some kinetic energy as it lands on the floor, but then it comes to a stop upon impact, without even rebounding. What happened to all of its kinetic energy? It was converted entirely to thermal energy, via the friction involved in the deformation of the lump as it made impact. If you had the right equipment you could actually verify for yourself that the lump was warmer after it landed than before.

2. Electromagnetic Radiation

All bodies (pieces or regions of matter, in the physicist's parlance) at temperatures above absolute zero radiate energy in the form of ***electromagnetic waves***. The subject of electromagnetism goes to the heart of physics, and it's much too complicated for a full treatment here, but all of you are familiar with the existence and effect of electromagnetic waves, in various ways: the radiant heat you receive from your fireplace or electric heater; the light you perceive with your eyes; radio, television, and radar waves; x rays.

Electromagnetic waves vary over an enormous range of wavelengths, from the very short wavelengths of cosmic rays and x rays, through the intermediate range of ultraviolet and visible light, down through the infrared to the long wavelengths of radio waves and the like (Figure 1-1). The shortest wavelength visible to the eye, at the edge of the ultraviolet, is about 0.36 μm; the longest wavelength, at the edge of the infrared, is about 0.76 μm.

Of the various physical laws governing the emission and absorption of electromagnetic radiation, the one that's most relevant for us here is ***Wien's law***: *the product of the absolute temperature and the wavelength of the most intense radiation is a constant*. That means that a body with a relatively high surface temperature, like the sun, radiates energy at relatively short wavelengths, and a body with a relatively low surface temperature, like the earth, radiates energy at relatively long wavelengths. The maximum intensity of the Sun's radiation is at about 0.5 μm, whereas the maximum intensity of the earth's long-wave radiation is at about 10 μm, well into the infrared range.

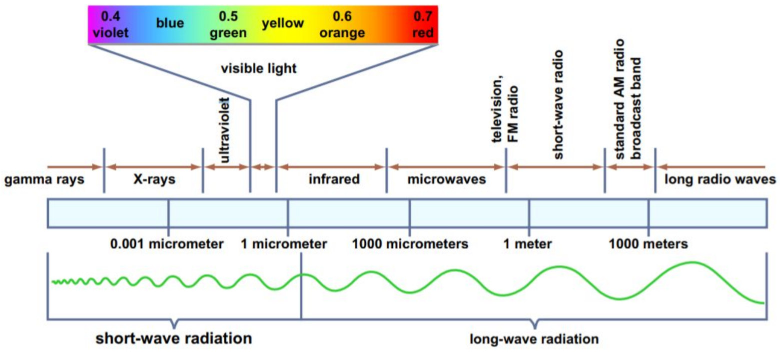


Figure by MIT OCW.

Figure 1-1. The electromagnetic spectrum. (From Lutgens and Tarbuck, 2004.)

But bodies like the Sun and the Earth don't radiate all of their energy at a single wavelength: they radiate *over a wide range of wavelengths*. Figure 1-2 is a graph of the intensity of solar radiation as a function of wavelength. We see from Figure 1-2 that *the sun radiates most of its energy in the visible and ultraviolet part of the spectrum, but a substantial part also in the infrared part of the spectrum*. A similar graph for the Earth's radiation back to space, also shown in Figure 1-2, would show a similar shape but shifted far toward the long wavelengths.

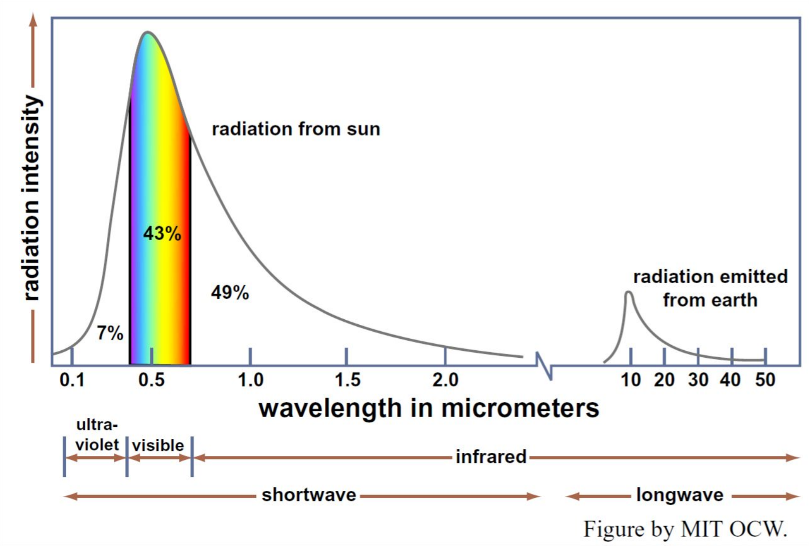


Figure 1-2. Intensity of short-wave solar radiation and long-wave terrestrial radiation as a function of wavelength. (From Lutgens and Tarbuck, 2004.)

The next time you are out in your yard or garden, or in the park or the countryside, I want you to think about long-wave radiation: you can't see it (although you *can* feel it), but it's operating everywhere, all the time, and it's as important as sunlight in terms of its effect upon local climate (although obviously not in terms of plant growth, or suntanning, or skin cancer).

Temperature and Heat

All of the atoms or molecules of matter undergo inherent motions. In gases, these motions consist in long, straight trajectories of the atoms or molecules, called ***free paths***. If the gas is enclosed in a container, the particles collide with the walls of the container. Macroscopically we sense the aggregate of such collisions as the pressure that the gas exerts on the walls of the container. Now and then, the particles also collide with one another as they fly about. The average distance traveled by a particle between collisions with other particles is called the ***mean free path***. The mean free path of air molecules at atmospheric pressure and a temperature of 0°C is about 2 x 10-5 cm (a very small distance, equal to about one thousand molecule diameters). The collision frequency is about 5 x 109 per second, which is an extremely short time. Each molecule makes about five billion collisions per second!

In solids, the atoms or molecules are fixed in place, either in the regular array of a crystal structure or in the irregular arrangement of an amorphous (noncrystalline) solid. You will learn more about that in Chapter 2. But the particles do not stand still: they vibrate in place, with a certain amplitude and frequency. In liquids, the atoms or molecules are in close association with one another, and they exert strong forces on one another, but they too undergo the same small-scale random motions that are characteristic of gases and solids.

What we are dealing with here are called thermal motions, or thermal vibrations. These motions are an inherent feature of all matter. The moving or vibrating particles have some energy of motion, called kinetic energy. The hotter the matter (that is, the higher the temperature of the matter), the more vigorous or energetic are the thermal motions. In fact, what physicists call the ***temperature*** of the matter is just *a measure of the average kinetic energy of all of the constituent atoms or molecules*. As the temperature of a body of matter is lowered, a point is reached at which the thermal energy of the atoms and molecules is at an absolute minimum. That temperature is called ***absolute zero***, about -273°C.

All of you know about the *Fahrenheit* scale of temperature, which we here in the United States use in our everyday lives. You probably also know that it has been supplanted in most other parts of the world, and by scientists everywhere, by the *Celsius* scale of temperature (earlier called the *centigrade* scale), with the freezing point of water at 0°C and the boiling point of water (at sea-level pressure) at 100°C. What may be less familiar to you is the *Kelvin* scale of temperature, in which the degrees are the same as in the Celsius scale but zero is fixed at absolute zero. In the Kelvin scale, the degrees are called kelvins (not "degrees Kelvin").

It’s important for you to understand the distinction between temperature and heat: they are related, but they are not the same! ***Heat*** is *the total internal energy of all of the constituent atoms or molecules in a given body of matter*. (By internal energy here, I mean all of the kinetic energy of the atoms and molecules, which is discussed in the preceding paragraphs, together with the energy associated with the forces of attraction among the individual atoms and molecules.) Other things being equal, the bigger a body of matter, the more heat it contains. That’s in contrast to the ***temperature*** of the body, which, as you learned above, is *a measure of the thermal energy of the individual atoms or molecules*.

In contrast to heat, the temperature of the body of matter is independent of how big the body is. Here is an everyday example of this: a small pot of water at nearly 100°C on the stovetop is much hotter (that is, has a much higher temperature) than the water in your almost-full bathtub at 40°C, but the total heat content of the water in the bathtub is much larger than the total heat content of the water in the pot.

Various units are used to describe heat. The two that are in common useinphysicsarethe*calorie*,whichisdefinedas*theamountofheatthat’sneeded to raise the temperature of one gram of water by one degree Celsius*. Incidentally, there is a lot of potential confusion about calories. What I defined above is sometimes called the *gram calorie*, abbreviated as “cal”. There is also the *kilogram calorie* (the amount of heat it takes to raise one kilogram of water by 1°C), abbreviated as “Cal”. There are one thousand gram calories (“small calories”) in one kilogram calorie (“large calorie”). The calorie that’s used to measure the energy content of foods—that is, the dieter’s calorie—is the same as the kilogram calorie.

Heat is also measured in *joules*, the unit of energy in the mks (meter, kilogram, second) system of units in physics. The joule is named in honor of James Prescott Joule, a British physicist (1818–1889), who was the first to demonstrate the equivalence of heat and mechanical energy by means of a classic experiment in which he increased the heat content of water in a large vat by adding mechanical energy to the water by stirring it with a paddle. One gram calorie is equal to about 4.18 joules. Just to confuse you further, in everyday life it’s common to measure heat in terms of *British thermal units* (Btu). One Btu is *the amount of heat needed to raise one pound of water one degree Fahrenheit* (technically, from 63° to 64°). One Btu is about equal to 250 gram calories.

Finally, the concept of heat capacity is an important for anyone who thinks about the Earth’s surface. The ***specific heat capacity*** of some given material is *the amount of heat that has to be added to a unit mass of that matter to raise its temperature by one unit*. This could be expressed, for example, as joules per kilogram or as calories per gram. The importance for the Earth’s surface is that the rocks, soil, and waters of the surface all have their own specific heat capacities, so it takes certain amounts of incoming solar energy to raise their temperature, and certain amounts of outgoing terrestrial long-wave radiation to decrease their temperature. Water has a much greater (I’m tempted to say spectacularly greater) specific heat capacity than the multifarious solid materials of the surface. That causes regions that are dominated by the oceans to have much more equable climate, other things being equal, than regions far removed from the oceans.

How Heat Is Transported

So far we’ve dealt only with radiation. Heat can also be transported by *conduction*, *convection*, and *advection*.

Conduction of heat through a solid is familiar to all: the handle of the pan on the stove heats up by conduction from the hot body of the pan. The same holds true for liquids and gases, although in that case conduction is often overwhelmed by the more efficient processes of convection and advection, by which the fluid, carrying its heat, is moved bodily from place to place.

The principle behind heat conduction is simple. The heat of a body is a manifestation of the thermal agitation of the constituent atoms and molecules. The hotter the material, the more energetic are the thermal motions of the atoms and molecules. When a difference in temperature from one place to another is imposed upon the body, the differing “intensity of jiggling” of the atoms and molecules from place to place is evened out, thus equalizing the temperature. We say that heat has been transported from the region of higher temperature to the region of lower temperature.

Convection is the term used for *the circulation of a fluid driven by differences in density brought about by differences in temperature*. The familiar example is a room warmed by a radiator on a cold winter day (Figure 1-3): the heating of the air by the radiator decreases its density, causing it to rise, and as it is cooled by the cold walls of the room, it sinks and takes a return path to the radiator, again to be heated. Convection cells of this kind are extremely important in atmospheric processes, on both small and large scales.

The horizontal motion of fluid from one region to another, driven by horizontal differences in atmospheric pressure, is called ***advection***. One can sense the importance of advection on a gloomy winter day when a warm wind blows from the south, melting much of the snow cover. Of course, what a local observer senses as advection might just be one very small part of some gigantic convection cell.

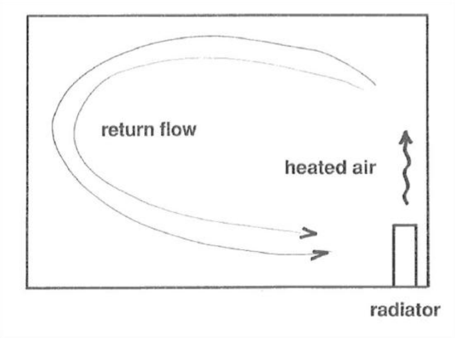


Figure 1-3. A convection cell.

1.1: Some Basic Physics Relevant to The Earth's Surface is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.2: Heading and Cooling of the Earth's Surface

HEATING AND COOLING OF THE EARTH’S SURFACE

1.2.1 Introduction

There is a branch of climatology, called ***microclimatology***, that deals with *the climate at a particular locality on the Earth*. One important part of microclimatology has to do with what could be described as “the climate near the ground”. Does that not seem like a suitable topic for a course on the environment of the Earth’s surface?

Most of the standard weather observations made at weather stations are taken a standard height above the ground that’s a little greater than the height of the average person. At most times, *the conditions right at the ground surface are rather different*, in terms of wind speed, temperature, and humidity. The climate near the ground, and its effect on the uppermost layer of the soil, is of obvious importance not just for agriculture but also for civil and environmental engineering, architecture, planning, and even medicine.

1.2.2 The Seasons

I often find myself wondering how much the average person understands about why there are seasons, and why sun height and day length change over the course of a year. Because I don’t know much about that (my basis for judgment is only anecdotal), I risk being either too elementary or too advanced for you in this brief section on the seasons.

First, some facts about the geometry of the Earth–Sun system. The e\Earth revolves around the Sun, once in a full year (that’s how a year is defined!), in a plane passing through the earth’s orbit, called the ***plane of the ecliptic***. The earth’s orbit is nearly, but not quite, a circle; it’s actually an *ellipse* with a slight eccentricity. To give you an idea of the degree of non-circularity of the orbit, the maximum Earth–Sun distance is about 152 million kilometers and the minimum distance is about 147 million kilometers.

The only easy way to detect by direct observation that the earth is revolving around the sun is astronomical: *view the stars at the same time every night throughout the year, and see how their position changes*. They make one complete circuit through the heavens in the course of a year.

You know, of course, that the earth rotates about its axis once a day, and that’s what causes the difference between night and day. (!) If the Earth’s axis of rotation were exactly perpendicular to the plane of the ecliptic, the length of the day and the height of the sun above the horizon at a given time of day (local noon, say) would not change through the year, and things would not be nearly as exciting as they are in reality. (By “day” here I mean the time during the 24-hour day that the sun is above the horizon. The word “day” is rather ambiguous in the English language; think of this aspect of “day” as “daylight day”)

The earth’s axis of rotation is in fact inclined to the plane of the ecliptic by about 23° 27', usually rounded off to 23-1/2°. This has far-reaching consequences for conditions at the Earth’s surface. As you can see from Figure 1- 4, there are two times during the year, called the ***equinoxes*** (one in spring, on 20 or 21 March,and one in autumn,on 22 or 23 September),when *the earth’s axis of rotation is exactly perpendicular to the line between the earth and the sun*. At those times of the year, and *only* at those times, the daylight day is exactly twelve hours long at every point on the earth.

The earth’s north pole and south pole represent special cases, in the sense that at the time of the equinoxes the sun is seen to lie right on the horizon and make a complete circle around the sky every 24 hours!

At times midway between the equinoxes, called the ***solstices***, *the earth’saxisofrotationliespreciselyintheplanethatpassesthroughboththesun and the earth and is perpendicular to the plane of the ecliptic* (Figure 1-4). At those times, one of the Earth’s poles points closest to the sun and the other pole points farthest from the sun. At the time of the summer solstice in the Northern Hemisphere, on 21 or 22 June, the sun is highest in the sky at a given time of day, as at noon, and the daylight day is longest. Conversely, at the time of the winter solstice in the northern hemisphere, on 21 or 22 December, the sun is lowest in the sky at a given time of day, and the daylight day is depressingly short, even nonexistent.

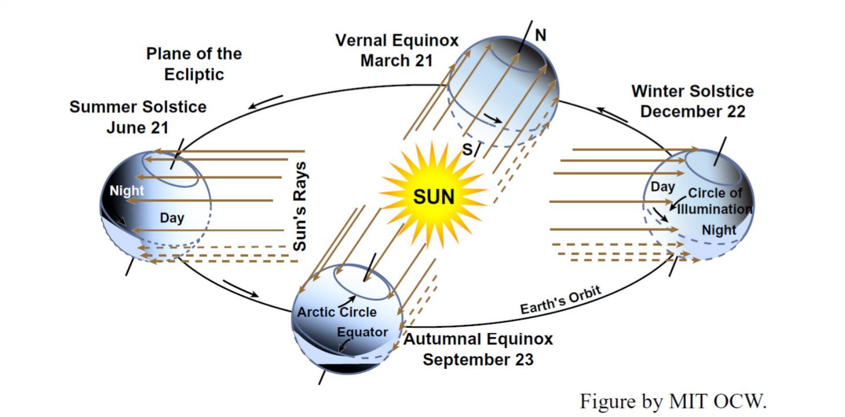


Figure 1-4. The yearly path of the Earth in its revolution around the Sun, showing how the seasons result from the differences in the angle between the Earth's rotation axis and the Earth–Sun line. (From Strahler, 1975.)

(If you think that things around the time of the winter solstice are bad at the latitude of Boston, think about what they are like in the latitude of Scandinavian cities, or Nome or Fairbanks in Alaska. On the other hand, during the legendary "white nights" of St. Petersburg, in early summer, it is said that one can read a newspaper at midnight without artificial light.)

But the angle at which the sun rises above the horizon at sunrise, and the equal angle at which the sun sets below the horizon at sunset, remains the same through the year (Figure 1-5). Think carefully about the geometry of the Earth– Sun system to convince yourself of that fact.

Incidentally, can you think of a reason why the equinoxes and solstices do not fall on exactly the same days, in March and September and in June and December, respectively, every year? It's related to the existence of ***leap years***. The astronomical duration of the year—the time it takes for the earth to make on complete revolution around the sun—is close to 365-1/4 days, whereas the calendar year has 365 days, by definition. Every four years they add 29 February to the year, and it makes things come out almost okay, except not quite, so once every century they have to have an extra adjustment of one day, etc., etc. That causes the precise times of the year when the equinoxes and solstices occur to vary depending upon the position of the given year relative to the leap years.

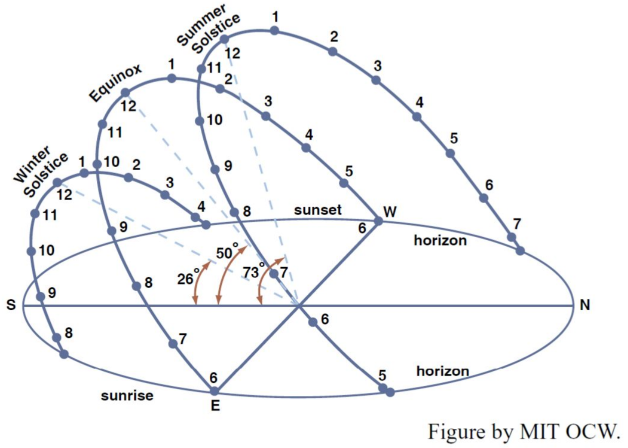


Figure 1-5. The Sun's path in the sky at the equinoxes and at the solstices at latitude 40° N. (Modified from Strahler, 1975.)

One final point: the time of *maximum* distance from the sun to the earth, called ***aphelion***, falls on 4 July, and the time of *minimum* distance from the sun to the earth, called ***perihelion***, falls on 3 January. Note (Figure 1-6) that (1) these times are out of sync with the seasons and (2) the earth is closest to the sun in the northern-hemisphere winter, by a factor of about 1.03 (152 x 106 km divided by 147 x 106 km), which because of the inverse-square decrease in the sun's radiant energy with distance from the sun translates to a factor of about 1.07 (or seven percent, in what I think is the conventional way of stating percentages). It's a good thing, too, except perhaps for ski fanatics: if the situation were the reverse, with the Earth farthest from the Sun in the northern-hemisphere winter, winters in New England would be even tougher than they are.

1.2.3 Sunlight

As you are probably aware, the Sun is fueled by thermonuclear fusion, by which hydrogen is fused to helium with enormous release of energy. The Sun's surface, at a temperature of about 6000°C, radiates energy in all directions over a wide range of wavelengths. The Earth intercepts a tiny fraction of that radiant energy, in the form of sunlight. Most of the ultraviolet part of the radiation is absorbed before it reaches the Earth's surface, although in recent times the degradation of the ozone layer as a consequence of certain man-made refrigerant gases in the upper atmosphere has led to greater ultraviolet radiation at the surface, especially at high latitudes. The effects of man-made gases on the ozone layer, and its consequences for ultraviolet radiation, is a field of active research by atmospheric chemists and physicists nowadays.

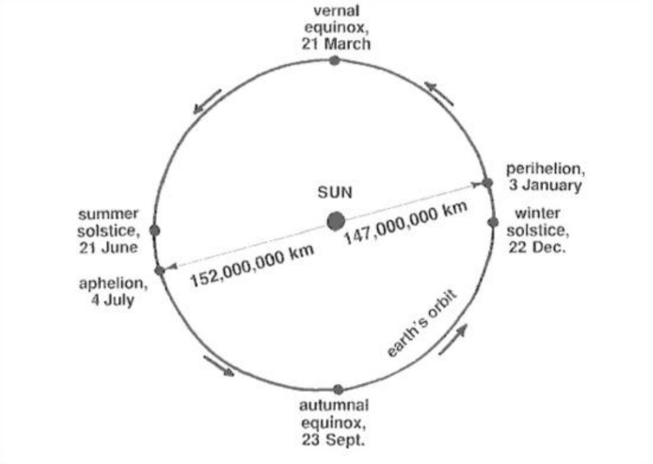


Figure 1-6. The Earth's orbit around the Sun, and the seasons. (Modified from Strahler and Strahler, 1992.)

How much of the Sun's energy reaches the earth? Think about one square centimeter of area, oriented perpendicular to the line between the Earth and the Sun and outside the Earth's atmosphere. The rate of delivery of radiant energy to that square centimeter is very nearly *two gram calories per minute*. This is often referred to as two langleys per minute, a ***langley*** being *one gram calorie per square centimeter*. This value is called the ***solar constant***, although in fact it's not quite constant: it varies slightly depending upon two effects:

- changes in the sun's output, connected with things like solar flares; and

- changes over the course of a year, because the earth's orbit around the sun is not a circle but rather an ellipse with a small eccentricity, as you have already seen.

How can I give you some feel for what the solar constant means in everyday life? First of all, I should point out that the energy flux (by that I mean rate of delivery of energy per unit area perpendicular to the direction of radiation) is at most *about half that just outside the atmosphere*, or about one langley, because of the inevitable absorption and reflection, even on the clearest of days. On cloudy days, or when the sun shines at a low angle through the atmosphere, the value is correspondingly much smaller.

All of you have some experience with how much energy it takes to heat water. You put a pan of water on the stovetop, turn on the burner, and wait and wait for the water to boil. As you will learn more in a later chapter, water has a very high specific heat capacity: it takes a lot of energy to raise the temperature of water by a given amount.

Suppose that you had a little cube, one centimeter on each side, filled with water and oriented with one face facing directly into the sunlight. The other five faces are perfectly insulated from their surroundings, and the face that catches the sunlight is completely transparent to the sun's incoming energy but can pass no heat back out to its surroundings. How long would it take to heat water contained in the cube from the freezing temperature (0°C) to the boiling temperature (100°C)? Answer: about 100 minutes.

1.2.4 Heating and Cooling of the Surface

The ground surface is the interface between the atmosphere and the solid and liquid material of the earth. The ground surface is where solar radiation is intercepted and converted to heat. It is also the source of outgoing long-wave radiation. It is where liquid water is evaporated and where incoming rainfall is stored as soil moisture and groundwater. First, in the following two paragraphs, are some very general things about heating and cooling. Then, after some more physics background, there is more detail on the heating and cooling of the Earth's surface.

Low-lying areas are commonly colder at night than higher ground nearby. On clear nights, the ground is chilled as its heat is radiated out to space. The cold ground then chills the air near the ground. The chilled air is slightly more dense than the overlying air, so it tends to flow slowly downhill, in the same way that water flows downhill. The cold air "ponds" in low areas. These are places where the first frosts of autumn are earliest and where the last frosts of spring are latest. If you ever have a chance to plant fruit trees, plant them on the highest ground around!

In hilly areas, north-facing slopes get less sunshine than south-facing slopes. Local temperatures on the north-facing slopes are colder than on south- facing slopes in both summer and winter. In areas with winter snows, the snow melts much later on north-facing slopes. Even over distances of a few meters, the difference in microclimate between a sunny, open area with low herbaceous vegetation and a nearby grove of tall trees can be spectacular.

Everybody knows that the Earth's surface tends to be warmed on a sunny day and cooled on a clear night. Let's look more deeply into how this happens. Many different and interesting effects must be taken into account.

If there were no atmosphere, things would be fairly simple: the Sun radiates energy to the Earth's surface at short wavelengths, most of it absorbed but some of it reflected, and the Earth would re-radiate that energy back out to space at longer wavelengths, and the temperature of the surface would become adjusted so that the outgoing long-wave radiation would balance the incoming short-wave radiation over a long time. (For that statement to make more sense to you, you need to be aware that *the intensity of emission of radiation from a body increases with the temperature of the body—to the fourth power!* That's called the ***Stefan- Boltzmann law***.)

A few additional words about the reflection of incoming solar radiation by the Earth's surface back into space are in order here. The local reflectivity varies greatly depending upon what's covering the surface. It ranges from as much as 95%, for fresh snow cover, to as low as just a few percent, for water surfaces at high sun. Figure 1-7 is a table giving approximate values of reflectivity for various kinds of surface. The overall grand average for the entire earth (that is, *the percentage of incoming solar radiation intercepted by the entire earth that's reflected back to space, on a long-term average*), is called the earth's ***albedo***; its value is about 30%.

fresh snow cover	75-95
dense cloud cover	60-90
old snow cover	40-70
clean firn snow	50-65
light sand dunes, surf	30-60
clean glacier ice	30-46
dirty firn snow	20-50
dirty glacier ice	20-30
sandy soil	15-40
meadows and fields	12-30
densely built-up areas	15-25
woods	5-20
dark cultivated soil	7-10
water surfaces, sea	3-10

Figure by MIT OCW.

Figure 1-7. The albedo of various kinds of surfaces. (From Geiger, 1965.)

In the real world, with a thick atmospheric cover, the situation at the ground surface is much more complicated, largely by virtue of two effects:

- *The atmosphere reflects and scatters some of the incoming solar radiation*. (***Scattering*** is the process whereby the atoms, molecules, and tiny particle in the atmosphere interact with passing electromagnetic waves whose wavelengths are approximately the same as the size of the particles, causing a part of the waves to be diverted in a wide range of directions.) What arrives at the surface is a combination of direct radiation and indirect, downscattered radiation. The atmosphere itself absorbs some of the incoming solar radiation, but, perhaps surprisingly, not much. (Incidental note: the shorter wavelengths are more susceptible to scattering than the longer wavelengths, so the indirect downscattered radiation we see at the earth's surface tends to be in the shorter- wavelength part of the visible spectrum—hence the *blue sky* on a clear day. By the same token, the *yellow or orange or even red sun* seen at sunrise or sunset is a consequence of the greater scattering of the shorter wavelengths during the long slanting passage of the Sun's rays through the atmosphere, leaving mainly the longer-wavelength radiation to come through to our eyes.)

- A lot of the long-wavelength back-radiation to space from the earth's surface is absorbed by the atmosphere. Some of the absorbed energy is re-radiated back to the earth's surface, and some is re-radiated out to space. The important effect here is that *there is a net radiation of long-wavelength energy to space, but the almost-equal magnitudes of back-and-forth long-wavelength radiation between the surface and the atmosphere is much larger than the net radiation to space*. This is the famous ***greenhouse effect***.

Certain atmospheric gases figure most prominently in the absorption of outgoing long-wave terrestrial radiation by the atmosphere, foremost among them being *water vapor, carbon dioxide, methane, ozone, and certain man-made gases like chlorofluorocarbons*. It's this strengthened greenhouse effect, caused by the increasing concentration of anthropogenic greenhouse gases, that is thought to be the cause of global warming (but the situation is in reality more complicated, in large part because of the still poorly understood potential changes in cloud cover; for such reasons, global warming has its responsible skeptics). Each greenhouse gas absorbs outgoing terrestrial long-wave radiation in different segments of the electromagnetic spectrum; they combine to leave only narrow windows of transparency (Figure 1-8).

Water vapor, because it is an effective absorber of long-wave terrestrial radiation and because it is present in the atmosphere in far greater concentrations than any of the other greenhouse gases, is the number-one greenhouse gas (but the others, most importantly carbon dioxide, humankind has some control over, which is not the case with water vapor). Beware of what you read in the news media about carbon dioxide being the most important greenhouse gas! I could also point out here that the greenhouse effect is much more our friend than our enemy: if it were not for the greenhouse effect, the Earth would be a frozen and lifeless planet.

You need also to be aware of several readily understandable processes specific to the earth's surface, which are important factors in the heat budget of the surface:

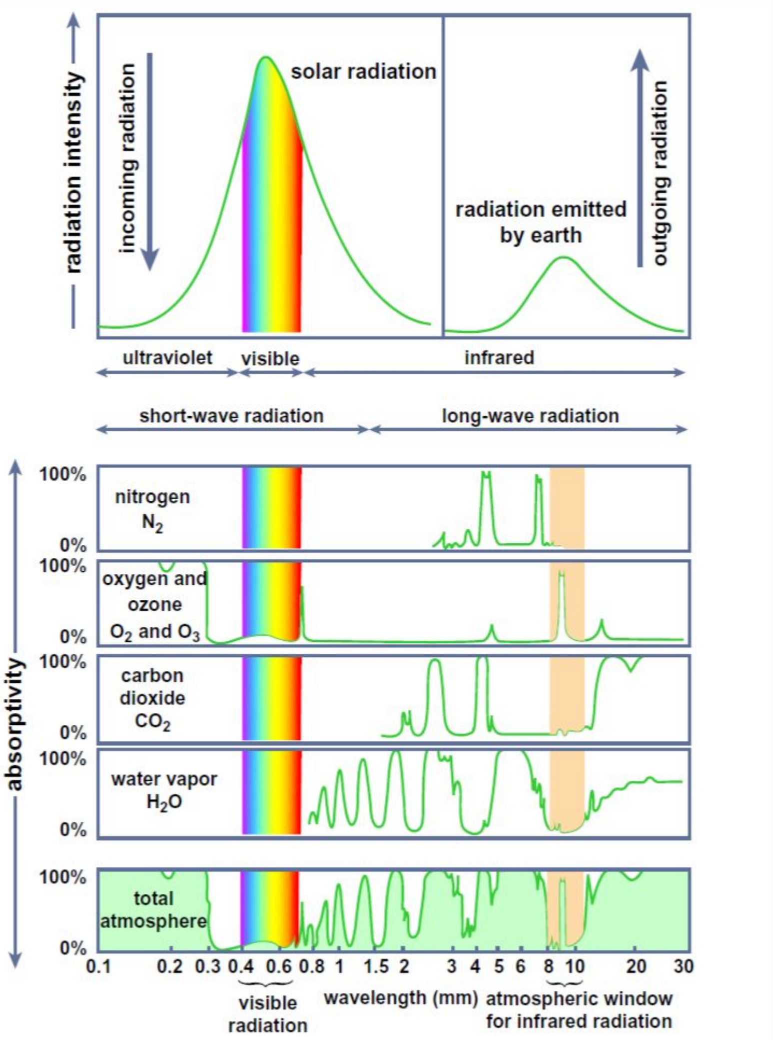


Figure 1-8. The absorptivity of selected greenhouse gases of the atmosphere, and of the atmosphere as a whole. (From Lutgens and Tarbuck, 2004.)

• When the surface is warmer than the immediately overlying atmosphere, heat is conducted from the ground to the atmosphere; when the surface is cooler than the air above, heat is conducted downward from the air to the ground.

• As the ground surface is warmed, heat is conducted downward into deeper levels of the soil; as the ground surface is cooled, heat is conducted upward from the deeper levels.

• Provided that the relative humidity is less than 100%, soil moisture present at the ground surface, or drawn upward to the ground surface by capillary action, is evaporated into the atmosphere, thereby delivering latent heat to the atmosphere. (As you will learn in Section 3 below, it takes a lot of heat energy to evaporate liquid water; that energy is released into the atmosphere when the water vapor condenses to liquid water.)

• Some heat is added to the ground surface by condensation of dew on clear nights, but that's not an important effect; what's more important is delivery of precipitation to the ground surface from above. If the temperature of rain is greater than that of the surface, the surface is warmed, and vice versa.

It's not easy to make any simple statement about the state of the ground surface in terms of heat, because all of the above effects and processes *vary with time of day, state of the weather, time of year, and climatic zone*. But the next time you are outdoors and about, looking at the ground surface, I want you to think about all of the foregoing processes, acting simultaneously to set the temperature of the surface. Just for a summary, look at Figure 1-9, which shows the various processes, and in a qualitative way their typical magnitudes, on a sunny summer day and on a clear summer night. The major effects are: incoming solar radiation absorbed by the ground (S), which is a combination of direct and downscattered radiation; incoming solar radiation reflected from the ground (R); incoming long-wave radiation from the atmosphere (LI); outgoing long-wave radiation to the atmosphere and outer space (LO); conduction of heat from the ground to the overlying air or from the overlying air to the ground (C); and loss (or minor gain) of heat from the ground by evaporation of soil moisture (V).

On average, over the course of a year, there is a net gain of heat by ***insolation*** (the term used to describe *the totality of solar radiation reaching the Earth’satmosphere*)atlowlatitudesandanetlossathighlatitudes.That imbalance has to be made up somehow, because the Earth’s average surface temperature changes only very slowly with time. That happens by net transport of heat from low latitudes to high latitudes by the Earth’s wind systems and ocean currents. One consequence of the existence of seasons is that the solar energy received by the Earth varies not only with latitude, as described above, but also with the seasons. Does it surprise you to learn that around the time of the summer solstice insolation at the north pole is even greater than at the equator at that same time, because the sun shines there twenty-four hours a day! Figures 1-10 and 1-11 show, in two different ways, how insolation varies as a function of both latitude and season.

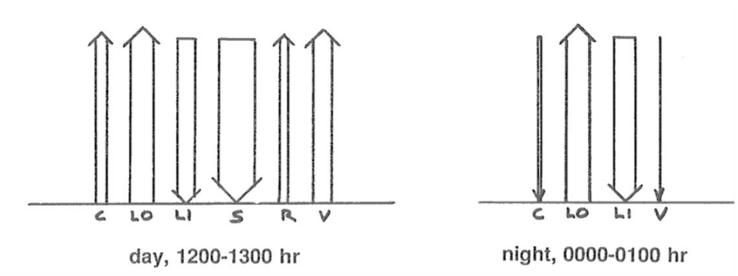


Figure 1-9. Elements of the heat budget of the ground surface. Left: clear summer day. Right: clear summer night. The thickness of the arrows show very approximately the relative magnitudes of the effect. (Modified from Geiger, 1965.)

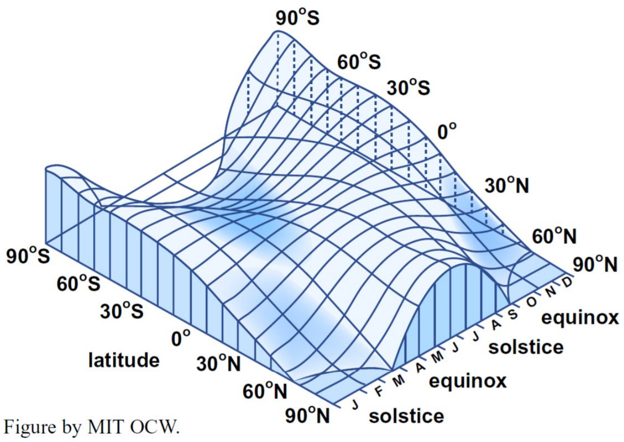


Figure by MIT OCW.

Figure 1-10. Insolation as a function of latitude and season. (From Strahler, 1975.)

What about the temperature of the soil (or, more generally, regolith and bedrock) beneath the ground surface? Despite songs about the cold, cold ground, the substrate is not necessarily colder than the ground surface. *The vertical distribution of temperature as a function of depth below the surface depends upon a multitude of factors*:

- ambient temperature of the air above the surface
- clarity of the sky
- state of the ground surface, especially moisture content and vegetative cover
- composition of the substrate
- moisture content of the substrate
- past history (because the temperature profile does not adjust instantaneously to changing surface temperature: there’s a long time lag)

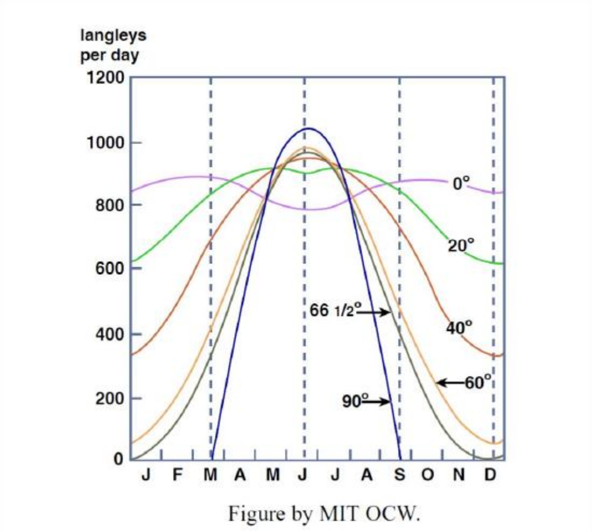


Figure 1-11. Insolation as a function of latitude and season in the northern hemisphere. (From Strahler, 1975.)

When the Sun shines on the Earth's cool land surface and heats it, that heat is conducted downward into the Earth, at a rate that depends on the temperature difference between the surface and the subsurface and also on the heat capacity and the thermal conductivity of the Earth's soil and rock materials. Likewise, when the Earth's surface cools by radiation of energy back out to space, heat from the still-warm subsurface materials is transferred by conduction back out to the surface. Such changes are mainly on time scales of day to night and of winter to summer seasons. Figures 1-12, 1-13, and 1-14 show some interesting results of actual measurement of the cyclic changes in temperature with depth.

Obviously the temperature near the surface is greatest in summer and least in winter, but there are two significant effects below the surface:

• The changes in temperature over one diurnal (that is, daily) cycle or over one yearly cycle decrease with depth in the Earth. Below a depth of about a meter, the day-to-day changes are negligible, and below a depth of about ten meters the temperature is almost the same year-round.

• The timing of the change in temperature at some depth lags behind the timing of the change in temperature at the surface. That's because it takes time for the heat to be conducted downward. The time lag increases with depth. At a certain depth, the cycle is out of phase with the surface by a whole half cycle: that is, the maximum temperature attained at depth happens at the time of minimum temperature at the surface, and vice versa!

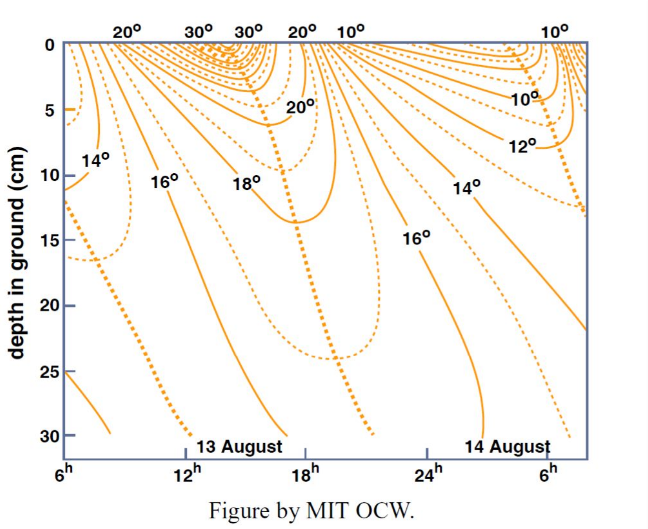


Figure 1-12. Penetration of the daily temperature wave (in degrees Celsius) into the ground on a clear summer day in Finland in mid-August. (From Geiger et al., 2003.)

In colder climates the effects would be qualitatively the same, but the temperatures would be lower. In such areas, as in New England, the upper layer of the substrate, to a significant depth, becomes frozen for quite some time in the winter. Building codes in cold regions are specified to ensure that footings for foundations are placed at a depth great enough to be below the deepest freezing level to be expected in the given climate.

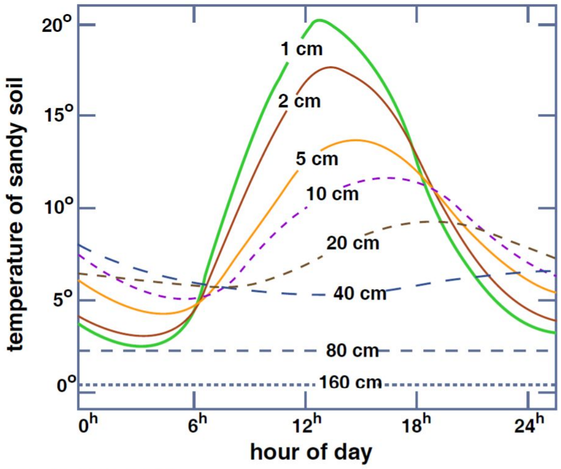


Figure by MIT OCW.

Figure 1-13. Daily sequence of temperature in sandy soil in May (ten-year averages). (From Geiger et al., 2003.)

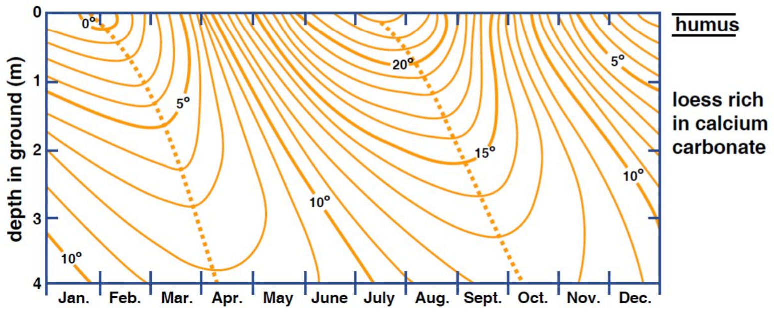


Figure by MIT OCW.

Figure 1-14. Annual sequence of soil temperature (in degrees Celsius). (From Geiger et al., 2003.)

1.3: Some Elementary Chemistry Relevant to the Earth's Surface

1.3.1. Atoms

Atoms are the building blocks of matter. Do I need to define for you what an atom is? One way of doing that is to say that an atom is the smallest unit of matter that reflects the chemical characteristics of the matter, in respect to its interactions with the atoms of other kinds of matter. In other words, atoms are units of matter that cannot be split up further without losing the chemical behavior of the matter, in terms of bonding with other atoms to form chemical compounds. If you're going to understand why various kinds of atoms bond together the way they do to form the important minerals, you have to know what atoms are like and why the atoms of the various chemical elements behave differently.

The simplest cartoon for an atom (Figure 1-15) has a number of electrons orbiting a nucleus consisting of closely packed protons and neutrons.

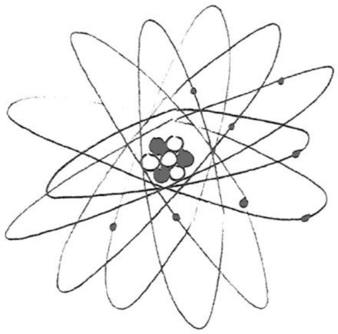


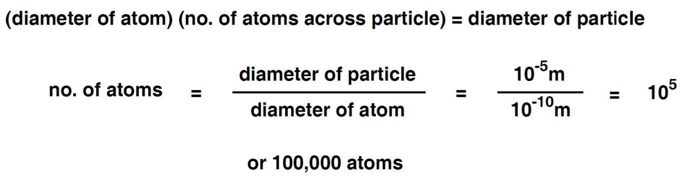
Figure 1-15. Extremely oversimplified cartoon of an atom.

Basic facts (don't worry about how we know them, and don't worry about remembering them for exam purposes):

- mass of proton: 1.7 x 10-24 g
- mass of neutron: 1.7 x 10-24 g
- mass of electron: about 1/1840 of the above electric charge on neutron: zero
- electric charge on electron (one "unit" of negative charge (smallest amount possible!)
- electric charge on proton: one "unit" of positive charge (just the opposite of that of the electron)
- size of atom: about 10-10 m size of nucleus: about 10-15 m

Note right away that for an atom to be electrically neutral is must have the same number of protons and electrons. So the nucleus is much heavier than the electron cloud around it, but also much smaller, by a factor of about 105.

What's the best way for you to visualize the actual size of an atom? Think about the very tiniest dust particle you can see with your unaided eye under the best lighting conditions. Assume this to be one hundredth of a millimeter, or 10-5 m. About how many atoms would stretch across that distance? See Figure 1- 16 for a computation.



(diameter of atom) (no. of atoms across particle) = diameter of particle

$$\text{no. of atoms} = \frac{\text{diameter of particle}}{\text{diameter of atom}} = \frac{10^{-5}\text{m}}{10^{-10}\text{m}} = 10^{5}$$

or 100,000 atoms

Figure by MIT OCW.

Figure 1-16. Computation to give you an idea how big an atom is.

What are electrons really like, and what are their orbits around the nucleus really like? These are complicated questions, difficult to answer in a simple way. To get a good idea you need to know something about quantum mechanics, a branch of physics, which has grown up since the early decades of the Twentieth Century, that deals with the behavior of very small particles.

Particles that are very small and very fast-moving behave in ways our common sense doesn’t cover. Electrons aren’t just little solid spheres, like billiard balls or mini-planets; they are both particles and little packets of waves at the same time, in that in some ways they behave like particles and in other ways they behave like waves. In this respect, electrons are just like light, which also shows this particle–wave duality. If this doesn’t make much sense to you, don’t worry about it. Also, we have no way of knowing where the electrons are in their orbits around the nucleus: all we can say is what the probability is that they will be found in a given place. So it’s a little better to cartoon the atom as a smeared-out “probability cloud” of electrons around the nucleus (Figure 1-17), although even that is still not very satisfactory.

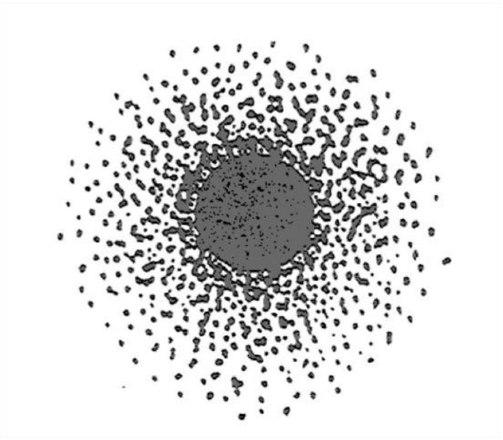


Figure 1-17. An atom schematized as a smeared-out probability cloud around the nucleus. The denser the shading, the more likely you are to find the electron there.

But with all this said, in order to get somewhere studying minerals we will make some simplifications about electrons and think of them as “ordinary” little particles traveling in “ordinary” little orbits around the nucleus.

1.3.2. The Chemical Elements

There are many different kinds of atom, depending upon the number of electrons, protons, and neutrons in the atom. In terms of the chemical behavior of atoms, which is what we’re interested in here, the important thing is the number of electrons (and the equal number of protons), because that’s what determines how the atom bonds (or fails to bond) with other atoms. Each distinctive kind of atom, in terms of the number of electrons and protons it has, is called a ***chemical element*** (or just an ***element***).

Let’s investigate the structure and behavior of the elements by starting with the lightest and simplest kind of atom: hydrogen. The nucleus of the hydrogen atom consists of one proton, and the electrically neutral atom has one electron orbiting the proton nucleus. The next lightest and next simplest atom is helium. Its nucleus consists of two protons, and the electrically neutral atom has two electrons orbiting it—et cetera. We can build up a picture of all of the elements in this way, by adding one proton (and therefore one electron) at a time—with some neutrons thrown into the nucleus also.

	1	2	3	4	5	6	7	8
1	H 1							He 2
2	Li 3	Be 4	B 5	C 6	N 7	O 8	F 9	Ne 10
3	Na 11	Mg 12	Al 13	Si 14	P 15	S 16	Cl 17	Ar 18
4	K 19	Ca 20						

Figure 1-18. The first part of the periodic table of the elements.

Figure 1-18 is a chart of the first twenty or so elements. This covers most of the important elements that go into the most important rock-forming minerals. Why have I arranged the elements in this way? You might remember from some chemistry course that this represents the first three rows of what is called the ***periodic table of the elements***, plus a little of the fourth row.

This grouping of the elements corresponds very closely to the chemical behavior of atoms of these elements, when it comes to combining, or bonding, with atoms of other elements to form chemical compounds. (Remember that minerals are just chemical compounds.)

This kind of grouping (more complete than shown here) was first perceived by the Russian scientist Dmitrii Mendeleev in 1869. His insight into the nature of organization of the elements in terms of their chemical behavior paved the way for people after him to actually discover new elements his table predicted. It was not until the early 1900s, with the advent of the quantum theory of matter, however, that people realized why the elements are arranged in this way.

Electrons are present in atoms in group of orbits called ***shells***. That is, electron orbits occur in groups. A filled shell (all the electron orbits of the shall are occupied by electrons) is a very stable configuration of electrons—stable in the sense that the electrons are tightly bound, and it takes a lot of energy to remove one, and also the atom has no tendency to take on electrons to try to fill a shell. Elements with filled-shell electron configurations are very inert and almost totally nonreactive.

The first shell has positions for two electrons. This corresponds to the first row of the periodic table. Hydrogen doesn't have a filled-shell configuration, so it enters into an enormous number of chemical compounds in its search for the extra electron, from some other atom willing to give it up, to attain the filled-shell configuration. But helium, which has two electrons already, is an inert gas. The second shell contains eight electrons, corresponding to the second row of the periodic table. Neon has both of the first two shells filled and, like helium, is an inert gas. The third shell also contains eight electrons, and corresponds to the third row of the periodic table. Argon is the inert gas with all three first shells filled.

1.3.3. Ions

How about the elements with unfilled shells? There is a strong tendency for an atom to end up with a filled shell, by either giving up electrons (to another atom) or taking on electrons donated from another atom (or by sharing electrons with another atom; see below). Atoms that have attained a filled-shell configuration by either giving up or taking on electrons are called ***ions***. The characteristic thing about ions is that have an electric charge.

Those elements (near the left side of the periodic table) with only a few electrons in their outer shell can fairly easily be stripped of those electrons so as to have a filled-shell configuration. Examples are sodium (Na), magnesium (Mg), potassium (K), and calcium (Ca). Ions like these have a positive charge, because they have lost electrons (which are negatively charged).

> Na atom minus 1 electron → Na+ ion (single positive charge)
>
> Mg atom minus 2 electrons → Mg2+ ion (double positive charge)

Figure 1-19 shows some examples of positively charged ions.

Elements that are near the right side of the periodic table, with almost enough electrons in the outer shell to produce a filled-shell configuration, can easily take on enough electrons to fill the shell. Examples are oxygen (O), fluorine (F), sulfur (S), and chlorine (Cl). Ions like these have a negative charge, because they have gained electrons.

> Cl atom plus 1 electron → Cl- ion (single negative charge)
>
> O atom plus 2 electrons → O2- ion (double negative charge)

Figure 1-20 shows some examples of negatively charged ions.

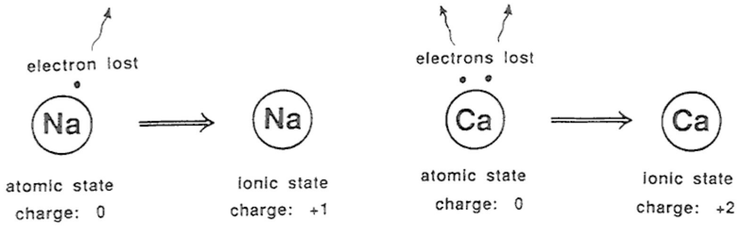


Figure 1-19. Examples of positively charged ions.

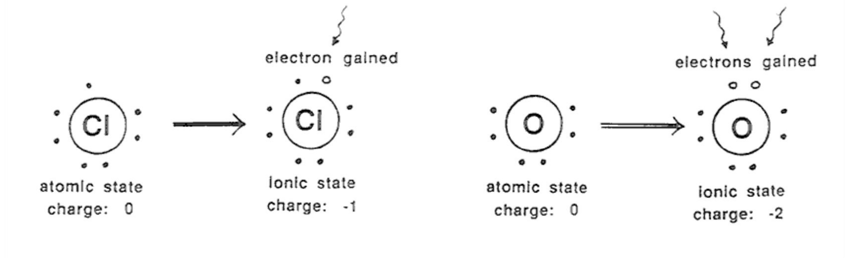


Figure 1-20 Examples of negatively charged ions.

For reasons that need not concern us here, positively charged ions are called ***cations*** and negatively charged ions are called ***anions***.

1.3.4. Ionic Bonds

All of you know that there are such things as electric charges and that these can be moved around from place to place. (Think about hair and clothes on dry days. Or get under the synthetic-fiber bedspread in the dark when the humidity is low and run your hands over its surface, and watch the sparks.) Benjamin Franklin recognized that there can be positive and negative charges (incidentally, the assignment of “positve” and “negative” is arbitrary) and that charges of the same sign repel one another and charges of opposite sign attract one another. This is expressed by ***Coulomb’s law***, one of the basic laws of physics:

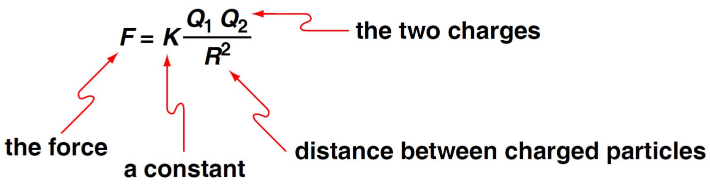


Figure by MIT OCW.

- If Q1 and Q2 are of the *same* sign, F is *repulsive*
- If Q1 and Q2 are of *opposite* sign, F is *attractive*

All of this applies to ions! Two ions of opposite charge can become bonded to one another by this Coulomb attractive force. This kind of bond is called an ***ionic bond***. Most minerals are held together by just this kind of bond. (Now we’re getting somewhere on the basic nature of minerals.) Crystals that are held together by ionic bonding are called ***ionic crystals***.

1.3.5. Ionic Crystals

The mineral halite (NaCl, sodium chloride, also known as table salt) is a good example of an ionic crystal—and an important rock-forming mineral as well. The structure of halite involves equal numbers of positive ions (sodiums) and negative ions (chlorides) arranged in a cubic array, each sodium surrounded by six chlorides and each chloride surrounded by six sodiums. (See Figure 1-21.)

Think of three mutually perpendicular lines, and along each line there’s a regular alternation of sodium and chloride ions. If you draw lines connecting all of the six chlorides that are the nearest neighbors to any given sodium, you get a regular geometric solid, called an octahedron, made up of eight equilateral triangles. For this reason we say that there is *octahedral coordination* of chlorides around sodiums. You can show for yourself that the same holds true for the six nearest-neighbor sodiums around any given chloride.

The picture in Figure 1-21 is in one sense a fake, because, for reasons of good visibility, I’ve drawn it with too much space between adjacent ions. To get the “true” picture you have to imagine that all of the spheres, of both kinds, expand (in such a way that the relative sizes of the two kinds stay the same) until they are all in contact; this is difficult for me to draw intelligibly.

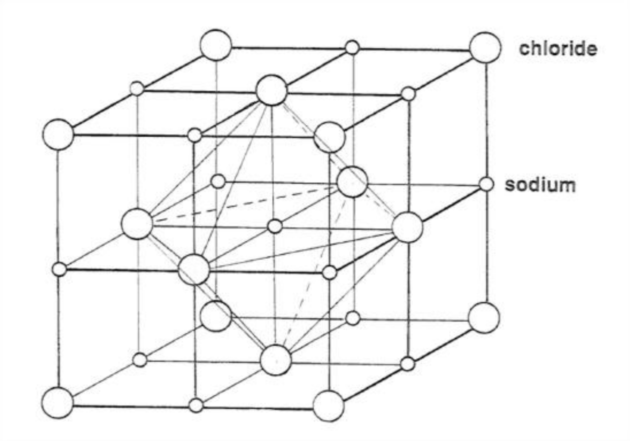


Figure 1-21. Blown-up sketch of the arrangement of sodium ions (small) and chloride ions (large) in the structure of halite.

It turns out that the relative sizes of Na and Cl ions are just about right for this kind of "cubic" packing—just right in the sense that they all come out to be just about touching. In fact, if they were of different relative sizes the packing would have come out to be of some different kind.

Look at the central sodium in the figure. The strongest forces on it are attractive, exerted by its six nearest neighbors, which are chlorides. (Keep in mind the strong, inverse-square dependence of the force on the separation distance in Coulomb's law.) The next strongest forces are repulsive, exerted by its twelve next-nearest neighbors, which are sodiums. And so on outward. The structure likes to form because the overall picture of forces, summed over all of the ions, near and far, surrounding that ion is attractive. Other structures with different kinds of atoms, which might seem okay at first thought, don't form because the overall picture of forces comes out to be repulsive. To get such a structure to hold together, you would have to deep squeezing it hard from all directions.

There are two other factors that determine whether a given mineral forms (i.e., is stable) in a given situation:

- All atoms are in a state of continual jiggling, because of their thermal energy. Only at absolute zero does all movement cease. This jiggling may be so strong that, even though the picture of inter-ionic forces is attractive, the crystal flies apart because of the thermal jiggling. This is why crystals (ice is an everyday example) melt as the temperature rises.
- The greater the concentration of Na and Cl ions in the medium surrounding the crystal, the higher the temperature at which the crystal melts, because ions from the surroundings tend to diffuse into the crystal and keep it together.

1.3.6. Covalent Bonds

There's another kind of bond that's especially important in silicate minerals, called a ***covalent bond***. It's not as easy to understand than the ionic bond. Atoms can fill their electron shells by *sharing electrons with other atoms*.

The simplest example is the bond in the H2 molecule (Figure 1-22). This is very schematic, and it doesn't give you a good picture that both electrons orbit both nuclei. It's as if each nucleus gets something for nothing. Figure 1-23 is a slightly better way of looking at it, in terms of the shape of the electron cloud.

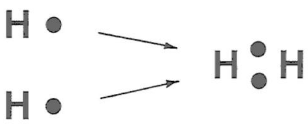


Figure 1-22. Very schematic representation of a hydrogen molecule.

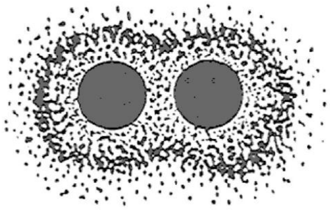


Figure 1-23. A slightly better way of representing the hydrogen molecule.

Comments:

- Covalent bonds (if they are favorable in the first place because of electron-orbit considerations) tend to be found between atoms of elements that are in the same or almost the same position in the periodic table, in terms of columns.
- Many bonds are purely ionic or purely covalent, but many are intermediate in character: there is partial sharing of electrons, but the density of electron distribution is shifted so that one component atom tends to be negatively charged and the other positively charged. This is called *partial ionic character* of the bonds.
- Covalent bonds are inherently more complicated to understand than ionic bonds, and I've been much more superficial about them. They depend upon certain intricacies of electron orbit structure that are understandable only on the basis of quantum mechanics. Covalent bonds actually have a directional property that's governed by the nature of electron distributions around the atoms, not just packing geometry as in the case of ionic bonds. It's as if the atoms had little hooks on them in certain positions.

1.3.7. Solids, Liquids, and Gases

Most chemical compounds can exist in all of three different *states of aggregation*—as a *solid*, a *liquid*, or a *gas*—depending on temperature and pressure. This was known, of course, long before the molecular theory of matter was developed in the late eighteenth century and the early nineteenth century. In the past hundred and fifty years, great effort has gone into developing theories to account for the nature of solids, liquids, and gases on the molecular level. (Here, I'll use the term *molecule* for atoms, ions, or molecules indiscriminately.)

You probably already know well the fundamental macroscopic distinctions among gases, liquids and solids:

gas:

- has relatively low density
- cannot withstand applied shearing forces without continuing deformation
- expands to attempt to fill its container uniformly

liquid:

- has relatively high density
- cannot withstand applied shearing forces without undergoing continuing deformation
- maintains a definite volume within a well-defined bounding surface 32

solid:

- has relatively high density
- withstands applied shearing forces without undergoing continuing deformation
- maintains a definite volume within a well defined bounding surface

On a more fundamental level, the basic picture of a **gas** is that its constituent molecules are widely separated relative to their diameter, except for occasional close encounters, and are moving in free trajectories with only occasional collisions with the walls of the container and with each other. If you could somehow travel on a gas molecule, most of the time you would be having a boring ride, because the molecule would be in a state of uniform straight-line motion at constant speed, with neighboring molecules so far away you'd have to use binoculars to see them. Every now and then, however, another molecule or the wall of the container would come rushing toward you, and your molecule would rebound elastically from the encounter.

This state of motion, with its well-defined elastic collisions and negligible mutual forces among molecules in the meantime, lends itself to successful modeling of the gaseous state—especially because many-body (that is, three-plus) collisions or interactions are so rare. Such modeling has led to the *kinetic theory of gases*, by which most of the properties of gases can be accounted for or predicted.

At the other end of the line, on a molecular level a **solid** is characterized by fixed and ordered positions of the constituent molecules, with packing so close that there are strong forces at all times between any given molecule and all of its near neighbors. At temperatures above absolute zero the molecules oscillate or vibrate in various ways about their fixed mean positions, but only rarely do molecules trade places in their ordered arrangement.

If you could occupy one of the molecules in the solid structure, you would feel as if you had been consigned forever to one of those vibratory-massage beds they used to have in motels, and if you looked over at the strictly regular pattern of neighboring molecules looming near you, they would be vibrating too, but maintaining their mean positions for extremely long times before trading places with a neighbor. And you would have the same picture for as far as you could see through the structure.

Despite the complexity of the mutual forces, the ordered and permanent arrangement in crystalline solids has allowed successful modeling, and the solid state is by now well understood.

Liquids, however, have presented much greater problems than gases and solids. As with solids, the molecules are always close enough together that they exert strong forces on one another, but the degree of ordering (or, more precisely, the time scale for the existence of ordering) is much less than for solids.

If you were to occupy a molecule in a liquid, your vibratory movement would be about the same as when you were in the solid, and your neighbors would loom almost as close. But the arrangement of your neighbors would stay the same only for a short time and then change to some different arrangement by a combination of just shifting in position and actually trading places. You might or might not get a sense of regular arrangement of the neighbors, but even if you did you wouldn't be able to follow that regularity very far outward, and anyway it would be transitory.

Various kinds of conceptual models of the liquid state, partly geometric and partly kinematic, have been devised. These models can be massaged, by supplying somehow the constants or parameters that must go along with them, to predict the properties of the liquid. None of these models comes close to accounting well for all of the important properties of liquids. Knowledge of the liquid state is therefore still far from perfect.

1.3.8. Chemical Reactions

In case your knowledge of chemistry is scanty, here is some basic stuff. Chemical reactions, like Reaction 1 above, are processes by which certain substances react with one another to produce other substances. (That includes situations in which a single substance undergoes a change by which it is transformed into one or more different substances, or by which two or more substances are transformed chemically into a single substance.) The beginning substances are called the ***reactants***, and the ending substances are called the ***products***. The substances involved might be solids (crystalline or amorphous), or ions or molecules dissolved in a liquid medium, or even just the atoms and molecules of a gas.

The details of chemical reactions involve usually complex interactions among the atoms that constitute the substances involved, and they need not concern us here. In the case of solid substances, you can think in terms of the atoms of the reactants being disengaged from the crystal structure of the reactant(s) and then added to the crystal structure of the product(s). Or, in the case of the ***dissolution*** of a solid substance in a liquid, what's going on is that the ions or molecules of the solid are being detached from the surface of the solid, to roam around freely in solution in the liquid. The reverse of a dissolution reaction like that is the addition of the dissolved ions or molecules onto the growing surface of the solid by removal from solution—a process called ***precipitation*** (not to be confused with the meteorological term, which is used for liquid or solid water particles falling through the atmosphere).

In the case of dissolution or precipitation, mentioned above, what's really going on is that ions or molecules are continually being added to and removed from the solid surface. Whether the solid is shrinking (dissolution) or growing (precipitation) depends on the relative rates of removal and addition of atoms or molecules.

Chemical reactions can go both ways. That's why Reaction 1, shown above, has the double arrow in the middle of it. If you pump up the concentration of the substances on the left side of the reaction, that drives the reaction to the right, producing the substances on the right (then called the products) at the expense of the substances on the left (then called the reactants). But if instead you pump up the concentration of the substances on the right, that drives the reaction to the left—and then you have to use the terms "reactant" and "product" in reverse!

1.3.10. Oxidation

The business about oxidation, and the difference between iron in the ferrous state and in the ferric state, is likely to be mysterious to you unless you have some sophistication in chemistry. Here's some background material on such matters. You probably have heard of *oxidation*, and you may also have heard of *reduction*. These two processes go hand in hand, and are often combined into the term ***redox***.

Many chemical reactions we deal with in and on the earth involve transfer of the actual nuclei of the atoms involved. The dissolution and precipitation of soluble salts like NaCl are good examples. A large and important class of chemical reactions, however, involve only *transfer of electrons*. This is the stuff of a field of chemistry called ***electrochemistry***. The practical applications of electrochemistry include batteries, electrical smelting of ores, and electroplating of one metal upon another.

A simple example should help to give you the basic idea. Suppose you mixed up an aqueous solution that contains ferric ions, Fe3+, and another solution that contains stannous ions, Sn2+. (Iron can exist in solution as two kinds of ions, with different electric charges: *ferric* ions, Fe3+, and *ferrous* ions, Fe2+; tin can exist in solution as two kinds of ions also: *stannic* ions, Sn4+, and *stannous* ions, Sn2+). If you then mix the two solutions together, the Fe3+ ions take electrons away from the Sn2+ ions, and in the process they become Fe2+ ions and make the Sn2+ ions into Sn4+ ions. (Remember that, by convention, electrons have a unit negative electric charge.)

The Fe3+ ions have stolen electrons from the Sn2+ ions. Why? A "deep" answer would require a lot of explanation, but an adequate ("shallow") answer here might be that the details of the electron structure of the two kinds of ions makes that "want" to happen, and so it does, spontaneously. The chemical reaction can be written, simply, as

$$2Fe3+ + Sn2+ \rightarrow 2Fe2+ + Sn4+$$

The reaction proceeds until all of one or the other reacting ions gets used up.

If you made up two separate aqueous solutions, one containing Fe3+ ions and the other containing Sn2+ ions, and then connected them with a conducting wire, electrons would flow from the one with the Sn2+ ions to the one containing the Fe3+ ions, to make the reaction happen, even though the two kinds of ions are not in contact! You could then measure the voltage on the wire, at the start of the process. The voltage is a measure of what is called the ***electrical potential*** between the two solutions. Think of voltage as an "electrical pressure", analogous to the water pressure that drives the flow of water in your home water pipes.

It would be nice to be able to measure, separately, the electrical potential associated with the removal of electrons from the Sn2+ ions and the addition of electrons to the Fe3+ ions. You can think of those two processes as "half- reactions", which add together to produce the whole reaction, written above. You could then use such results to figure out, directly, which substances steal electrons from which other substances.

Unfortunately, there's no way of doing that: all you can do is *measure the electrical potential associated with the whole reaction*. Chemists have gotten around that problem in a neat and convenient way: they characterize reactions of the kind described above that involve a wide variety of different substances *with reference to a solution that contains molecular hydrogen* (which could be bubbled, as a gas, through an aqueous solution). To make that concrete, you can imagine putting either the Fe3+ ions or the Sn2+ ions into a solution into which molecular hydrogen is added. The electrical potential of the reaction that takes place (called the ***redox potential***, denoted by Eh) can then be measured. Then they tabulate all of the half-reactions relative to the hydrogen standard.

Substances, like oxygen, that tend to steal electrons away from other substances are called ***oxidizing agents***; substances, like metallic iron, that tend to lose electrons to other substances are called ***reducing agents***. The word "oxidizing" here, however, is misleading, because a great many other substances can act as oxidizing agents, in the sense that they tend to steal electrons away from other substances. The rationale, I suppose, is that oxygen is the most important oxidizing agent in so many natural chemical systems. Likewise, there are many other reducing agents besides metallic iron, but because iron and its compounds are so important in natural chemical systems, it's natural to think in terms of metallic iron, or ferrous iron ions (Fe2+), as reducing agents.

Finally, here are some comments on oxygen vs. the two kinds of iron ions. In a solution containing oxygen, ferrous (Fe2+) ions are inevitably oxidized to ferric (Fe3+) ions, which have a strong tendency to combine with oxygen to form insoluble ferric iron oxides or hydroxides. If, on the other hand, the solution has no free oxygen (perhaps all of the available oxygen had already been used up in the oxidation of organic matter, which is a reducing agent), then the ferrous ions (Fe2+) remain happily in the solution.

1.4: The Nature of Water

1.4.1 Introduction

Human beings in general, and often even scientists in particular, are inclined to take water for granted: it’s almost literally everywhere, and we can’t live without it. It’s difficult to cite a process—physical chemical, or biological— operating on or near the Earth’s surface that doesn’t involve water in a substantial if not essential way. Because water seems so ordinary, we tend to judge the behavior of so many other substances with reference to it. In reality, however, *water is a highly unusual compound*, greatly different in many (but by no means all) of its properties and aspects of behavior than what one would expect in view of its apparently simple composition and low molecular weight. Just one example should suffice here: the *expansion when solid water thaws to liquid water*. This is *close to being unique*, and it has profound implications for processes and life on the Earth's surface.

This section attempts a survey of the nature and properties of water. Its goal is to show you the most important facts about the nature and relationships among water in its solid, liquid, and gaseous states, and how these can be explained on the molecular level. The latter endeavor will necessarily be somewhat superficial, because a fundamental approach to the behavior of water is heavily quantum-mechanical, and also because there is still much that’s not well understood.

1.4.2 The Water Molecule

The unusual properties of water are a direct consequence of the structure of the H2O molecule, so I’ll give you some facts about its *geometry*, *structure*, and *electrical state* before looking at the behavior and properties of water.

It is well established that the water molecule consists of two atoms of hydrogen and one atom of oxygen arranged as an isosceles triangle. The H–O–H angle is about 104.5°, and the O–H bond length is very close to 0.1 nm (nanometers; 10-9 meters). Assuming the effective atomic radius of the oxygen to be 0.14 nm and that of the hydrogens to be 0.12 nm, one way of schematizing the water molecule is shown in Figure 1-24. It’s known that this basic geometry is little affected by the changing environment of the molecule in the gas, liquid, or solid form over a wide range of temperatures and pressures.

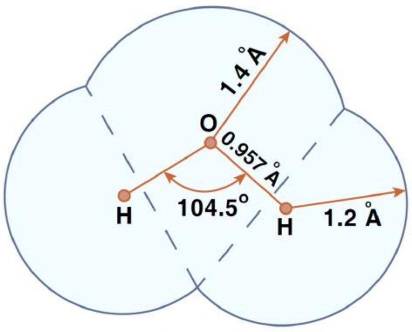


Figure by MIT OCW.

Figure 1-24. Cartoon of the water molecule, showing O–H distance, H–O–H angle, and effective radii of the hydrogen and the oxygen.

Keep in mind, however, that the picture in Figure 1-24 is something of a fake, because what really governs what the molecule looks like to the outside world is the electron distribution around the three nuclei, because that’s what determines how close you can get to the molecule from any given direction before enormous repulsive forces build up as you try to penetrate the electron field.

The three nuclei of the water molecule are always undergoing vibrations, of three kinds (Figure 1-25): the two hydrogens oscillate back and forth along the oxygen–hydrogen directions in phase (this is called the ***symmetric stretching vibration***); the two hydrogens oscillate back and forth along the oxygen–hydrogen directions a half cycle out of phase (this is called the ***asymmetric stretching vibration***); and the two hydrogens move in phase in opposite directions along circular arcs lying in the plane of the molecule (this is called the ***bending vibration***). These are the fundamental or normal modes of the vibration; the actual vibration is a linear superposition of these (in other words, they all add up together, without any one affecting either of the others). The periods of the oscillations are extremely small—of the order of 10-14 s!

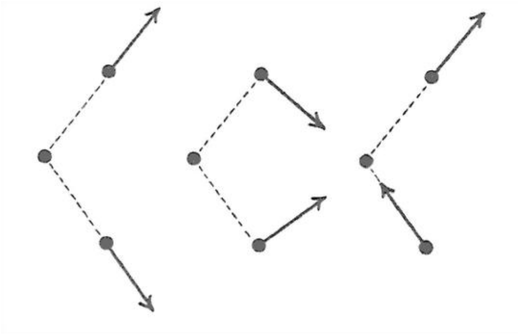


Figure 1-25. Modes of vibration of the water molecule.

The distribution of the electrons, in space, in and around the water molecule—the most important feature of the molecule in terms of bonding—is not what you might guess. Because oxygen is near the right side of the periodic table and hydrogen is at the left side, it should seem reasonable that the electrons tend to be more in the vicinity of the oxygen than in the vicinity of the hydrogens. The molecule is thus what is called a ***polar molecule***: the oxygen "end" or "side" has a negative electric charge, and the hydrogen "end" or "side" has a positive electric charge (Figure 1-26).

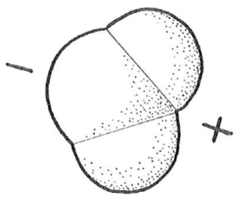


Figure 1-26. Overall distribution of electric charge on the water molecule.

But a more fundamental way of viewing the electron distribution of the water molecule (a way that is beyond the scope of this course) has to do with the quantum-mechanical structure of the electron "orbits", or, more properly, electron ***orbitals***. The six electrons in the outermost electron shell of the oxygen (those are the ones involved in bonding with other atoms) turn out to be in *equivalent* orbitals, and the way you can arrange four directions in space in the most equivalent or non-preferential way is *tetrahedrally*—that is, starting from the center of a tetrahedron, in the directions toward the four vertices of the tetrahedron (Figure 1-27). Two of these tetrahedral positions are associated with the bonding with the two hydrogens, and the other two are what I'll call very unofficially here the *open positions*. Figure 1-28 is a crude sketch of the geometry of the bonding directions.

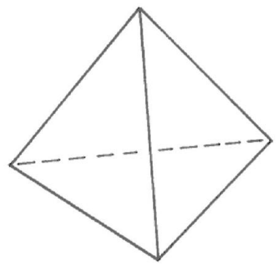


Figure 1-27. A tetrahedron.

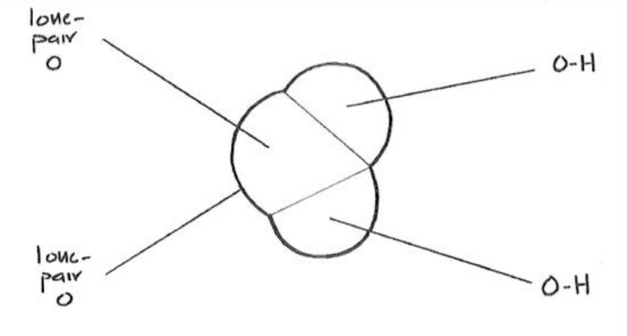


Figure 1-28. Geometry of bonding directions in the water molecule.

1.4.3 Ice

We'll look at the structure of ice before worrying about liquid water, because its structure is well known, thanks to the technique of x-ray diffraction. The ice that you and I are familiar with is known as ***ice I*** (or, more precisely, *ice Ih*). This is the form of ice that's found at room pressures and low temperatures. At higher pressures several other polymorphs of ice make their appearance, but they never occur under the conditions of the earth's surface and atmosphere. (Minerals that have the same composition but different crystal structure are called ***polymorphs*** of one another.)

The most striking and very unusual thing about ice Ih is *the openness of its structure*. This arises from the way water molecules are bonded to one another. Water molecules are bonded by what are called ***hydrogen bonds*** (Figure 1-29): one of the hydrogen atoms of one molecule is bonded electrostatically to the oxygen atom of the other molecule, in such a way that the hydrogen is attached to the oxygen at one of the open positions—of which there are two, remember, in tetrahedral relationship to each other.

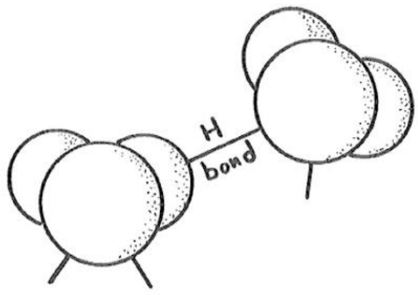


Figure 1-29. The hydrogen bond. (the atoms at the ends of the bond should be touching.)

The hydrogen bond is not nearly as strong as the very strong oxygen– hydrogen bonds within the water molecule itself, but it's strong enough to promote bonding to form a stable crystal structure (and, as you will see soon, liquid water too) at temperatures much higher than we have any right to expect.

In the ice structure, each water molecule is surrounded by, and hydrogen-bonded to, four water molecules, arranged tetrahedrally around that given water molecule. Two of these nearest-neighbor water molecules have one of their hydrogens bonded at the open positions of the oxygen of the given water molecule, and each of the other two nearest-neighbor water molecules have one of the open positions of its oxygen bonded to one of the hydrogens of the given water molecule.

Does that sound too complicated? Read the last paragraph through again, and then look at Figure 1-30, which is a blown-up view of this tetrahedral hydrogen-bonding arrangement, the best I can draw it for you. When looking at Figure 1-30, keep in mind that *all the water molecules should really be touching*.

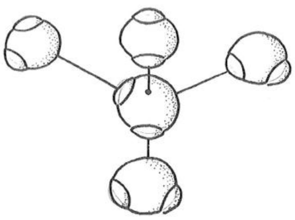


Figure 1-30. Blown-up view of the tetrahedral hydrogen-bonding arrangement in ice.

To build up the ice structure, I want you now to imagine starting with the top four of the five-molecule tetrahedral arrangement in Figure 1-30 and adding water molecules in all horizontal directions, to form a kind of *puckered sheet* (unofficial term). Then stack puckered sheet upon puckered sheet and bond them together by means of the vertically upward-directed and downward-directed fourth bonds extending from each molecule in each puckered sheet. Figure 1-31 is a sketch of just one little segment of two adjacent puckered sheets, viewed from the side.

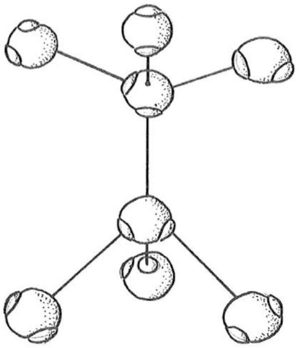


Figure 1-31. Blown-up sketch of one small segment of two adjacent “puckered sheets” in the ice structure.

If you stretch your powers of three-dimensional visualization a little more, you’ll see that *the puckered sheets have hexagonal symmetry when viewed normal to the sheets*. Figure 1-32 shows schematically the arrangement of the water molecules in one of the puckered sheets. The water molecules in the sheet are alternately high and low, and are so labeled *H* and *L* in Figure 1-32.

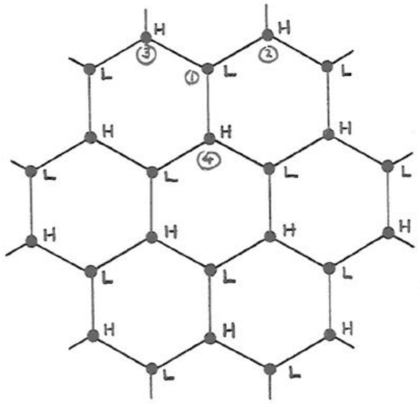


Figure 1-32. Schematic arrangement of water molecules in a “puckered sheet” in the ice structure. The circled numbers correspond to those shown in Figure 1-31.

So the ice structure is one of hexagonal symmetry, with big columns of empty space down the middles of the hexagons. A view from the side (Figure 1- 33) suggests (but does not show too well in the figure) *open tunnels running normal to the hexagonal axis* also. The hexagonal symmetry axis is called the ***c axis***, and the plane parallel to the puckered sheets is called the ***basal plane***.

What is striking about this hexagonally symmetric ice structure is its *openness*. If the hydrogen bonding did not happen, and the water molecules were close packed, each would be surrounded by *twelve* nearest neighbors instead of *four*, and the structure of ice would be much denser than it actually is. This is the fundamental reason why *the solid phase, ice, is less dense than the liquid* *phase*. (That seemingly simple little statement has profound implications for water on the Earth.) Only a handful of other substances, germanium and silicon among them, show this behavior.

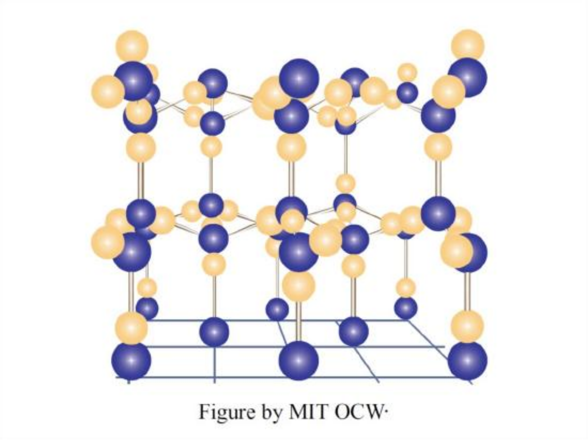


Figure by MIT OCW·

Figure 1-33. Side view of the ice structure.

1.4.4 Liquid Water

I said in an earlier section that the liquid state is complicated and relatively poorly understood. This is true even more so for water, *because of the existence of those unusual hydrogen bonds*. Here I will concentrate on some of the unusual properties of liquid water and make some qualitative comments on how they might be explained.

First off, think about *the density of water as a function of temperature and pressure*. The pressure dependence of the density is not out of the ordinary: water has a certain small compressibility, not out of line with other liquids. But the temperature behavior is astounding (Figure 1-34): at constant pressure *the density of liquid water actually increases with increasing temperature from 0°C to about 4°C and then slowly decreases with increasing temperature above that magic 4°C temperature*.

Keep in mind also that *the density increases in the transition from ice to liquid water upon melting at room pressures*, from about 0.92 g/cm3 to 1.00 g/cm3. When most solids melt to their liquid phase, the structure opens up a little, as the molecular ordering decreases and the molecules become partially free to wander relative to their nearest neighbors. But the opposite is true for water: *the packing of the molecules in liquid water is actually greater than in ice*.

The best way to explain the behavior of water density with temperature is to postulate that the melting of ice involves breakage of some of the hydrogen bonds, so that there is a partial collapse to closer packing. Models of the liquid water state tend to assume that there are transient and local ice-like structural regions with complete hydrogen bonding (called, expressively, “flickering clusters”), and other transient and local regions in which the water molecules are more closely packed. As the temperature of the liquid increases, a greater and greater percentage of the molecules are in a close-packed state, leading to the amazing increase in density from 0°C to 4°C.

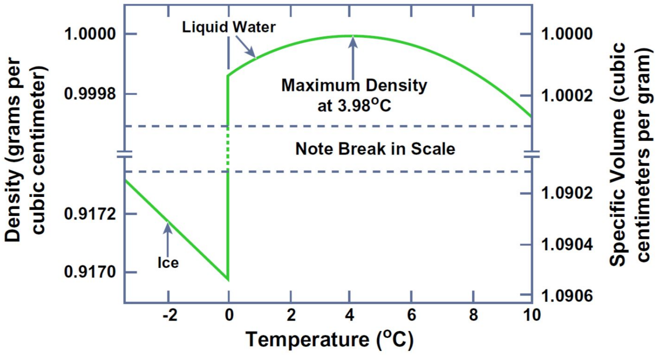


Figure by MIT OCW.

Figure 1-34. Graph of density against temperature for ice and pure liquid water. (From Gross, 1990.)

Counteracting the hydrogen-bond-breaking tendency is the normal tendency for overall increase in the separation distance of all the molecules in the aggregate due to increase in thermal agitation of the molecules as the temperature increases. At temperatures greater than 4°C, this expansion effect more than counteracts the condensation effect due to breaking of hydrogen bonds.

Another outstanding feature of liquid water is its *extremely high heat of vaporization*, about 540 cal/g. This is almost an order of magnitude higher than what might be expected for normal liquids of the same molecular weight. Why so high? Again one can invoke the progressively smaller percentage of hydrogen bonds per unit mass of water. It takes energy to break the hydrogen bonds, and that's what shows up as additional heat of vaporization, in addition to the usual contribution that goes into detaching molecules of the liquid from the liquid–vapor interface, where they are held tightly by the attractive forces of the subjacent molecules.

The specific heat capacity of liquid water is extremely high, for the same reason: as heat is added to liquid water, not only does it go toward increasing the thermal energy of vibrations of the water molecules; it also has to go toward breaking more and more of the remaining hydrogen bonds, which takes energy.

Finally, it's worth pointing out that many important properties of liquid water are not out of the ordinary. Properties that involve molecular transport, like the coefficient of viscosity and the coefficient of diffusion are in line with what would be expected of a normal liquid.

1.4.5 Phase Equilibrium in Water

Now that we have surveyed the structure of solid and liquid water, it's time to look at *the equilibrium among the solid phase, the liquid phase, and the vapor phase*. The best way to do this is by means of a ***phase diagram***, which shows the ranges of pressure *P* and temperature *T* characteristic of each phase and the sets of pressures and temperatures for which two (or even all three) phases can coexist with one another at equilibrium. Figure 1-35 shows such a diagram for water—but only for very low pressures, much less than atmospheric. Figure 1-36 is a simplified "cartoon" version of Figure 1-35.

What are the important features of the phase diagram in Figures 1-35 and 1-36? Note first that way down near the lower left corner there is an unobtrusive ***triple point*** among the vapor, liquid, and solid phases. At that particular set of values of *T* and *P* (at a temperature of about 0.1°C and a pressure of about 4.6 mm Hg, or about 6 x 10-3 atmospheres), *all three phases can coexist at equilibrium*, each with a different value of density.

Another interesting thing about the *P*-*T* diagram in Figures 1-35 and 1- 36 is the narrowing and eventual termination of the vapor–water equilibrium cylinder as the phase boundaries join at what is called the ***critical point***. This means that the distinction between the densities of liquid and vapor, which is great at the familiar Earth-surface temperatures and pressures, diminishes with increasing temperature and increasing pressure and finally vanishes at the triple point! At even greater temperatures and pressures, water exists in a single fluid phase, rather dense, with no distinction between the vapor state and the liquid state. (This is not relevant to this course, but I couldn't resist including it, because it's so fascinating—to science nerds like me, anyway.)

With reference to Figures 1-35 and 1-36, you can see that it's possible to convert liquid water to water vapor continuously, without the catastrophic phase change involved in crossing the boundary directly by the shortest route. You could do this by changing the *P*–*T* conditions by a roundabout route like that shown in Figure 1-37. Along that route, all the properties of the water vary continuously from the liquid phase to the vapor phase.

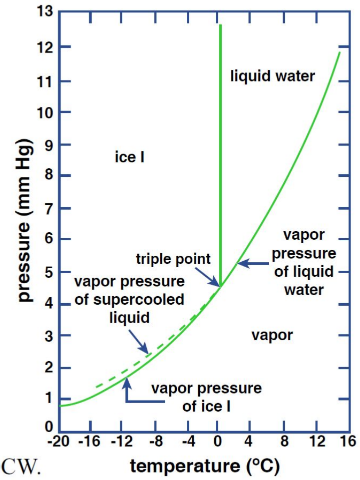


Figure by MIT OCW.

Figure 1-35. P–T diagram for water in the region of the triple point.

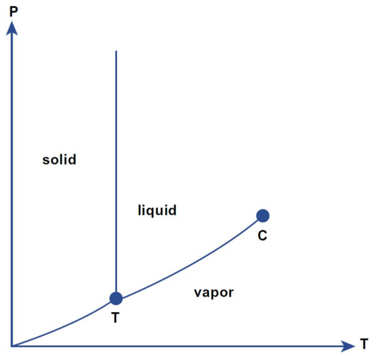


Figure by MIT OCW.

Figure 1-36. Cartoon *P*–*T* diagram with axes distorted to show qualitatively both the triple point and the critical point clearly. (Not to scale!)

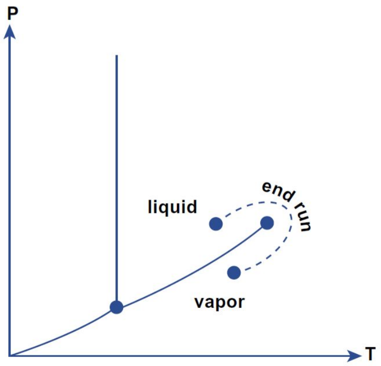


Figure by MIT OCW.

Figure 1-37. How you can pass from the liquid phase to the vapor phase, or vice versa, by continuous change in temperature and pressure around the critical point. (Not to scale!)

When you boil water, you are converting liquid water into water vapor. That process can be shown very simply on the *P*–*T* diagram for water (Figure 1- 38). You start with room-temperature water at atmospheric pressure (Point A in Figure 1-38). As you heat the water, it follows a path in the *P*–*T* diagram horizontally to the right, until it reaches the liquid–vapor phase boundary, whereupon it boils. The water then stays at the boiling temperature until all of the liquid water is used up.

You can also account nicely for the formation of dew and frost by use of the *P*–*T* diagram. In that case, however, you have to think more carefully about what to use for the pressure. The atmosphere consists of several gases, which together account for the atmospheric pressure, by virtue of their weight in the column of air that overlies the given locality. The contributions of each gas to the total pressure is called its ***partial pressure***.

Water vapor in the atmosphere ranges in concentration from considerably less than one percent, in the coldest and driest air, to more than four percent, in the warmest and most humid air. In the *P*–*T* diagram in Figure 1-39, start with a point (Point A) well to the right of the liquid–vapor phase boundary, at a temperature above freezing. On a clear night, as the ground surface cools, by loss of heat by radiation to outer space, the water vapor takes a path horizontally to the left in the *P*–*T* diagram. When it reaches the liquid–vapor phase boundary it condenses as dew on the cold surfaces.

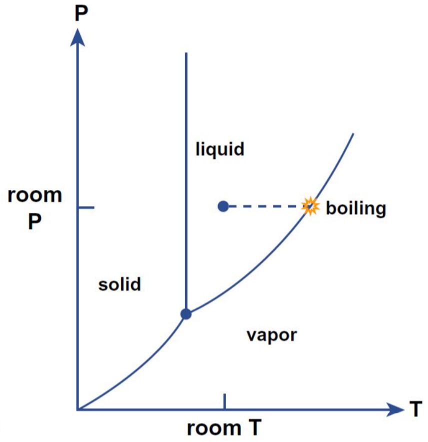


Figure by MIT OCW.

Figure 1-38. The boiling of water, as expressed in the *P*–*T* phase diagram for water.

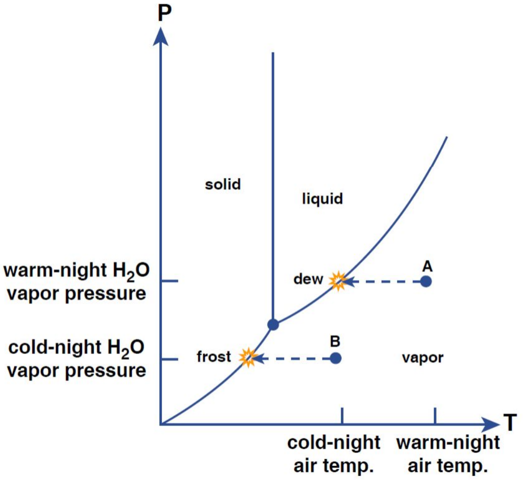


Figure by MIT OCW.

Figure 1-39. The formation of dew and frost, as expressed in the *P*–*T* phase diagram for water.

You can do the same thing for the formation of frost, by starting at a temperature below freezing. Then the path leads horizontally left, from, say, Point B in Figure 1-39, to reach the ice–vapor boundary, at which point little ice crystals grow on the cold surfaces, fed directly from the water vapor in the air. (You can conclude from this that frost is not just frozen dew—although dew itself can freeze, as the air temperature decreases during the night.)

1.4.6 The Chemistry of Water

The water molecules in liquid water have a slight tendency to dissociate into H+ ions and OH- ions. The H+ ions are called (naturally!) ***hydrogen ions***. They are basically just protons. (Actually, they tend to be hydrated by attachment to a water molecule, giving the formula H3O+, but conventionally they are written as, and treated as, just H+ ions.) The OH- ions are called ***hydroxyl ions***.

This dissociation of water molecules into hydrogen ions and hydroxyl ions is expressed by the following simple chemical reaction, with which you are probably all familiar:

$H2O \Leftrightarrow H+ +OH-$

Keep in mind that, as with all such chemical reactions, water molecules are all the time dissociating into hydrogen ions and hydroxyl ions, while at the same time hydrogen ions and hydroxyl ions are recombining to form water molecules.

The balance between dissociation and recombination is such that only a very small proportion of the water molecules, at any given time, are dissociated.

In pure water that is contained in a completely insoluble container and is not in contact with any gas, proportion of water molecules that are dissociated at any given time is one in 107—a very small proportion. Moreover, the numbers of hydrogen ions and hydroxyl ions are equal, simply because the dissociation of one water molecule produces one H+ ion and one OH- ion. The concentrations of both the H+ ions and the OH- ions is said to be 10-7. (But beware: I am being somewhat superficial here because I am not telling you the units of the concentrations. That would take a lot more explanation than this course permits.)

If, now, we allow some H+ ions from outside to be added to the water, somehow, there will be a corresponding and one-for-one decrease in the numbers of OH- ions. Conversely, if we allow some OH- ions to be added to the water, somehow, there will be a corresponding and one-for-one decrease in the numbers of H+ ions. That leads to the idea of *a scale of concentration of H+ ions*. That scale is called the ***pH scale***.

All of you probably have heard of the pH of aqueous solutions and know that it is somehow connected with the acidity of the water. The pH of an aqueous solution is defined as *the negative of the logarithm to base ten of the concentration of hydrogen ions in the aqueous solution*. (For those of you who are unfamiliar with the concept of logarithms, all you need to keep in mind that the base-ten logarithm of the number 107 is 7 and that the base-ten logarithm of the number 10-7 is -7). So the pH of pure water, in which, as discussed above, the hydrogen-ion concentration is 10-7, is 7. Aqueous solutions with greater concentrations of hydrogen ions have smaller pH values and are said to be *acidic*; aqueous solutions with lesser concentrations of hydrogen ions have larger pH values and are said to be *alkaline*.

1.5: Some Essential of the Chemistry of the Surface Zone of the Earth

1.5.1 Introduction

I think that a good case can be made that the three outstanding characteristics of the chemical environment of the near-surface zone of the Earth are that

(1) it is pervaded with water;
(2) it is largely oxygenated, and
(3) it involves the chemical element carbon in various important ways.

1.5.2 The Role of Water

Almost all of the chemical reactions that are important in the surface zone of the Earth take place in an aqueous medium. Water is the great reaction medium for near-surface chemical processes on the Earth. (The important exceptions are certain gaseous reactions in the Earth's lower atmosphere, which will not concern us in this course. That brings in a whole other field of science: atmospheric chemistry.)

Water is often described as the universal solvent. More specifically, water dissolves far more substances than any other liquid, largely because the hydrogen ions and hydroxyl ions, as well as the polar water molecules themselves, can insinuate themselves into the ionic crystals of which most of the solid Earth is composed.

Chemical reactions of the kind that are important in the Earth's near- surface solid materials just don't happen in a waterless medium. All you need to do to convince yourself of that is to compare the rates of weathering processes on Earth (there will be a long section on that in the following chapter) with weathering processes on a dry planetary body like the Moon. All of the important reactions by which the primary minerals of igneous and metamorphic rocks are broken down into other minerals, which are sable under Earth-surface conditions, along with dissolved materials, are hosted by water.

1.5.3 The Role of Oxygen

The geologic evidence is good that the Earth's atmosphere has contained abundant oxygen for about the latter half of geologic time. At present, the Earth's atmosphere consists of about twenty percent oxygen, by volume. You know, of course, that oxygen is essential to animal life on Earth. What is probably less obvious to you is that the presence of that oxygen means that chemical compounds in the reduced state which become exposed to the Earth-surface conditions are very susceptible to oxidation. This is especially true of rock- forming minerals that contain ions ferrous iron (Fe2+). When such rocks undergo weathering at the Earth's sauce, about which you will learn in the next chapter, the ferrous iron is oxidized to ferric iron (Fe3+).

Also, organic material (the tissues of plants and animals), which consists of a great variety of chemical compounds made up largely of carbon and hydrogen, become oxidized back to carbon dioxide and water when they are exposed to the oxygen-rich waters of the Earth's surface. Only in places where ferrous-iron-bearing material and organic material are sealed off from the oxidizing environment can these materials survive in the reduced state. This happens, for example, in places like lake bottoms or the sea floor where the materials are rapidly buried under fine-grained sediments (muds), which are largely impermeable to the passage of surface waters. Under such conditions, all of the oxygen is consumed during oxidation, leaving variable amounts of the reduced materials intact. With regard to organic materials, the consequences for human society are enormous: that's the origin of all of the fossil fuels we burn in today's world!

1.5.4 The Role of Carbon

This is a good place for some comments on the important role of carbon dioxide in the Earth's surface environment. Carbon dioxide, a gas, CO2, is a minor—but extremely significant—constituent of the Earth's atmosphere. The average concentration of CO2 in the modern but pre-industrial atmosphere is (was!) about 280 parts per million (ppm), or about three percent, by volume. As a consequence of our burning of fossil fuels, it's now up around 370 ppm, and climbing rapidly.

Carbon dioxide has an essential role in the Earth's atmosphere: after water vapor, it is the most important greenhouse gas, as described in an earlier section of this chapter. (And, in contrast to water vapor, which is a dependent variable in the sense that its concentration in the atmosphere is itself a function of variety of atmospheric processes, the concentration of carbon dioxide is, nowadays at least, an independent variable, in the sense that humankind is acting to raise its concentration independently of atmospheric processes.)

Carbon dioxide, like the other atmospheric gases, dissolves in water. For a given concentration of carbon dioxide in the atmosphere, there is a corresponding equilibrium concentration of dissolved carbon dioxide in waters exposed to the atmosphere. Incidentally, that concentration depends on the temperature of the carbon dioxide–water system: the higher the temperature, the lower the equilibrium concentration. It also depends strongly on pressure, as anyone who has opened a bottle of champagne knows well, but that's not a major factor in the behavior of carbon dioxide in the Earth's surface environment because the atmospheric pressure varies over only a small range.

The carbon dioxide dissolved in water reacts with the water to form a weak acid, carbonic acid, H2CO3, according to the reaction

CO2 + H2O ⇔ H2CO3

In turn, the carbonic acid dissociates to form hydrogen ions and bicarbonate ions, according to the reaction

HCO3 ⇔ H+ + HCO3-

The consequence is that the Earth's natural waters in contact with the atmosphere are mildly acidic, with a pH between 5 and 6. That natural acidity plays a major role in the chemical weathering of rock, as you will see in the following chapter.

1.6: The Flow of Water

1.6.1 Introduction

Much of what happens by way of natural processes on the earth's surface involves the flow of water or air. Just think about it: the obvious examples are streams and rivers, ocean currents, and the wind. Less obvious, but extremely important, is the flow of groundwater beneath the Earth's surface. The topic of fluid flow is of central importance in this course, and I would be remiss in not subjecting you to some basic ideas in the dynamics of fluid flow. Of course, whole courses can be devoted to fluid dynamics, and a genuinely deep understanding necessitates much more math and physics than is appropriate for this course. I know, however, from my own experience in teaching fluid dynamics that most of the really important concepts can be made clear without heavy mathematics.

In this section we'll concentrate in a qualitative way on some of the most important physical effects, those you need to be aware of in dealing with any situation in which a gas or liquid is flowing past a solid boundary. This section will be cursory but not overly superficial.

1.6.2 Variables

Two of the most important physical properties of fluids are density and viscosity. The ***density*** of a fluid, denoted by , is *the mass of fluid per unit volume of fluid*. The density of water is very close to one gram per cubic centimeter (1 g/cm3) or one thousand kilograms per cubic meter (103 kg/m3). The density of air at sea level and room temperature is far smaller, by a factor of about 800.

The ***viscosity*** of a fluid, usually denoted by μ, is *a measure of the resistance of the fluid to deformation*. Here's a great kitchen experiment to introduce you to the concept of viscosity. Obtain two big sheets of plate glass, about the size of home window sashes. You will have to attach a couple of handles to the top surface of one of them, perhaps with suction cups of some kind. Lay one of them flat on your kitchen table or counter and slather it with a "thick" (read "viscous") liquid, like honey, molasses, corn syrup, or motor oil (your choice). Right away, before the liquid starts to ooze off the glass, place the other piece of plate glass, the one with the handles, on top of the surface of the liquid (Figure 1-40). Be sure there are no trapped air bubbles.

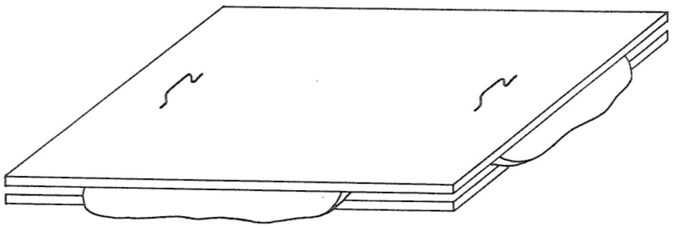


Figure 1-40. Shearing a viscous fluid between two parallel plates

Now grip the handles and shift the upper plate horizontally at constant speed parallel to the lower plate, being careful to maintain constant spacing between the two plates. You will easily sense that the greater the relative speed of the plates, the greater the force needed to maintain the shearing of the fluid. The viscosity is just the coefficient associated with the ratio between the rate of shearing and the force (measured per unit area in the planes of shear) that resists the shearing:

$$\text{viscosity} = \frac{\text{resisting force per unit area}}{\text{rate of shearing}}$$

Here, the rate of shearing is measured by the speed of the upper plate relative to the stationary lower plate, divided by the spacing between the plates (Figure 1-41). For most common fluids, air and water included, this ratio is the same whatever the rate of shearing; such fluids are called ***Newtonian fluids***.

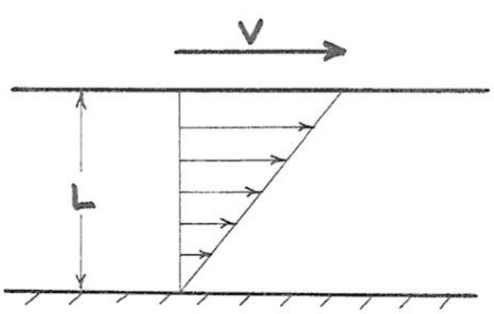


Figure 1-41. Cross-section view through the kitchen-table-top experiment on shearing of a viscous fluid, showing the spacing *L* of the plates and the speed *V* of the upper plate relative to the lower plate. Also shown is the distribution of local fluid speed you would measure along a line between the two plates.

One more fluid property, the ***specific weight*** (in other words, weight per unit volume), usually denoted by , is important when the fluid is in a gravitational field, as on the earth's surface. The specific weight is *the weight of the fluid per unit volume of fluid*. It’s not the same as the density of the fluid, which is the mass per unit volume. For us earthbound beings, the ratio of the density to the specific weight is almost exactly constant, because the two are related by virtue of the acceleration due to gravity, *g*: = *g*. You can “uncouple” the density and the specific weight only by taking a trip in a space ship to another planet!

Now I want you to concentrate on what the fluid dynamicists call an ***open-channel flow***. As the term implies, it’s just *a flow in some kind of channel that’sopentotheairabove,ratherthanbeingenclosedwithinapipeorconduit*as in your home plumbing system. To study open-channel flow you can play in the gutter after a brief heavy rainstorm, or park yourself next to a stream or river, or nail a channel together in your back yard and supply it by means of a hose. Such examples of channel flows, although rather different in scale and geometry, share most of their essential features.

Incidental but important note: It turns out that these open-channel flows differ from their pipe-flow cousins only in the presence of the so-called “free surface”, not in their essential behavior. Most of the flows we will touch upon in this course are open-channel flows, but when we deal with groundwater flow, you’ll see that pipe flow is the closer analogy.

There’s a wealth of variables you can use to describe the flow. Some are more concrete and easy to understand than others. On the one hand, you can think about the velocity of the flow, and on the other hand, you can think about the force per unit area the flow exerts on its boundary.

With regard to velocity, if you occupied any point in the flow that’s fixed relative to the boundaries and used a magic little velocity meter to measure the local fluid velocity, you would get some stable local ***time-average velocity*** if you continued your measurement for a fairly long time. This time-average local velocity varies continuously from point to point all over the cross section of the flow (Figure 1-42); more on what that distribution of local velocity looks like presently.

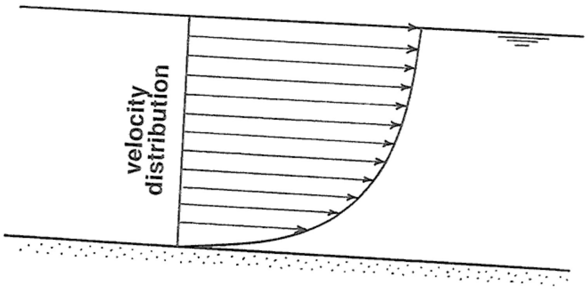


Figure 1-42. Time-average local velocity varies continuously from point to point all over the cross section of the flow

You can take the average of all the local velocities all over the flow cross section to get a ***mean flow velocity*** *U* (Figure 1-43). In real rivers this is how it’s actually done, but in laboratory channels, where it's easy to know the water discharge (the volume rate of flow), you can use the relationship $Q = UA$, where Q is the discharge and A is the area of the flow cross section, to find U.

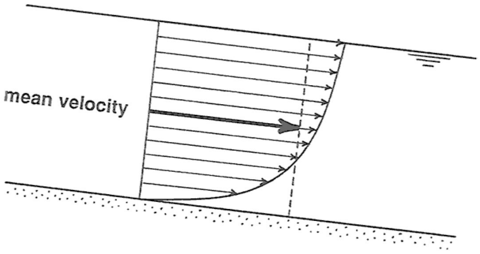


Figure 1-43. Averaging all the local velocities all over the flow cross section to obtain the cross-section mean flow velocity

You could also measure the surface velocity *Us* along the centerline

of the flow (Figure 1-44). This is greater than the mean velocity, by a factor of something like 1.2 or 1.3; it's related to *U* in a rather complex way that needn't concern us here. Incidentally, unless the channel is very wide compared with the depth of flow the position of maximum velocity on the cross section lies a little below the surface.

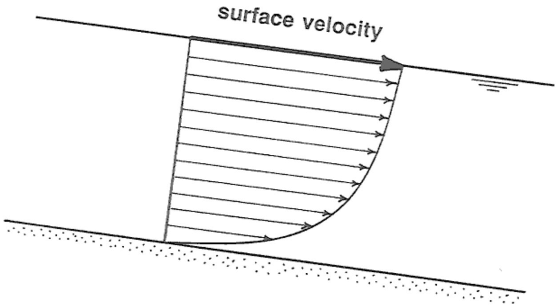


Figure 1-44. The surface velocity, shown along with the velocity distribution in the interior of the flow.

The average force per unit area the flow exerts on its boundary, called the boundary shear stress, denoted by τo (read "tau sub zero" or "tau

naught") is a little trickier to deal with than velocity, and is less concrete to visualize, but it's important for the obvious reason that it's what exerts forces on the sediment particles on the bed of a stream or river or on the surface of a sand dune under the wind.

Let me backtrack a little to point out something about the flow near the boundary that will probably seem very counterintuitive to you, but stands as a firmly established fact of observation: *fluid in direct contact with a solid boundary has exactly the same velocity as that solid boundary*. That's called the ***no-slip condition***. In the context of our open-channel flow, it means that the water right at the bed has zero velocity! But of course there is an increase in velocity upward from the bed, so all the water above the bed, no matter how small a distance above, generally has a nonzero velocity.

The reason I have brought in the no-slip condition here is this: the origin of shear stresses in fluids is just the existence of shearing itself. If that doesn't make sense to you at first thought, go back to your sticky, messy kitchen- table-top experiment on shearing of a viscous fluid. The stronger the shearing, the greater the shear stress that's generated, other things being equal. And foremost among those other things is the viscosity, introduced earlier.

So, back to the boundary shear stress τo. It's true that if you could get right down close to any point on the channel bottom, you could (with the right instrument, which doesn't actually exist!) measure a local shearing force per unit area the flow exerts on the solid boundary. That fits the term "boundary shear stress" well, but it's not what's conventionally meant by the boundary shear stress. What you have to do is take a spatial average of the local shear stress (that is, the average of a large number of "point values" of the local shear stress) over an area large enough to include a large number of representative "roughness elements" (sand grains, or boulders, or even larger protuberances, whatever "non-planarities" happen to be present on the channel bottom).

1.6.3 More on Forces Exerted by Moving Fluids

So far you have learned that a fluid moving past a solid body exerts frictional ("viscous") forces on the surface of the body. These are friction forces in the true sense of the term, even though the detailed nature of the force is not quite the same as the friction between a sliding brick and surface upon which it's sliding. Aerodynamicists have developed a very expressive term for such frictional forces: ***skin friction*** (by reference to the “skin” of an airplane wing). But there’s another important kind of force exerted by moving fluids: ***pressure forces***. These are forces exerted by the flow on what might be called blunt bodies or bluff bodies: those that present a “frontal surface” to the oncoming flow (Figure 1-45).

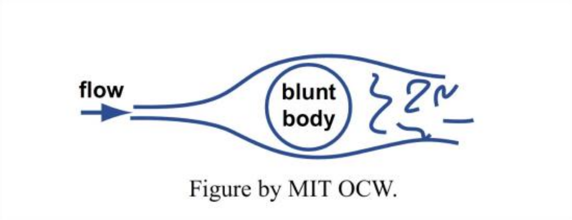


Figure 1-45. Flow around a blunt body.

It will probably seem obvious to you that the fluid pressure (by ***pressure*** I mean *force per unit area directed normal to a surface*) on the upstream side of a blunt body is greater on the upstream side than the downstream side. What is much less obvious, but nonetheless true, is that, in absolute terms, the pressure on the front is greater than the “free-stream pressure” (the pressure in the fluid at some point far away from the body) but the pressure on the back is less than the free-stream pressure. It would take a lot more background work than we have time for to satisfy you that that’s true; you will have to take my word for it.

If you are inclined to doubt the foregoing statement, here’s a home experiment which (in contrast to many of my proposed home experiments) would actually be easy to do. All you need is a barometer, of the kind they now peddle as part of “home weather stations”. You would need a windy day for this experiment, and a single-family home. Close all of the doors and windows tightly, so there’s no throughgoing wind in the house. Stand next to the upwind side of your house and read the barometer. Now go around to the downwind side of the house and take a reading there too. You would find, of course, that the local atmospheric pressure on the downwind side is noticeably less than on the upwind side. That’s the real origin of the wind force that tends to blow your house down. But now walk far away from your house and take a barometer reading in an open and unobstructed area (you are now in that “free stream” I mentioned earlier). That reading will lie between the first two—showing that the local atmospheric pressure on the downwind side of the house is indeed lower than the free-stream pressure!

The pressure difference, fore and aft, means that the fluid pressure is exerting a net force in the downstream or downwind direction on the bluff body; it’s just a matter of adding a larger downstream-directed force (call that the positive direction) and a smaller upstream-directed force (call that the negative direction); the result is a positive downstream force. This kind of force is very expressively called ***pressure drag*** or ***form drag***.

Except for very small-scale flows and very slow velocities, *the pressure drag is likely to be much larger than the skin friction*. Does that make sense to you? Probably not, because I don’t think it’s “intuitively obvious”. If not, then here’s still another experiment, another one that would be easy in real life, that you can do for yourself to demonstrate the truth of that proposition. It’s probably the simplest experiment I’ll propose all semester: all you would need is a pizza pan and a swimming pool! Stand in water up to your chest, and move the pan through the water in the direction parallel to itself. It slips easily through the water, right? Now turn it so that the surface of the pan is normal to the direction of movement. It takes a far greater force on your part to move the pan through the water in that orientation than before. That’s because the front-to-back pressure difference adds up to a much greater force than just the frictional forces on the two surfaces of the pan when it’s moved parallel to itself (Figure 1-46).

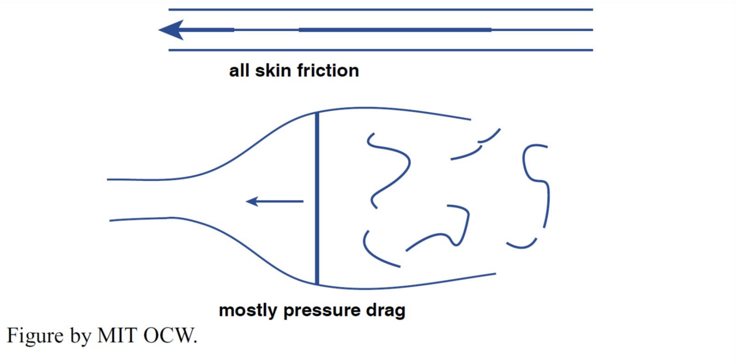


Figure 1-46. Skin friction and pressure drag on a pizza pan moved through still water.

1.6.4 Why Do Fluids Move?

The movement of fluids is part of the experience of our daily lives. But it's valuable to think for a moment about why fluids move. According to Newton's first law of motion, a body moves in a state of uniform motion (changing neither speed nor direction) unless it is acted upon by some force. Moving fluids are always acted upon by friction, so they come to rest unless some other force is available to offset the friction and keep them in motion.

Two important kinds of forces drive fluid motions:

- **The force of gravity.** In a flow down a sloping channel, the weight of an element of fluid has a downslope component (Figure 1-47). This downslope component of fluid weight counterbalances the frictional resistance exerted on the fluid by the bottom boundary.

- **Pressure gradients.** In a flow in a horizontal pipe or conduit, the fluid moves only if there is a downstream decrease in fluid pressure. To see why this causes the fluid to move, look at an element of fluid in the pipe (Figure 1-48). If there is a downstream pressure gradient (that is, the pressure decreases in the downstream direction), then the force on the upstream face of the element is greater than the force on the downstream face. The difference in forces on the upstream and downstream faces is a force directed downstream. This downstream force counterbalances the frictional resistance exerted on the fluid by the walls of the pipe.

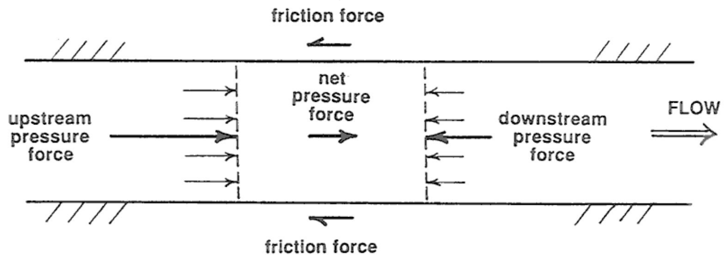


Figure 1-47. In a flow in a horizontal pipe or conduit, the fluid moves only if there is a downstream decrease in fluid pressure.

In a general fluid flow, both of these kinds of forces (downslope gravity components and downstream pressure gradients) can be present at the same time, and the fluid moves under their combined effect.

1.6.5 Channel Flows

Now let's think about what the flow really looks like in a large channel flow like a river. The basic picture is as shown in Figure 1-48: there's a balance between the downchannel pull of gravity on the fluid, which acts throughout the flow, and the resistance force the bed and banks exert on the flowing fluid, which is equal and opposite to the boundary shear stress we discussed above. A natural consequence of the action of these two different forces—gravity acting throughout the fluid and the resistance force acting only at the boundary—is that *the velocity increases upward in the flow*, from zero at the boundary, by the no-slip condition, to a maximum at or near the surface.

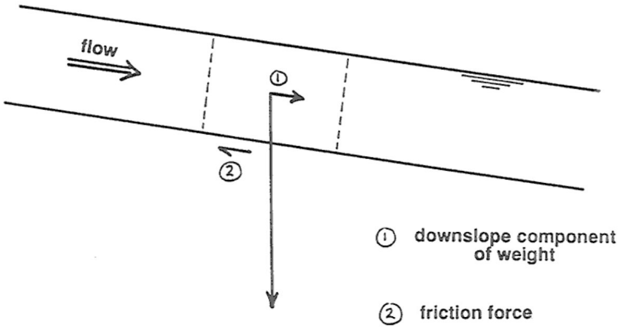


Figure 1-48. In a flow down a sloping channel, the weight of an element of fluid has a downslope component that's balanced by the friction force the boundary exerts on the flow.

That's only the crudest first-order picture, though: the question comes naturally to mind: what are the details of the distribution of time-average local flow velocity over the entire cross section of the flow? That's much too intricate a question for us to pursue here, so I will just hit the high points.

It's simplest to think about a river with a large width-to-depth ratio, which is usually the case. Then we can forget about the effect of the banks. A straightforward application of Newton's laws, which I'll have to skip over here, shows that the distribution of velocity should be parabolic (Figure 1-49A), with the vertex of one limb of a parabola located at the surface. (If you are inclined to pursue this matter further, look me up, and I will set the problem for you to work on.)

But if you went out and measured the local time-average velocity at a series of points along a vertical from the surface down to the bottom, you would find the actual distribution to be far from parabolic: the distribution would be much more uniform over most of the flow depth, and the necessary decrease to zero at the boundary would be in a very thin zone right next to the boundary. The distribution would closely approximate a logarithmic distribution (Figure 1-49B). The reason why the distribution is close to being logarithmic is beyond the scope of this course, but you can readily understand why the qualitative shape of the curve is the way it is, with a much smaller velocity change throughout most of the flow depth but a sharply greater velocity change right near the bottom. To see why, however, we have to do a little more preparatory work; read on.

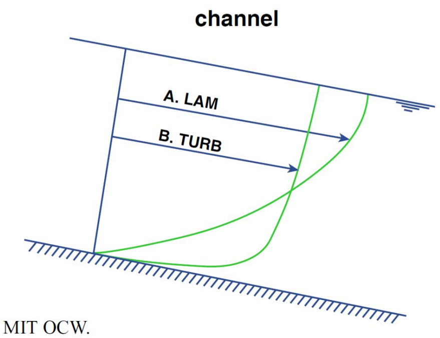


Figure by MIT OCW.

Figure 1-49. Laminar and turbulent velocity profiles in open-channel flow.

This difference in velocity profiles has to do with the difference between laminar flow and turbulent flow, which I am sure you have all heard about one way or another. Channel flows that are very shallow and/or very slow- flowing (like sheet flow or overland flow on the land surface during and after a rain) are characterized by the regular and locally rectilinear flow trajectories of ***laminar flow*** (Figure 1-50A), whereas channel flows that are deeper and/or faster- moving (almost all channelized flows fall into this category) show the irregularly sinuous flow trajectories characteristic of ***turbulent flow*** (Figure 1-50B).

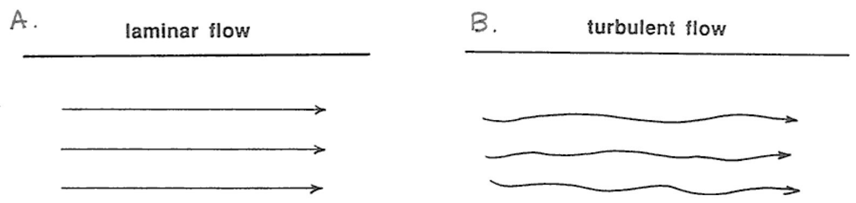


Figure 1-50. Paths of fluid elements in A) laminar flow and B) turbulent flow.

Turbulence isn't easy to describe, but because it's such an important aspect of fluid flows on the earth's surface I'll attempt to give you a qualitative picture of what turbulence is like, with channel flow as the example. One way of studying turbulence is to make a continuous measurement of velocity as a function of time at a fixed point in the flow. You would obtain a record that looks something like the graph shown in Figure 1-51: velocity fluctuations with a range of magnitudes and time scales would be present around the time-average velocity. Another way of seeing the turbulence is to release small neutrally buoyant markers from some fixed point in the flow and watch them move downstream (Figure 1- 52). The trajectories would be irregularly sinuous, with angles of deviation from the downstream direction amounting to no more than several degrees. Each trajectory would be different in detail from all of the others.

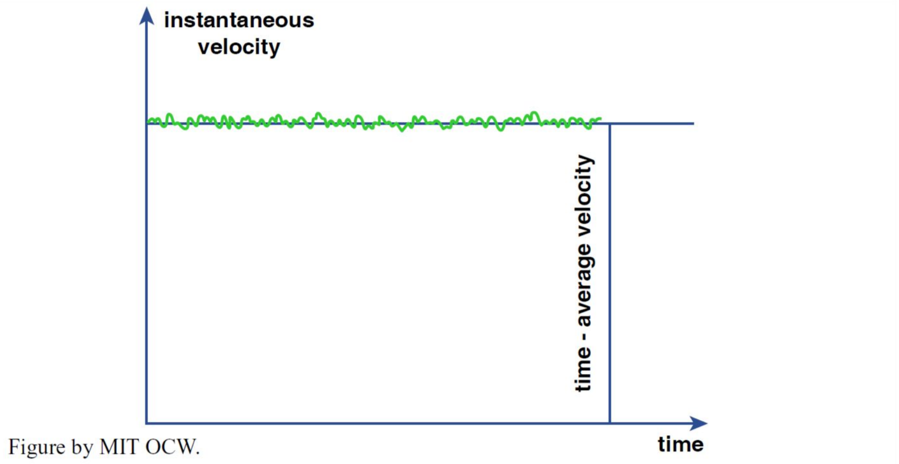


Figure by MIT OCW.

Figure 1-51. Typical record of a continuous measurement of velocity as a function

The best way to see the turbulence, however, is to sprinkle the flow with magic powder that allows you to see the ***turbulent eddies*** (Figure 1-53). You would see rotating swirls of fluid that grade continuously into one another and that change their shape continuously with time. Individual eddies retain their identity for a certain period of time (larger eddies live longer than smaller eddies). but any given eddy eventually becomes unrecognizable and is replaced with newly developed swirls. The flowing surface of a river carrying fine sediment in suspension shows eddies moving up to the water surface to spread out and flatten themselves. You might also watch smoke or steam issuing from a smokestack to see the production and decay of eddies as the plume of hot stack gases rises through the still ambient air.

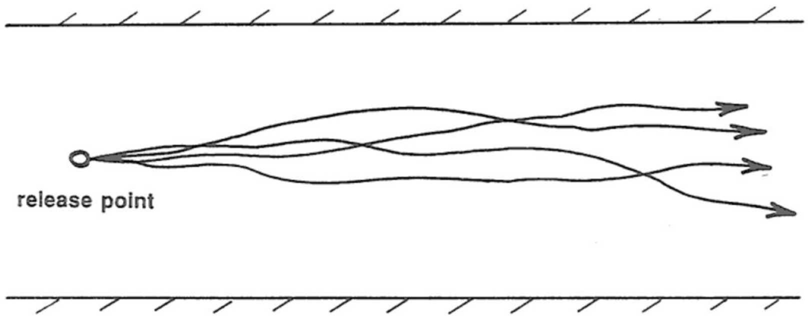


Figure 1-52. Trajectories of small neutrally buoyant marker particles released from some fixed point in a turbulent channel flow.

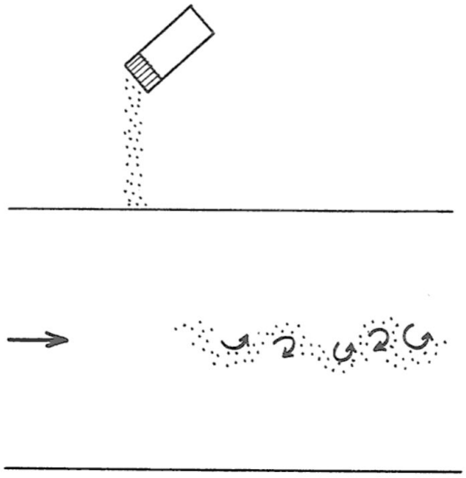


Figure 1-53. Sprinkling the flow with magic powder that allows you to see the eddies.

The maximum vertical scale of eddies in the flow is not much smaller than the depth of the flow, and the maximum cross-flow scale is even larger. There's a continuous range of sizes down to very small eddies, of the order of fractions of a millimeter. The smaller eddies are superimposed on the larger eddies. As the large eddies sweep along the boundary they cause periods of stronger flow and weaker flow. When you're standing on a broad plain or on the deck of a ship at sea, the wind gusts you feel are a manifestation of these large eddies affecting the flow boundary. Have you ever watched a broad field of tall grass or grain in a strong wind from a point high above? You would see a striking pattern of wind gusts moving along in the direction of the wind as they sway the grass.

Returning now to the important qualitative difference between laminar and turbulent velocity distributions, in turbulent flow the velocity distribution is much more uniform over most of the thickness of the flow but changes much more sharply very close to the boundary. It's easy to understand why this is so (Figure 1-54). In turbulent flow, over most of the flow depth the lateral motions of eddies tend to even out the differences in time-average velocity from layer to layer. Here's how this works (keep in mind the overall average upward increase in flow speed): upward-moving eddies tend to arrive at their new surroundings with a speed slower than their new surroundings, so they tend to slow those new surroundings down. Conversely, downward-moving eddies tend to arrive in their new surroundings with a speed faster than their new surroundings, to they tend to speed those new surroundings up.

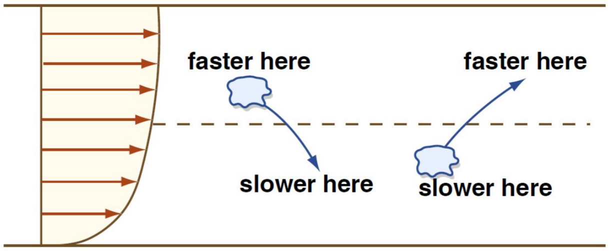


Figure by MIT OCW.

Figure 1-54. Why the velocity profile in turbulent flow is blunter over most of the profile than in laminar flow.

As we move closer and closer to the boundary, however, the fluid motions perpendicular to the boundary are more and more restricted—because remember that right at the boundary the fluid velocity must match that of the boundary, so there can be no component of velocity normal to the boundary there. In a thin layer next to the boundary, therefore, the turbulence can't even out the differences in velocity from layer to layer, so the gradient of velocity is very steep. This layer next to the boundary, where viscous shearing is more important than turbulence, could be called the ***viscosity-dominated layer*** (Figure 1-55). Its thickness is typically a very small percentage of the flow depth, no more than millimeters to at most a few centimeters. The next time you're sitting in a window seat in the jetliner, look out at the surface of the wing and think about that viscosity-dominated layer. The air speed, relative to the surface of the wing, decreases from its free-stream value of several hundred miles per hour to zero over a thickness of centimeters, and the viscosity-dominated layer is a small fraction of a millimeter. Incidentally, the cross-section shape of the wing (the "airfoil") is designed to minimize pressure drag, so in this case the body (the wing) is not really a blunt body at all! The reason the plane stays in the air is that the local air pressure on the upper surface of the wind is less than on the lower surface, but the aerodynamic reason for this would require a lot more work on your part.

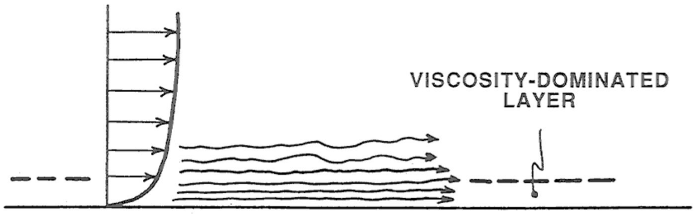


Figure 1-55. The viscosity-dominated layer at the base of a turbulent channel flow.

1.6.6 Oscillatory Flows

Everything in the preceding sections is relevant to unidirectional flow: a current that moves in one direction only, with velocity either unchanging with time (steady flow) or changing somehow with time (unsteady flow). Also important in the transport of sediments is oscillatory flow: a current that reverses its direction periodically with time. The approach in this section will be somewhat different from the approach in the preceding sections: I'll concentrate more on the origin and description of oscillatory flows than on their dynamics.

Except for reversing tidal currents, which technically are oscillatory flows with very long oscillation periods but which are better considered as reversing unidirectional flows, almost all of the important oscillatory flows in nature are caused by surface gravity waves which propagate on the surfaces of oceans and large lakes. Such waves are called ***surface waves*** because they are associated with the water surface, and they're called ***gravity waves*** because the dominant force involved in their propagation (the force which attempts to restore the deformed water surface to a planar condition) is the force of gravity.

The generation of wind waves is a complex process, even now not fully understood, that involves both the shear stress of the moving air on the water surface and on the unequal distribution of air pressure over the windward and leeward sides of geometrical irregularities on the water surface. I'll note here only that the size of the waves depends on three factors: the speed of the wind, the duration of time the wind acts on the water surface, and the downwind distance (called the ***fetch***) over which the wind acts on the water surface.

The best way to obtain an idea of the basic nature of the water motions produced by surface gravity waves is to watch the propagation of a single train of waves in a wave tank (Figure 1-56). You can imagine building a long tank with rectangular cross section, perhaps several meters long, open at the top. Put a wave generator at one end. A board hinged to the bottom of the tank can be rocked or oscillated back and forth to produce a train of waves. It's best to install some porous material at the other end of the tank to absorb the waves, so that they're not reflected to interfere with the propagating wave train.

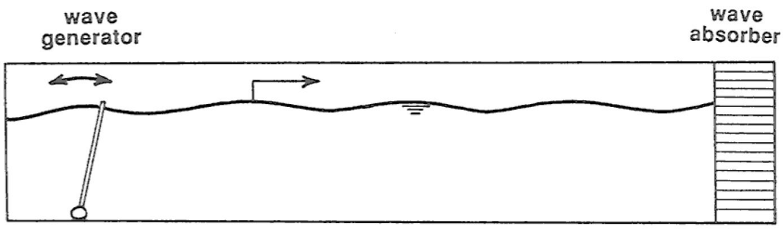


Figure 1-56. Making waves in a backyard wave tank.

Here are some terms used to describe the waves in a simple train of waves (Figure 1-57). The ***crest*** of the wave is the highest point on the wave profile, and the ***trough*** is the lowest point on the wave profile. The ***period*** of the wave is the time needed for one complete wave form to pass a point that is fixed relative to the bottom of the tank. Surface gravity waves in natural environments have periods that range from less than a second to almost twenty seconds. The ***wavelength*** of the waves is the distance from crest to crest or from trough to trough. The ***height*** of the waves is the vertical distance from trough to crest. The heights of large waves generated by major storms are commonly in the range of five to fifteen seconds

If you place a floating cork on the water surface and watch its motion through the wall of the tank as the waves pass by, you see that the cork travels in an almost perfectly circular orbit, making one circle during the passage of each wave (Figure 1-58). The sense of motion is such that the cork moves in the direction of wave propagation while the crest of the wave is passing, and opposite to the direction of wave propagation while the trough of the wave is passing.

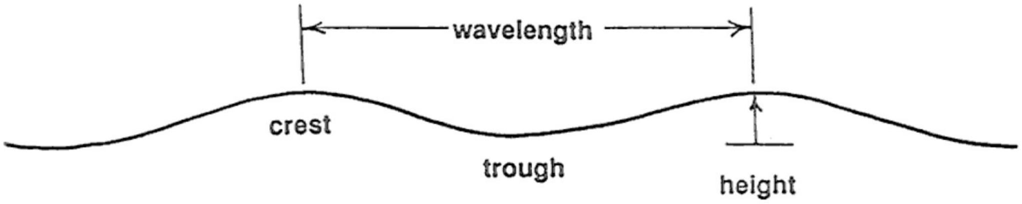


Figure 1-57. Some terms used to describe the waves in a simple train of waves.

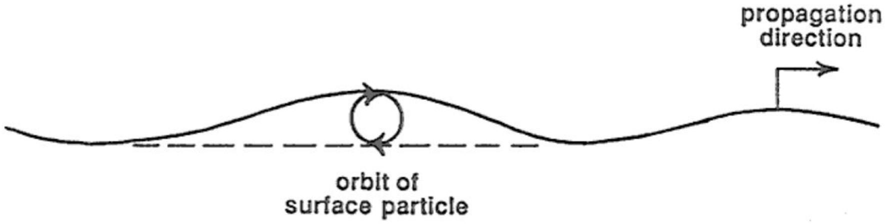


Figure 1-58. The circular orbit of a particle floating on the water surface as a wave passes.

The water below the surface also moves as the waves pass, but the size of the orbits of the water particles decreases very rapidly with depth below the surface (Figure 1-59). The size of the orbits at a depth equal to half a wavelength is only about ten percent of the size of the orbits at the surface, and the size of the orbits at a depth equal to one full wavelength is negligible. Such waves are called ***deep-water waves***. If the wavelength of the waves is much greater than the water depth, however, the size of the orbits is still appreciable at the bottom (Figure 1- 59). Such waves are called ***shallow-water waves***, and they are common during storms. Downward from the surface the orbits become flattened into ellipses with greater and greater eccentricity, until at the very bottom the orbits are back-and- forth straight lines. This should make sense to you, because at the bottom there can be no fluid motion perpendicular to the bottom—if the slight flow into and out of the porous sediment is ignored.

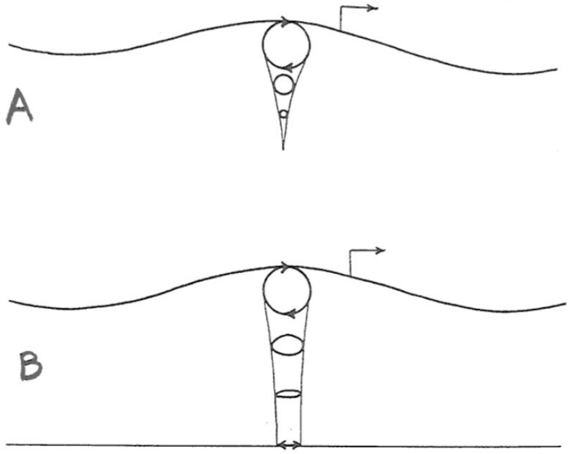


Figure 1-59. A) Decrease in orbital diameter with depth beneath a deep-water wave. B) Decrease in orbital diameter beneath a shallow-water wave.

Now examine more closely the nature of the oscillatory water motion at the bottom. If you measured the fluid velocity near the bottom as a function of time at a point which is fixed relative to the bottom, you would obtain a record that looks like the graph in Figure 1-60: the velocity varies sinusoidally with time. The nature of the motion can be characterized by its period T, or by the maximum velocity U_m that is attained during one oscillation cycle (this happens when the water is at the central point of its oscillation path), or by the distance d_o moved by a water particle during one full oscillation (this is called the *orbital diameter*).

It's part of the physical nature of waves that different trains of surface gravity waves can be superimposed on each other in such a way that the water-surface height and water motion at a given point is the sum of contributions from each superimposed wave train.

You might know from watching storms at sea or on large lakes that the sea state produced by a strong wind is complicated, because it's a superposition of a large number of separate trains of waves with slightly different periods, propagation directions, and heights. Also, when the wind direction changes as a storm passes, the waves adjust their direction of propagation in response. In such cases, the pattern of water movement at the bottom is no longer a simple back-and-forth movement. Even the relatively simple case of two regular trains of waves moving in different directions, the water movement on the bottom can be surprisingly complicated, to say nothing of the really complicated wave conditions produced during storms.

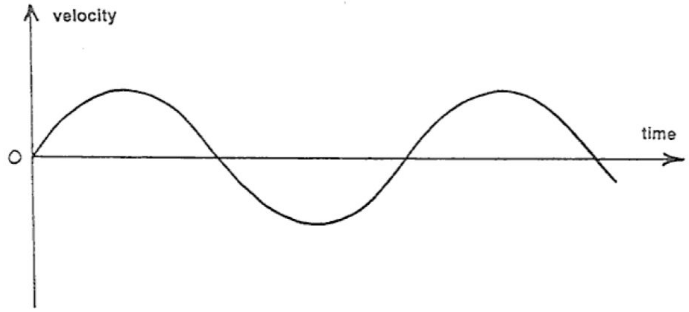


Figure 1-60. Fluid velocity near the bottom as a function of time at a point that's fixed relative to the bottom, as a shallow-water wave passes by overhead.

1.6.7 The Wind

The large-scale structure of the lower part of the atmosphere is obviously much different from that of even the largest river. The most obvious difference is the absence of a free surface in the atmosphere. An equally important difference has to do with the much larger scale of wind systems. This is important because the Earth's rotation plays a dominant role in determining wind direction (but this is not the course in which to pursue that very counterintuitive but fascinating matter) but is a minor aspect of even the largest of rivers. A third important difference is related to the easy compressibility of air, in contrast to the incompressibility of water. This means that the thermodynamics of compressibility plays an essential role in the behavior of the atmosphere.

Despite these far-reaching differences, it's fortunate for us as earthbound creatures that the structure of the wind in the near-surface layer, up to hundreds or even a few thousands of meters above the surface, is usually not greatly different from that of large-scale channel flows, in terms of such things as velocity distribution, turbulence structure, and boundary forces. Most of what I said in earlier sections of this chapter for channel flows of water is applicable to the wind. This is especially true of eolian sand transport and the shaping of sand dunes.

You might complain that the wind is so much more gusty than the ordinarily turbulent flow of rivers. I would counter that by saying that your conception of the gustiness of the wind is shaped by observations made for the most part down in among large and very irregularly shaped and irregularly distributed roughness elements like trees, hills, and buildings. To the extent that things like that occupy river bottoms (hills are certainly there, and we'll have to deal with them as a really important part of this course), and you sensed turbulence down at their level, flow in rivers would seem pretty "gusty" too. (There will be more on that later, when we talk about flow over subaqueous bed forms.) If you have ever experienced a strong wind at sea, you know that the wind unimpeded by large roughness elements is surprisingly steady.

CHAPTER OVERVIEW

2: Introduction and Geology

2.1: The Earth - The Ground Beneath Your Feet

To a great extent in this course, we will be dealing with processes that act on the solid materials at and near the Earth's surface. This chapter might better be called "the ground beneath your feet". This is the place to deal with the nature of the Earth's surface materials, which in later sections of the chapter I will be calling regolith, sediment, and soil.

I purposely did not specify any previous knowledge of geology as a prerequisite for this course, so it is important, here in the first part of this chapter, for me to provide you with some background on Earth materials.

We will be dealing almost exclusively with the Earth's continental surfaces. There are profound geological differences between the continents and the ocean basins, in terms of origin, age, history, and composition. Here I'll present, very briefly, some basic things about geology. (For more depth on such matters you would need to take a course like "The Earth: What It Is, How It Works", given in the Harvard Extension program in the fall semester of 2005– 2006 and likely to be offered again in the not-too-distant future.)

In a gross sense, the Earth is a layered body (Figure 2-1). To a first approximation, it consists of concentric shells: the core, the mantle, and the crust.

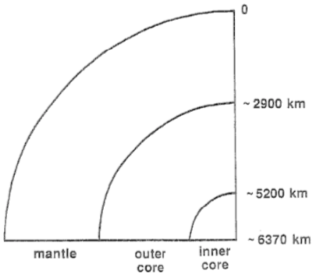


Figure 2-1: Schematic cross-section through the Earth

The core: The core consists mostly of iron, alloyed with a small percentage of certain other chemical elements. The outer part of the core is liquid, and the inner part is solid. There is no need here to deal with how that's known—or indeed why there is an iron core in the first place.

The mantle: The mantle, which constitutes the greater part of the mass of the Earth, has its lower boundary, with the core, at a depth of about 2900 km.
(The radius of the Earth is about 6400 km.) The mantle consists almost entirely of solid rock, except for certain places in the uppermost mantle where magma (= melted rock) is formed at certain times. (There is a deplorably widespread and thoroughly mistaken belief that the upper mantle is everywhere molten.) Saying that the mantle is solid rock, true as that statement is, is somewhat misleading, though, because the solid rock of the mantle can flow plastically in the same way that a very viscous liquid, like honey or corn syrup, flows. You will see the same kind of behavior in glacier ice, in a later chapter. "Silly putty" is another good example of such a material. The uppermost part of the mantle, however, is sufficiently cool that it behaves not as a very stiff liquid but as a rigid solid. This situation, whereby the uppermost rigid part of the mantle (which, together with the overlying crust, is called the ***lithosphere***) rides on the plastically flowing part of the mantle beneath (called the ***asthenosphere***), is the basis for what is called plate tectonics (Figure 1-2).

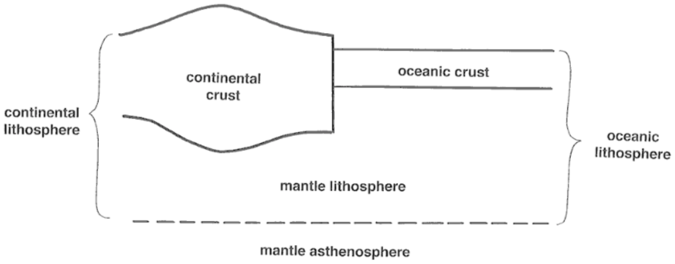


Figure 2-2. Schematic cross-section through the outer part of the Earth

The crust: The uppermost skin of the Earth, above the mantle, is called the crust (Figure 2-2). There are two basic kinds of crust, very different from one another in properties and origin: oceanic crust and continental crust. Oceanic crust is relatively thin, seldom more than seven to eight kilometers. Continental crust, on the other hand, is relatively thick, mostly thirty to fifty kilometers. The features of the upper surface of the continental crust, and the great variety of physical and chemical (and biological) processes and phenomena associated with that surface, are the focus of this course.

You probably know something about, or at least have heard about, plate tectonics. The lithosphere, the outermost part of the Earth, comprising the crust and the uppermost, rigid part of the mantle (Figure 2-2), is in the form of a number of segments, called ***lithospheric plates*** (or just *plates*), which are in contact with one another along what are called plate boundaries. The movements of these plates, and the consequences of those movements, especially in terms of what happens at plate boundaries, are collectively termed ***plate tectonics***.

We could get along fairly well in this course without reference to the core or the mantle, or, for that matter, even to the lower part of the continental crust. Deep subsurface processes, in the crust and in the underlying upper mantle, have a great indirect importance to this course, however: such processes (the nature of which is beyond the scope of this course!) cause *uplift and subsidence of the land surface of the continents*, both sharp and localized, in the case of active mountain belts like the western Cordillera of North America, and broad and gentle, affecting large areas of the stable continent of North America east of the Cordillera. Uplift and subsidence, in turn, are instrumental in erosion, transportation, and deposition of sedimentary materials and in shaping of the surface topography of the continent. We'll return to such matters in the final chapter.

2.2: Minerals

The technical definition of a mineral is elegant: *a naturally occurring crystalline solid*. That definition calls for a bit of commentary, though. By “crystalline” is meant that the atoms of the mineral are arranged in a regular three-dimensional array, called its ***crystal structure***. The qualification “naturally occurring” excludes the multitude of crystalline solids that can be synthesized in the laboratory but are not found in the natural environment. Also, there are naturally occurring solids, like glass or amber, that do not have a crystal structure and therefore technically are not minerals.

There are thousands of named minerals—but, fortunately for us in this course, only a few are common rock-forming minerals, and even fewer figure prominently in the unconsolidated material that mantles the continents.

Most of the common minerals are of a class called ***silicate minerals***. Silicate minerals have as their basic building blocks a polyatomic (five-atom) unit, called a ***silica tetrahedron***, that consists of one atom of silicon, relatively small, surrounded by four atoms of oxygen, relatively large, to give the shape of a tetrahedron (Figure 2-3). The five atoms are bonded very strongly together. The reason why silicate minerals are so common is that, in terms of abundances of the chemical elements in the crust, oxygen is the big number one and silicon is second.

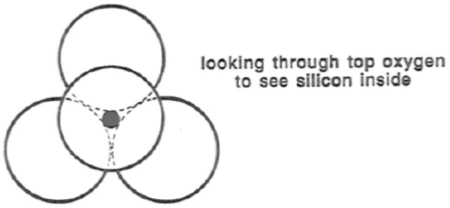


Figure 2-3. Schematic view of the silica tetrahedron, looking straight down on one of the oxygens. The small black dot represents the silica atom.

The atoms in mineral crystals are held together by strong forces called ***bonds***. An understanding of the nature of bonds necessitates some familiarity with atomic structure. All atoms consist of a positively charged nucleus around which a number of negatively charged electrons are in orbit. The fundamental nature of the electron orbits is complex, but for the purposes of understanding bonds in minerals it’s sufficient to know that electron orbits exist in distinctive groups known as ***shells***.

Chemical elements like helium, neon, or argon whose atoms have just enough electrons to fill all of the shells are chemically almost entirely inert. Elements whose atoms have their outermost electron shell almost filled have a strong tendency to take on one or a few electrons to fill the shell, and in so doing they acquire a negative electric charge. Similarly, elements whose atoms have only one or a few electrons in their outermost shell lose those electrons to revert to a filled-shell configuration, and in so doing they acquire a positive electric charge. Such atoms with electric charges are called ***ions***.

According to what is known in physics as Coulomb’s law, unlike electric charges attract one another and like charges repel one another. Certain combinations of positively charged and negatively charged ions can become packed together in a regular three-dimensional array in such a way that the sum of all the forces, attractive and repulsive, among all of the ions in the structure is attractive, meaning that the structure is a stable one and can exist as a mineral. Almost all minerals are of this nature, and are called ***ionic crystals***; the bonds in such a crystal are called ***ionic bonds***. Ions can also be formed from two or more atoms that share their electrons in such a way that each atoms effectively has a filled-shell configuration. The bonds among such atoms are called ***covalent bonds***. Covalent bonds are much stronger than ionic bonds, and the covalently bonded atoms form a single ion that is available for ionic bonding with other ions.

It turns out that in terms of electric charge the silica tetrahedron has a net negative charge. The silica tetrahedra are bonded to various positively charged atoms, mainly of the elements aluminum, iron, magnesium, calcium, potassium, and sodium. These ionic bonds are weaker than the strong covalent bonds within the silica tetrahedra. When a mineral is broken, it breaks along these ionic bonds rather than through the bonds within the silica tetrahedra.

There is an important further complication to silicate minerals, which makes them highly diverse (and almost unique in the universe!): individual tetrahedra can be joined together by sharing one or more (two, three, or even all four) of their oxygens. The phenomenon is called ***polymerization***. By polymerization, zillions of tetrahedra can be joined into chains or sheets that stretch all the way through a macroscopic mineral grain. In some minerals, all of the oxygens of the tetrahedra are shared, resulting in a complex three-dimensional network of shared tetrahedra.

There are other important minerals besides silicate minerals. Oxide minerals are perhaps the most important kind, in the context of this course. Oxide minerals are those in which negatively charged oxygen atoms are bonded with various positively charged ions. The most important oxides are *iron oxides* and *aluminum oxides*. These oxides of iron and aluminum are generally unspectacular in their appearance, but they are voluminous and important in Earth-surface environments. Carbonate minerals are also important near-surface non-silicate minerals. In carbonates, a negatively charged carbonate ion, consisting of one carbon atom covalently bonded with three oxygen atoms, is ionically bonded with calcium, magnesium, and/or iron ions; calcite and dolomite are common carbonate minerals.

Conclusion

To conclude this section, I'll list just the very most important kinds of minerals in the Earth's upper crust. It's important to know something about these minerals, because the mineralogy of solid Earth materials is of great importance in determining how those Earth materials are affected by their physical and chemical environment.

- *ferromagnesian silicates ("femags"):* This collection of minerals (olivines, pyroxenes, amphiboles, biotite, and certain metamorphic minerals) contain iron and/or magnesium, in various proportions, as their main positive ions. They tend to be dark in color. They are abundant in igneous and metamorphic rocks—but not in sedimentary rocks, because they tend to weather rapidly.
- *micas:* Micas consist of sheets of polymerized silica tetrahedrons, each with three of their four oxygens shared, with various positive ions between the sheets. The main micas are muscovite (a potassium mica), which is common in igneous, metamorphic, and sedimentary rocks, and biotite (a ferromagnesian mineral, with iron and magnesium rather than potassium). Biotite is common in both metamorphic and igneous rocks, but it weathers rapidly, so it's uncommon in sediments and sedimentary rocks.
- *feldspars:* Feldspars are the most abundant kind of mineral in the Earth's crust. They are network silicates with either potassium (these are called potassium feldspar, or just Kspar for short) or varying proportions of sodium and calcium (these are called plagioclase) in their structure. Feldspars are the major minerals of igneous rocks and are common in many metamorphic rocks. Potassium feldspar is common in sediments and sedimentary rocks, but plagioclase is not, because it weathers readily.
- *quartz:* Quartz, with the simple formula SiO2, is a network silicate mineral with a three-dimensional network of silica tetrahedra, with all of the oxygens shared with adjacent tetrahedra. It is common but not abundant in igneous rocks, is variably abundant in metamorphic rocks, and is very abundant in sediments and sedimentary rocks.
- *iron oxides:* Iron oxides, of which there are several forms, are common only in sediments and sedimentary rocks. In soils, there are several poorly crystallized hydrous iron oxide minerals.
- *aluminum oxides:* In poorly crystallized form, these are characteristic of soils formed by deep weathering of aluminum-bearing rocks under intense weathering conditions.
- *carbonates:* The calcium carbonate minerals calcite and aragonite, along with the calcium–magnesium mineral dolomite, are the main minerals in chemically precipitated carbonate sedimentary rocks, and they are common as well in sediments in the warm and shallow oceans.
- *clay minerals:* Several minerals with sheet-silicate structures are common as very fine-grained plate-shaped particles, some even smaller than a micrometer (one-thousandth of a millimeter; too small even to be seen with a powerful optical microscope). Clay minerals are the main constituent of most fine-grained muds.

2.3: Rocks

Rocks, in the form of what is called ***bedrock***, are exposed at the Earth's surface over wide areas, and they everywhere underlie the unconsolidated surficial materials, called ***regolith***, you will learn about later in this chapter. Moreover, owing to their physical and chemical breakdown under the conditions of the Earth's surface environment they are the source of that surficial material. There's no room in this course for a systematic and detailed treatment of rocks—that's the province of an introductory course in geology—but you should know at least some general things about them.

Rocks are naturally occurring aggregates of minerals. Again, that succinct definition is not very revealing of the nature and variety of rocks. I suppose that it verges on common knowledge that there are three kinds of rocks: *igneous*, *sedimentary*, and *metamorphic*. These three categories are for the most part distinct, although there is some fuzziness to the boundaries.

Igneous rocks are those that form by cooling and solidification of ***magma***. Magma is the term for melted rock in the Earth's interior. At certain times and in certain places in the Earth's shallow interior, down to a hundred or so kilometers, the rocks of the upper mantle or lower continental crust melt, to form magma. The melting is only partial, commonly up to fifteen to twenty percent of the rock, but the magma collects and then moves upward owing to its buoyancy (it's slightly less dense than the surrounding rock). It either becomes parked in subsurface spaces, called magma chambers, there to cool to form coarse-grained igneous rocks like granite or gabbro, or it rises all the way to surface to be ejected from volcanoes, either as flowing liquids or as already solidified particles.

Most of the igneous rocks you are likely to encounter as bedrock outcrops on the continents are of three general kinds (although this is a great oversimplification of the wide range of igneous rocks):

- coarse-grained and light-colored rocks like ***granite***, which consist mainly of feldspar with some quartz;
- fine-grained and dark-colored rocks like ***basalt***, which consist mainly of ferromagnesian minerals and feldspars; and
- rocks, called ***tuff***, that form by lithification of volcanic ash, the solid particles that are ejected explosively from certain kinds of volcanoes.

Sedimentary rocks are those that form from sediments that are deposited at the Earth's surface and become lithified (turned into solid rock, by a variety of processes) as they are buried more and more deeply in the earth's crust. Sedimentary rocks vary widely in composition and origin. There are two major realms of sediments and sedimentary rocks: *siliciclastic* and *chemical*. Siliciclastic sediments and rocks are those that consist of particles of minerals or rock that are derived from weathering of bedrock on the continents. Chemical sediments and rocks are those that are precipitated from natural surface waters, mainly in the ocean. By far the most important chemical sediments are carbonate sediments. Most carbonate sediment is produced by carbonate-secreting organisms, like clams or corals, although some is precipitated inorganically, in somewhat the same way that rock candy grows in a beaker on your windowsill.

Most of the sedimentary rocks you are likely to encounter are of the following kinds. (Again, this is a highly simplified list, in large part because these major sedimentary rock types grade imperceptibly into one another.)

- *sandstones:* Sandstones consist of sand-size particles of various minerals, mainly quartz but also potassium feldspar, as well as sand-size fragments of fine-grained rocks.
- *shales:* Shales consist mostly of very fine-grained mineral particles of quartz and clay minerals.
- *limestones*: Limestones consist mainly of the calcium carbonate mineral calcite. Also abundant are dolostones, which consist mainly of the carbonate mineral dolomite.

Metamorphism refers to changes in mineral composition or rock geometry that occur in solid rocks with increasing temperature and pressure. These changes produce a rock called metamorphic rock. It is important to remember that *changes producing metamorphic rocks take place while the rock is solid*. Any part of the rock that is melted eventually cools to form an igneous rock.

The basic idea behind changes in mineral composition during metamorphism is that each mineral has associated with it a range of temperature and pressure in which it is stable, in the sense that it remains in existence indefinitely. When a mineral is out of its stability range, it is liable to be converted into, or replaced by, one or more other minerals that are stable at those temperatures and pressures. A simple way of phrasing this is that *new minerals grow at the expense of the original minerals*. These changes in mineral composition act very slowly, because the rock remains solid and atoms can be exchanged between adjacent minerals grains only by atomic diffusion—although very small percentages of pore solutions along grain boundaries can expedite the transport of atoms.

Deformation (change in shape) is an important factor in the development of many metamorphic rocks. For most kinds of metamorphic rocks, change in mineral composition is accompanied by strong shearing deformation. Strong shearing tends to obscure or even obliterate original geometrical features like stratification in a sedimentary or volcanic rock. Also, most metamorphic rocks show some degree of what is called foliation: development of planar features in the rock, like layering or a tendency for the rock to split along parallel planes.

Below are brief descriptions of the most common metamorphic rocks. As with the sedimentary rocks, there are all gradations among these types.

- *slate:* Slate is produced by low-intensity metamorphism of shale. Individual crystals are too small to see, but the rock has a strong tendency to split along parallel planes, as a result of growth of new mica crystals of sheet-silicate minerals all parallel to one another.
- *phyllite:* Phyllite is produced when slate undergoes further metamorphism. Phyllite typically has a silky sheen.
- *schist:* Schist is formed by metamorphism of many igneous and sedimentary rocks. It has clearly visibly flakes of sheet-silicate minerals (muscovite, biotite, and/or chlorite), and it tends to split along parallel planes in the same way as slate and phyllite.
- *gneiss:* Gneiss consists mainly of granular minerals like feldspar, quartz, and ferromagnesian minerals. Gneiss is often prominently layered.
- quartzite: As the name implies, quartzite consists mainly of quartz. It forms mostly by metamorphism of quartz-rich sedimentary rocks.
- *marble:* Marble consists mainly of calcite. The calcite crystals tend to be large enough to see with the unaided eye, having grown larger during metamorphism without change in overall mineral composition.

2.4: Bedrock

By ***bedrock*** I mean *solid rock, exposed at the Earth's surface or buried at shallow depths, that is connected continuously with rock at greater depths in the* crust (Figure 2-4). A great many geologists spend most of their professional lives studying the bedrock of the Earth's continents.

In some places on the continents, bedrock is exposed at the surface, in the form of what are called ***outcrops*** or ***exposures***. There are natural exposures (to which the term *outcrop* is typically applied) and artificial exposures. Natural outcrops range in area from small, human-scale patches peeping out from the surrounding blanket of regolith, to whole mountainsides. Favorite places to find good outcrops are stream beds, hill slopes, mountain tops, and sea cliffs. These days, highway road cuts make the best artificial exposures; in times past, railroad cuts were prime exposures. Building foundations are good but temporary.

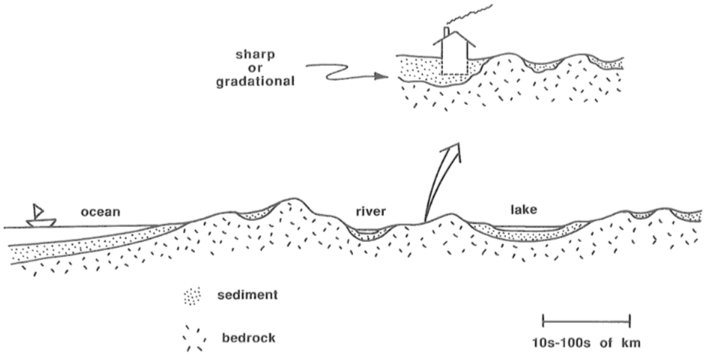


Figure 2-4. The bedrock of the continents

The percentage of continental surface where bedrock is exposed varies greatly, from nearly one hundred percent, in regions with rigorous climate and steep slopes, to zero, in regions with warm humid climates and little topographic relief. (The term ***relief*** is used for *the magnitude of differences in elevation of the land surface in a given region*.) The extent of exposure of bedrock depends mainly on the competition between two opposing processes: the rate of production of new regolith by weathering of exposed bedrock, and the rate of removal of existing regolith by the erosive action of running water or moving glacier ice or by downslope movement of regolith by the pull of gravity. Climate is important because the rate of production of regolith by weathering is much greater in warm and humid climates than in cold and arid climates. In some regions with a warm, humid climate and low relief, one needs to dig or drill tens or even hundreds of meters down to find fresh bedrock!

2.5: Weathering

2.5.1 **Weathering Introduction**

Weathering is the term used for the *chemical decomposition and physical disintegration of bedrock at and just below the earth's surface*. Weathering acts upon all bedrock near the surface, although with greatly varying nature and rate depending upon a number of factors.

Rock at depth tends to be brought upward toward the surface, by gradual unroofing by weathering at the surface and removal of the weathering products. It's a matter of relative motion: the bedrock at depth may be either motionless (relative to, say, the center of the earth) or moving slowly upward by uplift or downward by subsidence, but in any case *it's moving upward relative to the earth's surface if material at the surface is being stripped away*. Astherock nears the surface, it tends to be subjected to what's called *near-surface alteration*: it finds itself at much lower pressures and temperatures than at greater depth, and surface-derived waters, which tend to be oxygenated and carry dissolved CO2, have the opportunity to cause certain chemical reactions in many constituent minerals, like ferromagnesian silicates and feldspars. In a sense, you can think of such near-surface alteration as a precursor or forerunner to weathering in the stricter sense. But despite such near-surface alteration, the rock is still solidly bedrock, and nowhere near a condition that could be described as regolith.

Weathering, in the usually accepted sense, acts upon rock that is close enough to the surface to be affected not only by active circulation of surface- derived waters but also by such things as temperature changes and plant growth. But in thinking about weathering you don't have to restrict yourself to the actual rock surface, upon which the sun shines!

Weathering is traditionally subdivided into two sets of processes:

- *chemical weathering*, the chemical decomposition of some or all of the constituent minerals or the bedrock.
- *physical weathering*, the mechanical breakdown or disintegration of the bedrock itself into particles large and small.

This is a useful distinction, but here's a very important fact to understand: *chemical weathering and physical weathering tend to act simultaneously and reinforce one another*.

2.5.2 Chemical **Weathering**

Chemically, *the earth's surface is a very reactive place*. To begin with, in most places and at least at certain times *the surface is an aqueous environment*, and water is not only a very effective solvent in itself but also a necessary medium for a great many chemical reactions.

Owing to the high content of oxygen in the atmosphere, in most places *the earth's surface is a well oxygenated environment*, as in the shallow subsurface wherever the concentration of organic matter is not so great and the permeability so low (as is common in bogs and swamps) that the oxygen is consumed in oxidation of organic matter before it passes downward to deeper depths. (If you're unsure about the nature of oxidation, there will be some background material on it later.) Because so many of the common rock-forming minerals contain ferrous iron (that is, iron in the reduced state, and thus susceptible to oxidation to the insoluble ferric state), *oxidation reactions are an important aspect of chemical weathering*.

In the same way, because the atmosphere contains a substantial percentage of carbon dioxide (CO_2), *surface waters invariably contain appreciable dissolved* CO_2. You know already that CO_2 readily dissolves in water. That dissolved CO_2 reacts with the water itself to form a weak acid, *carbonic acid*, according to the reaction

$$\mathrm{CO_2 + H_2O \rightleftharpoons H_2CO_3} \tag{2.5.1}$$

At equilibrium, about one percent of the dissolved CO_2 is in the form of carbonic acid. Most minerals are less stable in an acidic medium than in a neutral medium, so the dissolved CO_2 enhances the weatherability of many minerals.

Specifically, what's going on with Reaction 1? A molecule of carbon dioxide, which consists of a carbon atom tightly bound up with two oxygen atoms, combines with a molecule of water, which consists of an oxygen atom tightly bound up with two hydrogen atoms, in such a way that a single, new, larger molecule, with the chemical formula H_2CO_3, is produced. In the aqueous solution, this process is going on all the time, and at the same time the H_2CO_3 molecules are being changed back into CO_2 and H_2O molecules. Where the balance is struck depends on the imposed concentration of CO_2, which in this case comes either from the atmosphere or from the chemical breakdown of organic matter. (You probably know that, by photosynthesis, plants take up water and carbon dioxide and convert it into complex carbohydrate molecules, and when the plant dies and the plant tissues are exposed to an oxygenated aqueous environment the reverse happens and the organic matter is broken back down into carbon dioxide and water.)

Carbonic acid is not the only acid common in natural near-surface waters: in most land areas, organisms (both plants and animals) living on or in the soil (that is, the uppermost part of the layer of regolith) use carbon dioxide and various nutrients to produce organic matter, which in its decay back to the constituent compounds tends to produce *a variety of weak organic acids*.

In the rest of this section, several kinds of chemical reaction will show the range of weathering reactions that are important in chemical decomposition of bedrock minerals. They involve *solution, oxidation, and carbonation*, in various ways. Keep in mind that although these particular reactions are important, *they are just representative reactions*: details vary depending upon the assemblage of minerals that are being weathered, and upon their particular compositions (which, remember, vary widely because of the wide range of ionic substitutions in silicate minerals).

Solution

Some minerals simply *dissolve in water*. Halite (sodium chloride) and gypsum (hydrated calcium sulfate) are good examples:

$$\mathrm{NaCl(s)} \rightleftharpoons \mathrm{Na^+ + Cl^-} \tag{2.5.2}$$

$$\mathrm{CaSO_4 \cdot H_2O} \rightleftharpoons \mathrm{Ca^{2+} + SO_4^{2-} + H_2O} \tag{2.5.3}$$

This is the simplest kind of weathering reaction to understand, but not the most important. A more important, and more complicated, kind of solution involves calcite, the constituent of limestone:

$$\mathrm{CaCO_3(s) + CO_2 + H_2O} \rightleftharpoons \mathrm{Ca^{2+} + 2\,H^+ + CO_3^{2-}} \tag{2.5.4}$$

The reason this reaction is more complicated is that it involves the atmospheric carbon dioxide dissolved in the water. The CO_2 has to be involved because of the CO_3^{2-} ion both in the structure of the mineral and as a product of the solution and then dissociation of CO_2 in water. (Incidental note: this reaction can go in either direction, depending upon the concentration of CO_2: if CO_2 is added to the system, as in weathering environments, calcite is dissolved; if it's extracted from the system, as in certain shallow warm ocean environments, calcium carbonate sediment is precipitated. Think about that the next time you're lying on a white carbonate-sand beach in the Caribbean.)

carbonation of silicate minerals that don't contain aluminum or iron:

Most silicate minerals that don't contain aluminum contain both Fe^{2+} and Mg^{2+} ions in various proportions. Rather than writing a single weathering reaction for such minerals, it's neater to look at separate reactions for those with only magnesium and those with only iron. Such minerals, although they exist, are not common, but their weathering reactions are instructive. First look at the weathering of olivine, which is a common mineral in igneous rocks, and one that's very susceptible to chemical weathering. Most olivines in igneous rocks contain some iron in addition to magnesium, but let's assume, for simplicity, that we are dealing with a pure magnesium olivine:

$$\mathrm{Mg_2SiO_4 + 4\,CO_2 + 4\,H_2O \rightarrow 2\,Mg_2^{+} + 4\,HCO_3^{-} + H_4SiO_4} \tag{2.5.5}$$

(I've used the single arrow to the right, rather than the double arrow, here and below because the reverse of these reactions is not actually how the minerals get built under deep Earth conditions.)

Note that "silica" (that is, SiO_2 in some form or other) is extracted from the original silicate mineral. In most natural waters, the dominant species of silica in solution is H_4SiO_4, a weak acid called *silicic acid*. It's in the form of a molecule that consists of four hydrogen atoms, four oxygen atoms, and one silicon atom, all tightly bonded together.

carbonation of silicate minerals that don't contain aluminum but do contain iron:

Now look at a ferromagnesian mineral with all Fe and no Mg. (We'll assume that the mineral is a pyroxene mineral; no details here.) In our oxygen- rich near-surface environment, it's a two-step process. The first step is like that expressed in Reaction 5 above, but the numerical values of the coefficients in front of the reactants and products are somewhat different, because we started out with a mineral with a different chemical formula. Qualitatively, the result is the same: *the iron-bearing silicate mineral is broken down in the presence of water and carbon dioxide to put iron and silica in solution*. The dissolved iron is in the ferrous ("plus-two") form, which is freely soluble in pure water. It is converted in the presence of oxygen to ferric iron oxide and silica in solution:

$$\mathrm{FeSiO_3 + 2\,CO_2 + 3\,H_2O \rightarrow Fe_2^+ + 2\,HCO_3^- + H_4SiO_4} \tag{2.5.6}$$

But that’s not the end of the story, because the ferrous iron in solution reacts with

oxygen dissolved in the water, to precipitate an iron oxide mineral, Fe_2O_3:

$$\mathrm{2\,Fe_2^+ + 3\,O_2 \rightarrow 2\,Fe_2O_3} \tag{2.5.7}$$

The iron in the Fe_2O_3 mineral is in the ferric (“plus-three”) form, which is almost totally insoluble in pure water.

The representation on Reaction 7 is still a bit of a fake, however, because ferric oxides produced at the earth’s surface are commonly hydrous or hydrated: one could describe them as hydrous ferric oxides. There are several closely related minerals of that kind; the most common are goethite, FeO(OH), and ferrihydrite, 5Fe2O3.9H2O). They all have colors that range from yellow through orange to brown; that’s why weathered surfaces on bedrock are so commonly of those colors. In a sense, ferrous-iron-bearing rocks *rust* when exposed at the earth’s surface.

If we had started with a real pyroxene, with a combination of Fe^{2+} ions, Mg^{2+} ions, and Ca^{2+} ions, the latter two would have been freed to solution while the first would have found its way into insoluble oxides; that’s because magnesium and calcium are readily soluble in water but ferric iron is almost completely insoluble.

weathering of aluminosilicates:

Now to the biggest show of all in chemical weathering. The example here will be what is certainly *the single most important weathering reaction on earth: hydrolysis and carbonation of feldspar*. We’ll use potassium feldspar as the example; weathering of plagioclase feldspar is similar.

$$\mathrm{2\,KAlSi_3O_8 + 2\,H_2CO_3 + 9\,H_2O \rightarrow Al_2Si_2O_{25}(OH)_4 + 4\,H_4SiO_4 + 2\,K^+ + 2\,HCO_3^-} \tag{2.5.8}$$

Comments:

1. The first product on the right is the clay mineral *kaolinite*. (There will be more material on clay minerals in a later section.) Other kinds of clay minerals are possible as solid products of weathering of feldspar as well, depending upon the details of the chemical environment.
2. The products in solution are silicic acid, as before, and potassium ions.
3. If the feldspar had been plagioclase, Na^+ and Ca^{2+} ions would have gone into solution rather than K+ ions.
4. The reaction could have been rewritten slightly to start with CO_2 rather than with H_2CO_3. The differences are immaterial. Those of you who are well versed in chemistry can try to rewrite the reaction by adding to it the following reaction: $2CO_2 + 2H_2O \Leftrightarrow 2H_2CO_3$.

In conclusion, if you don’t remember anything else about chemical weathering, remember this:

when feldspar is weathered, by reaction with water and carbon dioxide, the result is a clay mineral, positive ions (K^+ and/or Na^+ and/or Ca^{2+}) in solution, and silica, mostly in the form of H_4SiO_4, in solution.

In intense weathering situations, in warm and humid environments, weathering of aluminosilicate minerals like feldspar can go a step further, to the ultimate weathering product:

$$\mathrm{2\,Al_2Si_2O_{25}(OH)4 \rightarrow 4\,Al(OH)3 + 4\,H_4SiO_4 + 5\,H_2O} \tag{2.5.9}$$

The kaolinite, which itself was the weathering product of potassium feldspar, is stripped further of all of its silica, to produce a totally insoluble aluminum hydroxide mineral and more silica in solution. This is the ultimate weathering residue! There are several hydrous aluminum oxide minerals that come out of a reaction like this; the mineral *gibbsite* is the most common. Surficial materials that are especially rich in such insoluble hydrous aluminum oxides are called ***bauxite***. As you probably know, such weathering-residue mixtures of hydrous aluminum oxides are the world’s supply of aluminum ore. Depending upon the composition of the original rocks, such ultimate weathering residues are commonly mixed with insoluble hydrous iron oxides as well.

One final point, which will be important for our later consideration of sediments and soils: all of the solid weathering products in the reactions above— clay minerals and various oxides of aluminum and ferric iron—are commonly *of finer particle size than the minerals from which they are produced during weathering*.

What are the controls on chemical weathering? Viewed on the large scale, there are four major controls on chemical weathering:

- *climate:* The intensity of chemical weathering increases with precipitation and temperature. The reason it increases with precipitation is that weathering requires an aqueous medium for the reactions. The greater the percentage of time a given mass of weatherable rock is saturated with surface water, the more chemical weathering it is likely to experience. The reason that the intensity of chemical weathering increases with temperature is that, as with all chemical reactions, the rate of the reaction increases with temperature. That's basically because the vigor of the thermal agitation of the atoms and molecules increases with temperature.
- *rock type:* Some kinds of rocks are more susceptible to chemical weathering than others. Rocks that contain minerals that were formed under conditions of the high temperatures and pressures of the deep crust or upper mantle are much more weatherable than rocks that contain abundant minerals like quartz, calcite, or muscovite, which are stable, or at least only slowly weatherable, under most Earth-surface conditions. For that reason, most igneous and metamorphic rocks are generally more susdceptible to weathering than most sedimentary rocks.
- *relief:* Regions with high relief usually also have steep slopes. Gravity-driven mass movements (more on that in a later chapter) then are effective at stripping away newly weathered material, thereby exposing fresher bedrock to continued chemical weathering. In regions with low relief, however, the layer of regolith thickens to the point where the underlying bedrock is largely sealed off from further contact with the surface environment of chemical weathering.
- *vegetation:* As mentioned earlier, when plant materials decay they release carbon dioxide and weak acids into the regolith environment. These substances acidify the surficial material, thereby tending to drive the various chemical weathering reactions discussed above to the right.

There's another way of looking at the controls on chemical weathering: from the standpoint of *what a small mass of weatherable material actually senses*. This is the small-scale view, in contrast to the foregoing largescale view. It boils down, largely, to four more specific physicochemical factors:

- mineral composition of the rock
- chemical composition of the water
- rate of passage of water through the rock • temperature

2.5.3 Physical Weathering

Most rocks are strong: you need to pound on them with a hammer to break them. The strongest rocks, like tough quartzites, yield only reluctantly to even the most vigorous hammer blows. Other rocks, like soft sandstones, you can break with your own hands. (In fact, there is a continuous gradation between loose and unconsolidated sediment and strongly lithified sedimentary rock, depending in large part upon depth of burial.) Rock strength is a complicated topic, which is important for the construction industry, as well as for geoscientists who try to understand rock fracture that causes large earthquakes.

Let me a bit more specific about the nature of the strength of solids like rocks. There are three ways of breaking solids: by *compression*, by *tension*, and by *shear* (Figure 2-5). In rocks, the compressive strength and the shear strength are both large, but the tensile strength is much less: generally, it's much easier to break a rock by pulling it apart than by squeezing it. (That's true of concrete, an artificial rock, as well.)

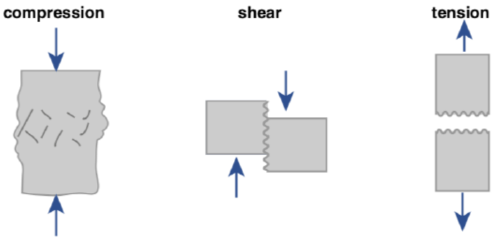


Figure by MIT OCW.

Figure 2-5. Three ways of breaking solids.

The easiest way to break a rock is to *wedge open a preexisting crack*. That's partly because of the low tensile strength, but also because of an effect known as *stress concentration* at the tip of the propagating crack (for reasons that are beyond the scope of this course). A simple analogy can serve well here: it's much easier to "break" a sheet of paper in tension by starting a cut and then pulling the edges of the cut apart than by grasping the sheet at opposite edges and pulling on it uniformly. When a rock fractures in tension, either rapidly as a result of a blow or an explosion, or slowly, by a variety of natural weathering processes, the crack enlarges as the tip of the crack propagates into the non-fractured rock.

Breakage of polycrystalline aggregates like rocks (or metals) is more complicated than breakage of homogeneous materials like glass. That's because *the crystal grains of the rock are joined together at the grain boundaries*, so the strength of the grain contacts needs to be taken into consideration as well as the strength of the mineral grains themselves. When a rock is stressed, extremely fine cracks tend to develop first along the grain boundaries.

Rocks at the surface tend to be broken down (disaggregated, disintegrated) by a variety of processes, whose importance varies greatly from place to place and from time to time. These physical processes (you might also call them mechanical processes) have traditionally been cited as important in physical weathering, but definitive quantitative studies of the various effects are discouragingly scarce. I'll set them out for you here without making much of an attempt to rank them in importance, except for some concluding comments at the end of this section.

sheeting: Sheeting is *a style of rock fracturing that develops sets of fractures just below the bedrock surface and concordant with it*. The reason such fractures tend to form is that as rock is brought up to the surface it is *unloaded*: the weight of overlying rock is reduced. The rock tries to expand slightly (rock has small but definite compressibility), and *it can expand only upward, not sideways*. This leads to development of fractures that are roughly parallel with the rock surface. Sheeting is most common in otherwise massive and non-fractured bedrock. The sheeting joints decreases rapidly in spacing downward from the surface, usually disappearing within several meters. The next time you see a photo of the mountain slopes in Yosemite, think in terms of sheeting fractures.

exfoliation: Exfoliation (Figure 2-6) is the *separation, during weathering, of successive shells, a few millimeters to some tens of centimeters thick, from massive bedrock*. The mechanical effect is somewhat like that of sheeting: as the thin near-surface layer of bedrock undergoes chemical weathering, the weathering-product minerals tend to be of lower density that the original minerals, which means larger volume, so the thin uppermost zone of the rock tries to expand. It can expand only upward, so fractures parallel to the rock surface are formed. In this situation, chemical weathering and mechanical weathering act together to disintegrate the bedrock. A related process leads to what's called ***spheroidal weathering***: the exfoliation and spalling is slightly more effective at edges and corners than on rock faces, so in an outcrop cut by sets of planar fractures in various orientations (such fractures are called ***joints***; they are ubiquitous in bedrock outcrops) *the edges and corners are rounded off* to produce a very spheroidal or "blobby" look to the outcrop. In advanced cases the spheroids are completely freed from the outcrop, and lie around on the surface or roll downslope.

Figure by MIT OCW.

Figure 2-6. Disintegration of a surface boulder by exfoliation. (From Hunt, 1972.)

granular disintegration: Granular disintegration is another process that is in part a mechanical effect of chemical weathering. Many if not most rocks consist of a mixture of minerals, some susceptible to chemical weathering and others not. As the weatherable minerals expand upon incipient chemical weathering, while the non-weatherable minerals retain their original size, intergranular forces are set up, which tend to break the bonds between mineral grains. The mineral grains fall from the rock surface, one by one, to leave an overall smoothly rounded but granular-rough surface. The mass of loose grains at the base of the outcrop is called ***grus***. This kind of weathering is most common in crystalline rocks like granite.

thermal expansion and contraction: Expansion and contraction in response to changes in temperature can by itself aid in propagation of fractures in bedrock ranging in scale from intergranular fractures to widespread joints. Such a process is most effective in arid regions where the day-to-night temperature differences are great. It's unclear whether the process is effective in thoroughly dry rocks, or whether the presence of water is an important element in the process. Experiments on heating and cooling of rock samples, over a great many cycles to simulate many decades or even centuries of the natural environment, have not been successful in producing non-negligible fracturing, but maybe the conditions of the simulation have not been representative. It's well known, however, that fire causes breakdown of rock by sudden expansion of the surface layer exposed to the extreme heating.

frost wedging: Frost wedging (Figure 2-7) is important in climates where surface water is abundant and the day-to-night temperature range spans the freezing point many times through the year, as in New England in the late fall and especially the early spring. Water expands upon freezing, to the tune of about 9% in volume, an extremely large change. When water confined deeply in a thin crack in bedrock freezes, it can exert enormous side forces on the walls of the crack. The maximum possible force, which has been computed to be over 2000 kilograms per square centimeter (!), is not anywhere close to being achieved, however, because the ice does not seal the crack as strongly as the rock walls themselves, but the process is nonetheless very effective. Mountaintops with expanses of bedrock subjected to freeze–thaw cycles tend to be a jumble of frost-shattered rock; hikers in the Presidential Range in New Hampshire know that well.

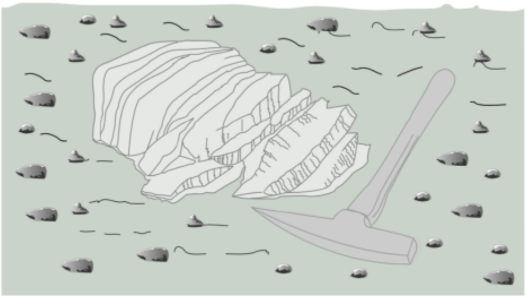


Figure by MIT OCW.

Figure 2-7. Disintegration of a surface boulder, probably by a combination of freeze-thaw, thermal expansion and contraction, and chemical weathering. (From Hunt, 1972.)

hydrofracturing: It's thought by some investigators that the frost-wedging effect is more complicated than just simple expansion upon freezing. Water as thin films in very small cracks may stay liquid even at very low temperatures, for subtle thermodynamic reasons. As the freezing front progresses inward from the rock surface, water can be forced under high pressures into microscopic cracks, causing propagation of fractures. The phenomenon is known as hydrofracturing.

"boulder lifting": It's often said that boulders "grow" in New England fields. Frost action provides a good explanation for the effect (Figure 2-8). As the water-saturated ground freezes from the top down, freezing reaches the base of a large boulder sooner than it reaches the surrounding soil, because the rock of the boulder has a much lower heat capacity than the water-rich soil around it and is therefore cooled faster. Freezing of the water beneath the boulder lifts or pushes the boulder upward a short distance. Upon thawing of the ice beneath the boulder, the boulder does not fall as far as it was raised, because of the tendency for collapse and flowage of soil and finer rock fragments into the space beneath the boulder. Given enough cycles, the boulder makes its appearance at the surface.

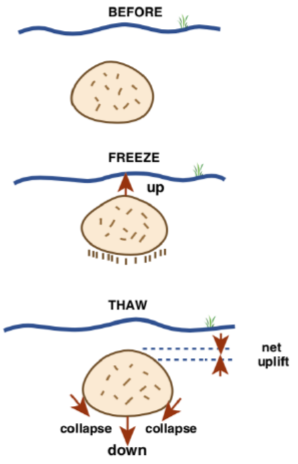


Figure by MIT OCW.

Figure 2-8. Uplift of a boulder by freezing and thawing of surrounding watersaturated regolith.

root growth: Root growth can be very effective at widening preexisting cracks in bedrock near the surface. All of you have seen plants growing heroically in cracks in rock. In many cases the plant just occupies an already large crack, but in other cases it's clear that the growing root has opened the crack.

Which of these various processes are the more important, and which are the less important? I can do no more than make an educated guess. (It depends, of course, on the particular setting of bedrock and climate, in any case.) In the right circumstances, *frost wedging* is clearly of great importance, but *granular disintegration*, and perhaps *exfoliation*, are probably of more general importance. Thermal expansion and contraction alone is generally considered to be less important than the other processes, except perhaps in particularly favored environments.

2.5.4 Weathering Profiles and Weathering Fronts

You've probably had the experience of picking up a broken stone and noticing that the outer part is colored reddish or brownish while the inner part looks a fresh gray. What's going on is that the stone, having been derived from some outcrop as a fresh piece and then perhaps rounded at least somewhat by transport, rested somewhere at the Earth's surface for a time long enough for the outer zone to have experienced some chemical weathering. The boundary between the outer, weathered zone and the inner, non-weathered zone is called the ***weathering front***. There's commonly a very abrupt change in the degree of weathering at the weathering front. In many cases the characteristic difference in color arises from the conversion of ferrous iron in minerals of the fresh rock to ferric iron oxides in the weathered zone.

The same effect tends to be present on a much larger scale when a bedrock landscape in an area of low relief is subjected to protracted weathering and the weathering products remain where they were formed. To see the effect, you need to have a vertical or steeply dipping cut down through the weathered rock, all the way down into fresh bedrock, which may be many meters or even tens of meters below the surface. Above the weathering front is a gradation from the most weathered material, at and near the land surface, down to the least weathered rock, at the weathering front. The succession of zones of degrees of weathering, from the surface down to the weathering front, is called a ***weathering profile***.

Weathering profiles vary greatly in their nature, depending upon the factors listed in an earlier section (largely a matter of rock composition and climate). Figure 2-9 shows a common weathering profile, of the kind that tends to develop on deeply in-place-weathered coarse-grained igneous rocks like granite. The material just above the weathering front, called ***saprock***, consists of partly chemically weathered minerals together with yet-unweathered minerals. The saprock preserves all of the original textures and structures of the parent rock, but its composition has been somewhat changed and its strength has been reduced. The saprock grades upward into material called ***saprolite***, which is more altered chemically than the saprock below but still retains much of the textures and structures of the parent rock. You might describe such rock as "rotten rock" (and that is in fact the etymology of the term saprolite). Above the saprolite is even more strongly altered material, which has lost even more of the coherence of the parent and shows little of the original textures and structures. Above this, where the material has lost its coherence entirely and could be described as "loosed", it is called ***regolith*** (see later section). The uppermost layer is the ***soil*** (see later section).

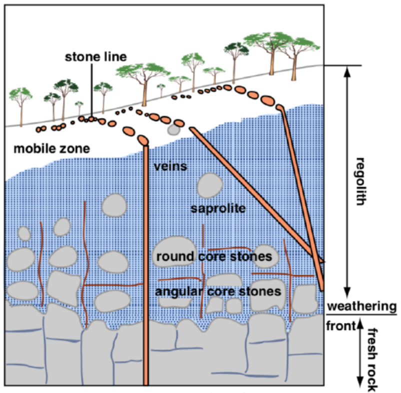


Figure by MIT OCW.

Figure 2-9. A typical weathering profile developed by deep, in-place weathering of coarse-grained granitic igneous rock. (From Taylor and Eggleton, 2001.)

Note in Figure 2-9 that *the geometry of the weathering front is strongly controlled by the presence of joints in the fresh rock*. The weathering can make better progress along the joints than in the non-jointed intervening rock. In the process, there's a tendency for non-weathered masses to be left behind as the weathering front progresses downward. These rounded, irregular masses are called ***corestones***. They are similar to the weathering spheroids mentioned earlier, but they differ in being formed well below the land surface, whereas weathering spheroids develop at the rock surface as weathering proceeds at the same time that the weathering products are removed from the outcrop.

2.6: The Mineral Particles of the Earth's Surface Materials

2.6.1

It seems a good idea, at this point, to describe the most common kinds of mineral particles in the ground under your feet—the materials that in the sections to follow are called regolith, sediment, and soil (regolith being the most inclusive term, as you will see). I think it’s accurate to say that *most regolith consists mostly of mineral particles*, coarse and fine. A wide variety of kinds of minerals can be found in typical regolith, but only a very few are of any abundance.

2.6.2

In terms of relatively coarse mineral particles, at the top of the list is **quartz**. Quartz is common in most of the source rocks from which surface materials are derived, and it’s largely immune to chemical weathering. This is a first-order fact: *most of the mineral grains in most of the regolith you’re ever going to see (with the notable exceptions of fine muds) consist of quartz*. Quartz grains are mostly equant (meaning approximately equidimensional) in “overall” shape, and usually in between subangular (that is, a bit more rounded than a classically jaggedy angular particle) and well rounded in “local” shape. The degree of roundness depends partly on the shape of the particle upon being freed from the parent rock by weathering and partly on the extent of mechanical wear the quartz particle has experienced, and perhaps also on some dissolution if the quartz particle is exposed to acidic environments (because the solubility of quartz in water, although extremely low in pure water, increases markedly with increasing acidity of the water).

2.6.3

Potassium feldspar is also important as coarse particles in regolith, depending upon the composition of the parent rock. Potassium feldspar is more susceptible to weathering than is quartz, but commonly some or even much of it escapes weathering before it is deposited as sediment and passes out of the reach of near-surface processes. Survival of potassium feldspar to be seen in regolith is especially common in situations where weathering is curtailed, either by a cold and rigorous climate or by rapid stripping of the weathered material from steep slopes. Commonly, potassium feldspar in regolith is partly or mostly weathered to kaolinite, and under magnification has a “punky” or “chalky” appearance.

2.6.4

Rock fragments are also common in regolith. This is especially the case for sediments. Imagine a raging river eating its way into fractured or partly weathered bedrock: it can entrain fragments of even fresh rock, large and small, and transport them to sites of eventual deposition. Glaciers can do the same thing even more effectively. Identifying large rock fragments is easy; also important, however, are sand-size fragments of various common fine-grained rocks. It takes careful microscopic work to identify such fragments.

2.6.5

Muscovite (the most common form of mica, but not the only one) as coarse particles is common in regolith as well. Micas are sheet-silicate minerals. Sheet silicates are those in which the silica tetrahedra are polymerized by sharing of three of the four oxygens to form a sheet structure, with hexagonal symmetry, and with all of the unshared oxygens pointing in the same direction. Figure 2-10 is a drawing of the sheet structure, looking down on the side with the unshared oxygens. There’s more to the mica structure than that, however: the basic sheet unit consists of two such silica sheets, each with their unshared oxygens pointing inward, with aluminum ions in between the two silica sheets. (If you want more detail on mica structure, see the clay mineral “advanced topic” below.)

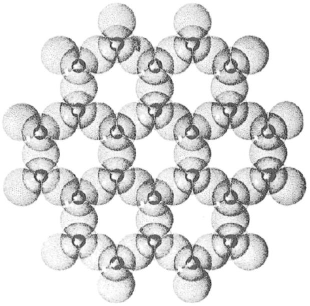


Figure 2-10 Polymerization of silica tetrahedra to form the sheet structure, a basic building block of sheet silicates.

2.6.6

The mica crystal is very weak in the direction parallel to the sheets of the structure, so in regolith it's usually in the form of plates or flakes. When it's coarse, so it can be seen with the unaided eye or with a hand lens, it is easy to identify. In sediments it's usually noticeably coarser than the equant particles like quartz, because its platy shape makes it easier for flowing fluid to transport it

2.6.7

In terms of relatively fine mineral particles in regolith, at the top of the list is the group of **clay minerals**. These are a group of minerals rather than a single mineral. Clay minerals are mainly sheet-silicate minerals. Some, called illite or smectite, have the same structure as muscovite, and nearly the same composition. Owing to various substitutions of some ions for others, though, the range of compositions of illite and smectite is wide. In fact, sedimentologists who have to deal with clay minerals often just refer to *illite—smectite* (or I–S for short). Kaolinite and chlorite, with yet different sheet structures, are other common clay minerals. For more detail on the nature of clay-mineral structures, see the following optional "advanced topic". The proportions of the various clay-mineral types in a given sample of regolith vary depending upon two major factors: *source-rock composition*, and *weathering environment*

2.6.8

An important thing for you to know about clay minerals is that *clay- mineral particles are all very small*, mostly less than a few micrometers. They range in size down to very small fractions of a micrometer, into what is called the colloidal size range. There will be more on colloids in the later section on soils, where they are especially significant.

Advanced Topic: Clay Minerals

I. Introduction

The word "clay" is used, somewhat ambiguously and confusingly, in three different but largely overlapping senses:

(1) As a *mineral* term, for a group of minerals, mostly but not entirely sheet silicates, that are produced from weathering of aluminosilicate minerals (mainly feldspars). They are almost always very fine, from the sub-micrometer range up to several micrometers, and they are mostly platy in shape, although some are curvy and curled. The principal kinds are *kaolinite*, *smectites*, *illites*, and *chlorites*. (I've used the plural for the last three because there's a very wide compositional range in each of them.)

(2) As a *size* term. According to the official grade scale for sediment sizes (see the later section on sediment), all particles finer than 1/256 of a millimeter, or about 4 micrometers, are termed clay.

(3) As a *material* term ("clay material", or just "clay"), which has a sticky and tenacious consistency. It's the stuff that potters use.

Most clay-mineral particles are of clay size; most clay-size particles are clay-mineral particles; and most clay material consists of clay-mineral particles of clay size!

As you can imagine, *clay minerals are not easy to study*: even in the coarser size range it's difficult to see them microscopically. Standard techniques for study include electron microscopy, which allows you to see the particle shapes beautifully, and x-ray diffraction, which allows you to identify their mineralogy, after a fashion. More sophisticated instrumental techniques have been developed more recently. Clay mineralogy is worth a whole course in itself. What I'll do here is just give you a cursory account of the most important clay minerals and their occurrence.

II. Structure

Two structural units are involved in clay-mineral structures:

octahedral sheets: two planes of closely packed oxygens and hydroxyls (OH^-) with a plane of Al, Fe, and/or Mg between. The Al, Fe, and/or Mg are in octahedral coordination with the O and OH; that is, each of the former is equidistant from six of the latter, which are arranged as an octahedron around the former. Figure 2-11 shows a rough sketch of the arrangement. Figure 2-12 is a more detailed view of the geometry, perpendicular to the sheet. The oxygen-to-oxygen distance is 2.94 Å, and the thickness of the sheet is 5.05 Å. If Al is in the octahedral positions, only two-thirds of the positions are filled; if Mg and/or Fe are in the octahedral positions, all of the positions are filled.

tetrahedral sheets, in which silica (SiO4) tetrahedra are polymerized by sharing of three of the four oxygens to form a sheet structure with hexagonal symmetry. Figure 2-13 shows two diagrammatic representations. The oxygen-to-oxygen distance is 2.55 Å, and the thickness of the sheet is 4.65 Å.

These two layers, or sheetlike arrangements, fit together, one on top of the other, in such a way that the vertices of the tetrahedra in the tetrahedral sheet point toward the octahedral sheet. The unshared oxygens of the tetrahedral sheet and the oxygens of the octahedral sheet are the same; they are shared for double duty. The hydroxyls that form part of the octahedral arrangement fit into the hexagonal holes of the tetrahedral sheet.

The dimensions of the two sheets are almost but not quite right for this shared arrangement. There has to be some strain in the lattices to make them go together to share oxygens. This seems to be accommodated by alternate cocking of the tetrahedra, to contract the tetrahedral sheet a little, and stretching and thinning of the octahedral sheet to make it a little bigger. I'll schematize this two-layer structure as shown in Figure 2-14.

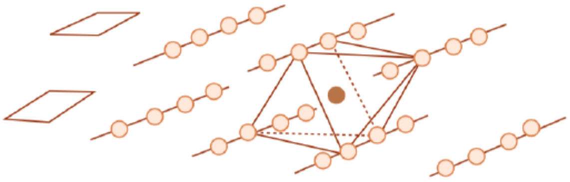


Figure by MIT OCW.

Figure 2-11. Sketch of the arrangement of atoms in the octahedral sheet of clay- mineral structures.

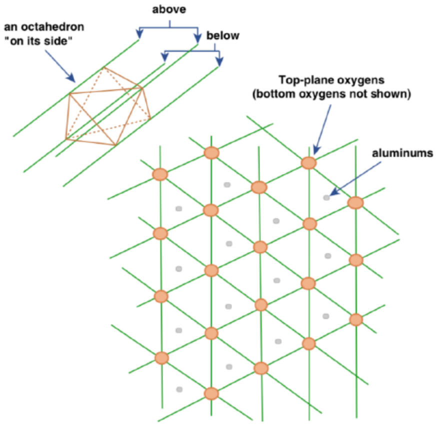


Figure by MIT OCW.

Figure 2-12. A more detailed view of the geometry of the octahedral sheet of clay-mineral structures.

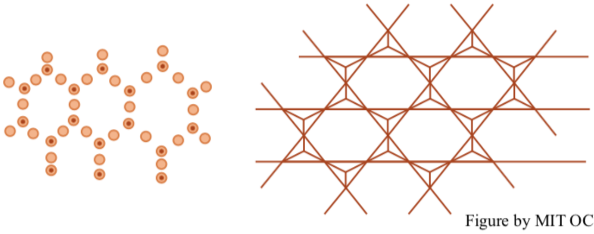


Figure by MIT OCW.

Figure 2-13. Diagrammatic representations of the arrangement of silica tetrahedra in the tetrahedral sheet of clay-mineral structures.

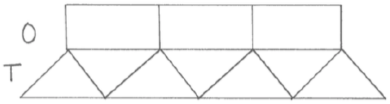


Figure 2-14. Highly diagrammatic sketch of the combination of octahedral and tetrahedral sheets in two-layer clay-mineral structures.

It's also possible to construct a three-layer arrangement in which a tetrahedral layer is sandwiched between two tetrahedral layers. Figure 2-15 shows the highly schematic way I'll represent this three-layer structure.

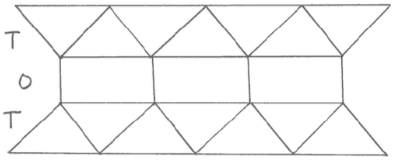


Figure 2-15. Highly diagrammatic sketch of the combination of octahedral and tetrahedral sheets in three-layer clay-mineral structures.

III. The Major Clay Minerals

Kaolinite:

Kaolinite (Figure 2-16) is the main two-layer mineral. It's fairly simple chemically: $Al_4(Si_4O_{10})(OH)_8$, with no isomorphous substitutions in either the tetrahedral sheet or the octahedral sheet; it's the purest of the clay minerals. Only two-thirds of the octahedral positions are filled, by aluminum; minerals like that are called *dioctahedral*. There's no net charge on the tetrahedral–octahedral layers; the layers are held together only by a kind of weak bonds called van der Waals bonds. The repeat distance normal to the layers is 7.2 Å, so this group of minerals is sometimes called *seven-Ångstrom clays*. The crystals can get relatively large, because there's very little stress in the structure; kaolinite is the coarsest of the clay minerals. Kaolinite is formed mainly by weathering of feldspars; production of kaolinite is favored by acidic and warm environments.

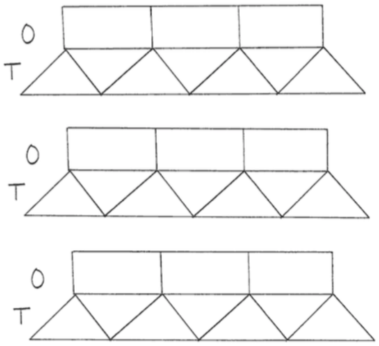


Figure 2-16. Highly simplified sketch of the structural organization of kaolinite.

Illite:

Illite (Figure 2-17) is a hydrous muscovite-like clay mineral. It has the same three-layer structure as muscovite, but less than the one-in-four replacement of Si by Al that's characteristic of muscovite (the figure is between one-in-five and one-in-six). So there isn't the same charge deficiency as in muscovite, and not as many K+ ions in the interlayer positions. H3O+ ions are also present in the interlayer positions. Also, substitutions in the octahedral layer are greater and more random; there's lots of variability. Illite is never found well crystallized, only as clay-size particles. The particle size is typically less than one micrometer. This is because there's considerable stress in the layers. Illite is closely similar in structure to the common muscovite you see in sedimentary, metamorphic, and igneous rocks; it's just much finer-grained. Illite is formed mainly by weathering of potassium feldspar in temperate weathering environments. With increasing intensity of weathering, kaolinite is formed instead.

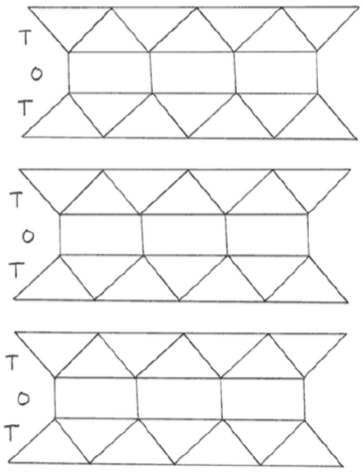


Figure 2-17. Highly simplified sketch of the structural organization of illite and smectite.

Chlorite:

Chlorite (Figure 2-18) is a kind of mixed-layer clay mineral. Its structure is an alternation of trioctahedral TOT layers, with Mg^{2+} and Fe^{2+} in the octahedral positions, and layers with the composition $(Mg^{2+}, Fe^{2+})_3(OH)6$, giving a repeat spacing of about 14 Å. There's a great range and diversity of ionic substitutions in each of the three ion sites, as well as differences in the geometry of stacking of the sheets Chlorite is produced in abundance only in relatively mild weathering environments, where it survives from the source rocks, although in the form of finer particles. So it's characteristically a high-latitude product.

2.6.9

Various oxides of iron and aluminum are especially important in the highly weathered soils of warm and humid regions, where (as you saw in the section on weathering) silicate minerals are broken down further to oxides. The main aluminum oxide is **gibbsite**. The formula for gibbsite, $Al(OH)_3$, shows that it is a hydroxide mineral. Gibbsite is formed by further weathering of silicate clay minerals, in which all of the silica is stripped out of the clay minerals; see Reaction 9 in the section on chemical weathering, above. The main iron oxide mineral in soils is **goethite**, FeOOH. It's goethite that imparts the characteristic yellow-orange color to weathered iron-bearing rocks and regolith. Recall from the section of chemical weathering that ferrous iron is freed from ferrous-iron-bearing silicate minerals and then oxidized to minerals like goethite. These aluminum and iron oxides (there are others, as well) in regolith are in the form of extremely small particles, in the colloidal size range. (See the later section on soils for more on colloids.)

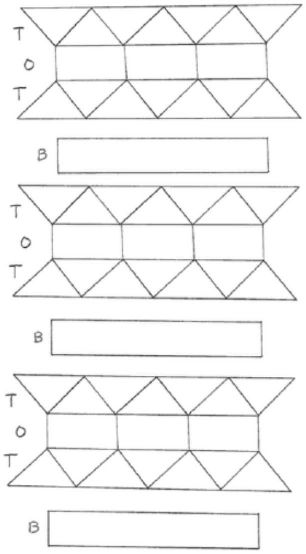


Figure 2-18. Highly simplified sketch of the structural organization of chlorite.

2.7: Regolith

The Nature of Regolith

What is regolith? The term ***regolith*** is used for *the layer or mantle of fragmental and unconsolidated rock and mineral material, whether residual or transported, that rests on bedrock*. One might quibble that the definition excludes fragmental material that happens to be locally cemented by surficial processes but is otherwise closely related to nearby unconsolidated material. (You'll be hearing more, much later in the course, about one such kind of material: caliche.) The definition is not specific about age; most regolith is young, geologically and even by human standards, but some is much older, not having been subjected to consolidation, for one reason or another, for geologically long times.

For most of us geologists, continental bedrock is our bread and butter: we inspect it minutely, hammer on it, sample it, photograph it, and think about it day and night. But in most parts of the world, finding bedrock to inspect is not easy, and in many places almost impossible: it is covered by a layer of regolith. When was the last time you walked across an outcrop of bedrock? When we are out and about in the natural world, away from the ubiquitous pavement that's so characteristic of our civilization, we are mostly walking on regolith. If only for that reason alone, study of the Earth's regolith should be high on the list of desiderata for Earth scientists. A good case can be made that *the Earth's mantle of regolith is for humankind by far the most important component of the solid Earth beneath our feet*: it feeds us, in the form of soil. In fact, however, there is a tendency for regolith studies to be low on the priority list. (I think that there lurks in the minds of many geologists the notion that bedrock studies hold far more interest, and glory, than regolith studies.)

Regolith originates by the complex of processes collectively termed weathering, which was the subject of the preceding section. What happens to regolith once it is produced? You can detect from the foregoing definition that regolith is either *residual* or *transported*. The terms are largely self-explanatory, are they not? In areas of low relief and with a climate conducive to deep weathering, a thick mantle of regolith, in some places many tens of meters thick, lies where it has been produced. In most places, however, the regolith we see has been moved from its place of origin, nearby or far away, by various agents of transportation.

Keep in mind, however, that it's usually not a matter of an episode of production and then a later episode of transportation: generally, both production and removal are going on at the same time. The thickness of the mantle of regolith depends in great part on the relative magnitude of the rate of production and the rate of removal. In some situations, as in regions with steep slopes, high relief, and rigorous climate, regolith is stripped away as fast as it is produced; in areas with gentle slopes and climates conducive to deep weathering, the layer of regolith is very thick.

Agents That Mobilize Regolith

Regolith resting on a sloping surface is pulled inexorably downward by the force of gravity (or, more accurately, by *the downslope component of the force of gravity*) (Figure 2-19). Such movement of the regolith may be imperceptibly slow or at speeds in excess of a hundred meters per second. An entire later chapter will be devoted to the subject of downslope movements of near-surface material by gravity, collectively termed ***mass movements*** (absolutely no relation to sociopolitical mass movements).

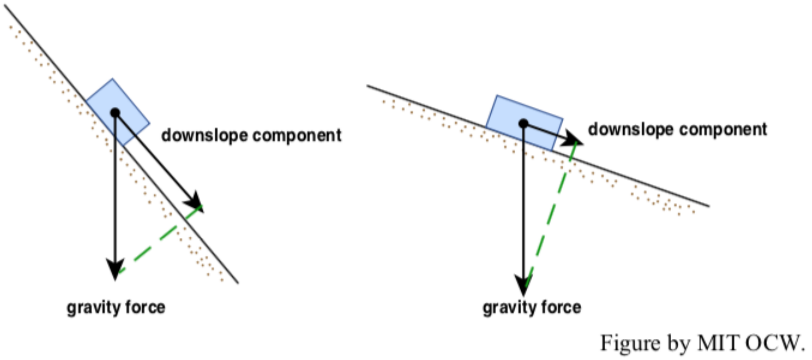
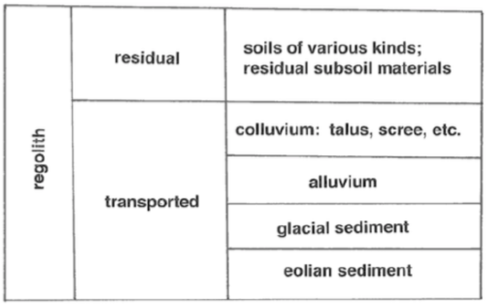


Figure 2-19. The downslope component of gravity acting on the layer of regolith. A) On a steep slope. B) On a less steep slope.

The principal agent for the transport of regolith, aside from the direct pull of gravity, is *the flow of water in streams and rivers*. Of course, river flow is itself a consequence of the downslope pull of gravity—but regolith transport by flowing water involves physical processes and effects that are fundamentally different from the direct pull of gravity. The mobilization and transportation of regolith by flowing water is the subject of part of the later chapter on streams and rivers. Movement of regolith by the action of land-based glaciers is of great importance in certain Earth-surface environments as well. Wind can also be a significant transporting agent, as described in the final chapter.

Kinds of Regolith

The main distinction in kinds of regolith is between ***residual regolith*** (also called ***sedentary regolith***), on the one hand, and ***transported regolith***, on the other hand (Table 2-1). With respect to transported regolith, several kinds are recognized, depending upon the agent of transportation. Transported regolith is collectively termed ***sediment***. There will be much more material, later in this chapter as well as in later chapters, on sediment, and how it is eroded, transported, and deposited. Colluvial deposits (called ***colluvium***) are those transported down steep slopes by the pull of gravity. Mainly, these are talus deposits (also called scree deposits) on steep slopes, and avalanche and landslide deposits of various kinds. Alluvial deposits (called ***alluvium***) are those transported and deposited by rivers and streams. In addition, there are glacial deposits and eolian (wind-blown) deposits.



regolith	residual	soils of various kinds; residual subsoil materials
	transported	colluvium: talus, scree, etc.
		alluvium
		glacial sediment
		eolian sediment

Table 2-1. Classification of regolith. (After Merrill, 1897.)

2.8: Sediment

When regolith is mobilized or entrained, by flowing water, or by the wind, or by moving glacier ice, it is called ***sediment***, because, as the term implies (the Latin verb *sedere* means to sit), it eventually comes to rest again as a deposit. Here I will touch upon only a few aspects of sediments. You will hear much more about sediment movement, and sediment deposits, in later chapters, especially the chapters on rivers, on glaciers, and on coasts.

One of the characteristic things about sediment is that *it tends to become fractionated, by size, during transport*. The basic reason is that a given transporting agent, like water or the wind, moving at a given speed, can move sediment particles up to a certain size but no larger—although another important effect is that the water or the wind moves finer particles faster than coarser particles. The end result of such fractionation is that the sorting of the particles in the deposit is usually much better than in the original material that was the source of the transported material. (The ***sorting*** of a sedimentary deposit describes the spread of particle sizes around the mean particle size: in a well sorted sediment the particle sizes cluster closely around the mean, whereas in a poorly sorted sediment the particle sizes range widely around the mean.)

For convenience and clarity in communication, sedimentary geologists have adopted a series of terms for sediment particles over the entire range of particle sizes (Figure 2-21). Such a subdivision is called a ***grade scale***. It's based on powers of two, so it's geometric (or, what comes out mathematically to be the same, logarithmic), rather than arithmetic, for reasons that are obvious but not easy to justify: for example, the difference in behavior between a 1 mm particle and a 2 mm particle is bound to be far greater than the difference in behavior between a 101 mm particle and a 102 mm particle, right? You can see from Figure 1-20 that in geoscience such things as mud, silt, sand, and gravel are not just qualitative concepts but precisely defined entities!

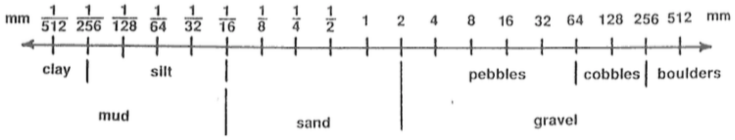


Figure 2-20. The Udden–Wentworth grade scale for sediment particle size.

The logic behind the grade scale in Figure 2-20 is fine for well-sorted materials, but how about poorly sorted materials? Various classifications have been proposed, but the one that's in most common use (with some variations) among regolith specialists is shown in Figure 2-21. Here you see what "loam" (so commonly pronounced "loom" locally!) really is, in the technical sense.

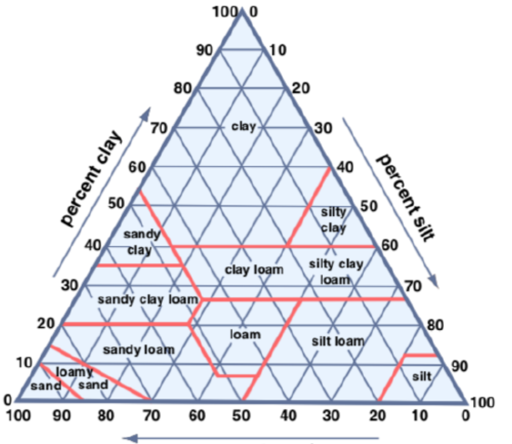


Figure by MIT OCW.

Figure 2-21. Conventional system for classification and terminology of sediment mixtures. (From FitzPatrick, 1986)

Advanced Topic: Measuring Sediment Size

1. Scientists or engineers who are responsible for studying surface deposits usually have to measure or characterize, in some way, the distribution of particle sizes. For large particles, in the gravel size range, it's possible to measure the size of each piece separately, with calipers, although that's a challenging endeavor unless the particles is very well rounded, and even then the matter is not entirely straightforward (as it would be if the particles were perfect triaxial ellipsoids). For smaller particle sizes, in the silt to fine pebble range, the classic way of measuring particle size is by ***sieving***.

2. Here is how particle size is measured by sieving:

- make up a stack of special sieves, from coarsest on the top to finest on the bottom, with a lid on the top and a pan on the bottom;
- put a small sample of deposit in the top sieve;
- shake the stack until the particles have found their size bin;
- weigh each of the fractions on a balance;
- compute, for each sieve size, the total weight percent finer than that size;
- plot the results on graph paper with size along the horizontal axis, logarithmically, and, along the vertical axis, weight percent finer than, from 0% to 100%, arithmetically.

An example of the result, called a ***cumulative distribution curve*** (common in many other applications of statistics as well), is shown in Figure 2-22.

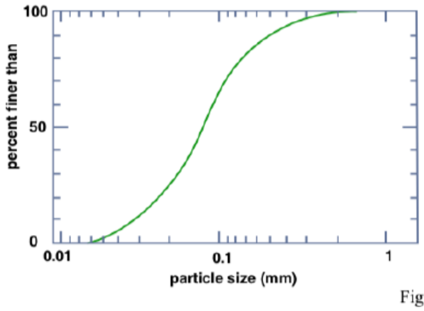


Figure by MIT OCW.

Figure 2-22. A typical cumulative distribution curve of sediment size.

3. The problem with sieving is that it works well only for sizes down to about 0.05 mm. For sediment finer than that, there are various techniques that involve suspension of the sample in water, settling of that suspension, and sampling during settling. The basic principle is that the coarser sizes settle faster than the finer sizes. Devices that are used to measure fine particle sizes in this way are called ***settling tubes***. The governing principle is that each particle has a terminal fall speed that is a function of its size. In recent years, a variety of more sophisticated instruments have been developed to measure particle size, either optically or sonically, by sensing each particle as it passes through a thin tube in suspension. Such instruments are now obtainable "off the shelf". There is still the tricky matter of intercalibrating between results from such instruments and traditional settling tubes.

2.9: Soils

2.9.1 Introduction

Along with oxygen and water, soils are essential to human existence. Almost all of our food comes, directly or indirectly, from crops grown in the soil. The only significant exception, of course, is food from the oceans. Hydroponically grown crops are an insignificant exception—and the infrastructure and cost involved in hydroponic agriculture are likely to doom it forever to a minor role in human food supply. All of our lumber, most of our paper products, and much of our clothing ultimately come from the soil. Our dependence on soils will continue, but the world stock of soils cannot increase. In fact, it is *decreasing*, because of soil erosion, as well as urbanization (and suburbanization). As our reserves of fossil fuels continue to shrink, energy resources from biomass (living plant material) are likely to increase.

The meaning of the term “soil” varies greatly depending upon the field of endeavor in which it is used. *Agriculturalists* use the term narrowly to mean the uppermost layer of surface deposits, in which plants grow and agriculture is practiced. The *soil scientist* has a somewhat wider view: soil is the uppermost layer of surface deposits, almost always extending well below the agriculturalist’s soil, which has developed upon underlying material by the influence of various soil-forming processes acting over long times. The *geologist* has an even broader view of soils as surface materials developed by weathering processes that have broken down bedrock into regolith, and the *civil engineer* (who practices the discipline of soil mechanics, and who is responsible for the siting and design of structures with foundations) takes a similarly wide view of soil as all materials at the surface that can be excavated without blasting. In the following sections, we’ll look at the nature of soils mainly from the standpoint of the soil scientist (what is the nature of soils, how do they develop, and how are they classified?).

The study of soils as natural bodies is called ***pedology*** (and its practitioners call themselves soil scientists, or pedologists). You can also think in terms of the ***pedosphere***. (There is a tendency among science educators nowadays to think in terms of “spheres”: geosphere, hydrosphere, biosphere, atmosphere, cryosphere. These terms are useful, in a broad and loose sense, to describe zones or domains of related materials and processes. I think that the suffix “*-sphere*” began to be used because in a very general sense these domains are like spherical shells around the Earth. I myself think that “sphere-ology” tends to be overdone on the part of the “spherophiles”, but it remains a useful pedagogical device.)

Soils can be thought about and studied on a wide range of scales: from large regions of continents, in the context of major soil types, to the distribution of soils in a local area of varied topography and substrate composition (Figure 2-23A), to the vertical profile of soil at a given point on the land surface (Figure 2-23B), to the texture and structure of the soil at a point in such a profile (Figure 2-23C), to the details of physics, chemistry, and biology of the soil on the colloidal and atomic scale (Figure 2-23D).

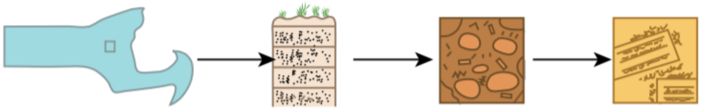


Figure by MIT OCW.

Figure 2-23: Thinking about soils on a wide variety of scales. (Inspired by Brady and Weil, 2002.)

Soils have a great variety of constituents. Here is a list of the most important kinds of soil constituents:

- mineral particles
- colloids
- organic matter
- microorganisms
- soil solutions
- soil gases

Most *soils consist mainly of mineral particles*. Such soils are often called ***mineral soils***. Some soils, however, consist mostly of organic matter. In general, soils contain the same kinds of mineral particles as do sediments; go back and review the section on sediments above. Most of the soils you are likely to encounter consist mainly of quartz, plus variable percentages of potassium feldspar and rock fragments, in the coarse fraction, and mainly particles of clay minerals, iron oxides, and aluminum oxides, in the fine fraction. The remaining constituents in the list above are described in the following sections.

2.9.2 Colloids

Colloids are *mixtures of materials in which one substance exists as extremely small particles, called colloidal particles, which are dispersed in a second substance, which you can think of as the medium*. In the case of colloidal particles in a uniform liquid or gaseous medium, the term colloidal solution is commonly used. In the context of soils, we're dealing with small particles consisting of minerals or organic matter that are dispersed in the aqueous pore solutions in the soil or are aggregated together in great numbers, or cling to the larger soil particles. Colloids are important in a great many materials, both natural and manufactured.

Most clay-mineral particles have a platy shape, and they have a tendency to stick together in aggregates of many particles, held together by certain short-range electrostatic forces. Such forces are small, but so are the particles— and it's important to realize that because *the ratio of surface area to volume increases with decreasing particle size*, clay-mineral particles are in a sense "almost all surface and very little volume". The state of aggregation of the clay- mineral fraction is widely variable, and the macroscopic strength and consistency properties of the deposit vary greatly corresponding to that. Also, as you will soon see, the small size of the particles, together with the electrical charges characteristic of their surfaces, make the tiny particles chemically extremely reactive.

Colloidal particles are in a kind of "no person's land" between true solutions and undoubtedly solid particles. At one extreme are ions and molecules in an aqueous solution. The smallest ions are of the order of atomic dimensions, about 0.1 nanometers (that is, about 10-10 meters). Many molecules that consist of a large number of atoms are much larger, but they still behave like smaller ions in true solutions. At the other extreme are particles larger than a few micrometers, which is large enough to be seen with light microscopes. Such particles are large enough to settle in the surrounding fluid under the influence of gravity.

Typical soils contain several important kinds of colloids. Here's a briefly annotated list of the most important:

- ***clay-mineral particles:*** There are several types; see the earlier background section on clay mineralogy. These types differ in their physical and chemical properties, partly because of their differing size but mainly because of their differing surface charges and their differing susceptibility to take on or release positive ions and water molecules to and from their internal surfaces between the TOT layers.
- ***iron and aluminum oxides:*** These are especially important in the highly weathered soils of warm and humid regions, where silicate minerals are broken down further to oxides. The main aluminum oxide is *gibbsite*, and the main iron oxides are *goethite* and *ferrihydrite* (no details of mineralogy here).
- ***humus:*** These are especially abundant in the upper parts of soils, nearest the surface. Humus colloids are not crystalline structures: they are complex molecular chains and rings of carbon atoms bonded to hydrogen, oxygen, and nitrogen atoms. They are among the smallest of colloidal particles in soils. They have an enormous capacity to adsorb water onto their surfaces.

The great importance of soil colloids arises from their chemical reactivity. There are two reasons for that: *surface area*, and *surface charge*. The following elaborates on those points.

- ***surface area:*** Chemical reactions between a solid and the fluid with which it is in contact take place at the surface of the solid, at the interface between the solid phase and the liquid phase. The rate of such a chemical reaction depends upon a number of important factors, aside from the nature of the reaction itself. Temperature is one of those important factors: in general, the higher the temperature, the faster the reaction. That's because the energy of thermal vibration of the atoms and molecules in both the solid phase and the liquid phase increases with temperature. Also important, however, is *the area of the surface at which the reaction takes place, per unit volume of the reacting system*. The extremely small size of colloidal particles means that their surface area, per unit bulk volume of material, is spectacularly large.
- ***surface charge:*** At any given tiny region inside a mineral particle, the electric charges associated with the ions of the structure balance out. At the surface of the particle, however, the picture is different—because the ions right at the surface are bonded only to the ions below them. That means that there is a kind of "unsatisfied" electric charge at the surface of the particle. With regard to clay-mineral particles, the surface charge is generally negative on the broad faces of the exposed TOT sheets and positive along the narrow edges of the plates. For macroscopic mineral particles, like sand grains, these surface charges are almost entirely negligible, but when the particles are extremely small their enormous area-to-volume ratio makes these surface charges very important.

Background: Surface Area

Take a nostalgia trip back to secondary-school math and think about the formulas for the volume and the surface area of some regular geometric solid. A sphere is a good example. The volume of a sphere is $(1/6)\pi D^3$, where D is the diameter of the sphere. In scientific parlance, *the volume goes as the cube of the diameter*. The surface area of the sphere is πD^2, so the surface area goes as the *square* of the diameter.

What's important here is not the particular formulas, or the values for volume and area of a particular sphere they give, but *how the ratio of surface area to volume varies with the diameter of the sphere*. You can easily see from the two formulas given above that the ratio of surface area to volume of any sphere is $[\pi D^2]/[(1/6)\pi D^3]$, or, simplifying, $6/D$. Ignore the factor 6 there; what's important is the $1/D$ part. The ratio of surface area to volume goes as $1/D$. That means that *as the diameter decreases, the ratio of surface area to volume gets bigger*.

Now for a revealing example of this $1/D$ effect. Think about a sphere 10 cm in diameter—about the size of a softball, a grapefruit, or a ball of pizza dough. It has a surface area of about 0.0314 square meters (or, in scientific notation, 3.14 x 10^{-2} m^2). Now compare that with a sphere 1 mm in diameter. That sphere has a surface area of only 3.14 x 10^{-6} m^2, far smaller, by a factor of ten thousand. But suppose you took the entire volume of the 10 cm sphere and converted it into an equivalent volume of 1 mm spheres. With some further simple math, and using the formula for the volume of a sphere, you would determine that the 10 cm sphere is equivalent to 106 (that's one million) of the little ones. Now the total surface area of the little ones is (106)(3.14 x 10-6) m^2, or 3.14 x 100 m^2. That doesn't sound like a lot, but compare it with the surface area of the original 10 cm sphere, which was 3.14 x 10^{-2} m^2. We see, then, that the aggregate area of the little spheres is a hundred times the area of the volume-equivalent big sphere! This is a concrete example of the $1/D$ effect: we reduced the diameter of the sphere by a hundred (ten centimeters to one millimeter), and the surface area of the equivalent volume increased by a factor of a hundred!

The combination of surface area and surface charge is a potent one. We need to deal with two different but related effects: coagulation, and ion exchange. Figure 2-24 is a cartoon that shows the various phenomena discussed in the following paragraphs.

coagulation: Clay-mineral particles, with their negative charges on faces and positive charges on edges, tend to clump together by electrostatic attraction, in an edge-to-face arrangement. (Keep in mind that like charges repel and unlike charges attract.) This tendency for coagulation is one of the important factors that make clay-rich soils sticky and tenacious.

ion exchange: As you will see in a later chapter, on water, the water molecule has a positively charged "side" and a negatively charged "side". Because of that, water molecules tend to be attracted to the charged surfaces of the colloidal particles. Picture a swarm of water molecules, perhaps up to several molecules thick, at the clay-particle surface, jiggling around and jostling one another, now and then trading places with other water molecules from the solution beyond. That layer of water molecules acts as a kind of "cushion" for the particle. Included in the layer of loosely bound water molecules are positive ions, which are attracted to the negatively charged faces of the clay particles. Also, smectite clays take up and give off water and positive ions to and from their internal surfaces. These loosely bound water molecules are the last to be extracted by plant roots as a soil dries between rains. You know that potassium is one of the most important plant nutrients. Its continued abundance in the soil is largely the result of its attraction to colloids, together with recycling in organic matter. Because of this combination of attraction to colloids and recycling by plants, under natural conditions the storehouse of available potassium, released by chemical weathering of primary minerals, is maintained at a workable level for plant growth in the face of continuing loss in solution to downward-percolating water from the surface.

2.9.3 Organic Matter

The uppermost parts of surface deposits—the soil of the soil scientist and especially of the agriculturalist—usually has an abundance of organic matter as well as inorganic mineral matter. It's not easy to specify what's meant by organic matter. It comprises a wide and complex variety of chemical compounds that a chemist would call organic compounds. Organic compounds are those that consist of carbon atoms, bonded to one another and to atoms of other elements, mainly oxygen and hydrogen but including such elements as nitrogen and sulfur as well, in a great variety of ways. Organic compounds are both natural and laboratory-synthesized.

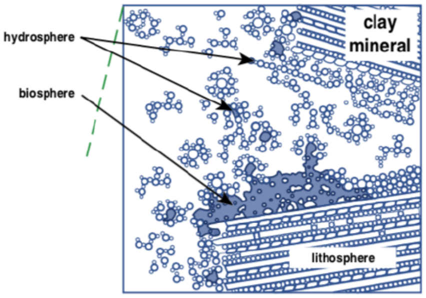


Figure by MIT OCW.

Figure 2-24. Submicroscopic view of a clay-mineral colloid particle in its aqueous environment.

The organic matter of soils is *ultimately derived from plant growth*. (Living plants and animals are not conventionally considered to be organic matter: you have to wait until the organism dies!) Nonliving plant tissues themselves are organic matter, as are the great variety of organic compounds that result from partial decay of that plant material. The tissues of dead animals and their decay products also represent soil organic matter, although not nearly in the same abundance as plant-derived organic matter.

A deep understanding of soil organic matter would necessitate a much greater familiarity with organic chemistry (the branch of chemistry that deals with organic compounds) than is appropriate for this course. Here we can only touch upon some of the most important considerations and kinds of materials.

2.9.3.2 The Role of Soil Organic Matter in the Carbon Cycle

The element carbon is a key substance in the workings of the Earth's near-surface environment. That comes about because of its important role in living organisms, soil organic matter, the atmosphere and the oceans in the form of gaseous and dissolved carbon dioxide, and carbonate sediments. There will be a more detailed treatment on the Earth's carbon cycle in a much later chapter on geochemical cycles. It must suffice here to point out that the soil plays a major role in that carbon cycle. It's been estimated that the Earth's soil contains almost twice as much carbon as the stock of living vegetation and the atmosphere combined.

Plants take up carbon dioxide gas from the atmosphere and convert it, by sunlight-powered photosynthesis, into the organic compounds that constitute plant material. When the plants die, some of the dead plant material is decomposed back to carbon dioxide and water by oxidation. In part that decomposition involves non-biological chemical reactions, but in great part it is mediated by the metabolic activity of soil microorganisms. The end result, however, is the same: production of CO_2. Some of that CO_2 remains dissolved in soil water and some is released into the atmosphere.

2.9.3.3 Kinds of Soil Organic Matter

There are many kinds of organic matter. These compounds range from easily decomposable to very difficultly decomposable. A typical plant consists of easily transformed compounds like polysaccharides and lipids, stiffened with the much more resistant materials cellulose and lignin, the principal components of woody plant tissue. Lignins, of which there are various kinds, are the slowest plant tissues to decompose. Only a few kinds of organisms, mainly fungi, can break down lignins.

The ultimate residue left behind by breakdown of plant materials is called ***humus***. What's called humus comprises a number of very stable organic compounds. Humus is always dark in color, the reason being that the chemical structures in the humus absorb nearly all wavelengths of visible light. Humus is so fine grained that its particles (which in reality are large molecules) are at the fine end of the colloidal size range. These compounds, whose basic building blocks are carbon hexagonal ring structures (called aromatic compounds, in organic chemistry) with large molecular weights (that is, containing very large numbers of polymerized atoms in their structures). They are the organic materials that are the most resistant to microbial decomposition. The water-holding capacity of humus, mentioned in an earlier section, is several times greater than that of clay-mineral particles.

Decomposition of humus is a slow process. The typical half-life of humus (that is, the time it takes to decompose half of an initial mass of the material) is measured in many decades, or even a few centuries! An important effect that slows the decomposition is complexing of the molecules of the humus with inorganic clay-mineral particles. By poorly understood processes, this association protects the humus from decay.

2.9.3.4 The Role of Organic Matter in Soils

With respect to plant growth, humus is an essential soil component. It's common knowledge among gardeners that plants grow better in soils rich in humus. There are various reasons for that, both direct and indirect. It's not just a matter of supply nutrients: humus is not particularly rich in nutrients (despite what many organic gardeners have been led to believe). Most of the nutrients needed by plants are inorganic ions dissolved in aqueous soil solutions, although humus does play some role in the nutrient supply for plants.

It's the indirect effects of humus on soil properties that are the most beneficial aspects of humus. Humus can be viewed as a soil conditioner. These effects involve physical, chemical and biological properties of the soil In terms of *physical effects*, humus reduces the stickiness and cohesion of clay-rich soils, and it promotes the formation of granular texture, so prized by agriculturalists because it enhances the tillability of the soil. In terms of *chemical effects*, humus is even more effective at adsorbing and holding onto nutrient ions than are clay-mineral particles. In terms of *biological effects*, humus is the most important substrate for the life activities of soil microorganisms; it provides most of the food for detritus-feeding soil organisms.

If you are willing to be overwhelmed by the multitude of ways that organic matter affects soil properties, plant productivity, and environmental quality, look at Figure 2-25.

2.9.4 Soil Organisms

The last really important component of the deposit comprises soil organisms: the plants and animals that spend part or all of their lives in the soil. The soil is one of the Earth's most important ecosystems. (For a brief brush-up on ecology and ecosystems, see the background section below.) Every handful of soil contains literally billions of organisms—almost all of them microorganisms, too small to be seen with the unaided eye. The microbiology of soils is a fully developed and very important discipline in itself, and I can only touch upon it here.

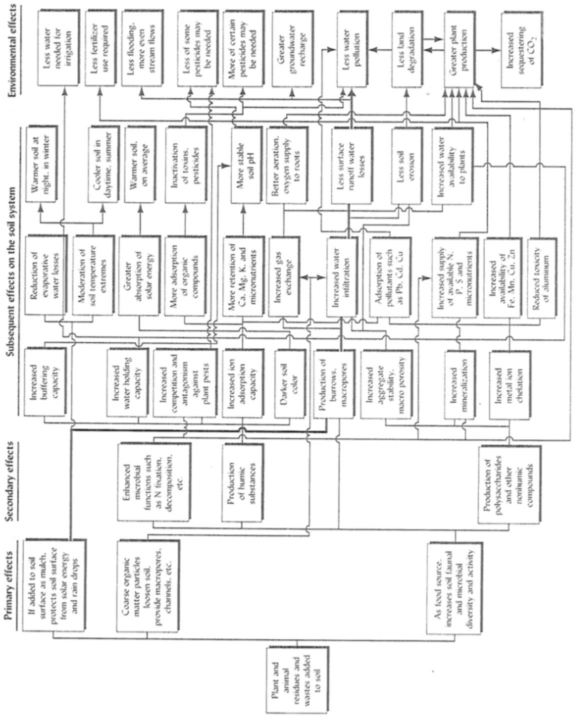


FIGURE 12.15 Some of the ways in which soil organic matter influences soil properties, plant productivity, and environmental quality. Many of the effects are indirect, the arrows indicating the cause-and-effect relationships. It can readily be seen that the influences of soil organic matter are far out of proportion to the relatively small amounts present in most soils. Many of these influences are discussed in this and other chapters in this book. The thicker line shows the sequence of effects referred to in the text in Section 12.7. (Diagram courtesy of R. Weil)

Figure 2-25. How organic matter affects soil properties, plant productivity, and environmental quality. (From Brady and Weil, 2002.)

Background: Ecology and Ecosystems

1. The term ***ecology***, which tends to be used too loosely in the popular media these days, is the study of *the interrelationships between organisms and their environment*. The largest entity studied in ecology is the ***ecosystem***, which consists of *some*

> *chosen part of the physical and chemical environment and all of the organisms within it*. An ecosystem involves all of the physical, chemical, and biological processes and interactions that operate within the given part of the environment. An ecosystem can be as large as the entire biosphere or as small as a puddle of water at our feet.
>
> **2.** What are the elements of an ecosystem? A ***habitat*** is *the local environment in which a given organism lives*. Ecosystems usually involve a number of different and distinctive habitats. A similar but different concept is that of an ***ecological niche***, which is *the position of the organism in its habitat*, including its way of life and its role in the ecosystem. Most habitats are occupied by several species, each with its own ecological niche. Each species in an ecosystem is represented by a number (usually, but not always, large) of individual organisms, called a ***population***. A population of two or more species that occupy a given habitat is called a ***community***. There may be more than one community in a given ecosystem.
>
> **3.** Within any ecosystem, there typically are many kinds of interactions among constituents, both living and nonliving. These interactions can be viewed in terms of flows of matter and energy through the ecosystem. Organic compounds are synthesized from the environment by ***producers***, which in all but the most specialized ecosystems are photosynthesizing plants. The producers are in turn consumed by plant-eating animals, called ***herbivores***. Some of the herbivores are in turn consumed by ***carnivores*** (or by ***omnivores***, which each both plants and animals). Other elements of the ecosystem are ***parasites***, which feed on living organisms without killing them, and ***scavengers***, which feed on dead organisms.
>
> **4.** The sequence of species ranging from the producers, at one end, to carnivores that no other carnivores eat, at the other end, is called a ***food chain***. Often the term ***food web*** is more appropriate, because the actual situation is more complicated than a simple linear arrangement of species. The tissues of producers, herbivores, and carnivores that are not consumed by species higher in the food chain are broken down by organisms called ***decomposers***, which are usually bacteria.
>
> **5.** When we think of ecosystems, we tend to think of tropical rainforests, or broad expanses of grasslands on which large mammals roam around and get photographed by ecotourists. In a real and important sense, however, *a shovelful of natural soil is an ecosystem*, and a very complex one at that.

One way of classifying soil organisms is into plants and animals. There’s more to that than it might seem, however, because although we all learned, when we were small, that life is divided into an animal kingdom and a plant kingdom, most biologists these days recognize *five* kingdoms. But that’s another story we need not deal with in this course. There are two more useful ways, for us, of classifying soil organisms:

- *heterotrophs vs. autotrophs:* ***autotrophs*** are organisms that generate their own tissues from inorganic raw material, by use of (mainly) solar energy. Plants are autotrophs. ***Heterotrophs*** are organisms that use organic matter produced by other organisms as food.
- soil organisms can be classified as ***macroorganisms*** (organisms with size ranging down to that of, say, earthworms), ***mesoooorganisms*** (organisms ranging down to a small fraction of a millimeter), and ***microorganisms*** (organisms of microscopic size, smaller than about a tenth of a millimeter).

The numbers of kinds of organisms per unit volume of a soil increases strongly with decreasing organism size. A typical soil might contain several species of vertebrates and several species of earthworms, along with as many as several tens of species of insects, dozens of species of mesoorganisms, but thousands of species of microorganisms. Moreover, the numbers of organisms per unit volume increases spectacularly with decreasing size: although, obviously, numbers vary widely depending upon the type and condition of the soil, numbers of macroorganisms and mesoorganisms are large but not overwhelming, whereas the numbers of microorganisms, per cubic meter, say, might be in the trillions!

A whole chapter should be devoted here to the nature and significance of the life activities of soil organisms. Interesting as they are, the activities of what have been called ***ecosystem engineers*** (vertebrate burrowers, earthworms, ants, termites, and the like) in modifying the physical nature of the soil pale in comparison with the life activities of soil microorganisms. Microorganisms play the essential role of decomposing soil organic matter (mainly plant residues) and converting it to carbon dioxide, water, dissolved ions, and the ultimately resistant residues called humus earlier in this section.

Several major kinds of microorganisms are well represented in soils. The most important are *algae*, *bacteria*, and *fungi*, and various kinds of *protozoans*. Algae, which are photosynthetic plants, are abundant only in the uppermost parts of deposits, very near the sunlit surface. Fungi and protozoans are important but restricted to the shallow zones. Bacteria, on the other hand, range widely downward to great depths.

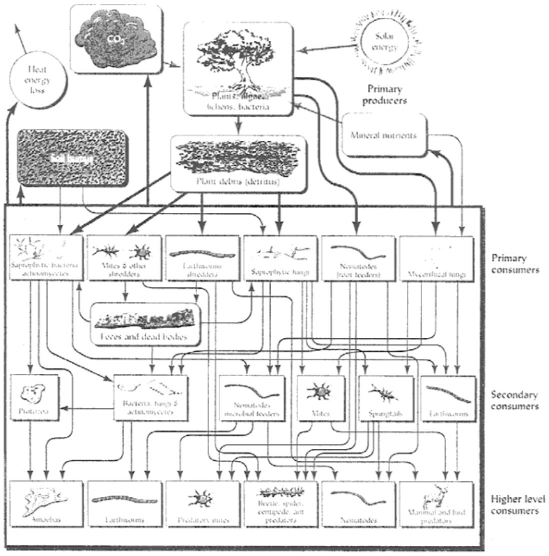


FIGURE 11.1 Greatly simplified and generalized diagram of the soil food web involved in the breakdown of plant tissue, the formation of humus, and the cycling of carbon and nutrients. The large shaded compartment represents the community of soil organisms. The rectangular boxes represent various groups of organisms; the arrows represent the transfer of carbon from one group to the next as predator eats prey. The thick arrows entering the top of the shaded compartment represent primary consumption of carbon originating from the tissue of the producers—higher plants, algae, and cyanobacteria. The rounded boxes represent other inputs that support, and outputs that result from, the soil food web. Although all groups shown play important roles in the process, some 80 to 90% of the total metabolic activity in the food web can be ascribed to the fungi and the bacteria (including actinomycetes). As a result of this metabolism, soil humus is synthesized, and carbon dioxide, heat energy, and mineral nutrients are released into the soil environment.

Figure 2-26. The soil food web. (From Brady and Weil, 2002)

Figure 2-26 shows a diagram of the soil food web. I've preserved the original caption for you. Note that it's described as "greatly simplified and generalized"!

2.9.5 Soil Solutions and Soil Gases

All soils are porous to some extent (that is, they contain empty spaces, called ***pore spaces*** or ***void spaces***, among the various solid constituents). It stands to reason that the pore spaces must contain either a gas or a liquid, or some of both. Whether the pore spaces are occupied by solely by liquid or whether they are occupied mainly by a gas depends upon where the region of soil lies relative to the groundwater table: below the groundwater table the soil is fully saturated with an aqueous pore solution, whereas above the groundwater table the pore spaces are occupied mainly by gas.

The reason I have used the term "aqueous pore solution" rather than just "pore water" is that the water in the pore spaces contains a great variety of dissolved substances, in varying concentrations. Just after a rainstorm, as the rainwater percolates downward to replenish the water table, the concentrations of dissolved materials is very low (except for dissolved atmospheric gases), but as time goes on some of the constituents of the soil particles are leached out to go into solution. The chemistry of such soil solutions is a important part of soil science.

Above the groundwater table, the pores spaces are filled mainly with gas. The composition of soil gases is not greatly different from that of the atmosphere, with one major exception: owing to continuing decay of organic matter, the concentration of carbon dioxide is typically substantially higher than in the overlying atmosphere.

Even in the soil zone above the groundwater table, the soil is seldom entirely dry: some residual water clings tenaciously to the soil particles. I will defer discussion of that until the chapter on groundwater.

2.9.6 Processes of Soil Development

The development of a soil (in the agriculturalist's and soil scientist's sense) depends upon five major (and interacting) factors, as shown in Figure 2-27. Of these, you might think that the parent material is the most important. It's generally believed, however, that climate is even more important than parent material. The relative importance of climate and parent material depends to a great extent upon time of development: in the initial stages of soil development, parent material is certainly the more important, but, as time goes on, soils developed upon different parent materials in a given climate tend to become more and more similar. Obviously, important differences as a function of parent material must remain: if for one soil the parent material is rich in quartz, and for another soil the parent material is a limestone or a mafic igneous rock like basalt, the soils can never converge.

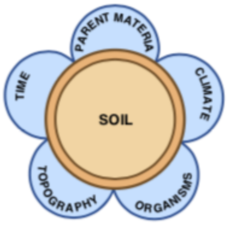


Figure by MIT OCW.

Figure 2-27. The five factors of soil development. (From FitzPatrick, 1986.)

Topography is generally of secondary importance; its role lies largely in governing the downslope creep of already developed soil, which tends to "freshen" the uppermost layer of the surface deposit by exposing less weathered material beneath.

The importance of organisms lies in their ability to add the all-important organic matter to the soil and to process that organic matter back into inorganic compounds, including nutrients for yet more plant growth.

I might also point out that *the five factors are not entirely independent of one another*; for example, organisms depend strongly on climate, and topography depends quite strongly upon climate as well, and also upon parent material and time.

This is a good place to point out that *some soils develop directly upon bedrock, but in many if not most cases, soil develops upon surface deposits, called sediments earlier in this chapter, that were derived from bedrock somewhere else and transported to the site of soil formation*. In some places, the surface deposit is grossly "out of place". That's true of almost all glaciated areas, New England being no exception. It's also true of areas of deposition from streams and rivers. In such areas, soil formation is usually nowhere near the "ultimate" or "equilibrium" soil, of the kind you would find in a low-latitude, non-glaciated area of low relief, where bedrock is weathered very deeply over a very long period of geologic time and the soil is highly developed.

(But just because a soil has had a long time to develop doesn't mean that it will be rich and deep: in the rain forests of the humid tropics, intense leaching removes nutrients released by chemical weathering so rapidly that the soil is thin, and has only a thin layer rich in organic matter at the surface; the lush stand of plants is in tenuous existence, recycling nutrients very efficiently. When such vegetative cover is removed, the soil rapidly becomes irreversibly barren.)

A variety of physical and chemical processes act to develop soil from fresh regolith. One of the most important of these we have already dealt with: weathering of preexisting bedrock and/or further weathering of relatively fresh and unweathered surface deposits. Another important process is vertical transport of matter in either dissolved or colloidal form by upward or downward flow of soil water. Such processes come under the soil scientists' term ***translocation***. Translocation also includes vertical mixing or churning by macroorganisms, most importantly earthworms and certain kinds of insects. Biological processes, most importantly ***humification*** of plant material, and chemical and organism-aided processes of ***cycling*** of nutrients and other compounds (most importantly nitrogen and carbon) are further soil-forming processes.

The nature of *water flow in the vertical direction in the soil*, which is largely a function of climate (intensity and frequency of rainfall; the effect of temperature, humidity, and vegetative cover upon evaporation from the soil), determines whether matter in solution undergoes net downward transport or net upward transport in the soil. This is the aspect of climate that's most important in soil development, along with the nature and rate of growth of vegetation.

2.9.7 Soil Profiles

Soil scientists classify soil on the basis of the sequence and nature of vertical zones or layers, called ***horizons***, in the soil. Each of the enormous variety of soil types has *a characteristic succession of such horizons*, each with its thickness, composition, physical properties, and transitions to overlying and underlying horizons. The particular succession of horizons is called a ***soil profile***.

To have a good picture of a soil profile, you need to go into your backyard and dig a trench, carefully, about a meter deep and wide enough for a good side view, with one side planar and almost vertical and shaved off carefully for good viewing with minimal smearing of soil material on the viewing face. Excavations for foundations, or even big holes for planting of trees or shrubs, can serve the purpose well. For the best and most representative results, you need to be sure to dig at a spot that hasn't been filled or excavated in recent times, by humans, either before or after the time of European settlement, and hasn't been cultivated, either. In New England, woodlands that served as pasture land rather than as tilled fields are the best places to see undisturbed soil profiles.

It's dangerous for me to describe one or a few soil profiles for you, because there is such a multiplicity of important soil types (see the next section), but for concreteness, and a good start in thinking about soil horizons and classification, here's *a description of a* *typical New England* soil (Figure 2-28). The two conditions such a soil experiences that are the most important factors in its development are (1) *strong net downward movement of water through the profile*, and (2) *acidic waters*. In the traditional classification of soils, such a soil would be called a ***podzol*** soil.

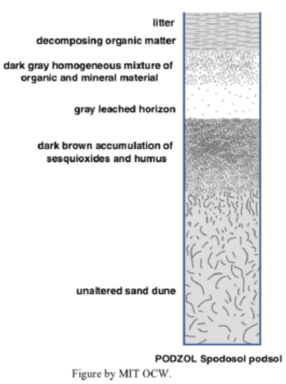


Figure by MIT OCW.

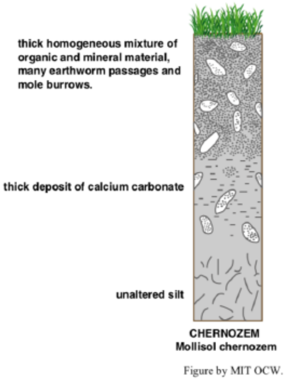


Figure by MIT OCW.

Figure 2-28 [left]. A vertical profile through a typical New England soil. (From. Figure 2-29 [right] . A vertical profile through a typical North American mid-continent soil. (From FitzPatrick, 1986.)

Right at the surface is a layer of forest litter: leaves and twigs that have fallen recently, within the last few years. That grades down into a layer of organic matter in a progressively more advanced state of humification with depth. Together these uppermost layers, called the *H horizon* if usually water-saturated or the *O horizon* if not usually water-saturated, are the part of the soil that's by far the richest in organic matter and living organisms—and most conducive to cultivation, in which case the original organic-rich layer is tilled and mixed downward to form the best medium for plant growth. Its thickness commonly ranges from just a few centimeters upward to ten to twenty centimeters. It passes rather abruptly into a layer consisting mainly of mineral grains, usually mostly quartz (if the soil is well developed and the parent material had a fairly high percentage of quartz), with a characteristic light gray color. Much, or most, or in extreme cases even all, of the other silicate minerals have been leached out of this layer by decomposition by net downward passage, from the surface, of waters rendered acidic by (1) dissolved CO_2 from the atmosphere and (2) organic acids produced in the uppermost organic-rich layer. This horizon is an example of what's called the *A horizon*. The A horizon passes downward, again fairly abruptly, into a layer, called the *B horizon*, with a characteristic color, usually combinations of yellow or orange or red or brown. Here is where iron leached out of the material of the A horizon is reprecipitated as hydrous iron oxides within the pore spaces among the existing mineral grains, and fine clay-mineral particles produced by weathering reactions in the A horizon have been deposited. (The general term for mobilization and removal of such things as iron and aluminum from a soil horizon is ***eluviation***; their deposition in a given horizon is called ***illuviation***.) The B horizon passes gradually into underlying material that's relatively little affected by soil processes; it's called the *C horizon* (but not universally considered to really be a soil horizon!).

The precise nature of the physicochemical processes that mobilize, transport, and deposit iron and, especially, aluminum are not yet entirely well understood. Iron, when weathered from parent minerals, is probably complexed with organic compounds and then precipitated as insoluble oxides lower down in the soil profile where the carrier organic compounds are degraded. The aluminum presents more of a problem. The aluminum is not actually put into true solution, but neither can it make its way down the profile as large clay-mineral particles. It probably travels either as a colloid (i.e., particles so fine as to be in the range between true solution and fine mineral particles) or physically as very fine, submicron-size clay-mineral particles.

Where the average water table lies far below, the B horizon tends to be diffuse; where the average water table is not far below, the B horizon tends to be more concentrated, and its colors stronger. (In areas where the average water table lies close to the surface, the B horizon tends instead to be gray from deposition of compounds containing iron in the reduced, ferrous state, as well as clay minerals.)

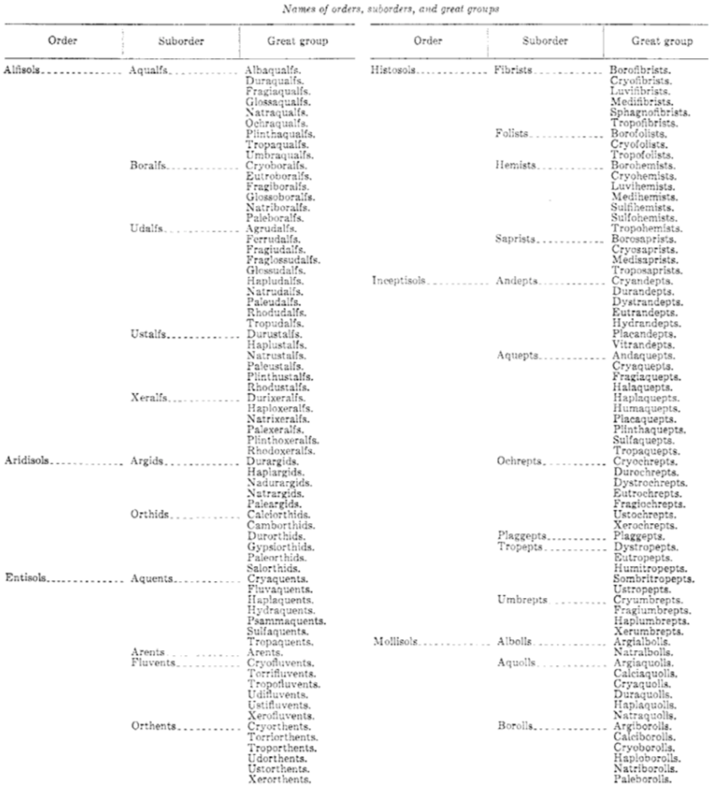
Now for something entirely different: a capsule description of a typical soil from the mid-continent of North America (Figure 2-29) Such soils are common in a wide area including western Kansas, western Nebraska, eastern Colorado, most of the Dakotas, and on into south-central Canada. In the traditional terminology, such soil would be called ***chernozem soils*** (from the Russian for "black soils"), ***chestnut soils***, or ***brown soils***. In such areas, with a semiarid (rainfall 15–25 inches) and strongly seasonal climate and dominantly grassland vegetation (now largely converted to agricultural use, especially in the eastern parts of the area), the balance between precipitation and evaporation is such that there is little net downward movement of water in the soil, with two important consequences: (1) the soil is largely a *closed system*, in the sense that chemical constituents produced by weathering remain in the soil rather than being carried away in solution in groundwater, and (2) the pH of soil waters tends to be *alkaline* rather than acidic. In such soils, upwards of a meter thick, the upper layer (A horizon) of the soil, beneath the surface organic layer, are a deep, homogeneous, and friable mixture of organic matter and mineral particles, not strongly leached, which is kept thoroughly mixed by the activity of burrowing organisms. This layer grades downward into a lighter-colored material with less organic matter, and then passes downward into a layer (the B horizon) in which calcium carbonate has been precipitated by mobilization of calcium ions in the upper layers and precipitation at deeper levels. The carbonate is precipitated as soil moisture is drawn back up toward the surface. In areas where the parent material contains carbonate minerals, the carbonated layer is better developed, but carbonate minerals tend to be deposited in the lower part of the soil profile even in the absence of carbonate minerals in the parent material. Why? Because of the common presence of calcium ions in parent silicate minerals, and the presence of carbon dioxide in the surface waters, to supply the carbonate ion.

2.9.8 Classification of Soils

Now, on to the troublesome and confusing matter of soil classification. Early classifications, before the middle of the 1900s, relied heavily upon terms and concepts developed in the late 1800s by Russian soil scientists, who were true pioneers in the scientific study of soils. These terms are common still, but in recent decades a systematic and very detailed classification system, developed and continually refined by soil scientists at the U.S. Department of Agriculture, has come into widespread use, despite its complexity and almost absurdly non- intuitive terminology. Also, other newer classifications are in use in other parts of the world.

(The "almost absurdly" in the preceding paragraph a value judgment, I know, but it's a widely shared view. Just for laughs, I am appending, as Figure 2-31, from the thick tome on soil taxonomy published by the Soil Survey Staff of the Department of Agriculture, the official names of the "great groups", which are the third stage in the taxonomic hierarchy. Each of the 185 great groups is in turn subdivided into subgroups. There are nearly a thousand subgroups! This classification is a marvel of rational and synthetic classification and terminology. Also in Figure 2-30 is a typical page from the same source, giving a description of some of the subgroups of one of the great groups. Another other interesting point about this classification is that, in contrast to any other classification I know of, in any field of science, they insist upon capitalizing every soil name: Durorthidic Xerorthents, rather than durorthidic xerorthents.)

I think that what would be most valuable for you at this point is to have soil classification stripped to its barest essentials. I will do that for you in two ways. First, Figure 2-31 gives the soil "orders" of the USDA classification system. Even this most boiled-down presentation needs a few words of explanation. A ***pedon*** (rhymes with "head on") is the basic soil unit: the column of soil, with its various horizons, in a small representative area of the order of a square meter. The ***epipedon*** is just the surface horizon of the pedon; it's usually rich in organic matter. Second, Figure 2-32, in three parts, gives (A) a simple classification of the earth's climatic zones, and corresponding to that, (B) the major vegetation types and (C) the major soil groups. The typical New England soil, described in the preceding section, is a podzol developed in a climate with severe winters (!) and with spruce–fir forest. (Minor note: the climate of southern New England is gradational into the colder end of the "humid" box in Figure 2-32 and the box for gray brown podzolic soil in Figure 2-32. The native forests in southern New England are a mixture of conifers and deciduous hardwoods.)



Names of orders, suborders, and great groups

Order	Suborder	Great group
Alfisols	Aqualfs	Albaqualfs.
		Duraqualfs.
		Fragiaqualfs.
		Glossaqualfs.
		Natraqualfs.
		Ochraqualfs.
		Plinthaqualfs.
		Tropaqualfs.
		Umbraqualfs.
	Boralfs	Cryoboralfs.
		Eutroboralfs.
		Fragiboralfs.
		Glossoboralfs.
		Natriboralfs.
		Paleboralfs.
	Udalfs	Agrudalfs.
		Ferrudalfs.
		Fragiudalfs.
		Fraglossudalfs.
		Glossudalfs.
		Hapludalfs.
		Natrudalfs.
		Paleudalfs.
		Rhodudalfs.
		Tropudalfs.
	Ustalfs	Durustalfs.
		Haplustalfs.
		Natrustalfs.
		Paleustalfs.
		Plinthustalfs.
		Rhodustalfs.
	Xeralfs	Durixeralfs.
		Haploxeralfs.
		Natrixeralfs.
		Palexeralfs.
		Plinthoxeralfs.
		Rhodoxeralfs.
Aridisols	Argids	Durargids.
		Haplargids.
		Nadurargids.
		Natrargids.
		Paleargids.
	Orthids	Calciorthids.
		Camborthids.
		Durorthids.
		Gypsiorthids.
		Paleorthids.
		Salorthids.
Entisols	Aquents	Cryaquents.
		Fluvaquents.
		Haplaquents.
		Hydraquents.
		Psammaquents.
		Sulfaquents.
		Tropaquents.
	Arents	Arents.
	Fluvents	Cryofluvents.
		Torrifluvents.
		Tropofluvents.
		Udifluvents.
		Ustifluvents.
		Xerofluvents.
	Orthents	Cryorthents.
		Torriorthents.
		Troporthents.
		Udorthents.
		Ustorthents.
		Xerorthents.
Histosols	Fibrists	Borofibrists.
		Cryofibrists.
		Luvifibrists.
		Medifibrists.
		Sphagnofibrists.
		Tropofibrists.
	Folists	Borofolists.
		Cryofolists.
		Tropofolists.
	Hemists	Borohemists.
		Cryohemists.
		Luvihemists.
		Medihemists.
		Sulfihemists.
		Sulfohemists.
		Tropohemists.
	Saprists	Borosaprists.
		Cryosaprists.
		Medisaprists.
		Troposaprists.
Inceptisols	Andepts	Cryandepts.
		Durandepts.
		Dystrandepts.
		Eutrandepts.
		Hydrandepts.
		Placandepts.
		Vitrandepts.
	Aquepts	Andaquepts.
		Cryaquepts.
		Fragiaquepts.
		Halaquepts.
		Haplaquepts.
		Humaquepts.
		Placaquepts.
		Plinthaquepts.
		Sulfaquepts.
		Tropaquepts.
	Ochrepts	Cryochrepts.
		Durochrepts.
		Dystrochrepts.
		Eutrochrepts.
		Fragiochrepts.
		Ustochrepts.
		Xerochrepts.
	Plaggepts	Plaggepts.
	Tropepts	Dystropepts.
		Eutropepts.
		Humitropepts.
		Sombritropepts.
		Ustropepts.
	Umbrepts	Cryumbrepts.
		Fragiumbrepts.
		Haplumbrepts.
		Xerumbrepts.
Mollisols	Albolls	Argialbolls.
		Natralbolls.
	Aquolls	Argiaquolls.
		Calciaquolls.
		Cryaquolls.
		Duraquolls.
		Haplaquolls.
		Natraquolls.
	Borolls	Argiborolls.
		Calciborolls.
		Cryoborolls.
		Haploborolls.
		Natriborolls.
		Paleborolls.

Figure_2-30.1.png" />

Figure 1-30. A) Orders, suborders, and great groups of the soil classification system developed by the Soil Survey Staff at the US Department of Agriculture. B). A random page from the system of soil taxonomy of the US Department of Agriculture. (From Soil Survey Staff, 1994)

B

360 SOIL TAXONOMY

States, but the subgroup is provided for use where needed.

Epiaquic Tropohumults.—Soils in this subgroup have some yellowish brown or olive brown colors in the upper horizons and have redder mottles inside the peds. The hue becomes redder with depth. The rainfall is seasonally high, and the upper horizons become saturated. The reducing conditions are commonly but not everywhere too mild to produce mottles of low chroma, but the hue in the upper horizons approaches that of Aquults. These soils are in the areas of high rainfall in Puerto Rico but are not extensive. Their slopes are mostly strong, and the soils are mostly cultivated.

Epiaquic Orthoxic Tropohumults.—These soils have some yellowish brown or olive brown colors in the upper horizons and have redder mottles inside the peds. The hue becomes redder with depth. The reducing conditions are commonly too mild to produce mottles of low chroma. The clay in these soils has low activity. The soils have strong slopes and are in the areas of high rainfall in Puerto Rico, but they are not extensive.

Humoxic Tropohumults.—These soils have a clay fraction that has low activity. They are common in the United States in Hawaii, where they formed on steep slopes or fans from basic rocks. They are largely used for pasture, but where slopes are favorable some of them are cultivated.

Orthoxic Tropohumults.—These soils have a clay fraction that has low activity, and they tend to be warmer than most of the Typic Tropohumults. They are not extensive in either Puerto Rico or Hawaii. Their slopes are mostly steep, and the soils are used mainly for pasture or tree crops.

Ustic Tropohumults.—These soils have an ustic moisture regime but are otherwise like Typic Tropohumults in defined properties. They are not known to occur in the United States, but the subgroup is provided for use if needed elsewhere.

Ustoxic Tropohumults.—Soils in this subgroup have an ustic moisture regime and a clay fraction that has low activity. They are not extensive in the United States.

Udults

These are the more or less freely drained, humus-poor Ultisols (plate 9D) in humid climates in mid or low latitudes that have well-distributed rainfall. Most of them have light-colored upper horizons, commonly a grayish horizon that rests on a yellowish brown to reddish argillic horizon. A few that developed from basic rocks have a dark brown or reddish brown surface horizon that rests on a dark red or dusky red argillic horizon. Some have a fragipan or plinthite, or both, in or below the argillic horizon.

The Udults developed in sediments and on surfaces that range from late Pleistocene to Pliocene or possibly older. Many are cultivated, either with the use of soil amendments or in a system in which they are cropped for a very few years and then are returned to forest to allow the trees to regather in their tissues the small supply of nutrients. Most of these soils have or had a forest vegetation, but some have a savanna that probably is anthropic. The moisture regime is udic. The soil temperature regime is mesic, isomesic, or warmer.

Definition

Udults are the Ultisols that
1. Are never saturated with water or have a redder hue or higher chroma than Aquults;
2. Have <0.9 percent organic carbon in the upper 15 cm of the argillic horizon and have <12 kg organic carbon in a unit volume of 1 square meter to a depth of 1 m exclusive of any O horizon; and
3. Have a udic moisture regime.

Key to great groups

FCA. Udults that have a fragipan in or below the argillic horizon.

Fragiudults, p. 360

FCB. Other Udults that have plinthite that forms a continuous phase or constitutes more than half the volume in some horizon within the upper 1.25 m of the soil.

Plinthudults, p. 366

FCC. Other Udults that have an argillic horizon that has <10 percent weatherable minerals in the 20- to 200-micron fraction in the upper 50 cm and have a clay distribution such that the percentage of clay does not decrease from its maximum amount by more than 20 percent of that maximum within 1.5 m of the soil surface, or the layer in which the percentage of clay is less than the maximum has skeletans on ped faces or has 5 percent or more plinthite by volume.

Paleudults, p. 364

FCD. Other Udults that have
1. An epipedon that has a color value, moist, of less than 4 in all parts; and
2. An argillic horizon that has a color value, dry, of less than 5 and not more than 1 unit higher than the value, moist.

Rhodudults, p. 367

FCE. Other Udults that have an isomesic or warmer iso temperature regime.

Tropudults, p. 367

FCF. Other Udults.

Hapludults, p. 362

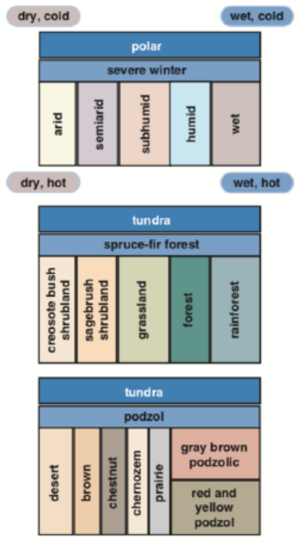
Fragiudults

These are the Udults that have a fragipan in or below the argillic horizon. They were formed mainly in loamy sediments or in residuum. The fragipan normally has an upper boundary about 50 to 75 cm below the soil surface. Perched ground water is present above the pan at some period during the year, and there are commonly thick gray skeletans at the top of the pan.

The Fragiudults in the United States are principally on gentle slopes throughout the southeastern states. Their temperature regime is mesic or thermic. Their vegetation in the United States has been forest, either of conifers or of broadleaf deciduous trees. Some were considered Planosols and others Red-Yellow Podzolic soils in the 1938 classification or the 1949 revision.

Figure_2-31.png" />

Figure 2-31. Orders of the US Department of Agriculture soil classification system. (Modified from Winegardner, 1996.)



Highly simplified relationship of vegetation type and soil type to major climatic zones. (Top) distribution of climate types. (Middle) Distribution of vegetation types on the climate base. (Bottom) Distribution of soil groups on the climate base.

Figure by MIT OCW.

Figure 2-32. Highly simplified relationship of vegetation type and soil type to major climatic zones. (Top) distribution of climate types. (Middle) Distribution of vegetation types on the climate base. (Bottom) Distribution of soil groups on the climate base. (From Hunt, 1972.)

References

Brady, N.C. and Weil, R.R. 2002, The Nature and Property of Soils: Prentice Hall, 960 p.

FitzPatrick, E.A., 1986, An Introduction to Soil Science: Longman, 255 p.

Merrill, G.P., 1897, A Treatise on Rocks, Rock Weathering and Soils: Macmillan, 411 p.

Hunt, C.B., 1972, Geology of Soils; Their Evolution, Classification, and Uses: Freeman, 344 p.

Soil Survey Staff, 1994, Keys to Soil Taxonomy: US Department of Agriculture, Soil Conservation Service, 306 p.

Taylor, G., and Eggleton, R.A., 2001, Regolith Geology and Geomorphology: Wiley, 375 p.

Winegardner, D.L., 1996, An Introduction to Soils for Environmental Professionals: Lewis Publishers, 270 p.

CHAPTER OVERVIEW

3: Topography

3.1: Latitude and Longitude

You probably know already that the basic coordinate system that's used to describe the position of a point on the Earth's surface is *latitude and longitude*. In this system (Figure 3-33), the Earth is imagined to be cut by a series of planes that pass through the north–south axis of rotation. The intersection of such a plane with the Earth's surface is called a line (really a curve) of ***longitude***, or a ***meridian***. Longitude is measured in degrees, from zero to 360. One meridian (the one that passes through Greenwich, England) is called the ***prime meridian***, and longitude is measured 180 degrees to the west of that and 180 degrees to the east of that. The opposite meridian, 180 degrees around the world from the prime meridian (and the intersection of the longitude plane with the other side of the world) lies about in the middle of the Pacific Ocean.

One consequence of this definition of longitude is that *the spacing between two meridians gets smaller as you go north or south from the equator*. Think about this the next time you fly west in a jetliner: You would have to move awfully fast to keep up with the sun, and land at the same time of day you took off, if you're flying along the equator, but if you're flying east to west in the far north or far south on the earth, you could easily arrive at your destination a lot earlier in the day than you took off!

The other element of the coordinate system is a series of latitude circles (Figure 3-33). These ***latitude circles*** are *small circles that are perpendicular to the earth's north-south axis*. (A ***small circle*** is *the intersection between the surface of a sphere and a plane that cut through the sphere but does not pass through the center of the sphere*. That's in contrast to a ***great circle***, which is formed by *the intersection between the surface of a sphere and a plane that cuts through the sphere and passes through the center of the sphere*.) These small circles are formed by passing planes parallel to the equatorial plane through the earth. By convention, the ***equator*** (*the curve on the earth's surface that's formed by passing a planet through the center of the earth and perpendicular to the north–south axis*) is at zero degrees latitude, and the north pole and south pole are at 90° latitude.

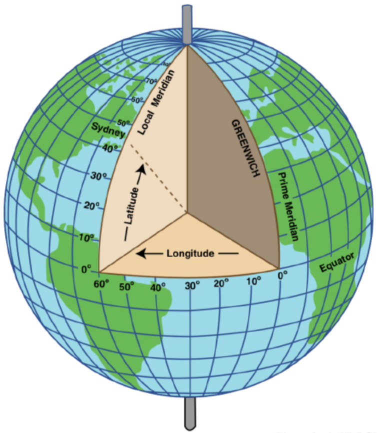


Figure by MIT OCW

Figure 3-33. Latitude and longitude. (From Greenhood, 1964.)

Big maps of the earth's surface that are bounded by latitude and longitude lines (as most maps are) are *not rectangular*: they narrow to the north in the northern hemisphere, and they narrow to the south in the southern hemisphere. Maps of smaller areas, of the kind geologists use when they are mapping in an area that's a small part of a state, are so close to being rectangular that one can usually ignore the effect of narrowing of longitude lines.

3.2: Maps

Map

Earth science is a very map-oriented discipline, because geologists are always having to view and think about the disposition of rock bodies across the landscape. This section provides just a little elementary material about maps. ***Cartography***, a well developed discipline in its own right, is *the study of maps as maps*.

People have been drawing maps to show the relative disposition of things on the earth's surface for a very long time. The accuracy of maps has improved as the ability to locate things relative to one another (which in one way or another always involves what could be termed *surveying*) has improved.

One unavoidably troublesome aspect of map-making has been (and continues to be) the need to transform spatial relationships on the surface of a nearly spherical body, the earth, into spatial relationships on a flat piece of paper, the map (Figure 3-34). The only way around this problem is to *draw the map on a small globe*, but the obvious impracticality of carrying around and using a globe-shaped map make that expedient not very useful. Various kinds of ***projections***, or *ways of systematically distorting a segment of a spherical surface to make it fit onto a plane*, have come into use. Each such kind of projection has its advantages and disadvantages. For geologic maps of fairly small areas you don't need to worry about the problem of projections, but for maps of whole countries and continents, or even of the states of the U.S., you do.

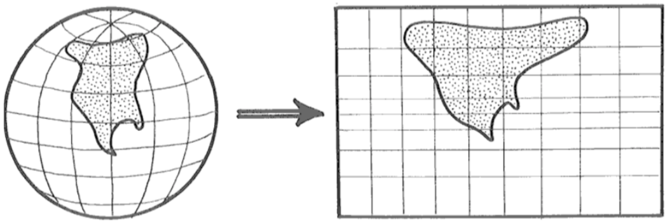


Figure 3-34. Trying to transform an area on a sphere into an area on a plane.

Aside from the fundamental requirement of representing accurately the spatial relationships in accordance with a given kind of projection, there are just a few elements essential to all maps:

- A *scale*, which expresses the ratio of a given horizontal distance on the map to that same horizontal distance on the actual land surface. This can be expressed as a numerical ratio, or it can be drawn as a labeled scale bar on the map, or (preferably) both.
- A *north arrow*, an arrow somewhere on the map, usually in the margin, that shows the direction of true north.
- A *key* or *legend* that explains to the user of the map all of the various symbols that show features on the map.

Map Scales

All maps have to have a scale. The ***scale*** of a map is *the ratio of the distance between any two points on the map and the actual distance between the corresponding points on the Earth's surface*. It's usually expressed as a ratio, called the ***representative fraction***, one divided by some number, like 1:25,000 or 1/25,000. It can also be expressed as a *verbal statement*: for example, one inch equals one mile, or one centimeter equals one kilometer. Most maps also show the scale graphically, by having a *horizontal bar at the bottom of the map with tick marks along the bar* labeled with the distances represented by the positions of the tick marks along the bar.

Here's something really tricky to remember about maps scales: A ***large-scale map*** is one that uses a relatively *large* distance on the map to represent a given distance on the Earth's surface

A ***small-scale map*** is one that uses a relatively *small* distance on the map to represent a given distance on the Earth's surface.

This is confusing, because *a small-scale map tends to cover a relatively large area of the earth's surface*, and *a large-scale map tends to cover a relatively small area of the earth's surface*!

3.3: Topographic Maps

Almost all of the area of the United States has been represented on topographic maps at various scales. For many years there were two standard map scales:

- Maps, called fifteen-minute quadrangle maps, that spanned 15' of longitude and 15' of latitude at a scale of 1:62,500, which is exactly one inch to the mile.
- Maps, called seven-and-a-half-minute quadrangle maps, that spanned 7-1/2' of longitude and 7-1/2' of latitude at a scale of 1:24,00.

In recent times those quadrangle maps have been supplanted by seven-and-a-half-minute quadrangle maps that are about the same size at the old 1:24,000 maps but at a slightly different, but more rational scale of 1:25,000. In addition, there is a series of maps with a scale of 1:250,000, called *quarter-million maps*. These latter, newer, maps use the metric system for map distances and contour intervals.

The 1:25,000 maps are what geologists typically use as a base for mapping the geology of a local area. The resulting maps are called ***geologic quadrangle maps***. But often the detail of mapping that the geologist desires calls for a larger scale: that is, the map scale is such that a smaller land area is represented on a given sheet of paper. Detailed geologic maps are often at scales of between 1:10,000 and 1:5,000. *The scale of map you work with depends on the level of detail you want to show on the map*.

Topographic Contours

Many maps of land areas have series of curved lines, called contours, that represent the topography of the area. Such a map is called a ***topographic map***. A ***contour*** is *a horizontal curve that's the locus of all points on the map with the same elevation* (Figure 3-35). A good way to understand the concept of contours is to choose a series of elevations and imagine passing a corresponding series of horizontal planes through the landscape (Figure 3-36). The contours are formed by the intersection of these planes with the land surface.

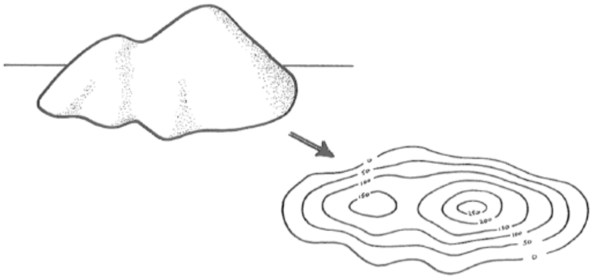


Figure 3-35. Topographic contour lines.

Adjacent contours touch one another only where the land surface is vertical, and they cross one another only where one part of the land surface hangs over another part (and in such overhanging areas, by convention the contours are dashed or dotted to indicate the overhanging relationship, or just omitted altogether).

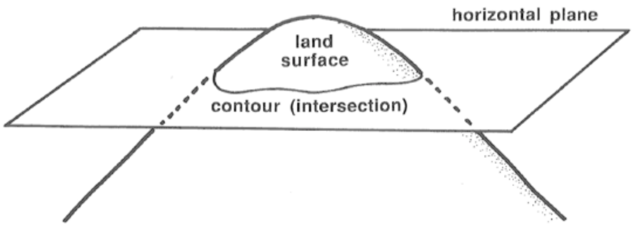


Figure 3-36. Viewing contours as the loci of horizontal planes with the land surface.

Topographic maps are useful to both technical people (geologists; environmental engineers) and to everyday people, like hikers. A geologist often uses a topographic map as a base on which to plot geology in making a geologic map.

Interpreting Contour Lines on Topographic Maps

To be a sophisticated user of a topographic map, you have to be able to *visualize the topography or "lay of the land" by examining the contour lines*. Or, even if you can't do a good job of visualizing the topography, you at least have to understand how to interpret it from the contour lines. Here are some basic pointers. The best way to learn how to visualize topography from contour lines is to practice with a real map—and you will be doing that in the take-home exercise.

For a map at a given scale, *the more closely spaced the contour lines, the steeper to slope of the surface*: closely spaced contour lines mean steep slopes, and widely spaced contour lines mean gentle slopes (Figure 3-37).

You can easily compute the local slope of the land surface, by:

1. measuring the distance between two points on the map,
2. using the map scale to find the real distance on the land surface,
3. finding the difference in elevation between the two points from the contour lines (at each point you usually have to interpolating between adjacent contour lines),
4. forming a right triangle with one leg the horizontal distance and one leg the vertical distance, and
5. finding the slope angle by trigonometry.

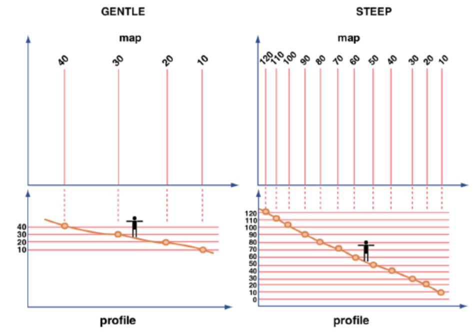


Figure by MIT OCW

Figure 3-37. Contours on gently sloping surfaces vs. contours on steeply sloping surfaces.

I suppose I need not insult your intelligence by pointing out (Figure 3-38) that hills or ridges are located where the contour lines on either side of the hill or ridge increase in elevation toward the top of the hill, and valleys are located where the contour lines on either side of the valley decrease in elevation toward the bottom of the valley.

Valleys usually slope in one direction or another because they are occupied by streams, which, as you know, flow downhill. Such a valley shows up on the topographic map as a series of V-shaped kinks in the successive contour lines. Those "V"s point *up* the valley (Figure 3-39A). Conversely, ridges commonly have crests that slope downward in one direction or the other, and the contour lines in that case form "V"s that point *down* the ridge or spur (Figure 3-39B).

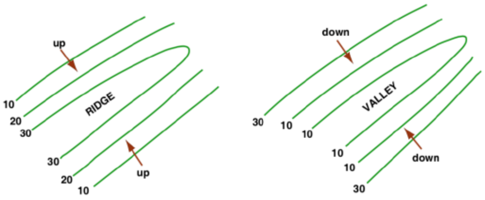


Figure by MIT OCW

Figure 3-38. Contour lines for ridges and valleys.

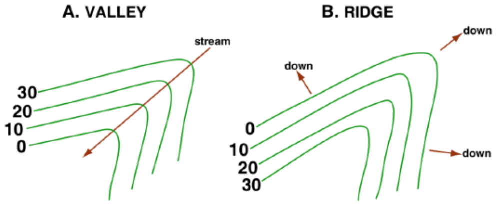


Figure by MIT OCW

Figure 3-39: A) The "V"s of the contour lines point up valleys. B) The "V"s of the contour lines point down ridges.

How do you know the elevation of a hilltop from a contour map? The answer is: you don’t, really, unless that information is supplied separately by being printed on the map! That’s because the tippity-top in general lies on the “incomplete” contour interval above the uppermost closed contour. But at least you can bracket the elevation of the summit to be within one contour interval (Figure 3-40).

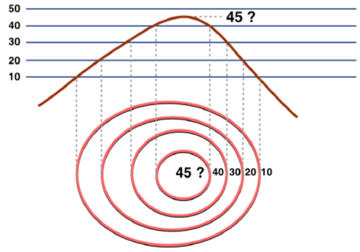


Figure by MIT OCW

Figure 3-40. Estimating the elevation of a hilltop from a contour map.

Closed depressions on the land surface are the opposite of hilltops, but in most areas they are uncommon or even nonexistent. Closed contours for which the elevation decreases inward are denoted in a special way, by putting little perpendicular tick marks on the *downslope* side of the closed contour (Figure 3-41).

Overhangs (also not common) are also treated in a special way: the obscured segments of contour lines are either ignored or drawn as dashed curves rather than as solid curves (Figure 3-42).

Now think in terms of riding saddles or curvy potato chips, or, for those mathematically inclined, hyperbolic paraboloids (Figure 3-11). In all of these cases, one is dealing with a curved surface such that the intersection between the surface and a vertical plane is convex up in one direction and concave up in a direction approximately at right angles to the first. Such features on topographic maps are called ***saddles*** or ***passes***. Saddles are common along ridges where stream valleys indent the ridge from two opposite sides.

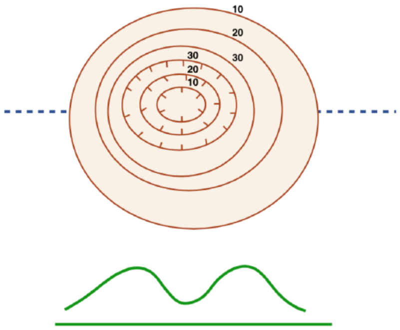


Figure by MIT OCW

Figure 3-41. How contour lines represent depressions.

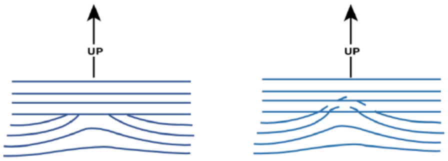


Figure by MIT OCW

Figure 3-42. How contour lines represent overhangs.

Locating Yourself on the Map

When you are working outdoors with a topographic map, the map is likely to be much more useful when you are able to *locate where you are on the map*. That’s a skill we won’t be able to practice in this course, unfortunately. Below are some considerations for you to keep in mind, in case you are called upon to work outdoors with a topographic map.

One way of locating yourself is simply to *choose to be somewhere that corresponds to a readily identifiable feature on the map*, like a road intersection, a stream confluence, or a mountaintop. But that’s overly restrictive: suppose you need to be somewhere not near such a feature?

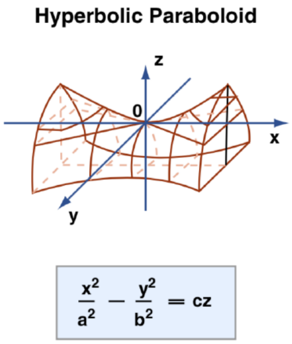


Figure by MIT OCW

Figure 3-43. A hyperbolic paraboloid (a kind of “saddle”) (From Burington, 1948.)

One technique, which is useful in arid and semiarid regions with substantial relief (by ***relief*** I mean *local differences in elevation from place to place*) is to look back and forth from map to land, several times, to relate the topography shown on the map to the topography you can see on the ground. It’s a knack one acquires, usually readily and without much difficulty, by practice.

Note

I have not been able to develop a laboratory exercise to make that technique real to students in a classroom, although for some years I have intended to put into practice an idea that might be useful in that regard. Perhaps one of you might be moved to pursue the idea. I would like to build a small table-top model of some small area with rugged mountain-and-valley topography by blowing up the topographic map of the area, cutting successive layers of a sheetlike material like the foam-core sheets used to mount prints, by tracing the outline of successive contour lines on successive sheets of the foam core, then stacking the sheets together, gluing them one by one, and covering the entire “terraced” mass with a moldable material like wallboard joint compound, and finally sanding the surface smooth. With such a model, one could do two instructive things:

1. mark particular points on the model and ask students to locate the points on the map, and
2. mark particular points on the map and ask students to show the corresponding location on the model.

It would be an extremely valuable teaching tool. If anyone, perhaps with some experience along that line, is interested in working on such a model, let me know.

The other technique, which is necessary in areas with low relief and/or heavy, view-obstructing vegetation, is *surveying*. You start from a known point on the map and run a survey line to the point of interest. This can be as crude as a compass-and-pace traverse or as sophisticated as a professional survey.

Visualizing the Lay of the Land

In working with topographic maps, one skill you need to develop is the ability to visualize the lay of the land from a topographic contour map. This involves picturing, in your mind’s eye, what the map area would really look like, in a view from a low-flying airplane on a clear day, just from the information provided by the topographic contours. To make full use of a topographic map, you need to be able to do that.

Visualizing topography from looking at the topographic contours on the map is a skill you have to develop for yourself, by thought and lots of practice. Some people can do it much more easily than others. It’s an exercise in using your “mind’s eye”, which is a useful skill in earth science, as well as in everyday life. (Think ahead toward when you might be wanting to remodel your home: can you picture the shape of your new kitchen, and how everything will fit together in it in three dimensions?)

To develop your skill in visualization, I will have a take-home exercise for you in which you will have to build a model of the earth’s surface, out of modeling clay, just by looking at a topographic map of the area.

Topographic Profiles

It is often useful to obtain a vertical cross-section view of the land surface along a line that extends from one point on the land to another point. It's easy to construct such a profile if you have a topographic map of the area already available. Here's how to proceed:

1. Pinpoint the two points at the end of the desired profile directly on your topographic map, and draw a faint pencil line between them.
2. Lay out, on a blank sheet of paper, a corresponding faint pencil line near the top of the paper than extends from the starting point on the land surface to the ending point. Establish a conversion ratio between the scale of the map and the scale of your cross-section line; then, as you pick points off the map, you can easily plot them along your cross-section line.
3. For a large number of points along the profile line on the topographic map that happen to fall on topographic contour lines, pick off the elevations, and transfer them to your cross-section line on the sheet of paper.
4. Connect the elevation points on your cross-section with a smooth curve.

3.4: Stream Networks, Drainage Basins, and Divides

3.4.1 Tracing Stream Courses

In most areas of the world, except in the driest of deserts (and beneath glaciers), one can trace fairly easily on a topographic map the system of main streams and their tributaries. In some places streams "expand" into lakes, but the principle is the same.

Permanent streams are always shown as thin blue lines on official topographic maps, and ***ephemeral streams*** (those that flow only after a heavy rain) are often shown as dot–dash blue lines. Many well-defined valleys, however, which presumably would have streams flowing in them briefly after a heavy rain, have no streams shown in them. These are usually located in the headwaters of larger streams, which *are* shown on the map. So it's just a matter of extending the streams shown on the map farther upstream, to where valleys are no longer defined and the land surfaces slopes uniformly. (Remember about the "V"s of the contour lines pointing upstream in valleys.) You yourself can trace the courses of such streams by recognizing the position and downslope direction of such valleys.

Here are some points that should help you in tracing stream courses where there are valleys with no stream shown in them.

1. Keep firmly in mind that *in a valley, the contours tend to "vee" up the valley*. You know that's happening by seeing that the "vees" point in the direction of increasingly high contour lines.
2. *When you draw the stream course, make it pass directly through the crotch of each contour "vee"*. Streams tend to be curvy in real life, so don't hesitate to end up with a curvy stream.
3. You are likely to find places where two tributary streams come together, at what's called a ***confluence***, to form a main stream. With a margin of error of the space of one contour interval, it's easy to locate such a confluence: just *look for places where a contour with one "vee" is succeeded upward by a contour with two closely spaced "vees"* (Figure 1-44). The confluence lies somewhere between those two contours.

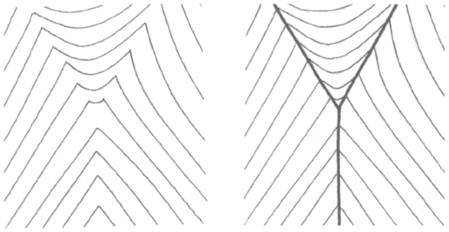


Figure 3-44. How you can locate the point of stream confluence by examining the "V"s of the contour lines.

3.4.2 Stream Divides and Drainage Basins

It should make sense to you that *the land area between individual streams is on at least slightly higher ground than the streams themselves*; if not, then the whole area would be a lake. (The exception to that last statement, fairly common in New England, involves low-lying wetlands bordered by or laced with well-defined streams.) Somewhere on that higher ground is a ***stream divide***: *a continuously curving locus of points on the map separating an area of the land surface with drainage into one stream from an area of the land surface with drainage into another stream* (Figure 3-45).

Stream divides partition a given area of the land surface into ***drainage basins***, each drained by a different main stream and its various tributary streams (Figure 3-45). Except in unusual situations, *the land area is partitioned exhaustively and non-overlappingly into drainage basins*.

If you are dealing with a fairly small area of the land surface, as represented, for example, by a 7-1/2' topographic map, you run into the uncertainty about whether a stream that runs off the edge of your map, along with its drainage basin, runs into one of the other streams that runs off the edge of your map, at a point somewhere outside the area of your map, or into some other stream that doesn't even show up on your map. To ascertain that, you have to examine adjacent map areas. In the context of drainage basins and their size, it's important to know that.

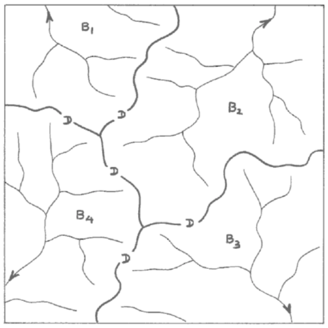


Figure 3-45. Drainage basins and drainage divides.

Just think in terms of traipsing around the land surface with buckets of water. If you pour the water upon the ground (and assume that it’s going to run off to a stream rather than soaking in right on the spot), *to which stream does it flow*? Or, if you prefer something messier and but probably more exciting, imagine that the entire land surface is coated with ultra- slippery mud, and you let yourself slide on your backside down the slope toward a stream channel: which stream do you end up in?

What’s going on is that the water, or you, are passing downward along what’s mathematically called the ***gradient***: *the route of steepest descent*. You can trace out such routes of steepest descent, from any given point on a sloping land surface as represented on a topographic map, by drawing curves that are *everywhere normal to (i.e., at right angles to) the topographic contour lines* (Figure 3-46). Doing this by eye is not too difficult, once you get the knack of it.

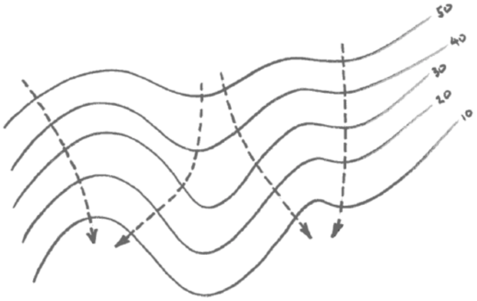


Figure 3-46. Gradient curves down a sloping surface.

Here are some considerations on divides. When two streams are separated by a well-defined ridge crest, locating the divide is easy (Figure 3-47A). When the ridge crest itself slopes (Figure 3-47B) or is broad and not well defined (Figure 3-47C), the job is not as easy. The divide between two streams that join together at a confluence at some point ends at that confluence point (Figure 3-47D). At their high ends, divides meet can meet at the “crotch” of a Y-shaped mountain slope (Figure 3-47E) or at the summit of a hill or mountain (Figure 3-47F).

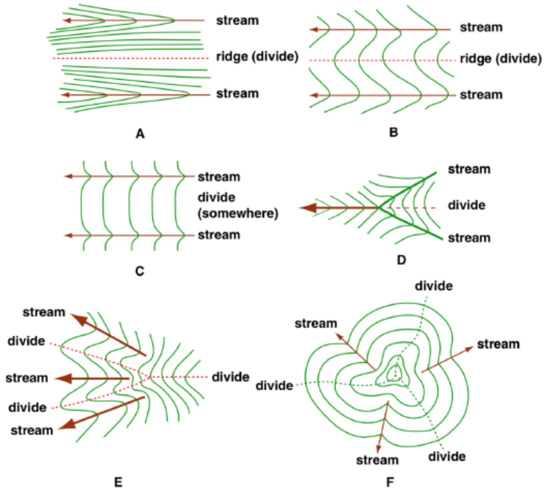


Figure by MIT OCW

Figure 3-47. Considerations on stream divides. A) Two streams separated by a well-defined ridge crest. B) Two streams separated by a sloping ridge crest. C) Two streams separated by a broad and ill-defined ridge. D) Stream divides in the vicinity of a confluence. E) Two stream divides that merge into a single divide upslope. F) Stream divides meeting at the crest of a hill.

What follows is a rather lengthy "home experiment" that should be useful to you if you are having trouble dealing with the concept of stream divides. Start with several cylinders, which could be tall soda bottles with tops and bottoms cut off, or fat mailing tubes (probably the best), or the cylindrical wooden posts from an old bedstead. (Mathematically, these are *circular* cylinders.) Make them all the same length, ideally several times the diameter of the cylinder. Cut each through lengthwise, along a plane parallel to the axis of the cylinder but offset a bit. Keep the bigger pieces and discard the smaller. These bigger pieces should look like fireplace logs that have been split down the middle but with imperfect aim. Now place them side by side on a rigid planar surface like a cutting board, with adjacent edges touching. The result should look a little like a giant washboard. (Washboards are a disappearing item; have you ever seen one, much less use one?) Put the cutting board with its cylinders in your bathtub, with one of slowly with water, in equal increments of depth, each time stopping to mark, with a permanent-ink felt-tipped pen, the water line on the surfaces of the cylinders. You will have to solve for yourself the problem of keeping the cylinders from floating away in the process. Now drain the water and view the bathtub from a point directly above, way up near the ceiling of your bathroom. What you want to try to see is a topographic map of your tilted-cylinder model. (You might have to squint a bit to help your imagination. If you are into photography, the best thing would be to put yourself high above the model and take a telephoto shot.) The result would look something like what is shown in Figure 3-16. (Incidentally, since the intersection of a circular cylinder with a plane that cuts the cylinder at some angle less than 90° to the axis of the cylinder is an ellipse, the contours on your "map" are segments of ellipses.) The model you've created is not a bad approximation to many real-life examples of valley-and-spur topography.

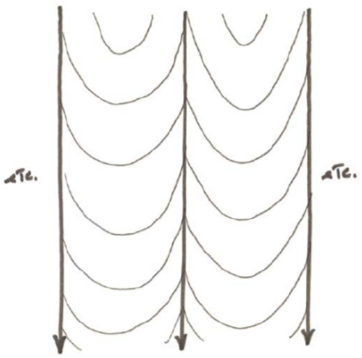


Figure 3-48. Contour lines in your bathtub cylinder-segment model.

3.5: Geologic Maps and Cross Sections

3.5.1 Rock Units

Field geologists who study and map the bedrock that underlies an area of the land surface attempt to recognize rock units, which they can then represent on geologic maps. Rock types are not randomly arranged in the Earth's crust but tend to exist in distinctive bodies called ***rock units***. Rock units are *large three-dimensional bodies of rock with compositions that are distinctive and different from adjacent rock units*. Rock types vary widely in size, shape, composition, and origin.

Rock units are formed by the action of some particular process or set of processes—for example, sediment deposition, magma intrusion or extrusion, or metamorphism. Rock units can consist of sedimentary, igneous, or metamorphic rocks. The defining characteristic of a rock unit is that certain processes operate for some period of time to produce a body or mass of rock with fairly uniform rock type, or perhaps a consistent alternation of two or more rock types.

Here are some examples of the kinds of rock units a geologist might recognize in the field:

- A succession of sedimentary layers of one or several rock types, deposited in some distinctive sedimentary environment. The thickness of such a unit might range from several meters to many thousands of meters, and the lateral extent might range from hundreds of meters to a few hundreds of kilometers.
- A succession of volcanic rocks with distinctive composition. Such units might be as thin as a single lava flow (which could be as thin as a few meters) or as thick as many hundreds of meters. Such units might be interbedded with sedimentary rock units as well.
- A single igneous intrusive body. Such intrusive units cut across other rock units that they intrude.
- A unit of metamorphic rocks of a particular composition. Such a unit might have extremely complicated geometry, owing to intense deformation accompanying metamorphism. The intensity of metamorphism might vary systematically from one area to another within the metamorphic rock unit.

The minimum dimensions of a rock unit can be as small as meters or even decimeters, in the case of thin igneous dikes, for example. Units that small would not ordinarily be represented on a geologic map unless the purpose of mapping is to display the geology of a very small area in great detail, as for example at a major construction site.

Rock units are in contact with each other across three-dimensional surfaces or relatively thin zones of transition or gradation. The lines that represent contacts between rock units on a geologic map are the lines of intersection between the actual three-dimensional contact surfaces and the land surface itself. Recognizing and interpreting the nature of contacts between rock units is central to geological fieldwork. It's largely by interpretation of the nature of such contacts that the geologic history of an area is worked out.

Many rock units receive formal names. The basic sedimentary rock unit (and also metamorphic and volcanic rock units) is the ***formation***. Formations have two-part names: the first part is a place name, like a town, a river, or a mountain, and the second part is either the word "Formation" or a rock term like "Sandstone". Volcanic and metamorphic rock units have similar two-part names. Formations can be subdivided into ***members***, which can have either formal names or just informal names. Related formations can be lumped together into larger units called ***groups***, which receive place names in the same way as formations. Intrusive igneous units, especially large units, can have formal names, but smaller units, even though they might be mappable, usually are not formally named.

3.5.2 Geologic Maps and Cross Sections

A ***geologic map*** is *a map that shows the distribution of bedrock that is exposed at the Earth's surface or buried beneath a thin layer of surface soil or sediment*. A geologic map is more than just a map of rock types: most geologic maps show the locations and relationships of rock units.

Each rock unit is identified on the map by a symbol of some kind, which is explained in a legend or key, and is often colored a distinctive color as well. Part of the legend of a geologic map consists of one or more columns of little rectangles, with appropriate colors and symbols, identifying the various rock units shown on the map. There is often a very brief description of the units directly in this part of the legend. The rectangles for the units are arranged in order of decreasing age upward. Usually the ages of the units, in terms of the standard relative geologic time scale, is shown as well.

All geologic maps convey certain other information as well. They show the symbols that are used to represent such features as folds, faults, and attitudes of planar features like stratification or foliation. They have information about latitude and longitude, and/or location relative to some standard geographic grid system. They always have a scale, expressed both as a labeled scale bar and as what is called a "representative fraction", 1:25,000 for example, whose first number is a unit of distance on the map and whose second number is the corresponding distance on the actual land surface.

All geologic maps (except perhaps very special-purpose maps that show all the details of an area that might be the size of a small room!) involve some degree of generalization. Such generalization is the responsibility of the geologist who is doing the mapping. Obviously, it is not practical to represent features as thin as a few meters on a map that covers many square miles: the width of the feature on the map would be far thinner than the thinnest possible ink line. The degree of generalization necessarily increases as the area covered by the map increases. You could easily see this for yourself if you have access to a geologic map of some small area together with the corresponding geologic map of the entire state: the detail of the small area on the state map would be much less than on the full map of that small area.

Most geologic maps are accompanied by one or more vertical ***cross sections***, which are *views of what the geology would look like in an imaginary vertical plane downward from some line on the land surface*. These cross sections are constructed by the geologist after the map is completed. Their locations are selected so as to best reveal the three-dimensional nature of the geology. Cross sections are constructed by projecting downward the geologic features and relationships that are observed at the surface. Constructing cross sections requires the geologist to be able to visualize the geology in his or her mind. The degree of certainty about the geology shown on the cross section decreases downward with depth below the surface.

CHAPTER OVERVIEW

4: Groundwater

4.1: Introduction- How Water Gets to Be Groundwater

Introduction

I suppose I don't need to tell you that ***groundwater*** is the term used for *the liquid water at's present beneath the land surface in the pore spaces in regolith and in cracks in bedrock* in virtually all areas except in very cold climates where the subsurface water is permanently frozen in the form of what's called permafrost.

Groundwater is invisible until it emerges at the surface from springs and wells. And there's nothing spectacular about groundwater and its movement, the way there is with, say, catastrophic landslides or floods. But much of the world's water supply is from groundwater wells, so it's a topic of great importance.

I think it's true that the general public is aware of the existence of freshwater underground. I suspect, however, that a clear understanding of the environment and movements of that water is not widespread. Where does groundwater come from? Where does it reside? How does it move? What happens to it? How deep down do we have to drill or dig to find it? How deep in the Earth does it extend? It's questions of this kind that we need to address in this chapter.

How Water Become Groundwater

The main way that groundwater is replenished is from the Earth's surface, by infiltration of surface water down through the soil to become groundwater. As a prelude to our study of groundwater, this section deals with this first step in the process of groundwater flow: infiltration of surface water through the uppermost layer of the solid Earth.

You all know that some of the rain that falls on the soil surface runs off into streams and rivers and some sinks down into the soil. I'll define ***infiltration*** as *the downward movement of water across the upper surface of the soil layer*. The ***infiltration rate*** is *the rate at which the surface water moves downward across the upper surface of the soil layer*. The infiltration rate is measured in depth of water per unit time, the same as precipitation. The ***infiltration capacity*** is *the maximum rate of infiltration at a given point on the soil surface and under a given set of conditions*.

Once the water infiltrates into the soil, it tends to continue its downward movement into deeper layers. The term ***percolation*** is used for *the continued downward movement of the water that infiltrates the upper surface of the soil layer*. Not all of it keeps on moving downward, however: if the regolith layer is sufficiently dry beforehand, some of the downward-percolating water is left behind in the pore spaces between the regolith particles, mainly in two forms: little fillets at grain contacts, and absorbed by porous materials like plant residues.

4.2: A Home Experiment on Infiltration

Experiment on Infiltration

Here’s a home experiment that would not be difficult for you to do if you are a home gardener. It’s a very realistic experiment on the controls on infiltration rate and infiltration capacity.

Set up an adjustable-rate lawn sprinkler on a very gently sloping and freshly spaded and raked plot in your garden (Figure 4-1). Measure the “precipitation” from the lawn sprinkler with a little rain gauge, and measure the surface runoff directly by collecting it or catching it at the downslope edge of the plot. You can compute the infiltration rate indirectly by the equation

$$S = P - I(1) \tag{4.2.1}$$

where S is surface runoff, P is precipitation rate, and I is infiltration rate. You have to express S in units that are the same as I and P: depth of water per unit area per unit time. You can easily spot the point of nonzero S when you’re watering your garden: water stands on the soil surface as a thin, smooth sheet instead of soaking in immediately.

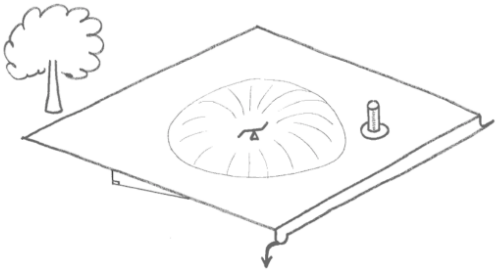


Figure 4-1. An experimental sprinkler plot for studying infiltration.

But there’s a little more to it than this. Figure 4-2 is a typical plot of infiltration capacity *Ic* against time, for a given soil. There’s a rapid decrease from very high *Ic* in the first few minutes, but then it levels off to a steady value. (It’s hard to measure *Ic* at the very beginning.) Why the initial decrease?

1. Some soil particles, like clay or organics, have a tendency to swell when they come in contact with liquid water, and this tends to close off the pore spaces. And your intuition should tell you that *the smaller the passageways, the more difficult it is to push water through them*.
2. The impact of raindrops pulverizes the soil structure at the surface and washes fine particles into the uppermost pore spaces, producing a kind of "micro- armor".

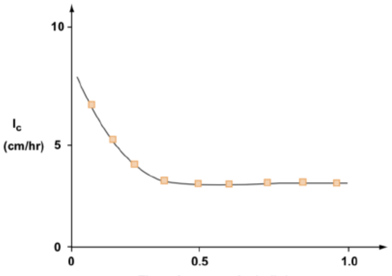


Figure by MIT OCW.

Figure 4-2. Plot of infiltration capacity against time for a given soil.

How is the infiltration capacity measured?

1. With an infiltrometer: a tube about one foot in diameter, driven about one foot into the soil. You maintain the water level in the tube at some height above the ground surface, and you measure the rate of input of new water needed to replenish the water lost to infiltration. The advantages of this arrangement are that it’s inexpensive and easy to use; the disadvantage is that it disturbs the soil.

2. With a sprinkler on an experimental plot, as in the home experiment above. The advantage of this is that it's accurate, realistic, and non-disruptive, but the disadvantage is that it's expensive and time-consuming, and you can't do it very well on irregular ground.

What is going on, physically, during infiltration and percolation? It's simple: a raindrop falls on the granular surface and then drains down through the interconnected pores of the soil by the pull of gravity. The passageways are diverse: they might be intergranular interstices, animal burrows, shrinkage cracks, or rotted rootways. If the precipitation rate is less than the infiltration capacity ($P < Ic$), then some of the pores are occupied by downward-draining water but others are occupied by air (Figure 4-3A). If the precipitation rate is greater than the infiltration capacity ($P > Ic$), then all of the pores are occupied by downward-draining water (Figure 4-3B). In both cases the downward-moving layer of water is marked by a blurred but recognizable front or lower boundary. You can see this nicely for yourself by excavating and examining a little vertical wall of soil in your garden after a heavy brief shower or a longer rain after a dry spell (Figure 4-4). After the rain stops, the infiltrated water keeps descending as a coherent layer.

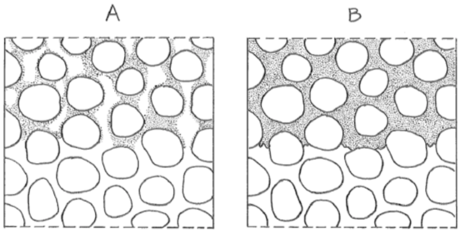


Figure 4-3. Sketch of grain-scale infiltration behavior. A) Pore spaces are occupied by both water and air. B) Pore spaces are occupied entirely by water.

You perform an experiment like this every time you water your house plants! Take a pot with established soil structure, dry it to total dryness in the oven, weigh it, and then water it very fast, to simulate the condition that P is greater than Ic. As the leading or lower boundary of the percolating layer passes the bottom of the pot, water starts pouring out onto the floor. Then, as the trailing boundary of the percolating layer passes, water stops coming out of the bottom of the pot rather abruptly. Now weigh the pot again, and convert the difference between final weight and initial weight to *volume of water per unit bulk volume of the soil*. This quantity is call the ***field capacity*** of the soil: *the maximum concentration of soil water that can be held by the soil against the pull of gravity*. The field capacity is substantially greater than zero in most soils. It's much smaller, but still nonzero, even in very coarse gravel soils.

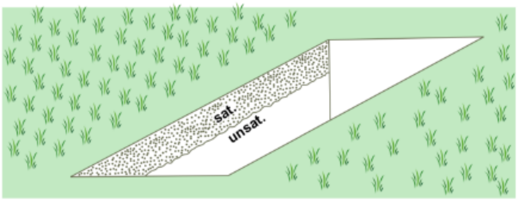


Figure by MIT OCW.

Figure 4-4. Studying infiltration after a heavy rain by means of a vertical-walled trench.

The water that remains in the soil after initial drainage is called ***soil moisture***. The water that constitutes this soil moisture is located in several kinds of places in the soil (Figure 4-5):

- as thin films on all particles
- as fillets at points of grain contact
- suffusing porous organic matter

Soil moisture makes up a very small percentage of freshwater, but it's of critical importance for life on the Earth's surface, because it's what plants use. Plants pull water out of the soil and transpire it, until the remaining soil moisture is held so tightly that plants can't take it up, and then they wilt. This level of moisture is called the ***wilting point***. The soil moisture is depleted in two other important ways as well:

- slow *residual drainage* downward
- evaporation at the surface, after *upward diffusion* as water vapor through the soil or by *capillary rise* as a liquid through the soil

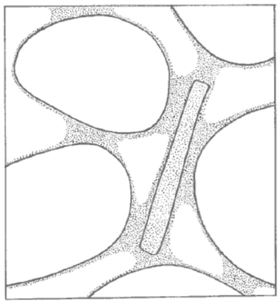


Figure 4-5. Sites of soil moisture.

Background: Surface Tension and Capillarity

Inside a liquid like water, each molecule is attracted by molecules all around it. A molecule at the surface of the liquid, however, is attracted by molecules below it and beside it but not above it, because none are there. The net effect of that asymmetry of forces is that the surface water molecules attract one another in the plane of the surface, and the result is analogous to the tensile force in a thin sheet of stretched rubber. That's why the effect is called ***surface tension***. It's as if the surface wants to "pull itself together". It's for that reason that globules of water, and soap bubbles as well, try to take on the shape of a sphere.

Now think about a small mass of water resting on a solid surface. If the attraction between the water molecules is greater than the force of attraction between the water molecules and the atoms or molecules of the solid, then the mass of water "balls up" on the surface. We say that the liquid does not wet the solid surface. If the force of attraction between the water molecules is less than the attraction between the water molecules and the atoms or molecules of the solid surface, then the liquid spreads out as a uniform thin film. We say that the liquid wets the surface.

If you put a hollow tube with a very small diameter into a liquid that wets the surface of the solid of which the tube is composed, the liquid rises up in the tube, above the liquid level outside the tube, because of the tendency for the liquid to spread itself upward onto the solid surface. The effect is called ***capillarity*** or ***capillary rise***.

The smaller the diameter of the tube, the higher the capillary rise in the tube. That's because the height of capillary rise is the outcome of two competing effects: the upward force of capillarity, and the downward force of gravity. The surface area within the tube goes as the square of the diameter, whereas the perimeter of the tube goes as the first power of the diameter. Because the capillary rise depends on the perimeter but the downward gravity force depends on the cross-sectional area, the smaller the tube, the greater the upward capillary force relative to the downward gravity force.

You may had noticed that in such a situation the water surface inside the tube is concave upward, with the level higher at the wall of the tube than in the middle of the tube. That curved surface is called a ***meniscus***.

4.3: The Physics of Groundwater Movement

Once surface water infiltrates below the surface of the soil and keeps on moving downward by percolation, it has become groundwater. At this point we have to deal with the physics of groundwater movement. This comes under a branch of fluid dynamics known as *flow through porous media*. The essential features of flow through porous media are in common with flow in pipes and channels, (Chapter 1), but certain aspects are distinctive to groundwater flow.

I'll say some things about the basic nature of groundwater movement by means of another rewarding home experiment (Figure 4-6). This one is quite simple to do and would be easy and inexpensive to set up in just a short time in your basement or in your backyard. For this experiment you will need to have a large vessel, open at the top, that will hold water without too many leaks. You could nail something together out of wood and make it leak-proof with a caulking gun, or you could borrow or buy a 55 gallon metal drum. A large plastic rubbish barrel would be good also. Mount the barrel on a stand above the floor or the ground, and attach a horizontal pipe to the wall of the vessel just above the bottom. Don't worry too much about the practical problem of how to attach the pipe to the barrel in a way that's leak-proof or nearly so. Tack a piece of fine metal screening over the entrance to the pipe as it leaves the barrel. From the downstream end of the pipe, stuff the pipe full of some natural granular material like the soil in your garden or the sand in your sandbox. When the pipe is full, attach another piece of fine metal screening to the downstream end of the pipe.

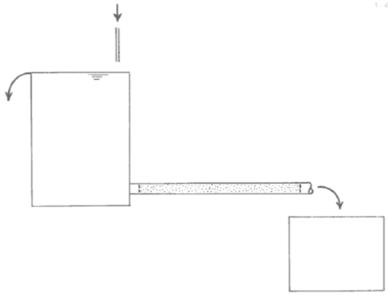


Figure 4-6. A home experiment on groundwater flow.

Run your garden hose into the vessel until water overflows, and then leave it running during the experiment to ensure a constant level of water in the vessel. What you want to do is measure the discharge (that is, the volume rate of flow) of water through the porous medium and out the downstream end of the pipe. You can easily measure this by using a catch tub at the downstream end of the pipe and measuring the volume of water caught in the catch tub with a measuring cup and measuring the time of flow with a stopwatch.

Think now about the nature of the porous granular medium in the pipe: it consists of *an enormous number of little solid particles, of more or less irregular geometry, each in contact with some small number of neighboring particles*. Look at some representative unit volume of the porous material. It's made up of two different sub-volumes (Figure 4-7): *solid granular material*, and *empty interstitial spaces*, which might be called ***void spaces*** or ***pore spaces***. Before you start the experiment, these pore spaces are filled with air. One of the important physical properties of the porous medium is the ***porosity***, defined as *the volume of pore spaces divided by the bulk volume of the material*. Loose sands or gravels typically have porosities of something like 20–30%, depending mainly on the size distribution of the material: a well sorted granular material has a larger porosity, other things being equal, than a poorly sorted granular material.

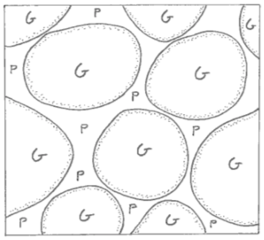


Figure 4-7. Geometry of pore spaces in a porous granular medium.

What I want you to study in your experiment is *the discharge of water through the pipe as a function of the height of water in the vessel*. You can easily vary the height of the water in the barrel by cutting holes in the side of the vessel so that the excess water runs out of the vessel at different levels above the entrance to the pipe.

Before we discuss the results of this experiment, let's think about *the nature of flow through the porous medium in the pipe*. You can't easily observe that flow, because even if you made the pipe out of a transparent material like glass or acrylic plastic you would still have a hard time observing the details of flow among the grains. But you could imagine shrinking yourself down to microscopic size and taking a submarine trip through the porous medium by drifting along with the flow.

The flow through the porous medium is like flow through a large number of tiny pipes with very irregular geometry, branching and rejoining, which narrow as they pass around grains and widen as they pass between grains. The narrow, constricted places among grains are called ***pore throats***. Although the patterns of flow are very complicated in detail, *the essential nature of the flow is not greatly different from the flow through a pipe*; it's just that the "pipe" is now narrow, tortuous (with lots of twists and turns, that is), and very intricate geometrically.

The question now arises: *Is the flow through the porous medium laminar or turbulent?* In laminar flow, the fluid moves along in regular straight paths, without mixing sideways. When you pour a stiff (that is, very viscous) liquid like oil or paint, you are observing laminar flow. Flow of water or air at all but the slowest speeds, however, is turbulent. By putting little floating tracers in the flow you can observe how flow paths are sinuous and irregular. An even better way to observe turbulence is to watch smoke as it rises from a chimney or smokestack: you can see the irregular swirls, called turbulent eddies. Go back to Chapter 1 for more details on laminar and turbulent flow.

Whether a flow of fluid is laminar or turbulent depends on four factors: the speed of the flow, the depth or width of the flow, and the density and viscosity of the fluid that's flowing. The deeper and faster the flow, and the greater the fluid density and the smaller the fluid viscosity, the more likely the flow is to be turbulent. In the case of flow through the porous medium in the pipe, the passageways are very small, and the flow velocity is in almost all cases rather small, so you should expect that *flow in porous media is laminar rather than turbulent*, except in unusual situations where the passageways are very large and the velocities are very great, as for example in very coarse open gravels, or in certain kinds of basalt flows with large connected tunnels or passageways, or in solution tunnels in limestones (Figure 4-8).

What drives the flow through the sand-filled pipe is *the downstream pressure gradient* (that is, the rate of decrease of fluid pressure with downstream distance in the pipe). That's the same thing that causes the water to flow out through your home piping system! It should make good sense to you that the speed of flow through the porous medium in the pipe depends upon the pressure gradient: the greater the pressure gradient (the driving force that causes the flow), the greater the speed of flow. But you should expect two other factors to affect the speed of flow as well: the size of the solid particles of the porous medium, and the viscosity of the fluid. The particle size is important because larger particles mean wider the pore throats, and therefore less frictional resistance to flow, because the friction arises from contact of the fluid with solid surfaces. (Go back to Chapter 2 and review how the ratio of surface area to volume increases with decreasing particle size.) The viscosity of the fluid is a measure of the resistance of the fluid to a deforming force: the greater the viscosity, the more difficult it is to make the fluid flow—as all of you know from trying to make the honey flow out of the bottle, compared to water.

Advanced Topic: Mathematical Analysis of Flow Through A Porous Medium

It should make sense to you that there must be some mathematical relationship between the various physical quantities involved in the flow described in the preceding paragraphs: the average speed of flow V in the pipe, the pressure gradient G, which drives the flow through the pipe, the grain size D of the particles that constitute the porous medium, and the viscosity μ of the fluid flowing through the porous medium. Provided that the flow is sufficiently slow, the accelerations of the fluid as it flows through the little passageways are very small, so we can neglect the inertial properties of the flow as embodied in the density ρ. So some function of V, G, D, and μ has to be equal to a constant:

$$f(V, G, D, \mu) = const_1 \tag{4.3.1}$$

Because the right side of Equation 4.3.1 is a constant number and has no physical dimensions, the variable or variables involved in the left side of the equation have to be dimensionless as well. You can easily demonstrate for yourself that the mass, length, and time dimensions of the quantity $GD^2/V\mu$ all cancel out, so it's a dimensionless quantity. So the functional relationship has to look like this:

$$\frac{GD^2}{V\mu} = \text{const }_1 \tag{4.3.2}$$

You can easily rearrange Equation 4.3.2 to show that the velocity \(V\) of flow through the porous medium is linearly related to the quantity GD^2/μ:

$$V = \text{const }_2 \cdot \frac{GD^2}{\mu} \tag{4.3.3}$$

(where const_2 is just the inverse of const_1). The value of const_1 (or const_2) depends on the shape and packing of the particles in the pipe.

If now you made a number of runs with different water levels in your supply tank, you'd get a graph of the results that looks like Figure 4-9: you'd see a *linear relationship* between the velocity V and the quantity $GD2/\mu$. (How do you measure G and V, you might be thinking. Well, G is just the difference between the pressure at the upstream end of the pipe, which by the hydrostatic equation is just ρgh, where h is the height of the pipe entrance below the water surface, and the pressure at the downstream end of the pipe, which is the same as the atmospheric pressure, divided by the length of the pipe. And V is equal to the discharge divided by the cross-sectional area of the pipe.) Experiments like this have been done many times, and they show that there's a function like this for all porous media. But you should expect that *the value of the constant is different for different porous media, because of the differences in particle shape and packing geometry*.

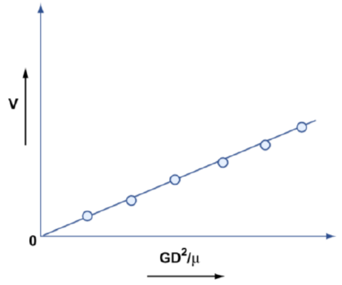


Figure by MIT OCW.

Figure 4-9. Graph of flow velocity V against GD^2/μ for flow through a porous medium.

4. Incidentally, the fact that you indeed find a linear relationship in your experiment shows you that we were correct in our assumption that the flow in the porous medium is laminar and that the density can therefore be neglected.

5. The results of the barrel experiment are a manifestation of a well-known law in the flow of porous media called ***Darcy's law***. Darcy's law states that *the flow speed through a porous medium is directly proportional to the pressure gradient through that medium, and directly proportional to the square of the characteristic size of the pore spaces of the medium, and inversely proportional to the viscosity of the medium*.

What's usually done with Equation 4 is to absorb the D^2 into the constant:

$$V = \text{const}_3 \cdot \frac{G}{\mu} \qquad (5)$$

The constant const_3 depends on the size distribution of the porous medium as well as on the particle shape and packing. It's called the ***intrinsic permeability*** of the porous medium, and it's usually denoted by small k. (It's misleading to call this quantity a constant. It's constant only for the particular porous medium we used in our home experiment! Each porous medium has its own value of intrinsic permeability.)

6. One final massaging of the relationship that started out as Equation 3 leads to another measure of permeability, called the ***hydraulic conductivity***, which is more commonly used in dealing specifically with groundwater flow. You have to make use of the concept of the ***hydraulic head***, which is *the level to which a column of water would rise if a tiny test column is inserted anywhere in the flow system* (Figure 4-10). This height h is related to the pressure p in the liquid by the hydrostatic equation,

$$p = \gamma h \qquad (6)$$

where γ is the weight per unit volume of the liquid. Remember that the pressure gradient G in Equation 3 should really be written $\Delta p/\Delta x$, where x is the direction down the pipe. Substitution of the expression for p in Equation 6 into $\Delta h/\Delta x$ gives $\gamma\ \Delta h/\Delta x$, and substituting this resulting quantity into Equation 5 gives

$$V = \text{const}_3 \cdot \frac{\gamma \frac{\Delta h}{\Delta x}}{\mu} \qquad (7)$$

where $\text{const}_3(\gamma/\mu)$ is the hydraulic conductivity. The hydraulic conductivity is usually denoted by capital K. You can easily see for yourself that the dimensions of K are *velocity*, because $\Delta h/\Delta x$ is a ratio of length variables, and therefore dimensionless; values of K are commonly cited in meters per day.

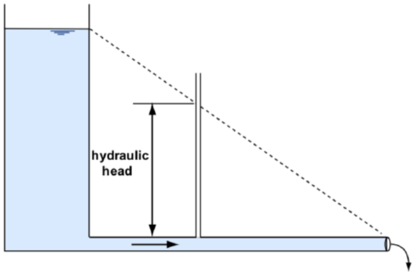


Figure by MIT OCW.

Figure 2-10. Sketch showing the hydraulic head in flow through a pipe packed with porous material.

What you would find, when you run the barrel experiment, is that the flow speed is directly proportional to the pressure gradient, and directly proportional to the size of the particles of the porous medium, and inversely proportional to the viscosity of the medium. This result is a manifestation of a well-known law in the flow of porous media called ***Darcy's law***. See the "advanced topic" above for details, if you are interested.

How the speed of flow depends on the pressure gradient, the particle size, and the fluid viscosity in this way is a reflection of a physical property of the medium that is known as the ***permeability*** of the medium. In a qualitative sense, the permeability is *a measure of how easy it is to force fluid through the porous medium by imposing a pressure gradient*. The permeability is related to the porosity (the permeability can't be very high unless there is substantial porosity), but it's not the same as the porosity. In fact, it's possible for the medium to have a high porosity but a low permeability, if the pore spaces are not sufficiently well connected; hence the concept of ***connected porosity***. The permeability of the medium is of great importance for groundwater studies, and it is also very important in the petroleum industry: you can't pump oil and gas out of deep sedimentary rock unless the both the porosity and the permeability are sufficiently great.

The home experiment on flow through a porous medium is realistic in all respects but one important one: the direction of flow is constrained to be straight down the pipe. That's not necessarily relevant to flow within a large volume of porous medium, as in the subsurface of the Earth. The critical question here is: *What determines the particular direction of groundwater flow within a large volume of a porous medium?* I can't pursue that question in any detail here, because it depends in a complex way on the dynamics of the water flow. There will be a bit more on this matter later in the chapter. If you would like to get some further insight into the patterns of flow, look into the following "advanced topic".

Advanced Topic: What Controls the Pattern of Flow of a Fluid

1. Think about the pressure of the water in some large tank like the supply tank for your home experiment when the water isn't moving. Think about a little unit area at the bottom of the tank. The pressure of the water at the bottom of the tank is equal to the weight per unit area of the water in the column overlying that small unit area, times the height of the column of water above that unit area. If the weight of the water per unit volume is γ and the depth to the bottom of the tank is h, then the pressure p at bottom of the tank is just γh. And by extension of that argument, the pressure of the water at any depth h *within* the tank is also equal to γ times that depth h. This water pressure in still water is called the ***hydrostatic pressure***.

2. The hydrostatic pressure within the still water in the tank is a manifestation of a balance between the weight of the water overlying a given point, which acts vertically downward, and the pressure gradient at that point, which acts vertically upward. So although there's a gradient of pressure in the tank, the water doesn't move, because that pressure gradient is offset by an equal and opposite force, namely the weight of the water.

3. Now suppose you took your ice pick (does anybody still have an ice pick?) and poked a hole in the side of the supply tank. Water would squirt out of the hole. In the interior of the tank in the vicinity of the hole the water is now in motion toward the hole. What you've done is impose a low pressure (namely, the atmospheric pressure) on the water at the hole, just the same as at the upper surface of the water in the tank. In doing so, you have disrupted the previously hydrostatic distribution of pressure near the hole-to-be and caused non-hydrostatic pressure gradients in the water near the hole—that is, pressure gradients that are no longer balanced by the weight of the water. That's the mechanical reason why the water flows from the tank!

4. The exact distribution of pressure in the vicinity of the hole and the resulting patterns of water motion are much too complicated for us to deal with here, but this example suggests that *the direction of water movement at any point is in the direction in which the difference between total pressure and hydrostatic pressure (a quantity called the* ***dynamic pressure****) decreases most rapidly*. It's differences in this dynamic pressure that cause fluid in any situation to move. So you can be confident that, whenever you are dealing with groundwater flow, the flow will always be in the direction of most rapid decrease in dynamic pressure. Figure 4-11 shows qualitatively the distributions of total pressure, hydrostatic pressure, and dynamic pressure in the tank once you've punched the hole and the water is flowing out. Note how *the flow lines are everywhere normal to the contours of dynamic pressure*.

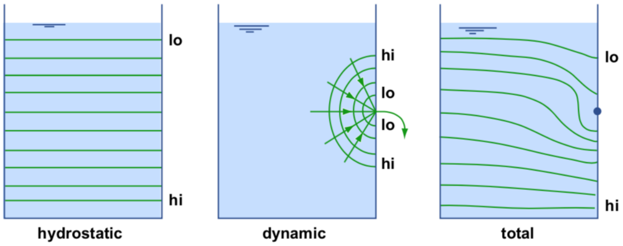
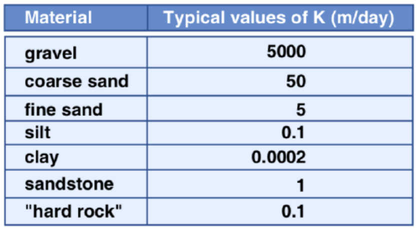


Figure by MIT OCW.

Figure 4-11. Distributions of hydrostatic pressure, dynamic pressure, and total pressure in outflow through a small hole in the wall of a barrel (qualitative).

5. Of course, the reasons for the distribution of dynamic pressure are always very complicated, and really beyond the scope of this course. Prediction of the spatial distribution of the dynamic pressure, and therefore the spatial pattern and speeds of water movement in the porous medium, are one of the major topics in the study of groundwater hydraulics. I'm not doing anything more than giving you the barest flavor of this endeavor.

I don't know what your intuition tells you about representative speeds of flow in your home experiment as a function of the nature of the porous medium, but Table 4-1 gives some representative values for various common kinds of porous media within the Earth. The values range enormously from coarse gravel, in which speeds are of the order of centimeters per second, to solid rock (which in reality is porous because of tiny spaces at grain boundaries and other miscellaneous rock fractures), in which speeds are of the order of a thousandth of a millimeter per second. Table 1 also gives corresponding values of hydraulic conductivity K, discussed in the following paragraph.



Material	Typical values of K (m/day)
gravel	5000
coarse sand	50
fine sand	5
silt	0.1
clay	0.0002
sandstone	1
"hard rock"	0.1

Figure by MIT OCW.

Table 4-1. Representative values of hydraulic conductivity K for various kinds of porous media.

Figure 4-12 shows a practical application of the principles we dealt with on the basis of the home experiment. There's a sloping land surface in which the deeper bedrock is covered by a fairly uniform but perhaps rather thick layer of loose and much more permeable material. High up on the slope is a source of pollutants, and farther down the slope, ten meters let's say, is your home or summer place, where you might have a water well. The groundwater flow is directly down the slope through the porous surficial layer at some speed that depends on the intrinsic permeability of the porous medium. If you know the slope of the ground and the intrinsic permeability of the material, you can compute the travel time of a pollutant tracer from the input point to the water well beneath your house. Assume for the sake of discussion that the slope of the ground is one in ten. The gradient in hydraulic head, $\Delta h/\Delta x$, within the porous medium is then 0.1. Using Equation 7 we find that V, the characteristic velocity of the groundwater, is $0.1K$. Using the representative values for K given in Table 4-1 you can obtain travel times for various kinds of porous medium (Table 4-2). You can see that depending on the permeability of the medium the grace period between the time of introduction of the pollutant and the time it pollutes your water well varies enormously.

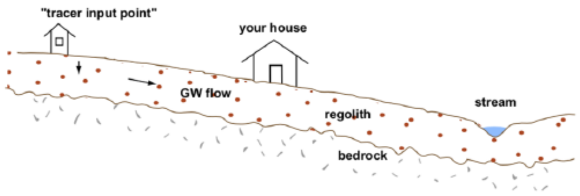
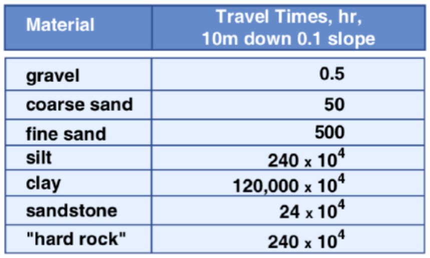


Figure by MIT OCW.

Figure 4-12. A tracer experiment in groundwater flow.



Material	Travel Times, hr, 10m down 0.1 slope
gravel	0.5
coarse sand	50
fine sand	500
silt	240×10^4
clay	$120{,}000 \times 10^4$
sandstone	24×10^4
"hard rock"	240×10^4

Figure by MIT OCW.

Table 4-2. Ten-meter travel times for the same kinds of porous media as in Table 4-1.

A Home Experiment on Groundwater Flow

Now it’s time to do a more realistic and ambitious experiment on the flow of groundwater (Figure 4-13). For this purpose you will need to build a very large shallow square tank in your backyard or use an entire spare room in your home. If you decide to use the spare room, you had better shore up the floor with some big timbers, because otherwise it’s likely to collapse.

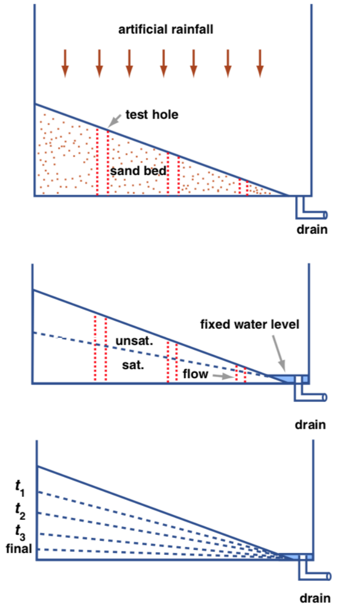


Figure by MIT OCW.

Figure 4-13. Another home experiment on groundwater flow.

Line the tank or the room with a large new polyethylene tarpaulin tucked in neatly at the corners, and fill the space with something like a meter of sand. Taper the layer of sand so that it has maximum depth at one side and zero depth at the other side, and along that latter side provide a drain or a sump pump at one corner. Put in several test holes or wells across the space, and line the holes with window-screen cylinders so that the sand doesn't fall into the holes. Now connect the hose to your kitchen sink and spray the sand surface to simulate a brief and heavy rainstorm.

Here's a summary of the results you would obtain:

- There will be water in the drain, and it will flow long after the rain stops.
- There will be water in all the test holes.
- The profile connecting the water level in the test holes will show the same sense of slope as the surface of the sand; this defines the groundwater table. The ***water table***, also called the ***groundwater table***, is *the locus of points where the water pressure is equal to the atmospheric pressure*. It's the top of the permanently saturated zone.
- The slope of the groundwater table is less than the slope of the sand surface, and it decreases with time (see the lowest part of Figure 4-13).
- The movement of groundwater is in the downslope direction, toward the drain. You can tell this by injecting food coloring in the uppermost test hole and seeing it appear in successive holes and finally at the surface, at the drain.
- If the rainfall is too heavy there will be some surface runoff down to the drain, but even in this case most of the water will infiltrate and become part of the groundwater flow.

This is a very realistic experiment in groundwater flow. The only problems with it are these:

- The scale is too small, and things happen too fast.
- The material is uniform and has a fixed and permeable floor. In real life there's usually a gradual downward decrease in both porosity and permeability—although in glaciated areas the structure of the subsurface is often just the same as in this experiment.

4.4: Qualitative Aspects of Groundwater and Groundwater Flow

Figure 4-14 shows a vertical cross section through a representative area of the Earth's surface, showing some of the important features of distribution of groundwater. The part of the subsurface lying *above* the groundwater table is called the ***vadose zone*** or ***aerated zone*** or ***unsaturated zone***. Most of the time, the pore spaces in the vadose zone are occupied mostly by air (plus locally generated gases); only right after a heavy rain are the pore spaces filled with downward- percolating water. The part of the subsurface zone lying *below* the groundwater table is called the ***phreatic zone*** or the ***saturated zone***. In the phreatic zone the pore spaces are always filled with water.

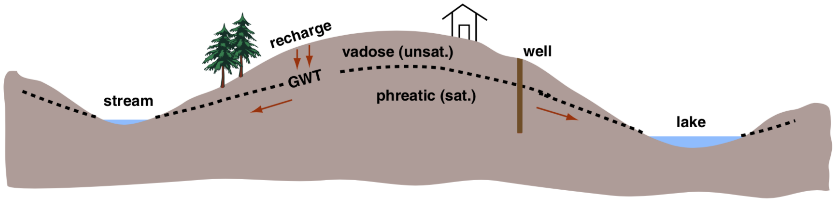


Figure by MIT OCW.

Figure 4-14. Vertical cross section through a representative hill-and-valley area of the land surface, showing features of distribution of groundwater.

In reality the interface between the vadose zone and the phreatic zone (that is, the groundwater table) is not a sharp and well-defined surface: it's a fuzzy zone of transition. That's because of capillary rise of water up into the pore spaces of the lowermost part of the phreatic zone. (See the background section on surface tension and capillarity earlier in the chapter.) The zone of partial saturation is called the ***capillary fringe***. Its thickness ranges from just a centimeter or two, in coarse gravels, to as much as a couple of meters, in fine silt-rich sediments. The reason that there is a gradation in the degree of saturation is that, owing to the spread of particle sizes, adjacent pore spaces vary widely in their effective size.

The topography of the groundwater table mimics the topography of the land surface itself: it's high under hills and low under valleys. But the topography of the groundwater table is more subdued than that of the land surface, because the depth to the groundwater table is greatest under hills and least under valleys. In fact, *the groundwater table intersects the land surface at rivers and lakes and springs*. This should not be surprising to you if you think of rivers and lakes as just the places where the permanently saturated zone emerges from beneath the mantle of porous medium!

There's no necessary relationship between the groundwater table (the surface of contact between the vadose zone and the phreatic zone), on the one hand, and the contact between bedrock and regolith, on the other hand. In areas with a relatively thin mantle of regolith and a relatively deep water table, the water table lies mostly within bedrock (Figure 4-15A). In areas with a relatively thick mantle of regolith and a relatively shallow water table, however, the water table lies mostly within the regolith (Figure 4-15B). In New England, both extremes are common, basically because of the highly variable depth of regolith resulting from glacial erosion, transportation, and deposition of regolith during the last Ice Age.

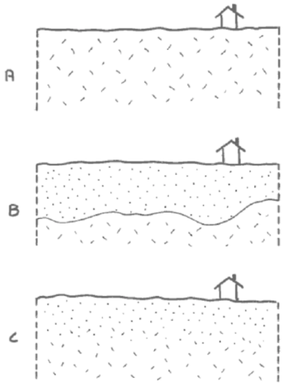


Figure 4-15. Groundwater table vs. position of regolith–bedrock contact.

Figure 4-16 shows a record of the level of the water table in a well located as shown in Figure 4-14 during a representative year. In this area, the water table fluctuates vertically by about ten feet in the course of the year. Fluctuations might be considerably smaller or larger than that, depending partly on the *variability of rainfall* but also very importantly on the *permeability of the subsurface material*. Obviously, the water table tends to be high during spring and fall rainy periods and low during summer droughts. But the water table tends to be *low during the winter* also, because the ground at the surface freezes to a depth of a meter or even more, preventing, or at least impeding, recharge, while the groundwater at greater depths continues to flow and thus lower the water table. (The term ***recharge*** is used to describe *replenishment of groundwater in a subsurface region from which groundwater was previously withdrawn*, either naturally or by human activities.) Another noteworthy thing about the record in Figure 4-16 is that *the water table rises faster than it falls*. This is because recharge involves percolation vertically downward through a relatively thin layer of porous material, whereas drainage involves slower groundwater movement through long distances down low gradients of hydraulic head.

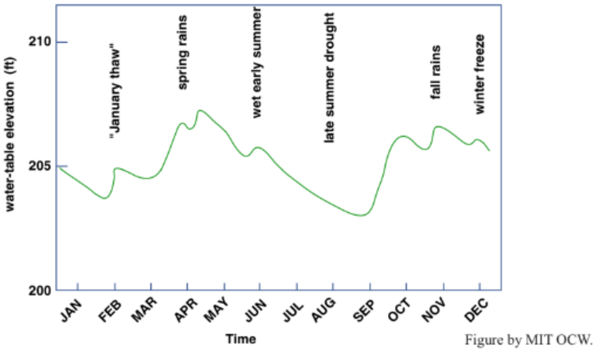


Figure 4-16. Representative record of the level of the groundwater table in a well.

Here's an incidental note on the depth of water wells. Usually a well is drilled far below the level of the local water table, for two reasons:

- to ensure that the bottom of the well stays below the water table even during severe droughts;
- to get below the level of polluted near-surface waters (but the problem is that even the deeper waters don't stay unpolluted forever!).

The prices one pays for deeper wells (aside from the dollar cost per foot of well) are:

- generally, the deeper the well the smaller the porosity and permeability of the medium, and so the lower the rate of flow into the well;
- The slower the flow into the well, the longer the residence time of the groundwater in the porous medium, so the longer the time available for uptake of ions from the medium, so the harder the water.

An ***aquifer*** is *any body of porous material in the Earth (rock or regolith) with sufficient volume and connected porosity to yield appreciable water to wells or springs*. Aquifers contain relatively large drainable porosity, relatively large volume, and relatively high permeability. The concept of an aquifer is a loose one, partly because of those sneaky words *appreciable* and *relatively*: The term is used for any volume of subsurface material that's a good producer, or a potentially good producer, of water, relative to surrounding volumes of subsurface material. The minimum lateral dimensions of what are called aquifers might be as little as tens of meters, and the minimum vertical dimensions might be as little as some meters. Large aquifers, however, might have lateral dimensions of hundreds of kilometers and vertical dimensions of many tens of meters.

The opposite concept is that of an ***aquiclude***: *any body of subsurface material through which water can move at only negligible rates*, or at least at rates much smaller than through adjacent aquifers. Also, the term ***aquitard*** is used for any body of subsurface material through which groundwater travels slowly, relative to some adjacent aquifer, but not so slowly as to be negligible, as in an aquiclude.

Aquifers can be classified into four kinds (Figure 4-17):

1. ***unconfined aquifer:*** an aquifer in which the groundwater is in direct contact with the overlying atmosphere through connected pore space.
2. ***confined aquifer:*** an aquifer overlain by an aquiclude so that it is not in contact with the atmosphere except in some upstream area of recharge.
3. ***leaky aquifer:*** a confined aquifer whose overlying aquiclude allows some non-negligible passage of groundwater into or out of the aquifer.

4. ***perched aquifer:*** an unconfined aquifer present above a shallow and laterally restricted aquiclude.

Aquifers and aquicludes can exist on regional as well as local scales. Figure 4-18 shows a cross section through such a system. A highly porous and permeable sedimentary formation, like a well sorted and poorly cemented sandstone, is overlain by a highly impermeable shale in a large region with very low angles of dip of the formations. Wells may tap the confined aquifer at distances of hundreds of kilometers from the recharge area, where the aquifer formation is exposed at the surface. Note that there's also a much shallower unconfined aquifer that derives its water much more locally.

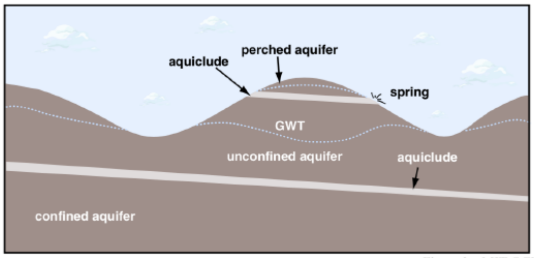


Figure by MIT OCW.

Figure 4-17. Schematic vertical cross section showing various kinds of aquifers.

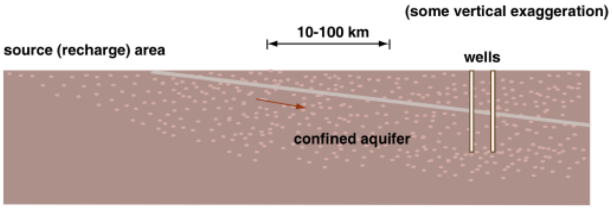


Figure by MIT OCW.

Figure 4-18. Schematic vertical cross section showing regional extent of aquifers and aquicludes.

If the surface slope of the region is greater than the slope of the surface of hydraulic head associated with the regional confined aquifer, a well drilled to the confined aquifer will produce *a water flow at the ground surface*, with no need for pumping (Figure 4-19). Such a well is called an ***artesian well***. Figure 2-20 shows a homey analogy in which you water your garden with a hose leading from an elevated tank of water. In this situation, the nozzle of the hose is a kind of artesian well. And, in a broad sense, if you live in an area with a central municipal water supply all of the faucets in your home or apartment are artesian wells!

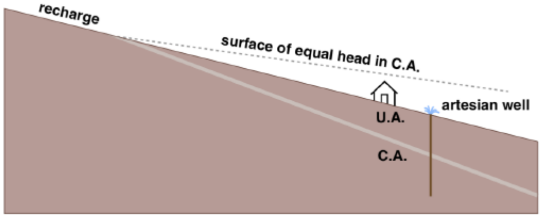


Figure by MIT OCW.

Figure 4-19. Schematic vertical cross section showing the nature of artesian wells.

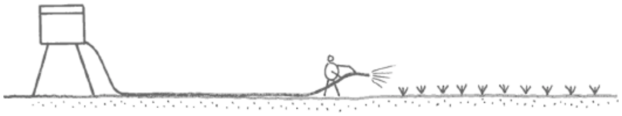


Figure 4-20. A home analogy for artesian wells.

Here's one more concept that's useful in dealing with groundwater supplies: The ***specific yield*** of an aquifer is *the ratio of the volume of water that drains out of the aquifer (when the groundwater table is lowered) to the total volume of the aquifer subjected to drainage*. The specific yield obviously depends on the *porosity* of the aquifer, but it also depends on the *size of the pore spaces*, because some water always adheres to the surfaces of the solid materials of the aquifer, both as thin films on surfaces and as fillets in reentrants, and the smaller the size of the pore spaces, the greater the percentage of the total porosity that remains occupied by this adhering water. This concept of specific yield is analogous to that of the field capacity of a soil.

I haven't said anything yet about the patterns of flow within a typical aquifer. The complexities lie in flow through deep unconfined aquifers, because shallow unconfined aquifers (as in Figure 4-12) or confined aquifers (as in Figure 4-18) can be treated approximately as flow in rivers or closed ducts. The patterns of flow in a porous medium are a response to the spatial distribution of gradients of dynamic pressure. No simple statement about the patterns of motion can be made without setting up the geometry of the problem and solving the equation (essentially Newton's second law written in the appropriate form) that governs the motion.

Figure 4-21 shows qualitatively the flow patterns in a typical deep unconfined aquifer in an area of irregular surface topography. Just below the water table in areas where the water table is sloping, the flow is as you might expect: downward and parallel to the water table. But near the crest of the water table the flow is nearly vertically downward, and directly beneath where the water table merges with rivers and lakes the flow is nearly vertically upward. At points deep within the aquifer, the directions of flow bear no obvious resemblance to the surface topography of the water table.

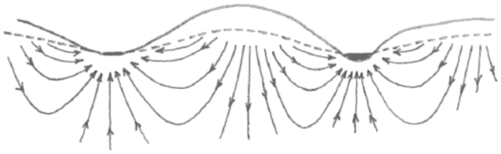


Figure 4-21. Qualitative flow patterns in deep groundwater.

4.5: Some Practical Things About Water Wells

A great many households in rural or semi-rural areas depend upon wells for their water supply. In earlier times, shallow wells were dug by hand to depths of a few tens of feet. In more recent times, wells have been driven by various techniques to much greater depths of many hundreds of feet to tap deeper groundwater. Most such driven wells are lined with a jacket or casing that's porous in the lower part of the well. This casing keeps the walls from collapsing or crumbling inward, without inhibiting the water flow. This section presents some aspects of the flow of water in the immediate vicinity of wells.

The first thing to keep in mind is that *the removal of water through a well disrupts the normal or background flow of groundwater in the vicinity by producing a locally lower water pressure in the lower part of the well*. The relatively strong pressure gradients near the well cause a flow of water radially inward toward the well. The difference in elevation between the depressed water level in the operating well and the elevation the water table would have if water were not being drawn from the well is called the ***drawdown*** of the well (Figure 4-22). The locally cone-shaped surface of the water table in the vicinity of the well is called the ***cone of depression***. In areas with closely spaced wells, the cones of depression of the wells may interfere with one another to produce a general lowering of the water table in the area (Figure 4-23).

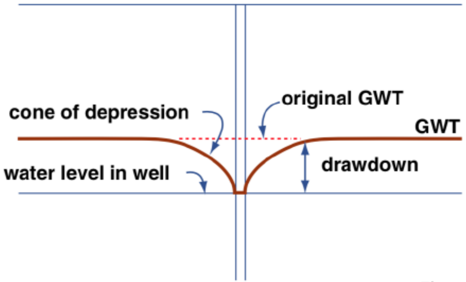


Figure by MIT OCW.

Figure 4-22. Geometry of the groundwater table around an active water well.

Your well becomes polluted from an upstream pollution source only if its intake lies in the path taken by the pollutant. A pollutant introduced at a point source reaches the groundwater table by downward percolation and then travels with the groundwater flow as a plume, narrow at first but gradually widening because of slow lateral diffusion (Figure 4-24). (A ***plume*** is a mass of fluid that moves through ambient fluid, as a consequence of a difference in density or just because that fluid is injected, over some period of time, into the flowing ambient fluid at a point, as is the case here. A good example of a plume of the former kind is smoke that rises out of a chimney or a smokestack.)

A well upslope of the point of pollutant injection can tap a pollutant plume, if the distortion of the natural flow pattern is so great as to produce *locally upslope flow toward the well*. The situation shown in Figure 4-25, in which all of the pollutant plume is drawn up by the well, is an extreme example.

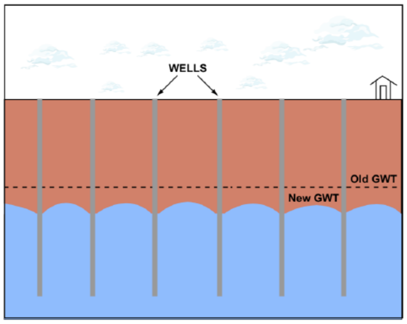


Figure by MIT OCW.

Figure 4-23. Geometry of the groundwater table in the vicinity of a number of closely spaced active water wells.

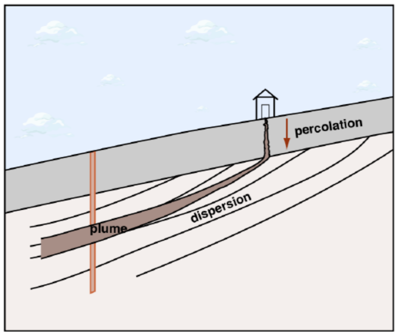


Figure by MIT OCW.

Figure 4-24. A well lying in the path of a pollutant plume.

If in the long term one or more wells in a given area remove water faster than it can be supplied by recharge by percolation from surface rainfall, the water table is lowered. In areas with shallow wells and abundant recharge, where natural flow rates are high, the water table responds rapidly, in hours or days. But in areas with deep wells and slow recharge, the depressed water table might take many years or even decades to readjust, even if no water at all is drawn from the wells. In a very real sense, groundwater is being *mined* by these wells, because recharge is on a time scale much longer than the lifetime of the wells. This is the case today in many areas of the arid and semiarid central and western parts of the United States.

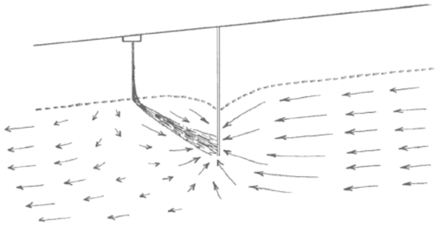


Figure 4-25. Locally upslope movement of a pollutant plume toward an active well.

4.6: Groundwater in Coastal Regions

So far we have been concerned with groundwater in inland areas well away from the seacoast. In coastal areas, *what's the interaction of groundwater with the sea?* You know that fresh water is less dense than salt water, by a small but important few percent. So where groundwater is in contact with seawater, the groundwater should tend to float upon, or rise through, the seawater.

Perhaps you can get the clearest picture by considering a small, porous, rainy island in the ocean (Figure 4-26). (Nantucket and Martha's Vineyard are good examples—although Chamber of Commerce members probably would not appreciate that description.) Assume that the seabed all around the island is saturated with seawater. Provided that the seabed consists of porous material, this is an excellent assumption. Then the body of fresh water that constitutes the groundwater under the island is itself floored by an even lower layer of denser seawater. The sharpness of the boundary between the groundwater and the seawater is determined by the relative importance of groundwater flow, which maintains the sharpness of the boundary, and diffusion of the salt from the seawater into the groundwater, which tends to blur the boundary. The boundary emerges at the seabed some distance from shore, and landward from that line groundwater flows out to the seabed and then rises and mixes with the water of the ocean.

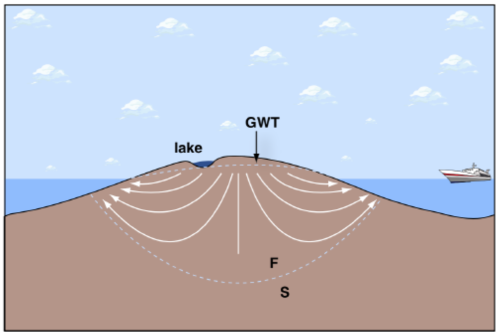


Figure by MIT OCW.

Figure 4-26. Pattern of groundwater flow in a coastal region.

How deep is the boundary between groundwater and seawater under the island? That depends on the elevation of the groundwater table above sea level within the island. *The depth of the boundary below sea level is just about 40 times the height of the groundwater table above sea level*—because the ratio of sea-water density to fresh-water density is about 41/40. It's a simple problem of buoyancy. If you are confused, think about floating a balloon full of fresh water in a tub of seawater. The ratio of submergent depth to emergent height of the balloon is governed by exactly the same principle.

Now suppose you drilled a well somewhere on the island. At first you are tapping only freshwater. But now you have to worry about two cones: not just the cone of depression that affects the groundwater table, but also an inverted cone that represents what might be called the "drawup" of the bounding surface between the groundwater and the seawater (Figure 4-27). If you pump fresh water out of the well so vigorously that the boundary is drawn all the way up to the base of the well, you start to pump seawater as well as fresh water from the well. This effect is called ***salt-water intrusion***. Salt-water intrusions of this kind are common in coastal areas underlain by porous materials, as on Cape Cod and Long Island. The problem is reversible, but there is only one remedy: *reduce the rate of pumping*.

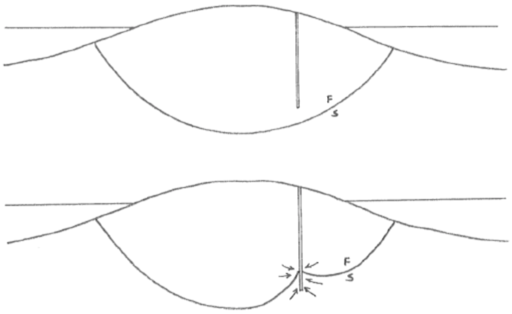


Figure 4-27. Effect of pumping on the position of the fresh–salt boundary beneath a coastal region.

READINGS

Easterbrook, D.J., 1999, Surface Processes and Landforms, Second Edition. Prentice Hall, 546 p. (Chapter 7)

CHAPTER OVERVIEW

5: Rivers

5.1: Introduction

Introduction

On the continents, except in the most arid regions, precipitation exceeds evaporation. Rivers are the major pathways by which this excess water flows to the ocean. Over the continental United States the average annual rainfall is about 75 centimeters. Of this, about 53 centimeters is returned to the atmosphere by evaporation and transpiration. The remaining 22 centimeters feeds streams and rivers, either directly (by landing in the channels or running off across the surface) or indirectly, by passing through the shallow part of the Earth as groundwater first. This 22 centimeters represents an enormous volume of water: 5.2 x 108 cubic meters per day (1.4 x 1011 gallons per day).

Rivers are also both the means and the routes by which the products of weathering on the continents are carried to the oceans. Enormous quantities of regolith are produced on the land surface by weathering, and most of this material is transported by rivers to the sea, either as particles or in solution. The other two principal agents that transport this material to the ocean, glaciers and the wind, are minor in comparison.

Rivers and ***streams*** (which term you use is a flexible matter of scale) are *channelized flow sofwater on the Earth's surface*. The term ***overland flow*** is used for *non-channelized flows of water*, usually less than a few centimeters deep but very widespread. There is a pronounced dichotomy between non-channelized flow and channelized flow. Have you ever walked up a small stream channel to see what happens to it? Its termination is almost always well defined.

Rivers are enormously *diverse*, in:

- **size**: varies by many orders of magnitude
- **geometry**: highly variable
- **substrate**: bedrock or sediment
- **sediment type:** sediment size ranges from mud to gravel
- **stage of development**: young, with rugged topography and rapid change, to old, with gentle topography and slow change
- **climate**: ephemeral and flashy to very steady

No two rivers look exactly alike, but we can talk about many things that most if not all rivers have in common, like

- how to analyze the hydrology of rivers
- the dynamics of turbulent open-channel flow • the dynamics of fluvial sediment transport

The classic areas of study of rivers are these:

- ***fluvial hydrology:*** the study of water *as water* in rivers
- ***fluvial hydraulics:*** the study of the flow of rivers
- ***fluvial sedimentation:*** the study of sediment movement and in rivers
- ***fluvial sedimentology:*** the study of sediments in rivers (overlapping strongly with the preceding field) and of fluvial sedimentary deposits
- ***fluvial geomorphology:*** the study of fluvial geomorphic processes (things like sediment movement, channel changes, broader river-valley processes, drainage-network development) and the long-term evolution of rivers, river valleys, and drainage systems

Some Basic Characteristics of Rivers

- Rivers have a *wide range in size* (as measured by either water discharge, sediment discharge, or length).
- Rivers have a *wide range of water discharge* and an *even wider range of sediment discharge*, as a function of time.
- Rivers are *curvy*; they are seldom straight for a long distance.
- Rivers don't stay in one place: they *shift laterally* in various ways and at various rates, so there's at least temporary *deposition* at many places in the river system.
- Most medium to large rivers can *keep pace with crustal subsidence or uplift* in some reach of the river by erosion or deposition.
- Rivers have a *long history*. (How does a river start?)

5.2: Fluvial Hydrology

Measurement of Streamflow

Two aspects of streamflow are typically monitored on all major streams and an enormous number of minor streams as a function of time on a regular basis (in United States, mostly by the U.S. Geological Survey):

Stage. The ***stage*** of a river is *the height of the water surface of the stream above an arbitrary datum*, usually either sea level or an elevation slightly below the channel bed (Figure 5-1). Stage is related to depth, but the two are not the same.

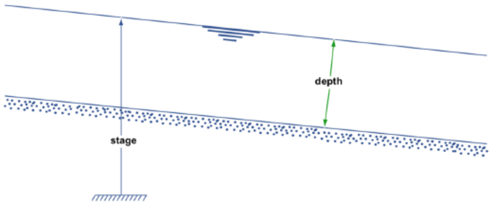


Figure by MIT OCW.

Figure 5-1. The stage and depth of a river.

The stage of a river is fairly easy to measure. Various kinds of *stream gauges* are in use. The simplest is a permanent vertical surface with vertical scale markings you read directly on a regular basis. What's more desirable, but also much more expensive, is an automatically and continuously recording gauge. There are various kinds of such gauges in use; most common is a float gauge, in a stilling well, connected to a strip-chart recorder (Figure 5-2).

Figure 5-2. A stream-gauging station.

Discharge. The ***discharge*** of a river is *the volume rate of flow past a given cross section*, measured in cubic feet per second, cfs (cusecs) or cubic meters per second, m^3/s (cumecs). It's not nearly as easy to measure discharge as it is to measure stage. Most measurement of river discharge makes use of a simple equation that relates discharge Q past a cross section to the area A of the cross section and the mean velocity U of flow past that cross section:

$$Q = UA \tag{5.2.1}$$

If Equation 1 doesn't make sense to you immediately, just imagine that at a given instant you could magically mark the water that's passing through the given cross section and then watch that marked surface drift downstream for a unit interval of time (Figure 5-3). The distance traveled by that surface is, on the average, equal to the mean velocity U, because velocity is distance per unit time. The volume of water between the given cross section and the marked surface after that unit time, which by definition is just the discharge Q, is the product of the cross-sectional area A and the distance traveled by the surface during that unit time, which is equal to the velocity U, times 1, the unit of time.

So if you measure cross-sectional area and mean flow velocity you can solve for the discharge. This is easier said than done, but it's what's done in practice. There's an elaborate set of practical guidelines for doing this in natural streams; the procedure basically involves

- taking a large number of positions across the stream,
- measuring depth-averaged mean velocity (and depth), at each position, and then
- computing and averaging.

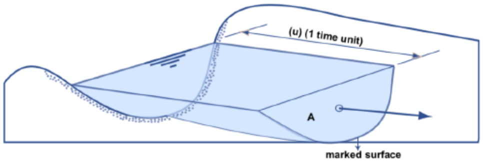


Figure 5-3. Seeing what is meant by Equation 1 for river discharge.

Obviously there's a certain unavoidable sloppiness to this, especially under difficult high-water conditions. It's done from bridges, cableways, or boats. The accuracy is something like 5% at best, 10–15% at worst. One tries to do it fast relative to change in stage and discharge.

Clearly, you can't measure the discharge continuously, or even often. But it's important to be able to know the discharge at times you don't (or can't) measure it. *How do you derive a continuous record of discharge?* The technique involves what is called a ***stage-discharge diagram***, often called a ***rating curve***. The rationale is that if downstream conditions don't change (for example, by aggradation or degradation occasioned by new structures built upstream, or the shifting of meander bends upstream) this curve is the same all the time, so once you have it you can find the discharge satisfactorily just by knowing the stage and going into the diagram. A lot of effort goes into deriving and checking rating curves. Rating curves are usually convex upward, as in Figure 5-4.

Hydrographs

It should seem natural to plot the results of streamflow measurements in the form of *a graph of stage vs. time or discharge vs. time*. The latter is most common and useful. Both kinds of graph are called ***hydrographs***. Time scales and discharge data used in hydrographs vary widely:

- To study individual floods, you need a "continuous" record of discharge for days or weeks (Figure 5-5A).
- A common longer-term hydrograph shows peak or mean daily discharge for a whole year (Figure 5-5B).
- Longer-term hydrographs show mean monthly or annual discharge over many years (Figure 5-5C). Hydrographs vary widely from river to river depending on climate and substrate. This reflects the circumstance that rivers can be *flashy* or *steady*.

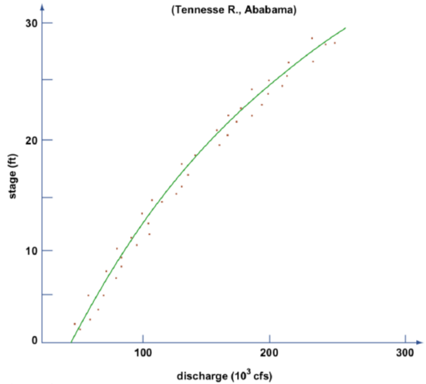


Figure by MIT OCW.

Figure 5-4. Example of a rating curve.

What is the characteristic or typical short-period hydrograph of a stream produced by a rainfall of given duration with given steady intensity? Refer to Figure 5-6.

AB: end of spell without rainfall; all surface runoff has ceased, and groundwater runoff is gradually decreasing.

B: surface runoff from a rainstorm reaches the channel.

BC: this is the ***rising limb*** of the hydrograph; surface runoff increases sharply.

C: this is the ***peak*** or ***crest*** of the hydrograph; surface runoff peaks.

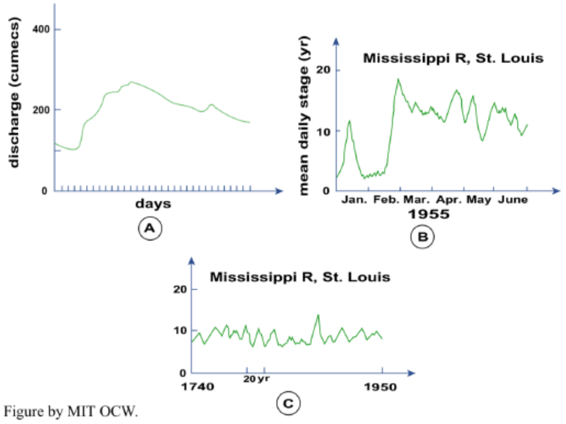


Figure by MIT OCW.

Figure 5-5. Examples of rating curves. A) Discharge hydrograph of n individual flood (conceptual). B) Stage hydrograph of the Mississippi River at St. Louis for an entire year. C) Stage hydrograph of the Mississippi River for over two centuries.

It's also possible to show both surface runoff and groundwater runoff on the same hydrograph (Figure 5-7). Note that sometimes groundwater flow actually goes negative, because of ***bank storage***: instead of the groundwater feeding the river, the river feeds the groundwater table.

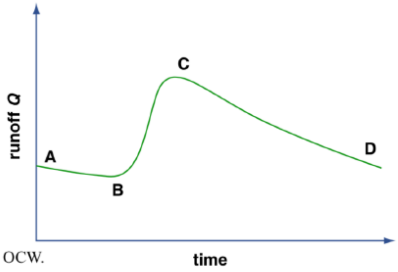


Figure by MIT OCW.

Figure 5-6. Characteristic short-period hydrograph of a stream, associated with a rainfall event of short duration and steady intensity.

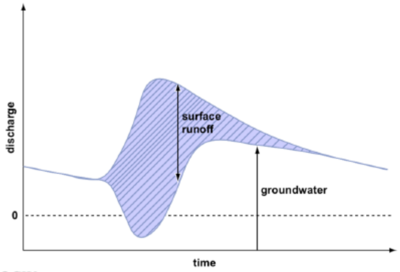


Figure by MIT OCW.

Figure 5-7. Showing surface runoff and groundwater runoff on the same hydrograph.

How does one account for the shape of the rising limb of the hydrograph? Think about a tiny "drop" of water falling on the watershed. Assume that it's not infiltrated or reevaporated but travels as surface runoff. At first it travels as ***overland flow***, and then later as ***channelized flow***. (In fact, watersheds can be classified as "small" or "large" depending on whether the ratio of time involved in overland flow to time involved in channel flow is large or small, respectively.) Eventually the drop passes a given station at the outlet of the drainage basin. This takes a certain average time. You can imagine a map of the drainage basin upstream of this point as being contoured by ***isochrons*** (curves of equal travel time) (Figure 5-8)—although it would be very difficult to compute or measure this in actual practice.)

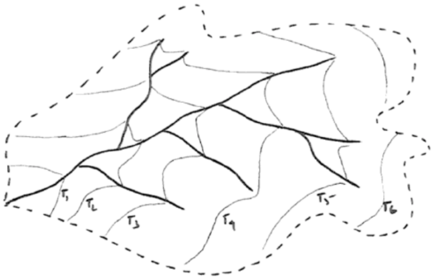


Figure 5-8. Map of a small drainage basin, showing isochrons of equal travel times of surface runoff to the outlet of the basin.

Assume a brief and uniform rainfall over the watershed (that is, the rain doesn’t last long, and the rate of precipitation is the same everywhere in the basin). Surface runoff at our station can be accounted for by looking at cumulative area as a function of time (Figure 5-9). For our brief and uniform storm, this should be equivalent to the rising limb of the hydrograph for this storm if we make the transformation

Runoff = (rainfall depth) times (area)

It should make sense to you that the time it takes to reach the peak of the hydrograph thus derived (that is, the time it takes until all of the watershed area is now contributing to the discharge at the station) gets longer as the watershed area gets bigger. We haven’t dealt with the falling limb of the hydrograph here, but, obviously, after the rain stops, less and less of the watershed area is contributing water, so the discharge past our station gradually decreases.

A hydrograph is this simple only for an ideal rainstorm in a very small watershed. Hydrographs of real rivers are invariably more complicated, for obvious reasons involving *unsteadiness* (varies with time) and *non-uniformity* (varies from place to place) of rainfall.

It’s difficult to say anything striking about hydrographs at this point, but the differences in discharge reflected in the hydrographs are of paramount importance in both the channel pattern and the sedimentary processes in streams.

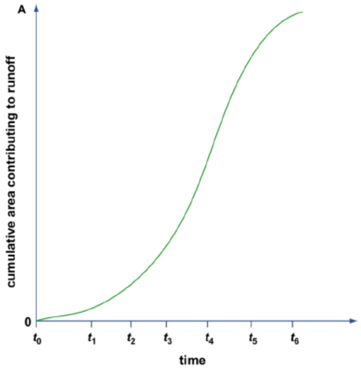


Figure by MIT OCW.

Figure 5-8.

In terms of *flashiness* (qualitative magnitude of change in discharge from time to time), one of the most important aspects of river behavior for sedimentation, you can plot a similar *diagram of discharge vs. percentage of time* (this is called a ***flow-duration curve*** (or a ***discharge-duration curve***) if you have a long record of discharge. Figure 5-10 shows such a curve for two medium-size streams in Ohio, of about the same size and discharge:

1. Sandy Creek, Sandyville, Ohio: A = 481 mi^2, underlain by surficial sand and gravel, good porous aquifers;
2. Rocky River, Berea, Ohio, A = 269 mi^2, glacial till and clay, highly impermeable material.

On the vertical axis is plotted Q/A instead of just Q, to normalize for the area of the drainage basin.

The curve for Rocky River shows greater peak flows and lesser base flows; this is described as *flashy* behavior. Sandy Creek, on the other hand, shows lesser peak flows and greater base flows; this kind of behavior is *steadier* and not as flashy.

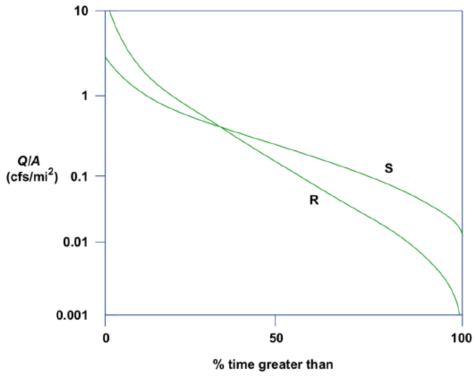


Figure by MIT OCW.

Figure 5-10. Discharge–duration curve for two small rivers in Ohio.

5.3: Open-Channel Hydraulics

This would a good place to go back and review the material on open-channel flows in Chapter 1. Recall that in open-channel flows the presence of the free surface means that the geometry of the flow can change in the flow direction not just by being constrained to do so by the geometry of the boundaries but also by the behavior of the flow itself. This means that the acceleration of gravity can no longer be ignored, because the forces of gravity help shape the free surface. For example, gravity waves can be generated on the free surface. But the effect of gravity is more far-reaching than just that. Babbling brooks and white-water rivers clearly have complex free-surface geometries governed by bed relief, expansions and contractions of the channel, and, less obviously, upstream and downstream conditions. But all open-channel flows, even broad, majestic rivers like the Mississippi, are subject to such effects of gravity.

Here are three basic concepts in the hydraulics of open-channel flow I think you should know about:

bed shear stress: The ***bed shear stress*** is *the force per unit area the flow exerts on the bed.* Actually the force per unit area varies strongly from point to point, depending on the details of the bed geometry, and the concept of boundary shear stress is built around the idea that you average over an area large enough to eliminate the effects of such things as sediment particles or local sediment topography. The importance of the bed shear stress lies in its role in moving the sediment particles resting on the river bed.

flow resistance: The resistance to flow, or ***flow resistance***, is the force or drag the boundary exerts on the flow. You should recognize that in uniform flow, where the flow isn't accelerating or decelerating, Newton's third law tells us that this is just the opposite of the bed shear stress. The importance of the flow resistance is less easy to state. It has to do with the role of flow resistance in determining the particular combination of flow depth and flow velocity (out of an infinite number of possible combinations) with which the imposed water discharge is passed through a given reach of a river (Figure 5-11).

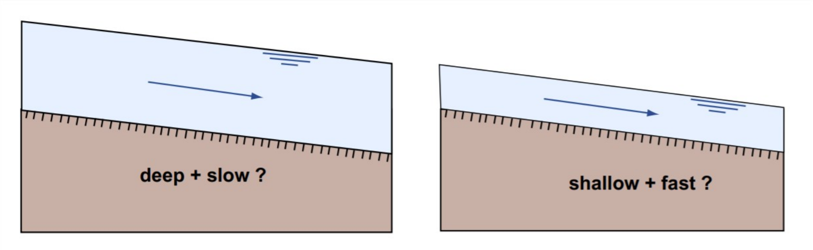


Figure 5-11. Which combination of depth and velocity to pass a given discharge through an open channel?

slope: If rivers were always straight, the definition of the ***slope*** would be a straightforward concept: it's the difference in water-surface elevation between two stations along the river, divided by the horizontal distance be stations (Figure 5-12). But if the river is curvy (as is usually the case, at least to some extent), then you have to measure the horizontal distance along the sinuous projection of the course of the river on a horizontal plane (Figure 5-13). The slope can be measured in feet per mile (as in the U.S.) or in some metric units like meters per kilometer. Recalling some trigonometry, you might recognize the slope as the tangent of a slope angle. Measuring the slope of a river is not easy: you have to do some surveying to establish elevations, and you have to worry about what, exactly, is the course of the river.

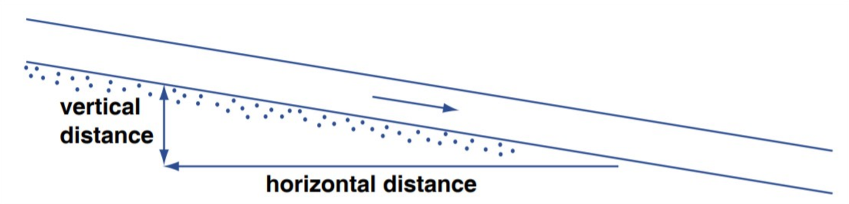


Figure 5-12. The slope of a river.

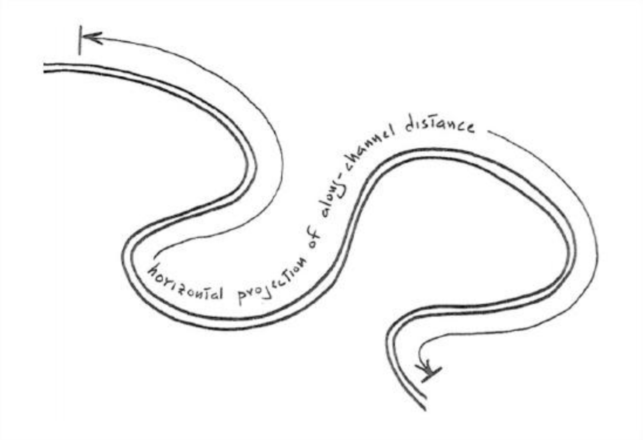


Figure 5-13. Measuring the horizontal component of the distance along a sinuous river.

Advanced Topic: The Resistance Equation For Open-channel Flow

1. It’s easy to derive a fundamental equation that relates the flow depth, the slope, and the bed shear stress of a river, if you are willing to assume that the flow in the river is about the same in cross-sectional shape and area at all cross sections (such rivers are said to have ***uniform flow***, which is often close to being the case).

2. Think about the water contained in a volume that’s formed by the river bed, the free surface, and two cross sections a unit distance apart (Figure 5-18). One of the classic ways to get somewhere in the analysis of a problem in dynamics (and we’re dealing with such a problem here) is to apply Newton’s second law, $F = ma$, where F is the force on some body of matter, m is the mass of the body, and a is the acceleration of that body under the action of that force) to an appropriately chosen part of the dynamical system. What I want you to think about here are the forces that act on the water in the volume I just defined, which from now on I’ll refer to as the “body”. Because the flow is uniform, and the river discharge varies only slowly with time, it’s a good assumption that the body is not accelerating. So Newton’s second law tells us that the sum of all the forces acting on the body in the streamwise direction has to be zero.

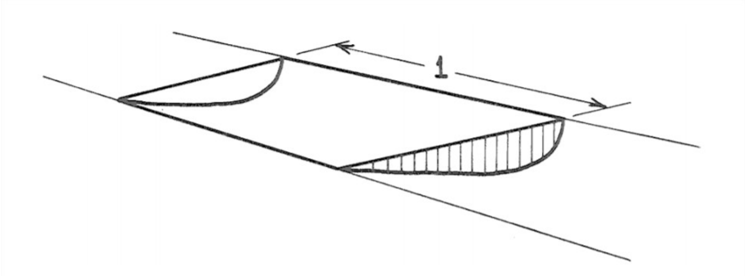


Figure 5-14. The water contained in a volume that is formed by the river bed, the free surface, and two cross sections a unit distance apart.

3. What are the forces acting on the body in the streamwise direction (Figure 5-15)? First of all there are hydrostatic fluid pressure forces on both the upstream and downstream faces of the body. We can forget about these, because they are the same upstream and downstream and they act opposite to each other. There’s the weight of the body—that’s a gravity force acting vertically downward—and a component of that weight acts in the downstream direction. It’s this downstream component of the weight that pulls the water down the channel. If the weight per unit volume of the water is γ , and the cross-sectional area of the flow is A, and the slope angle is α, then the downstream component of the weight is $(1)(A)(\gamma)\sin\alpha$. Finally there’s the upstream-directed frictional force exerted by the boundary on the moving body. It’s this upstream-directed frictional force that resists the downstream-directed gravity force. If the wetted perimeter of the flow (that is, the total distance along the line of contact between the flow and the bed, as viewed in a cross section normal to the river flow) is P, then the frictional force is $(1)(P)(\tau_o)$. Writing the balance between the frictional force and the gravity force, we have

$$(1)(P)(\tau_o) = (1)(A)(\gamma)\sin\alpha$$

or, doing a little rearranging,

$$\tau_o = \gamma \frac{A}{P} \sin \alpha$$

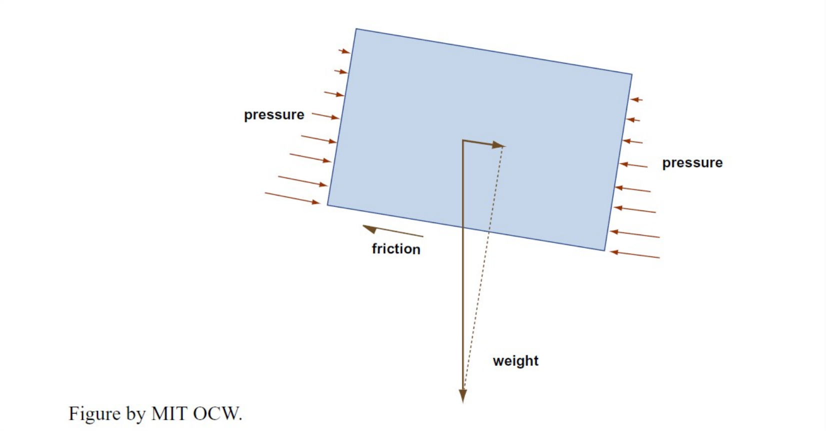


Figure by MIT OCW.

Figure 5-14.

4. A slightly different and more specific way of obtaining a relationship like this is to assume that the width of the river is much greater than its depth, which is often the case. Then, if you look at a body that's like the one used above but is rectangular volume a unit length long and a unit length wide (Figure 5-16), the downslope component of the weight of the body is (1)(1)(*d*)sin ��, where *d* is the flow depth, and the frictional force on the body is (1)(1)(��o), and the balance equation analogous to Equation 8.3 is

$$\tau_o = \gamma d \sin \alpha \tag{5.3.1}$$

These two relationships, Equation 8.3 or Equation 8.4, are called ***the resistance equation for open-channel flow***. A simple result, no? Not many fundamentally important relationships in fluid dynamics are so easy to derive. One useful practical application of the resistance equation is that it gives you a way of finding the bed shear stress once you know the flow depth and the slope—and that's difficult to do, otherwise.

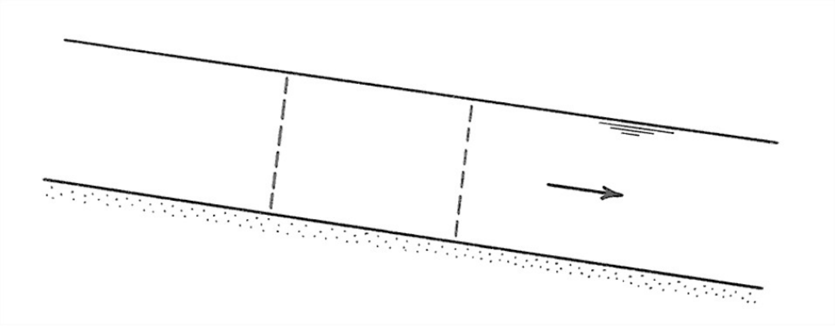


Figure 5-16. A body of water, in a river, that has unit length and unit width and extends from the river bed to the free surface.

Now I want to address two questions that get to the heart of how rivers actually operate. Neither of these questions is easy to deal with. The first is this: **What determines the slope of a river?** One way of answering this question is that the slope is determined by (1) the vertical and horizontal scales of broad crustal uplift that establish the topography that underlies the river system in the first place and (2) the subsequent general reduction in land elevation as the river wears down its drainage area. That's true, but there's more to the story, because rivers can meander within their valleys (as you will see in more detail later in this chapter) and thereby increase the length of their course without changing the elevation along their course. A meandering river has a considerably gentler slope than a straight river in the same river valley.

The other question is: **What determines the particular combination of flow depth and flow velocity associated with a given water discharge?** For any given discharge, there are an infinite number of such combinations; the river might flow fast and shallow, or slow and deep, and still transport the water discharge imposed from upstream. It's natural to pose this question at this point, but I think it's wise to postpone an attempt at an answer until we've dealt with bed configurations later in this chapter. See Section 8.7.

5.4: The Energy of Rivers

The content of this section is a little less grandiose than the title would suggest. I would like to introduce some basic ideas about the energetics of rivers and then supply a calculation to show how much energy is actually expended by a representative large river.

Remember the “home experiment” on dropping a lump of modeling clay to demonstrate the nature of energy, back in Chapter 1? A river is a falling body too, in a very real sense; its fall is just constrained to be at a very low angle, by the gently sloping bed of the river.

When viewed as an energy system, a river is a converter of mechanical (potential) energy to thermal energy. The potential energy of the river water is converted to thermal energy by internal friction within the water. The kinetic energy of the river, however, remains nearly constant, because the flow isn’t changing its speed much downstream. The nature of the internal friction is actually very complicated, because it depends on the details of the turbulence in the river.

When we obtain hydroelectric power from rivers, what we’re doing is locally arranging the river, by building a dam and making a lake, so that the conversion of potential energy to heat energy is suppressed along some stretch of the river, and we convert the potential energy directly to electrical energy by turbines and generators instead.

It might interest you to think about the power expended by a large river. Let’s make a very crude calculation of the rate of energy release by the lower Mississippi River, per square meter of the bed, as it flows downhill. One way of doing this is to think about a column of water above one square meter of the bed of the river, and how fast that column of water loses its potential energy as the river flows downslope. That loss of potential energy shows up as heat, via friction within the water column, owing to shear of the water, and at the bed of the river, as bottom friction. Think of this as the continuing degradation of the mechanical energy of the river into the thermal energy of the water. (Of course, the river doesn’t keep on heating up: it’s losing heat to its surroundings all the time at about the same rate that the heat is being produced by friction.)

You’re likely to get confused about units here. In the mks (meter– kilogram–second) system of units in physics, the unit of force (including weight, which, remember, is a force) is the **newton** (N). The unit of energy is the **joule** (J), which is equal to one newton-meter.

BACKGROUND: WORK AND ENERGY

What comes to your mind when I mention work? Maybe what you do for a living, or things you have to do that are the opposite of fun. In physics, however, work has a very specific meaning: when a body of matter is acted upon and thereby moved by a force, the ***work*** done by the force on the body is equal to *the product of the component of the force in the direction of movement, and the distance the body moves*.

In physics, work is equivalent to energy. You probably have heard of Newton’s second law of motion, mentioned in the background section on energy. It’s not difficult to show, with some math, that Newton’s second law can be recast into an equivalent form that says that *the work done on a body is equal to the change in kinetic energy of the body*. That’s why the joule, the unit of energy in the mks system of units, is equal to one newton-meter.

In its lower reaches, the Mississippi is about ten meters deep, as a very round number, and its mean velocity is as much as a few meters per second. Let’s assume, conservatively, one meter per second. The slope of the river is something like 10-4 (meaning that it drops about a tenth of a meter in one kilometer of downstream travel).

If our column of water is moving at one meter per second and drops a tenth of a meter in one kilometer of travel, it is losing elevation at a speed of 10-4 meters per second. (Think about that for a while, to convince yourself.) The weight of the unit-area column of water is equal to the weight of a cubic meter of water, times its height of ten meters. The mass of a cubic meter of water is (basically by definition!) one thousand kilograms. We have to multiply that by the value of the acceleration of gravity, about ten meters per second per second, to find its weight. Then we have to multiply by the height of the column, ten meters. The result is 105 newtons. That mass, with a weight of 105 newtons, is losing elevation at 10-4 meters per second, so the rate of loss of potential energy is ten newton-meters per second—or 10 joules per second, as per the definition of the joule in the background section above. That’s the rate at which the unit-area column of water in the river loses its mechanical energy. One joule per second is called a ***watt*** (abbreviation: W). The grand final result is *ten watts per square meter of river bottom*. That doesn’t sound like a lot (a ten-watt bulb is even dimmer than the classic dim bulb), but think of how many square meters there are on the bed of the Mississippi River (a few kilometers wide, and hundreds of kilometers long, even in just its lower reaches).

> That long and involved computation above has relevance to hydropower. What a hydroelectric station does is convert the mechanical energy of the river directly into electrical energy. The falling water turns turbines connected to electrical generators, with minimal friction involved, instead of slowly losing its potential energy to heat by friction as it flows downstream.

5.5: The Morphology of Rivers

Figures 5-17 and 5-18 are simplified flow-transverse cross sections through a representative single-channel alluvial river of medium to large size. Figure 5-17 shows the entire stream valley, and Figure 5-18 shows details of the river channel itself. In the following paragraphs I will elaborate upon the various features shown in these figures the diagrams. I'll defer a description of the planform features of the river (that is, what you would see from the air, above the river) until later.

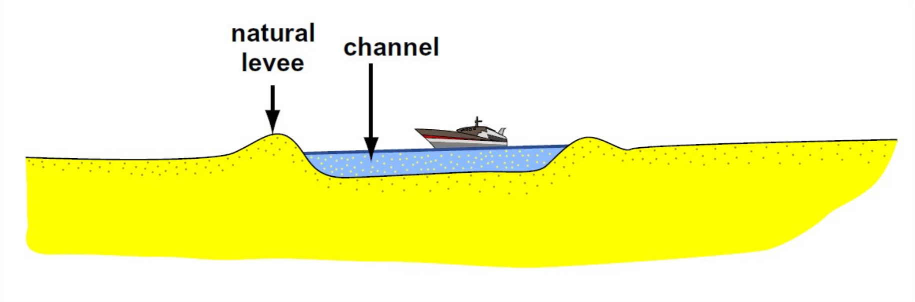


Figure 5-17. Simplified flow-transverse cross section through the valley of single- channel river of medium to large size.

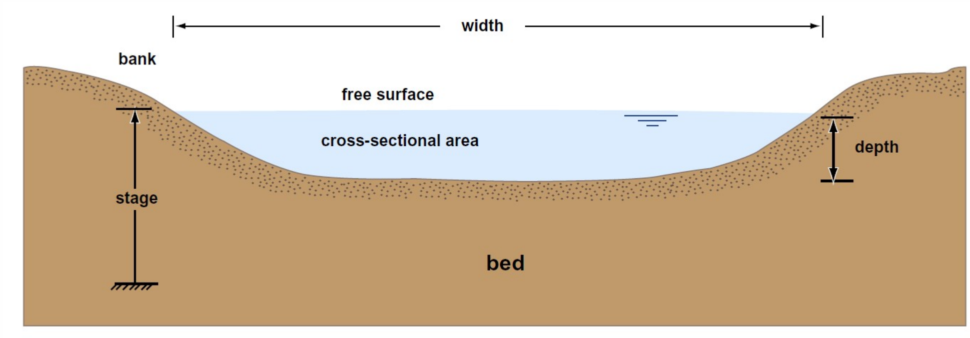


Figure 5-18. Simplified flow-transverse cross section through a representative river channel.

Most medium to large rivers flow on beds of sediment that they have deposited and can transport again; in a later section of this chapter, rivers of this kind will be called ***alluvial rivers*****.** The unconsolidated sediment in the river valley, lying above the bedrock "basement" of the river, is called the ***valley fill***. Its thickness ranges from just a veneer to hundreds or even thousands of meters. In the case of rivers flowing across areas of the crust that have undergone substantial and prolonged subsidence, the valley fill is buried so deeply that it is at least partly lithified, and the material grades over into what would be considered the "ancient sedimentary record" (the term geologists use for sedimentary rocks that are very old by human standards).

The ***floodplain*** of a river is *an area of low relief adjacent to the river channel, which is inundated at times of high river stage*. During floods the floodplain receives a layer of fine sediment that settles out of suspension as the flood waters spread over the floodplain and decrease in velocity. If the river is not undergoing net aggradation (see below for what I mean by that), then the floodplain builds up to a level at which the rate of removal of fine sediment by erosion back into the main channel at times of low water is great enough to strike a balance with the rate of addition of fine sediment from suspension during floods. Most river floodplains are heavily vegetated, and, depending upon climate, are often dotted with shallow lakes and swamps (called ***backswamps***). Floodplains are among the best areas for agriculture, because they continually receive fresh influxes of fertile soil.

Alongside many river channels are low ridges called ***natural levees***, formed by deposition of the finer fraction of suspended sediment from flood waters passing across the river banks when the river is above flood stage. There's preferential deposition because the flood waters decelerate as they leave the main channel flow.

The river channel itself can be characterized most fundamentally by its cross-section shape and cross-section area. The ***width*** is the distance, normal to the local trend of the river, from bank to bank; obviously the width depends strongly on the river stage as well as on the average size of the river. The ***depth*** of the river varies from point to point across the section. A good way of encapsulating the lateral dimensions of the river is to specify the ***hydraulic radius***: the ratio of the cross-sectional area to the wetted perimeter at a given cross section. (To figure out the ***wetted perimeter***, you would use one of those distance- measuring wheels you can rent or buy. Start at the water line on one bank and walk straight across the river to the water line on the opposite bank. Whether you could do that without underwater breathing gear depends on the depth of the river.) For a very wide channel with a nearly rectangular cross section, with an approximately level bottom and steep banks, the hydraulic radius is nearly equal to the flow depth. (You might try figuring that out for yourself; ,it would take some careful thought and a bit of math.)

Another significant aspect of river geometry is the vertical profile. Imagine traveling up the river, keeping track of two things: the elevation of the riverbed above sea level, and the map distance from the mouth of the river. Then plot a graph with the riverbed elevation on the vertical axis and the upstream distance on the horizontal axis. Pass a smooth curve through the points. The result is what is called the ***longitudinal profile*** (or ***long profile***) of the river.

The longitudinal profiles of most rivers are *concave upward*, as shown in Figure 5-19. The reason is not difficult to understand. In the downstream direction, one tributary after another joins the river, each adding discharge. As the river grows larger, the ratio of cross-sectional area to wetted perimeter increases. Because the slope of the river depends, in large part, on the relative magnitude of the downslope driving force of gravity, which is affected by the whole volume of the river, and the upslope resisting force of friction, which is affected by the area of the riverbed, the slope decreases downstream.

The ***base level*** of a river is *the elevation of the water surface of the water body, either the world ocean or a lake along the river course, into which the river flows* (Figure 3-20). The base level changes with time: lake levels fluctuate as a consequence of variations in precipitation in the watershed of the river or because the outlet of the lake is eroded downward, and sea level changes, for various reasons and often very substantially, over a great variety of time scales, ranging from decades to tens of millions of years.

Think about what happens to the river as its base level changes. The concept to keep mind is that *the river has some equilibrium longitudinal profile*, in the sense that if conditions of precipitation, sediment supply, and base level remain constant the longitudinal profile stays the same. If a different set of conditions is imposed upon the river, the river adjusts its longitudinal profile accordingly toward a new equilibrium.

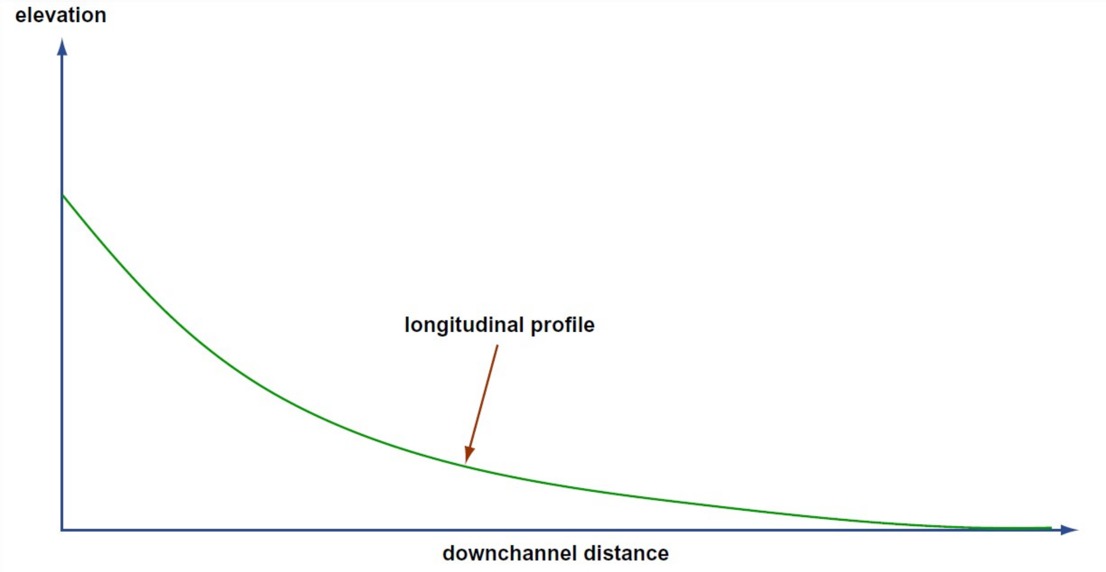


Figure 5-19. The longitudinal profile of a river.

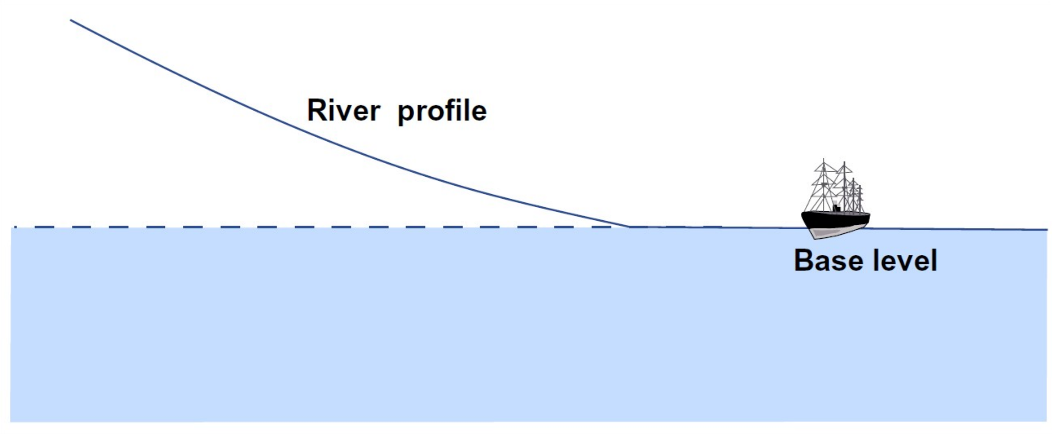


Figure 5-20. The base level of a river.

If base level rises, some of the sediment that's carried along by the river toward the river mouth is deposited along the way to raise the river bed, thereby establishing a new equilibrium longitudinal profile. If base level falls, the river erodes its bed to adjust toward a new, lower equilibrium profile.

There's more to be said, however, about what happens as the river erodes its bed as a consequence of a fall in base level. The erosion does not happen uniformly everywhere. all at the same time, but by upstream propagation of a point where the channel slope changes, from steeper downstream of the point to less steep upstream of the point. The point of change in slope is called a ***knickpoint*** (Figure 5-21). The position of a knickpoint is marked by a waterfall or rapids. Knickpoints migrate slowly upstream, thereby extending the new, lower longitudinal profile as the river eats its way upstream. If a floodplain has developed in the river valley, the old floodplain downstream of the knickpoint survives, for a long time, as a pair of terraces above the new, lower river channel Because the difference between old and new equilibrium profiles decreases upstream, other things being equal (the elevations of the highlands in the headwaters of the river are very conservative), the height of this knickpoint decreases as it migrates upstream. Often, if base level drops abruptly a number of times during some long period of time, more than one knickpoint is present along the river course, each slowly making its way upstream.

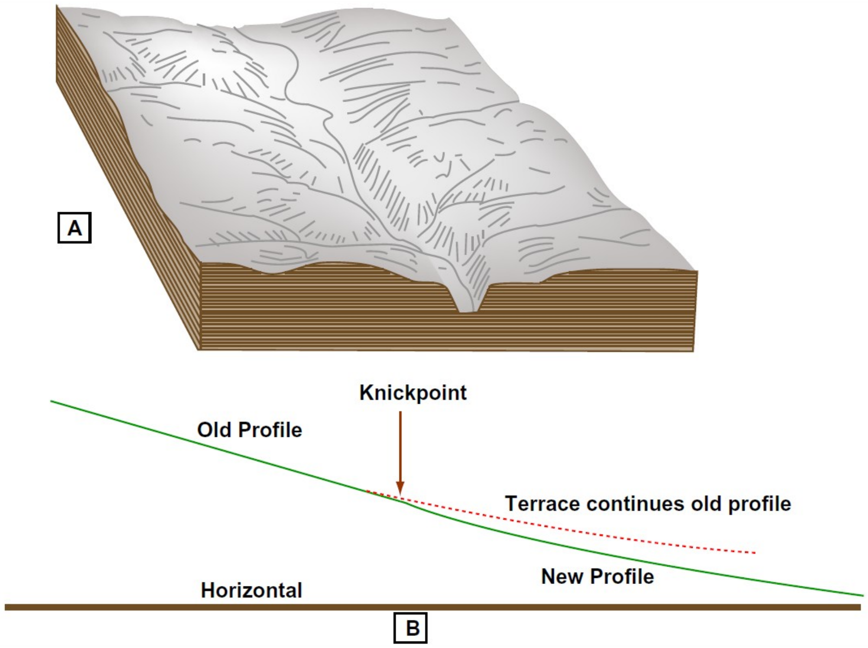


Figure 5-21. A knickpoint along a river course. A) Perspective view of a stream valley, showing a waterfall at a knickpoint. (From Thornbury, 1969.) B) Profile view of a stream in which a knickpoint is propagating upstream. (Modified from Holmes, 1965.)

As you will see in the later material on the plan-view features of rivers, rivers do not stay in one position but instead tend to shift laterally across their floodplains, by erosion at one bank and deposition at the other bank. (That's the basic reason why there are floodplains in the first place.) As the river lowers its bed in response to a fall in base level, and at the same time shifts its course laterally, it develops a new floodplain that’s entrenched below the level of the old floodplain. The result is a pair of flat-topped ***river terraces***, one on either side of the river. The slopes at the edges of the modern floodplain retreat without much change in their shape, because they are continually being undercut along their bases rather than wearing away over their entire surface. Sometimes there is more than one set of such terraces.

5.6: Classifying Rivers

Rivers are varied in so many ways that you should expect complexity in classification. Rivers can be classified in several ways:

- by the nature of their substrate
- by the percentage of time they flow
- by their relationship to the groundwater table
- by the kind of sediment load they carry
- by the dominant particle size of the bed sediment • by their morphology

In the following I'll make some comments on classification of rivers in each of these ways. (But I'll defer discussion of classification of rivers by sediment load until later, after I've talked about the sediment load.)

Nature of Substrate:

Some rivers, especially small rivers in mountainous areas, flow directly on bedrock. Such rivers are called ***bedrock rivers***, or ***non-alluvial rivers*** (Figure 5-22A). Other rivers, especially large rivers, flow on *a bed a sediment they have deposited and can continue to transport*. Such rivers are called ***alluvial rivers*** (Figure 5-22B). Of course, some rivers lie in between, in that they have bedrock beds in some reaches and alluvial beds in other reaches (Figure 5-23). Alluvial rivers have held most of the problems, fascination, and importance for fluviologists, but if you're a white-water canoeing enthusiast I suppose you're more interested in bedrock rivers.

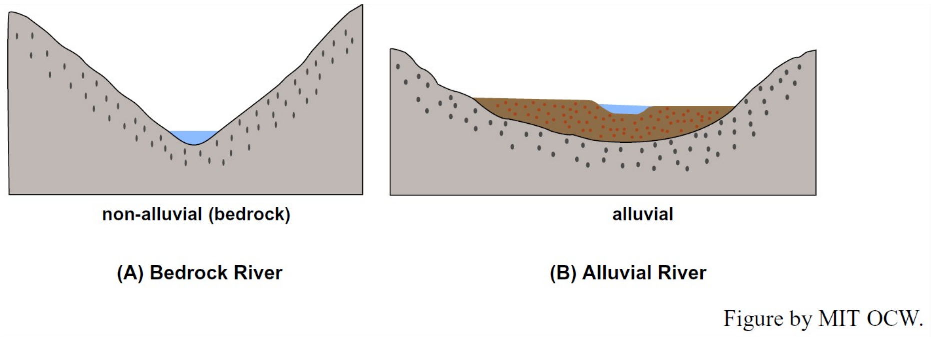


Figure by MIT OCW.

Figure 5-22. A) Bedrock rivers. B) Alluvial rivers.

Relationship to Groundwater Table:

Think about *the relative position of the water surface in the river and the local groundwater table*. If the water surface in the river lies *above* the local groundwater table in the river banks, then the river *loses water to its banks*. Such a river is said to be an ***effluent river*** (Figure 5-24A). On the other hand, if the water surface in the river lies *below* the local groundwater table in the banks, then the river *gains water from its banks*. Such a river is said to be ***influent river*** (Figure 5-24B). (These two terms are difficult to keep straight. It helps to think in terms of alternative terminology: an effluent river is also called a ***gaining river***, because it is gaining water from the adjacent substrate, and an influent river is called a ***losing river***.) Keep in mind that it's also possible for the groundwater table to lie entirely below the river bed (Figure 5-25). The river is still called an effluent river in that case.

Percentage of Time the River Flows:

Some rivers show a flow of water all the time, even long after the last rainstorm in the watershed. Such a river is called a ***perennial stream***. Other rivers flow for only a short time after a rainstorm, and for the rest of time, usually most of the time, their beds are dry. Such a river is called an ***ephemeral stream***. Some rivers lie between these two extremes: during the wetter part of the year they flow as a perennial stream, whereas during the drier part of the year they flow as an ephemeral stream. Such a river is called an ***intermittent stream***. Figure 5-26 shows cartoon hydrographs of a perennial stream, an ephemeral stream, and an intermittent stream.

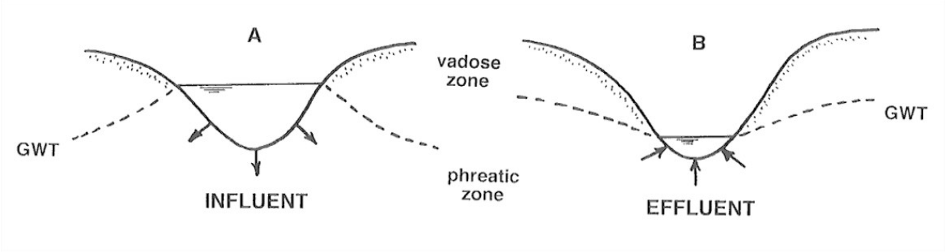


Figure 5-24. A) An influent river. B) An effluent river.

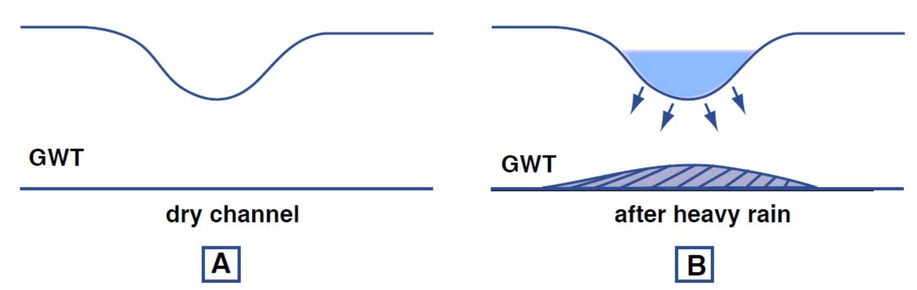


Figure 5-25. The groundwater table can lie entirely below the river bed. A) Before a heavy rain. B) After a heavy rain.

In an ephemeral stream, the water table always lies below the bed of the stream; the stream never receives any water from its bed or banks. In a perennial stream, the situation is more complicated. Think about the relationship between the river level and the groundwater table in some time period that starts in a dry spell, extends through a major rainfall event in the watershed, and ends during another dry spell. At the end of the first dry spell the river level lies below the groundwater table in the river banks (Figure 5-27A). After a heavy rainfall the river stage rises rapidly to lie well above the level of the groundwater table in the banks (Figure 5-27B). Groundwater is stored in the river banks, in the sense that the groundwater table is locally and temporarily higher there than in the surroundings. At the end of the rainy period both the river stage and the groundwater level are of about the same height and are about at their highest (Figure 5-27C). Then (Figure 5-27D) both the river stage and the groundwater table fall back to the dry-spell situation shown in Figure 5-27A. This sequence of events is called the ***runoff cycle***.

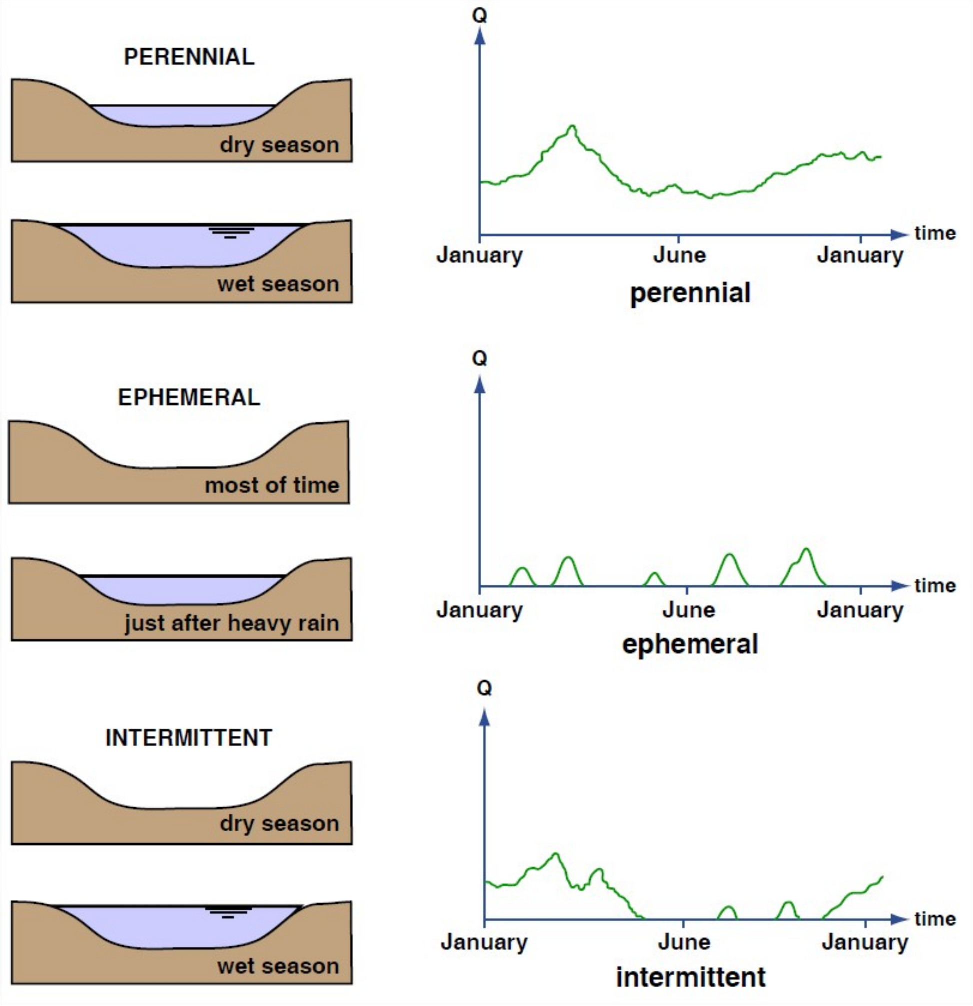


Figure 5-26. Perennial, intermittent, and ephemeral streams, and associated representative hydrographs.

Morphology:

The morphology of rivers, especially in plan view, varies enormously. The most common way to classify rivers is on the basis of their plan-view morphology. The morphology of rivers is bound up in a complex way with the nature of the sediment load, so a full appreciation of this section must await Section 8, on the sediment load of rivers.

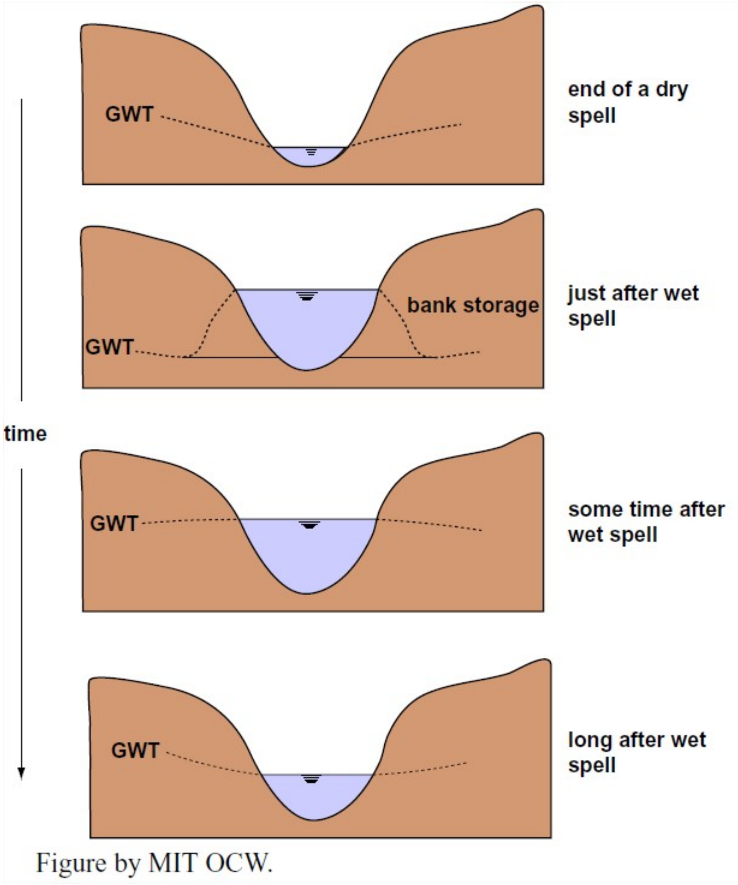


Figure 5-27. The runoff cycle.

Two characteristics are used in the classification of rivers by morphology: sinuosity and “multichanneledness”. ***Sinuosity*** can be defined with respect to two arbitrary points along the river as *the ratio of the along-channel distance between the two points and the straight-line distance between the points* (Figure 5-28). The minimum sinuosity, for a straight river, is 1; the more sinuous the river, the greater its sinuosity. Very sinuous rivers can have values of sinuosity approaching 4. ***Multichanneledness***, an awkward but useful word, reflects *the number of individual flow channels shown by a river in a cross-stream traverse across the entire river system*. Many rivers have only one channel, except perhaps where an occasional island divides the channel into two. Other rivers show a large number of channels, all of about the same size and nature, separated by numerous bars and islands. The individual channels of such a river are called ***anabranches***.

Sinuosity and multichanneledness are to a large extent independent of one another, so it’s natural to resort to a two-independent-variable pigeonhole classification with sinuosity along one axis and multichanneledness along the other axis (Figure 5-29). ***Straight*** rivers—those with sinuosity not much greater than 1—are surprisingly uncommon in nature. In fact, it’s hard to keep rivers straight: humans straighten them out for their own purposes, and the rivers try to become sinuous again, by erosion and deposition on the banks. Both ***braided rivers*** *(low-sinuosity, multichanneled)* and ***meandering rivers*** *(high sinuosity, single-channeled)* are very common; more on them later. ***Anastomosing rivers*** *(high-sinuosity, multichanneled)* are much less common.

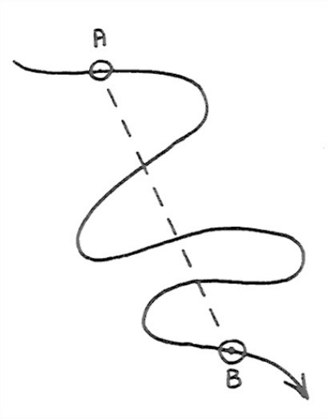
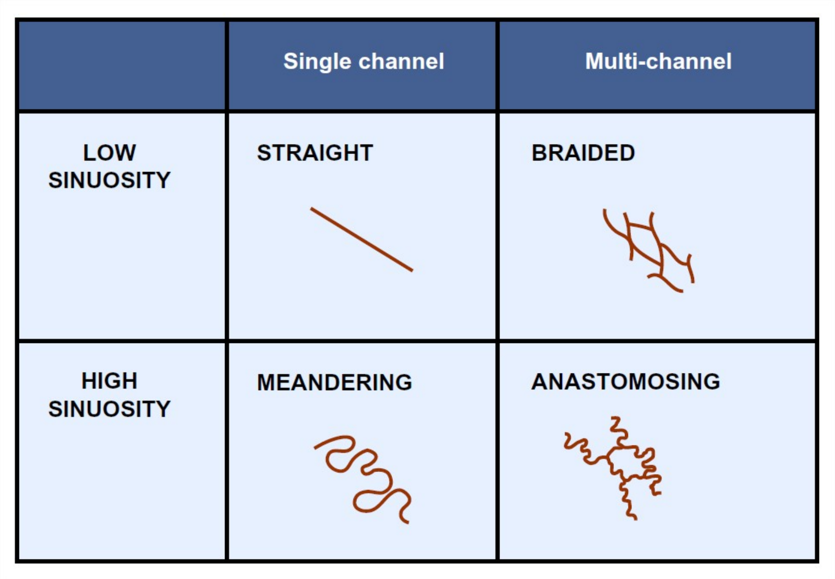


Figure 5-28. The sinuosity of a river.



	Single channel	Multi-channel
LOW SINUOSITY	STRAIGHT	BRAIDED
HIGH SINUOSITY	MEANDERING	ANASTOMOSING

Figure 5-29. One way of classifying rivers.

5.7: Variables Involved In Rivers

Think about the variables that describe the characteristics and behavior of rivers. These fall into the broad categories of flow, sediment, geometry, and other. Here's a fairly inclusive list of such variables:

flow:

- cross
- section
- stage
- discharge
- velocity
- turbulence

sediment:

- bed-material properties
- load
- sediment discharge
- bed configuration
- slope or profile
- base level

geometry:

- width
- depth
- cross-section shape
- plan pattern

other:

- chemistry
- biota

Of these, some can be considered to be ***independent variables***, in the sense that they are imposed on the river and the river has to live with them, and other can be considered to be ***dependent variables***, in the sense that the river adjusts their values in response to the independent or imposed variables.

Independent variables:

- temperature (almost entirely)
- biota (mostly)
- discharge (entirely)
- sediment discharge (approximately, in the long term)
- base level (entirely)
- chemistry (almost entirely)
- sediment characteristics (partly)
- slope (in the short term but not in the long term)

5.8: Fluvial Sediment Transport

Introduction

This section is a brief account of the nature of sediment transport in rivers.

Regolith is produced by weathering of bedrock on the continents and is then transported away from the site of production as sediment (transported regolith), in both particulate and dissolved form. The oceans can be viewed as the repository for this sediment, although much is stored in sedimentary basins on the continents and recycled into new continental bedrock on time scales that may be a large fraction of geologic time.

Rivers are by far the most important carriers of sediment on the continents, although glaciers have been even more important at certain times and places. So although the essence of rivers is that they are flows of water, one of their most outstanding characteristics is that they are also transporters of sediment. I think it's fair to say that most of the interesting things about rivers are connected in some way with the sediment-transporting nature of rivers, either directly or indirectly. Fluvial sediment transport is an important area in several different disciplines: fluvial engineering, geomorphology, and sedimentology.

A big problem in gaining an appreciation for sediment transport in rivers is that it's generally difficult to observe sediment transport in real rivers, unless they are very shallow and not carrying much sediment. A good way to get around this difficulty, in part, is to build an artificial river in your backyard and watch from close up how the river transports its sediment.

It's not difficult to build a simplified small river that reproduces many of the essential aspects of fluvial sediment transport. Nail together a large wooden channel, open at the top and with a rectangular cross section (Figure 5-30). Caulk the seams and joints so they won't leak intolerably. (All such channels leak a little sometimes, even in the fanciest of laboratories.) The channel might best be about a meter wide and a few tens of meters long, but I realize that's a big order for backyard construction; you could get away all right with a channel no more than ten to twenty centimeters wide and five to ten meters long.

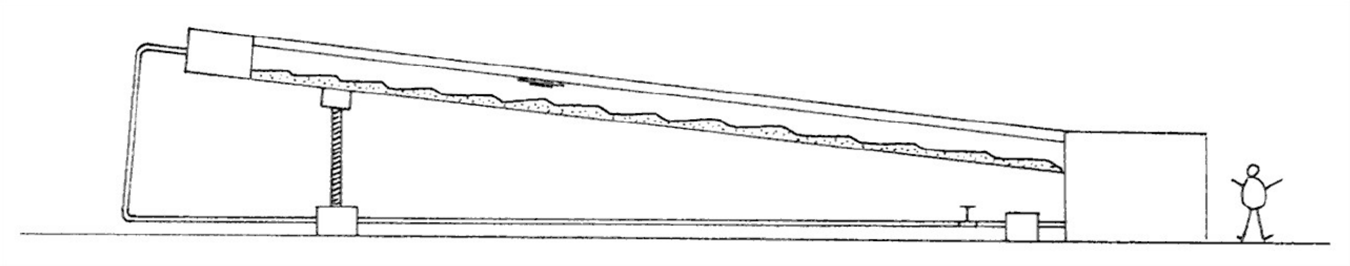


Figure 5-30. An experimental channel to model flow and sediment transport in rivers.

At its downstream end the channel should pass into a big open tank. Install a pump and some piping to take the water from the downstream tank and recirculate it to the upstream end of the channel. (This is by far the most expensive and, for most people, the most challenging part.) A valve in the return pipe lets you adjust the water discharge. You might mount the whole channel on a jack near the upstream end, so that you can change the slope of the channel easily, but that isn't really necessary. It would also be nice to make at least one sidewall of the channel out of glass or transparent plastic, for good viewing of the sediment transport. At the very least, install a few small subsurface portholes in the channel walls. Place a thick layer of sand on the floor of the channel. The mean size of this sand might be as fine as 0.1 mm or as coarse as 1 mm, but to see the widest range of phenomena and features use a sand with a mean size a little less than 0.5 mm.

Threshold of Movement

Now you are ready to make some experiments (flumologists— scientists and engineers who work with channels of the kind you just built in your back yard, which are called flumes— call them *runs*) in your channel. For each run, fill the tank and the channel with water to establish the depth of flow, turn on the pump, and adjust the valve for a certain water discharge and therefore (given the flow depth) a certain mean flow velocity. Arrange each run to have a flow depth as great as the flume will allow, ideally at least a large fraction of a meter.

The first thing you should do is study the beginning of sediment movement. (First, it would be good to smooth off the sand bed to be planar. That's not hard to do if you use an underwater scraper blade attached to a device that slides along the top edges of the channel walls.) Clearly, if the flow is too slow it won't move any sediment. As you gradually increase the flow velocity, however, at some point some sediment particles begin to move. That point is called the ***threshold of movement***. It may surprise you to find how difficult it is to define or locate the threshold, because there's weak movement or slight movement over a wide range of flow strengths. The problem of the threshold of movement reflects two questions that have long been recognized as important in fluvial sediment transport:

- How strong a flow is needed to initiate the transport of a given sediment?
- How coarse a sediment can be moved by a given flow?

The second question has to do with what's called the ***competence*** of a river. Just to give you some feel for the flow strength needed to move the sand in your channel, you would have to produce a current velocity of about 0.2 m/s to reach threshold conditions—depending on the sand size, of course, and on the water depth also.

Now make some runs with flow velocities greater than the threshold value, to study the modes of particle movement. *The aggregate of sediment particles which are transported by a flow at a given time* is called the ***sediment load***, or just the ***load***. The load can further be subdivided in three different ways: on the basis of

- *its physical nature;*
- *its presence or absence in the bed;* and
- *how it travels*.

On the basis of its physical nature, the sediment load of rivers is conventionally subdivided into ***particulate load*** and ***dissolved load***. Does it surprise you that in general *the dissolved load is of the same order of magnitude as the particulate load*? Keep in mind that the concentration of the dissolved load doesn't depend greatly on the discharge, so there's a lot of dissolved load even at times of low discharge. In fact, there tends to be an inverse correlation between discharge and dissolved load, for the understandable reason that at times of high discharge most of the water in the river has run off the surface without spending a long time in contact with regolith and bedrock, where, by the variety of weathering processes you learned about in Chapter 2, the solutes of the dissolved load come from.

The load can also be divided into ***bed-material load***, which is *that part of the load whose sizes are represented in the bed in non-negligible percentages*, and ***wash load***, which is *that part of the load whose sizes are not present in the bed in appreciable percentages*. The wash load, which is always the finest fraction of the load (mainly clay or finest silt size), is carried through a long segment of a river without any exchange of sediment between the bed and the flow. Of course, you don't have any wash load in your backyard channel, unless you choose to dump in some water containing a suspension of fine clay.

In real rivers, much of the very finest fraction of the load has such a small settling velocity that it travels for long distances before settling to the bed, and even when it reaches the bed it tends to be resuspended immediately. So the finest fractions of the sediment are represented almost not at all in the bed material in the river channel. Only in overbank areas during floods, where the water velocities are small, is the wash load deposited. Even there, storage tends to be temporary, because shifting of the river channel tends eventually to re-erode the fine sediment that was previously deposited on the floodplain.

Finally, the load can be divided into ***bed load***, which *travels in direct contact with the bed or so close to the bed as not to be substantially affected by the fluid turbulence*, and ***suspended load***, which is *maintained in temporary suspension above the bed by the action of upward-moving turbulent eddies* (Figure 3-31). I hope it's clear from these definitions that bed load is always bed-material load, and suspended load is likely to be partly bed-material load and partly wash load, although in particular cases it could be all wash load, or all bed-material load.

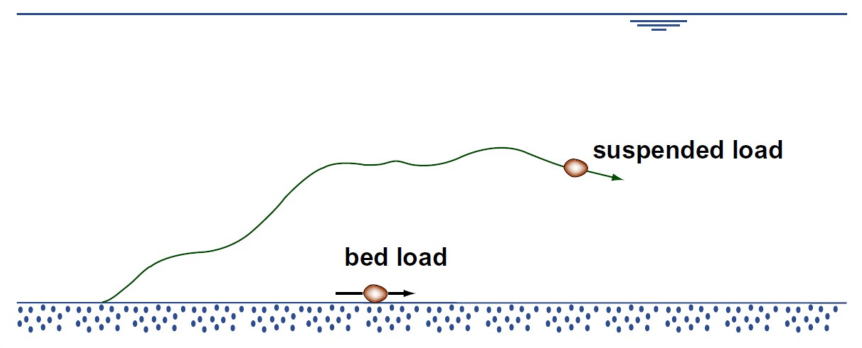


Figure 5-31. Bed load and suspended load.

This may sound confusing, but it makes sense. Figure 5-32 may or may not help.

The movement of bed load is sometimes called ***traction***. Bed-load movement can be by *rolling, sliding, or hopping* (Figure 5-33). It's not easy to observe bed-load movement in detail, but, if you're a good photographer and you made a high-speed close-up motion picture of bed load, you see that the particles characteristically take occasional excursions downstream, by rolling or hopping or bulldozing irregularly with brief stops along the way, and then come to rest for some time before being moved again. Once a grain is dislodged from a place of rest, it's susceptible to continued movement by the flow until it finally finds a rather sheltered position among the bed particles, at favorable sites called "pockets", and then it's not dislodged again until it's affected by an especially near-bottom eddy or until one or more of the bed particles sheltering it are themselves put into motion.

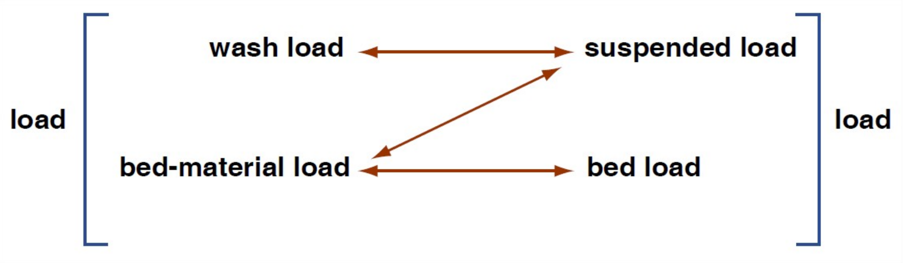


Figure 5-32. Relationships among bed-material load, wash load, bed load, and suspended load.

If you're a keen observer you would notice that *the set of particles that form the load keeps changing from time to time*, because particles are continually coming to rest and being set into motion again. Also, there's a problem in distinguishing between bed load and suspended load: **how far can a grain move up into the flow and still be considered bed load?** The standard criterion is *whether or not fluid turbulence has a substantial effect on the time and distance involved in the excursion*. Although the distinction between bed load and suspended load is a convenient one, there's no sharp break between bed load and suspended load. Also, a given particle can be part of the bed load at one moment and part of the suspended load at another moment, depending upon the time history of fluid forces and motions to which it is subjected. (And of course at still other times the same particle might not be moving at all.) Therefore, at any given time there's an appreciable overlap in the size distributions of the bed load and the suspended load, although clearly the suspended load tends always to be finer on the average than the bed load.

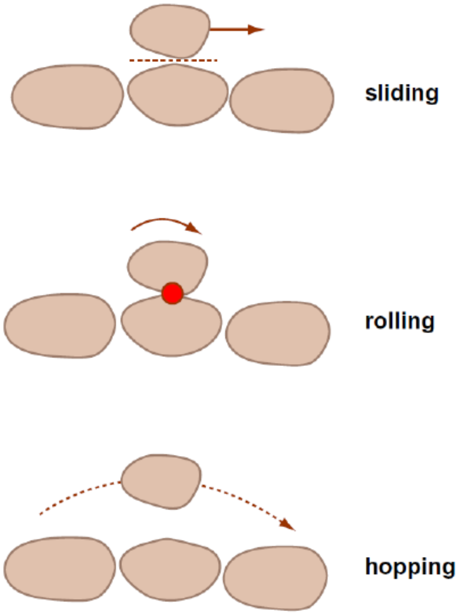


Figure 5-33. Modes of bed-load movement.

Particles moving as bed load are susceptible to being carried up into ***suspension*** when *the maximum vertical turbulent velocity fluctuations are greater in magnitude than the settling velocities of the particles*. If the conditions of the flow and the settling velocities of the particles fulfill that condition, then some of the moving bed-load particles occasionally find themselves caught in a strong upward-moving eddy, and the particle is carried for some distance above the bed. The particle is affected by a series of eddies as it moves downstream; depending on the motions of the individual eddies, the particle may rise only a short distance from the bed and travel only a short distance downstream before it settles back to the bed, or it may rise high above the bed, even almost to the water surface, and travel far downstream. Obviously, *the smaller the settling velocity and the stronger the turbulence, the greater the average height above the bed and the greater the distance of downstream travel by the particle*.

The sediment particles are not really suspended above the bed, in the way that a painting is suspended on a nail in the wall: they are always settling back toward the bed and will eventually return to the bed. Only particles of colloidal size, much finer than a micrometer, can be truly suspended. Such particles have such small mass that Brownian motions caused by the random collisions of molecules against the particle keep the particle in permanent suspension.

When the flow is relatively weak and/or the sediment is relatively coarse, the concentration of suspended sediment drops off rapidly upward, and the upper part of the flow may not have any suspended sediment at all (Figure 5-34A). When the flow is relatively strong and/or the sediment is relatively fine, however, suspended sediment is present throughout the entire depth of flow, and the concentration of suspended sediment drops off only slightly upward (Figure 3-34B).

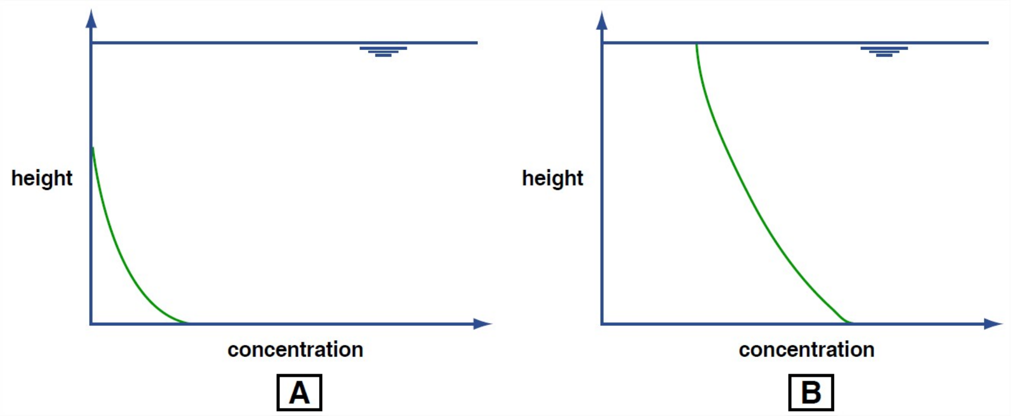


Figure 5-34. Vertical profiles of suspended-sediment concentration in a river.

Classification of Rivers by Sediment Load

It should seem natural to you that some rivers, especially those whose sediment sources consist mostly of coarse sand and gravel, carry mostly bed load. Such rivers are called, understandably, ***bed-load rivers***. On the other hand, the sediment sources of some rivers are mainly clay, silt, and fine sand. The sediment load of such rivers is mainly suspended load. Such rivers are called ***suspended- load rivers***. Finally, in a great many rivers both bed load and suspended load are important; such rivers are called ***mixed-load rivers***. Keep in mind that this classification is very loose, both because there's a continuous gradation in nature of sediment load from river to river and also because a given river carries its load in different ways, depending on its discharge at the given time.

You can also classify rivers by the dominant particle size of the bed material, into ***gravel-bed rivers***, ***sand-bed rivers***, and (much less common than the first two) ***mud-bed rivers***. (Of course, there are gradations among these three kinds.) In most sand-bed rivers, most of the bed-material sediment discharge on average is in the form of suspended load. In gravel-bed rivers, on the other hand, most of the sediment discharge is accounted for by bed-load transport.

Sediment Transport Rate

The rate at which sediment is moved past a cross section of the flow is called either the ***sediment transport rate*** or the ***sediment discharge***. It's related to the sediment load, but it's different, just because different fractions of the sediment load are transported at different rates. It can be measured in mass per unit time, or in weight per unit time, or in volume per unit time. For many decades hydraulic engineers have sought after formulas to predict the sediment discharge that's associated with some combination of flow conditions and sediment characteristics in some reach of a river. There are many such formulas, but none of them work gratifyingly well.

ADVANCED TOPIC: SEDIMENT DISCHARGE FORMULAS

To derive a sediment discharge formula, you try to think about the physics of sediment transport in a way that allows you to develop the form of some rational equation for transport rates, which contains within it one or more "adjustable parameters" whose values are assigned by analysis of selected data sets already at hand. It's a sad fact that the physics of sediment transport is so complicated that you can't develop an equation from the first principles of physics that contains no such adjustable parameters. It's almost no exaggeration to say that there are dozens of sediment discharge formulas in the hydraulic-engineering literature. The physical basis behind these formulas ranges widely, and none of them does a really good job.

Here's just the barest start at deriving a sediment discharge formula. Your common sense tells you that *the stronger the flow the greater the sediment transport rate*. And an important first-order fact of observation is that *the sediment transport rate is a very steeply increasing function of the flow strength*. Think about the simplest way to embody these important facts in a formula for the sediment transport rate per unit width of flow, usually written q_s. Perhaps the simplest approach to quantifying q_s is to write an expression like

$$q_S = A\tau_o^n \tag{5.8.1}$$

where A is a coefficient and n is an exponent much larger than one. Better yet, τ_o might be replaced with τ_o-τ_c, where τ_c is the threshold boundary shear stress for sediment movement. The trouble with Equation 5.8.1 is that it has no strong basis in the actual physics of the transport. The true situation must be much more complicated than Equation 5.8.1. But with A and n adjusted by use of observational data, Equation 5.8.1 can serve for very crude estimates of q_s.

Measuring the sediment transport rate is a notoriously difficult task. The suspended-load transport rate is usually measured by trapping small samples of the passing flow, with its suspended sediment, in a series of little catch bottles arranged vertically through the entire depth of the river, together with measurements of the local flow velocity along that vertical. Bed load is more difficult to measure. Usually bed load is measured with devices called *bed-load traps* or *bed-load samplers*. Think in terms of those pole-mounted dustpans that are used to sweep litter from paved surfaces—but with netting in the back of the pan to let the water flow through. They come in all sizes and geometries, but their purpose is to catch all the bed load that comes upon the trap from upstream— without catching too much or too little. Unfortunately there are many practical problems connected with bed-load traps.

Bed Configuration

Now you need to make a series of runs in your channel to study the bed configuration. By the term ***bed configuration*** I mean *the overall bed geometry that exists on the bed of the channel at a given time in response to the flow*. If you haven't had any experience with bed configurations, you may be thinking, "What's the big deal? Won't the sediment just move on a planar transport surface?" But it turns out that you would just as often be wrong about that than right. If the bed configuration is not planar, it's composed of *individual topographic elements* called ***bed forms***.

Make a series of runs with slightly increasing mean flow velocity above threshold conditions. In each run, let the flow interact with the bed long enough for the bed configuration to be statistically steady or unchanging. After that time the details of the bed configuration change constantly but the average characteristics remain the same. The time required for the flow and the bed to come into a new state of equilibrium might be as little as a few minutes to as long as several days, depending on the sediment transport rate, the size of the bed forms that develop, and the extent of modification of bed forms that were left over from the preceding run.

At low flow velocities, the bed becomes covered with ***ripples***: little ridges of sand with their crests and troughs oriented mostly transverse to flow but rather irregular in detail, with gentle upstream surfaces and generally angle-of- repose downstream surfaces (Figure 5-35A). Their spacings are 10–20 cm, and their heights are a few centimeters. They move slowly downstream, orders of magnitude slower than the flow velocity, by erosion of sand from their upstream sides and deposition on their downstream sides. Except at the highest speeds, you have to watch them closely to see them move. Ripples on real river beds look almost exactly like those you can produce in your backyard channel.

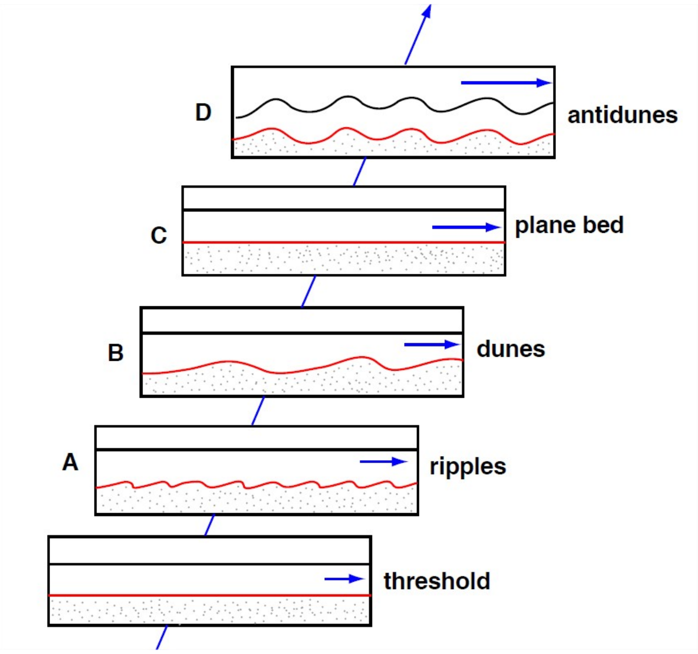


Figure 5-35. Sequence of bed configurations in open-channel flow over a bed of sand a few tenths of a millimeter in diameter.

At a flow velocity that's a moderate fraction of a meter per second, ripples are replaced by larger bed forms called ***dunes*** (Figure 5-35B). Dunes are fairly similar to ripples in geometry and movement, but they are at least an order of magnitude larger. The transition from ripples to dunes is complete over a narrow range of only a few centimeters per second in flow velocity. Dunes in large rivers can attain truly gigantic proportions: heights of over ten meters, and spacings of many hundreds of meters. Superimposed on such large dunes are one or more orders of smaller dunes, all the way down to little ripples.

As you increase the flow velocity further the dunes become lower and more rounded, over a fairly wide interval of flow velocity, until finally they disappear entirely, giving way to a planar bed surface over which abundant suspended load as well as bed load is transported (Figure 5-35C). Judging from the appearance of the bed after the flow is abruptly brought to a stop, the transport surface is strikingly planar: relief is no greater than a few grain diameters. But it's difficult to observe the mode of grain transport over the planar bed because the bed is obscured by abundant bed load and suspended load.

As you increase the flow velocity still further, subdued standing waves appear on the water surface, and the resulting pattern of higher and lower near-bed flow velocity causes the bed to be molded correspondingly into a train of waves that are in phase with the water-surface waves. Under certain conditions these coupled bed waves and surface waves increase in height and become unstable: they move slowly upstream and at the same time grow in height, until they become so steep that they break abruptly, throwing much sediment into suspension (Figure 5-35D). The bed and water surface then revert to a planar or nearly planar condition, whereupon the waves build again and the cycle is repeated. Because of their upstream movement these forms are called antidunes**.** Antidunes are important in shallow fast-flowing rivers but not in deep rivers. An excellent place to watch small antidunes in action is at the beach where a small stream, flowing fast and shallow, passes across the sandy beach to reach the sea.

More on Fluvial Hydraulics

Back in Section 4 I put off discussion of the following important question: **What determines the particular combination of flow depth and flow velocity associated with a given water discharge?** I can pick that up again, now that I've said some things about bed configurations.

The force the flow exerts on the bed of a river are of *two kinds*: skin friction and form drag:

- ***Skin friction*** (a term borrowed from aerodynamics of airplane wings) is the local friction force exerted on locally smooth areas of the river bed.
- ***Form drag*** is the force the flow exerts on a non-planar part of the river bed, like a ripple or a dune, by exerting higher fluid pressure on the upstream side of the form than on the downstream side.

Just think back to the swimming-pool experiment with the tray or pizza pan, described in Chapter 1. In almost all river flows, *the form drag is much larger than the skin friction*.

It’s easy to understand that the particular combination of depth and velocity for a given discharge in a river is mediated by the nature of the flow resistance: the greater the flow resistance, other things being equal, the deeper and slower the flow. And, as you can easily imagine given the ruggedness of ripples and dunes on a river bed, *the flow resistance is dominated by form drag whenever there are bed forms on the river bed*. So the most important factor in determining the combination of depth and velocity is the nature of the bed configuration.

This bed-configuration effect on the flow resistance has an important and beneficial consequence for the behavior of rivers during floods. As the discharge and therefore the flow velocity increases during a flood, the bed configuration is ripples and then rugged dunes, so the flow depth is relatively deep and the flow velocity is relatively low. As the velocity increases, however, eventually the dunes are washed out to a plane bed, and as the discharge continues to increase, the flow depth is then relatively low and the velocity is relatively high. (The word *relatively* in the last two sentences implies comparison with the same discharge but a different bed configuration.) The highest stage during the flood, when the bed is planar, is typically even greater than what it was at the time of transition from dunes to plane bed, but the important thing is that *the stage is much less than it would have been if this bed-configuration effect did not exist*—as in a bedrock river, for example.

The Sediment Yield

I need to introduce just one more thing about sediment in rivers. You should make a distinction between the sediment discharge and what’s called the *sediment yield* of a river. The ***sediment yield*** is *the sediment discharge divided by the total drainage area of the river upstream of the cross section at which the sediment discharge is measured or estimated*. The sediment yield measures the rate, per unit area, at which sediment is removed from the watershed. It’s important in studies of the long-term evolution of landscapes drained by rivers.

5.9: Morphology and Dynamics of Meandering Streams

The two most characteristic plan patterns assumed by rivers are *meandering* and *braided*. First of all, keep in mind that the tendencies toward meandering and braiding are complementary rather than mutually exclusive, in the sense that many rivers show elements of both at the same time and within the same reach.

First, some descriptive stuff on the geometry of meandering rivers. Figure 5-36 shows most of the elements of a meandering river system. The *meander bends* or ***meander loops*** are never perfectly regular, but they are often described as if they were (Figure 5-37), in terms of ***wavelength*** and ***amplitude***. One can also think about a ***radius of curvature*** in the bend, but that's likely to vary from point to point in the bend. The narrow part of the meander bend, between adjacent reaches of the channel above and below the bend, is called the ***meander neck***.

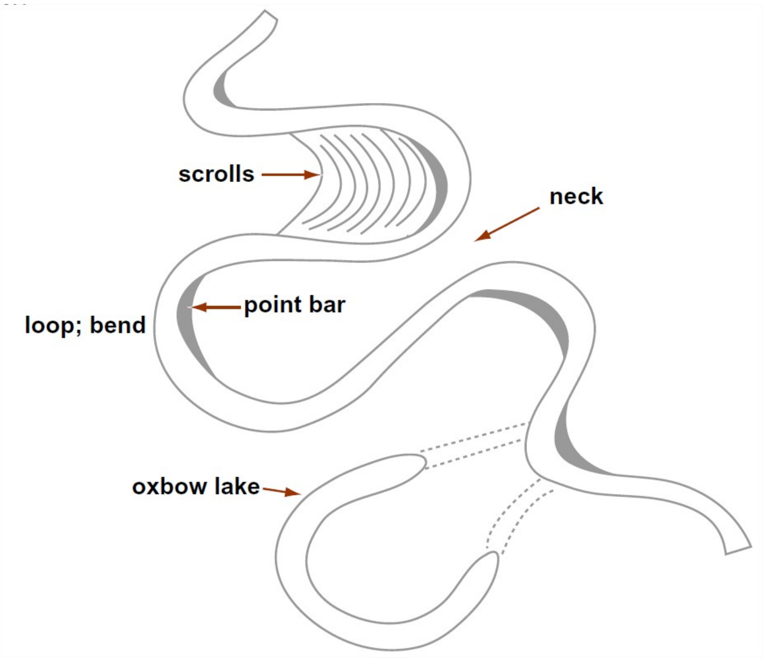


Figure 5-36. Elements of the plan-view features of a meandering river

On the inner, or convex, side of the meander bend is a fairly smooth and largely non-vegetated body of sand, sloping gently downward toward the center of the channel. This sand body, called a ***point bar***, lies everywhere below bank-full stage and is inundated by floods. Sediment is gradually accreted to the point-bar surface, causing the point bar to shift laterally. Point-bar deposits are for

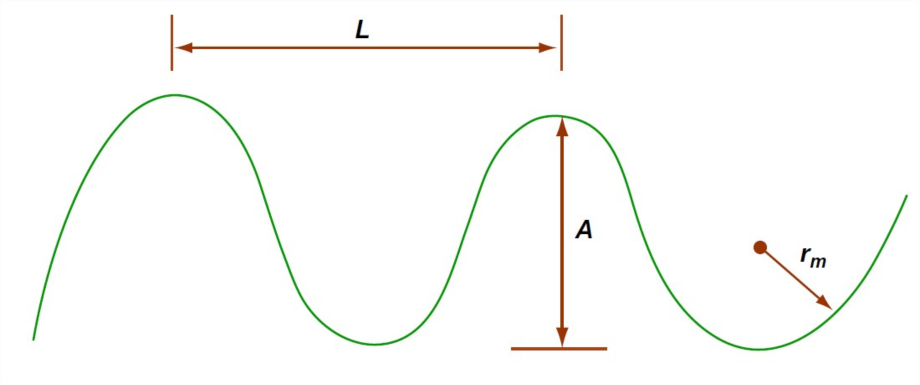


Figure 5-37. Idealized geometry of meander bends.

On the outer, or concave, side of the meander bend is a steep bank or cliff, often called a ***cut bank***, of consolidated or semiconsolidated sediment that is eroded during strong flows. The bank is eroded by various processes, the most important of which is undercutting low on the bank and collapse of large masses into the channel.

The flow-transverse profile of a meandering river is highly asymmetrical. The flow is deepest near the foot of the eroding outer bank, and it shoals gradually up the point-bar surface. The velocity of flow is highest in the region near the free surface and near the outer bank. Figure 5-38 is a cartoon of the flow-transverse profile, showing the characteristic pattern of isovels (lines of equal velocity). There's also a characteristic helical secondary circulation within the bend, such that the flow near the bed has a component inward toward the point bar, and the flow near the surface has a component outward toward the outer bank. This spiraling secondary circulation is readily explained by the differences in centrifugal force between the high-speed regions of the flow above and the low- speed regions of flow below. It's this secondary circulation that causes the accretion on the point-bar surface.

Meander bends generally have a tendency to change with time toward *larger amplitudes* and *narrower necks*. Eventually, the river jumps across the narrow neck during a flood. Such an event is called ***meander cutoff***, or ***neck cutoff***. The river thereby straightens itself spectacularly in quantum jumps to offset the gradual amplification of the meander bends. Sometimes there's a less catastrophic variety of meander cutoff, whereby the river occupies an old slough that marks an earlier position of the bend; then the amplitude decreases, but not nearly to zero. Cutoffs of that kind are called ***chute cutoffs*** (because the river comes to occupy a chute between the two sides of the bend).

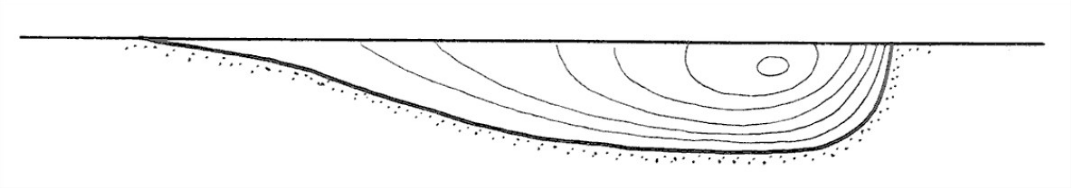


Figure 5-38. Flow-transverse section through a meander bend, showing isovels.

Point bars tend to be built episodically, during floods. The earlier positions of the top of the point bar are marked on the floodplain on the inner side of the meander bend by low and curving ridges arranged congruently within the meander bend (Figure 5-39). These curving ridges are called ***meander scrolls*** or ***meander scars***. One of the best kinds of evidence that the bend shifts in position as it grows is the truncation of earlier meander scrolls by the present channel.

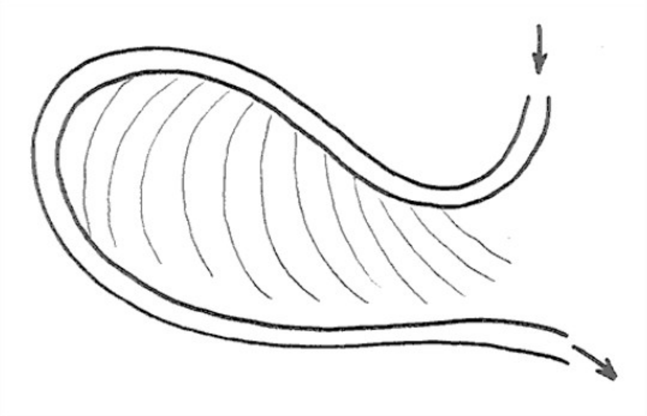


Figure 5-39. Meander scrolls.

The ends of an abandoned meander loop soon become plugged by fine sediment to form an ***oxbow lake*** (Figure 5-40). Oxbow lakes are filled in very slowly by deposition of fine sediment during overbank flows on the floodplain. Their outlines in the floodplain remain visible from the air long after they are filled, however, owing to slight differences in color and vegetation on the floodplain surface. The floodplains of meandering rivers show a complex pattern of several generations of truncated meander scars and partly or wholly filled oxbow lakes recording a long history of meandering.

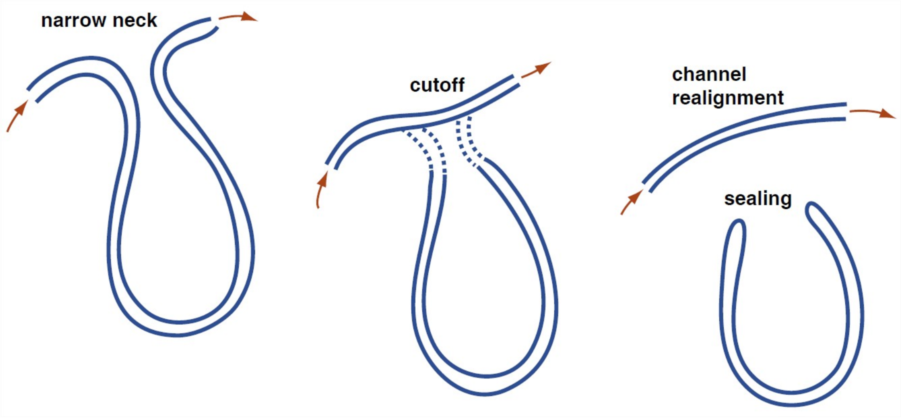


Figure 5-40. Development of an oxbow lake by meander cutoff.

Perhaps the best way to approach the problem of the dynamics of meandering and braiding is to think in terms of *the development of a meandering or braided pattern in a stream that is initially straight and regular and flowing within a homogeneous floodplain*. This kind of thing is easy to do in a wide flume in the laboratory (Figure 5-41): put the sediment in, homogenize it by hand, level it off, and then carve a regular channel in it. (You could fairly easily build apparatus in which to do this in your own backyard, which by now is getting rather crowded.) When a uniform sediment-moving flow is then started in the channel, the channel remains straight and regular for a long time, but eventually, after hours or even tens of hours, the inevitable little irregularities of the initial channel lead to an alternating pattern of erosion and deposition along the banks and therefore to increase in sinuosity of the channel.

If the banks of the laboratory stream are freely erodible (that is, if they consist of loose sand and gravel, rather than cohesive muds) then *the channel pattern is at first regularly sinuous but eventually becomes highly irregular and braided*. It's not nearly as easy to do this experiment in cohesive materials, but what's been done shows that in cohesive banks the braiding does not develop, but instead a statistically steady process of meander growth and cutoff develops.

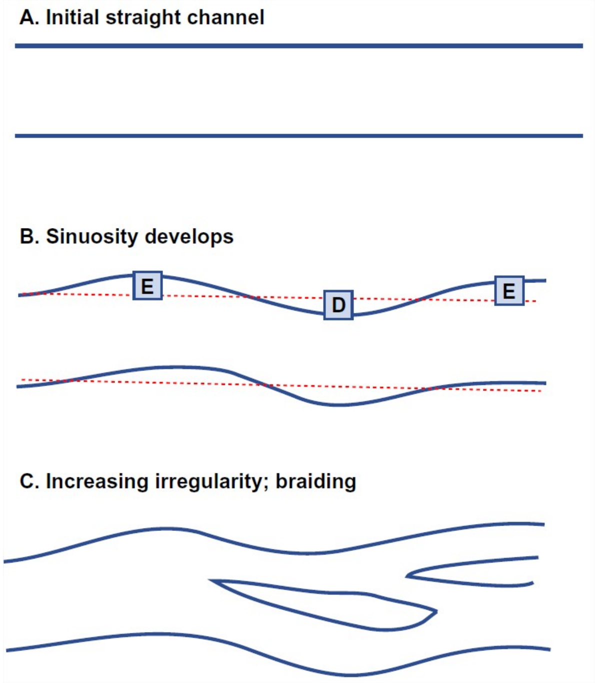


Figure 5-41. Development of braiding.

So you can see that a natural approach to the dynamics of meandering is to treat meandering as what the experts call a *stability problem*. In such an approach one attempts to capture the physics of the process, in terms of the interaction between the flow and the sediment transport, in the form of a set of equations, and then one attempts to linearize those equations in order to apply them to the behavior of a very-small-amplitude perturbation of an originally straight channel. *If the perturbation is damped, then the straight channel is stable; if the perturbation is amplified, then the straight channel is unstable.* The fastest-growing wavelength of the disturbance is the one that should specify the scale of the resulting meanders. Although the full equations may not be workable for determining the evolution of the meandering pattern once the disturbances have a finite amplitude, various other lines of attack have been brought to bear on particular aspects of finite-amplitude meanders, with considerable success.

On a more concrete note, it’s widely agreed that the most important factor in determining whether a river meanders is *bank stability*. A floodplain with a large percentage of fine, cohesive sediment in it is conducive to meandering, as is a dense cover of vegetation. If the sediment is noncohesive and freely erodible, and the climate is inconducive to vegetation, the river braids rather than meanders. Incidental note: before the Silurian, when there were no land plants, there is no evidence of meandering streams, only braided streams.

5.10: Drainage Networks

Introduction

Given an area of land on which rain falls, there are always streams, even if they are only ephemeral. The entire area can be parceled exhaustively and mutually exclusively into sub-areas such that each sub-area is the drainage area of a single stream (Figure 5-42). Such areas are called the ***drainage basins*** or ***watersheds*** or ***catchments*** of those streams. You learned some things about drainage basins in the earlier chapter on topography. Recall from Chapter 3 that the boundaries between adjacent drainage basins are called ***divides***. Just think of standing straddling the divide and pouring a glass of water from one hand into one drainage basin and from the other hand into the other drainage basin. Here the emphasis is on the network of streams in the drainage basin. The ***drainage network*** of a drainage basin is *the collection of all of the channels in a given drainage basin*: the main stream, often called the ***trunk stream***, and all of the tributaries, as well as the tributaries of tributaries, and so on.

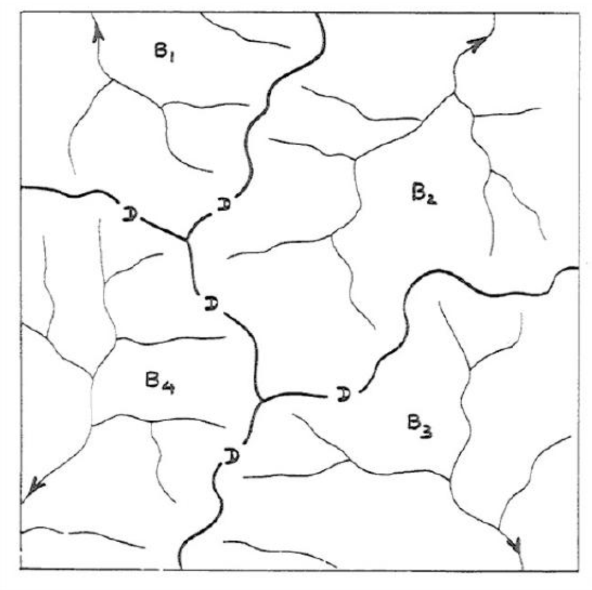


Figure 5-42. Drainage networks and drainage basins.

Drainage Patterns

By ***drainage pattern*** I mean *the spatial relationships of individual streams in the drainage basin*. The drainage pattern is a reflection of several factors:

- stage of development
- initial slope
- rock hardness
- rock structure
- recent tectonism

The pattern of each of the drainage networks shown in Figure 5-41 is what you would expect to see in a region that is underlain by fairly uniform materials—that is, materials that are about the same in their characteristics, in particular their susceptibility to erosion, from place to place.

Because an area responds to denudation (by ***denudation*** I mean *lowering of the land surface by weathering, erosion, and removal of the resulting regolith*) in a way that reflects the underlying geology, you can tell a lot about the geology of an area by looking at a map of the drainage pattern. Here are some of the common drainage patterns (Figure 5-43), with a few words on the bedrock controls:

Dendritic. This pattern, by far the most common, involves irregular branching of tributary streams in many directions and at almost any angle, though usually less than 90°. A dendritic pattern develops on rocks of uniform resistance, and without structural control. It is most likely to be found on flat-lying sedimentary rocks or massive igneous rocks. The drainage networks shown in Figure 5-41 could be described as dendritic.

Trellis. A system of subparallel streams, usually aligned along the strike of rock formations or between parallel or nearly parallel topographic features deposited by wind or ice. The main streams often make nearly right-angle bends to cross major ridges. The primary tributary streams are usually at right angles to the main stream, and the secondary tributaries are usually themselves at right angles to the primary tributaries, i.e., parallel to the main streams. Trellis patterns are most common in folded mountain belts like the Appalachians, where alternating folded weak and strong layers have been truncated by stream erosion.

Rectangular. Both the main stream and its tributaries show right-angle bends. This reflects control exerted by joint or fault systems. Where the joint or fault systems intersect at acute angles, an ***angulate*** pattern develops.

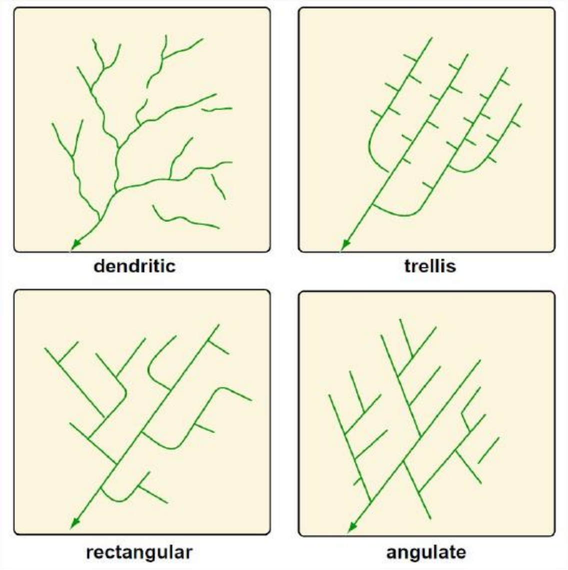


Figure 5-43. Common drainage patterns.

Some other patterns (Figure 5-44) are less common:

- **Centripetal.** Drainage lines converge into a central depression, like a sinkhole, crater, or other basin.
- **Radial.** Streams diverge from a central elevated tract: domes, volcanic cones, isolated conical hills.
- **Parallel.** Found where there is a strong slope control, leading to regular spacing of parallel or nearly parallel streams.
- **Angular.** Around maturely dissected domes which have alternating belts of strong and weak rock around them.

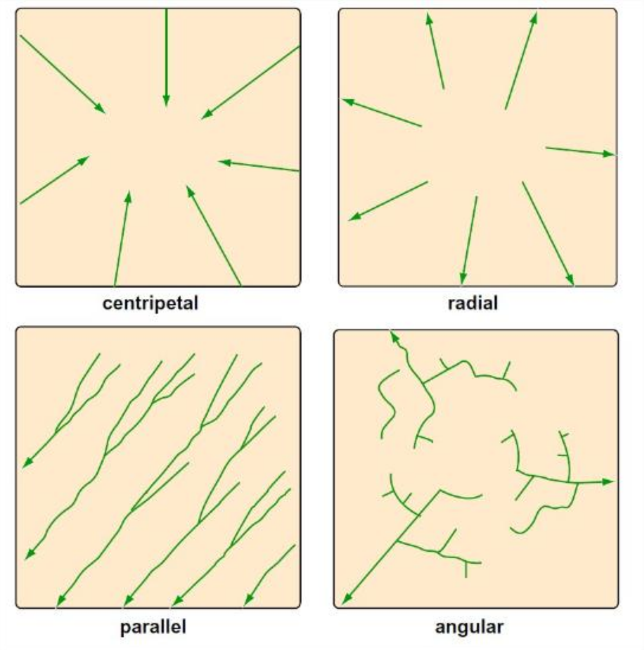


Figure 5-44. Other, less common drainage patterns.

ADVANCED TOPIC: FLUVIAL MORPHOMETRY

Fluvial morphometry is the term used to describe the analysis of the topological characteristics of drainage networks..

Here are some things you could measure about a drainage network:

- *configuration* or topological properties
- *linear* properties: analysis of a branching system of one-dimensional curves in plan.
- *areal* properties: analysis of areas of drainage basins, again in plan.
- *relief* properties: analysis of slope of channels or drainage areas.

Each one of these properties has been subjected to extensive study by the techniques discussed below. We'll examine only the first here.

First we need to define the concept of the ***order*** of a stream. To do that, look at a representative drainage network, and concentrate on its topology. The stream-channel network can be subdivided into segments lying between confluences, together with ***fingertip streams*** lying upstream of any confluences (Figure 5-45). One can develop a hierarchy of stream orders based on the tributary relationships of the channels, in the following way.

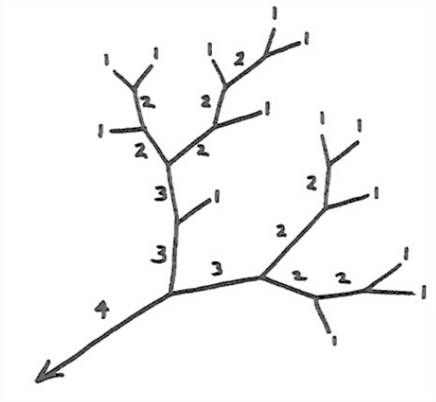


Figure 5-45. An example of stream orders.

The convention is that the fingertip channels (those with no tributaries) are called ***first-order streams***. The junction of two first-order streams produces a ***second-order stream***, and the junction of two second-order streams produces a third-order stream, and so on. Another convention is that the junction of a stream of given order with a stream of lower order doesn't change the order of a higher- order stream. The trunk stream of the drainage basin—the one by which the given drainage basin is actually defined, by choosing a given point on that stream—is the highest-order stream associated with that drainage basin.

The various relationships between stream order and length, area, and number properties of drainage basins have come to be known as ***Horton's laws***. Horton's law of stream orders states that

> The numbers of streams of different orders in a given drainage basin tend closely to approximate an inverse geometric series in which the first term is unity and the ratio is the bifurcation ratio.

(The ***bifurcation ratio*** is *the ratio of the number of streams of a given order to the number of streams of the next highest order*.) By the nature of a geometric series, this means that when you plot the logarithm of the number of streams of a given order against the stream order you tend to get a straight line (Figure 3-46). Graphs of this kind are called ***Horton diagrams***. Most of the curves in Horton diagrams tend actually to be slightly concave upward rather than being a straight line. Analysis of all the studies that have been published shows that this is a strong tendency. There are similar laws for other characteristics of the drainage network: the length of stream segments between confluences, and the drainage areas of stream segments.

Figure 5-46. Plot of the logarithm of the number of streams of a given order against the stream order, for a given drainage basin.

An important observation is that Horton’s law holds for almost all watersheds that have been studied, in all kinds of areas, with all kinds of climates and bedrock types. Although there seem to be some deviations caused by lack of isotropy and homogeneity of the bedrock structure and composition, the law of stream numbers on the whole is quite insensitive to such geologic controls. This suggests that there must be some very general basic cause.

Another interesting observation is that the bifurcation ratio also seems to be insensitive to geologic controls as well as to climate. It can be shown theoretically that the bifurcation ratio can never be less than 2; it’s always between 3 and 5, and usually around 4. It shows only a small range of variation from region to region and from environment to environment, except where powerful geologic controls dominate.

Much effort has been expended to try to interpret the meaning of Horton’s law of stream orders. Although there are dissenters, it’s widely accepted that Horton’s law is a manifestation of randomness in the topological development of the drainage network. In a classic paper, Shreve (1966) showed that the assumption that all topologically distinct possibilities for how streams of a given order combine to form a network are equally likely leads to the characteristically slightly concave-up curves on Horton diagrams.

5.11: Fluvial Deposits

Fluvial deposits are an important part of the ancient sedimentary record. The reason why is not obvious; after all, rivers drain areas of the continents that are undergoing erosion. Most rivers, except the smallest, are alluvial rivers: they have a bed, and a floodplain, composed of their own sediments. But in most cases this alluvial valley sediment is not very thick. Only in certain cases does the alluvial valley fill become thicker.

Two effects are conducive to deposition in rivers: *progradation* and *crustal subsidence*.

Progradation. As a river wears down the land and delivers sediment to the sea, the mouth of the river builds seaward. Because the longitudinal profile of a river is anchored by base level at the mouth, this means that there has to be a slight upbuilding in the lowermost reach of the river (Figure 5-47). This may not seem like a big effect, but even some tens of meters is a lot of sediment.

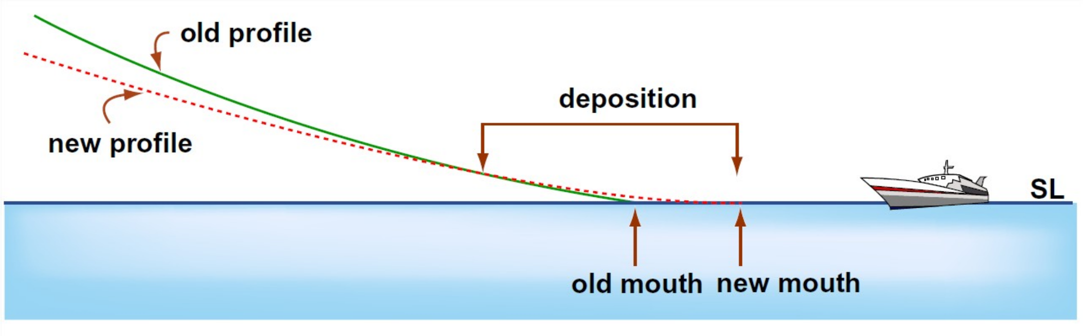


Figure 5-47. Deposition by seaward progradation in the lower reaches of a river system.

Crustal subsidence. The only way to get a really thick sequence of fluvial sediment is to *drop the continental crust beneath the river*. As this happens, slowly, along some reach of the river, there develops a very slight expansion of flow and decrease in flow velocity and therefore in sediment-moving ability. Just by simple bookkeeping, this must lead to storage of sediment along the river: if what comes into a given area of the bed is greater than what goes out, sediment is stored in that area, and the bed builds up. From an anthropomorphic standpoint, the river tries to maintain its longitudinal profile while the bottom drops out from under it, and it does so by leaving a little of the passing sediment to build up its bed (Figure 5-48).

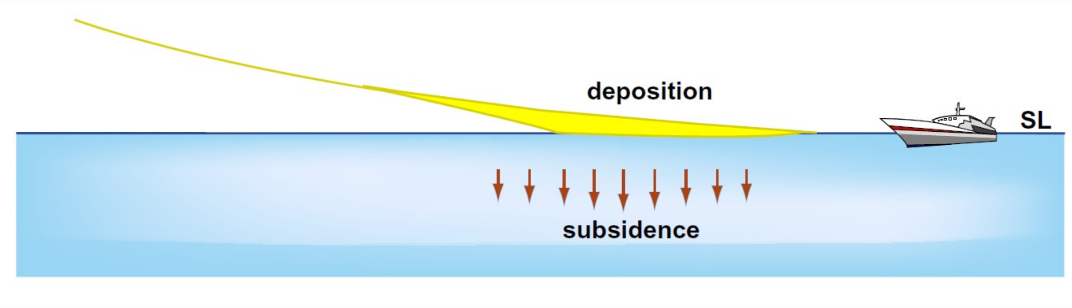


Figure 5-48. Deposition by crustal subsidence in the lower reaches of a river system.

5.12: Floods

The Definition of a Flood

Here's a simple definition of a river flood: *the occurrence of a flow of such magnitude that it overtops the natural or artificial banks in a reach of river channel*. If a floodplain exists, here's another way of defining a flood: *any flow that spreads out over the floodplain*.

The *relative* volume of flood water is not large: on average, water discharged in excess of channel capacity constitutes about 5% of the total annual discharge of the given drainage basin. But the *absolute* volume of flood water is staggering. Hypothetical example: a two-inch rainstorm over a one-square-mile catchment the soils of which can absorb 0.5 inches would supply almost 3.5 million cubic feet of water is a couple of hours! (That's a cube 150 feet on a side.)

A flood is a *wave*: it has a wavelike shape (although very subdued, with low amplitude relative to wavelength), and it propagates downriver at some speed, typically less than the mean velocity of flow of water in the river.

More on Floodplains

Here are two similar but not quite identical definitions of a floodplain, taken from the literature:

- the flat area adjacent to the river channel, constructed by the present river in the present climate and frequently subjected to overflow (Leopold, 1994)
- the flat area adjoining a river channel constructed by the river in the present climate and overflowed at times of high discharge (Dunne and Leopold, 1978)

As you learned in an earlier section, the existence of a floodplain is testimony to the lateral migration of the river channel. With time, the shifting channel sweeps the entire width of the floodplain, eroding the "front" bank and depositing the "rear" bank as it goes. If the climate, or the base level, changes in such a way that the river degrades (i.e., lowers its bed), the old floodplain is incised, forming a set of terraces, and a new, narrower floodplain develops. If the climate, or the base level, changes in such a way that the river aggrades, the floodplain typically (except in extreme cases) aggrades along with it.

What determines the height of the floodplain above the channel bed? That is, why does the river have the characteristic transverse profile we see? That has to do with the relationship between *effectiveness* and *frequency*. Low discharges are common but not effective in sediment transport and channel shaping, whereas very high discharges are very effective but very uncommon.
The discharges with greatest effectiveness in maintaining channel and floodplain geometry are at some intermediate value: high but not extremely high, and common but not extremely common. For most rivers, bank-full discharges (those that spill over onto the floodplain) occur once or twice a year, on average. (It turns out that the discharge that is responsible for the maximum sediment discharge, in a given time period like a year, called the ***effective discharge,*** is very close to being equal to the ***bank-full discharge*.**)

What are floodplains like?

• They range in width from several meters, on the smallest streams, to many tens of kilometers, on the largest rivers.

• They are not flat as a board: they have minor relief, in the form of sloughs and depressions left behind by the irregular shifting of the river channel.

• They are developed in their classic sense in connection with meandering rivers. Braided rivers work irregularly across the entire valley, such that in a sense the entire valley is "river bed".

• They are typically heavily vegetated in their natural state; in fact, the bank-stabilization effect of vegetation is one of the most important, if not *the* most important, factors in the existence of the floodplain in the first place.

• They vary in their "wetness": usually they are dry and well above the local water table for most of the year, and are thus conducive to human use, but in areas with consistently high rainfall and/or strongly aggradational river regime ,there can be extensive swampy areas on the floodplain.

Important Aspects of Floods

There are several important aspects of floods:

- *discharge:* discharge is important because of its close relation to flow velocity: how erosive is the flood flow in terms of banks, levees, and river structures?

- *stage:* the height of the water determines flood damage on the floodplain (This is what the public thinks about most in connection with floods.) Also, it relates to the water-passing capacity of bridges, culverts, dam spillways, etc.
- *volume:* this relates to how much water is available to be caught and impounded in reservoirs for water supply and flood control.
- *area inundated:* this is determined by stage together with river-valley topography.
- *flow velocity:* Although the flow velocity is much less over the floodplain than in the main channel, it can be great enough to cause damage.

Here are some important questions connected with river floods:

- governing effects?
- getting worse?
- how to control?
- how to predict?

What are the important effects or factors that govern the magnitude of a flood?

• **Rainfall** in the drainage basin upstream of the given point on the given river. This boils down to both rainfall *intensity* (depth per unit time) and rainfall *duration*, and also the *area* of the drainage basin covered by the rainstorm. The worst scenario is: high intensity for a long time over a large area of the drainage basin.

• Preceding condition of **soil moisture** in the area of rainfall: dry, which lessens the runoff, or saturated, which maximizes the runoff? The difference can be as much as half an inch of equivalent rainfall. That's a big effect for small floods but only a small effect for large floods.

• **Channel storage**. Think in terms of what happens when a flood discharge from an upstream river enters a large lake. The lake has a large surface area, so it takes a long time for the flood flow to raise the lake elevation and therefore increase the discharge out of the lake. This storage effect is also present to a non-negligible extent along the river channel itself: As the flood proceeds downstream, it has to fill a greater and greater volume of space in the river channel itself. This effect *attenuates* the flood in the downstream direction: in effect, it *stretches out* the hydrograph and lowers its peak (in terms of either the stage hydrograph or the discharge hydrograph).

• **Channel geometry**. The narrower the channel, for a given flood discharge, the more the stage rises. When the stage reaches bank-full, the water spreads out over the floodplain, and the stage then increases far less rapidly with increasing discharge.

• **Transition from dunes to plane bed.** As discussed in an earlier section, this tends to slow the increase in stage with increase in discharge: the channel can pass a given discharge at higher velocity and lower stage because of the reduced resistance to flow afforded by the plane bed.

Are Floods Getting Worse?

The answer to that question is certainly "yes", in terms of *dollar value of damage*, just because of greater human use of flood-prone areas near rivers. But the answer is yes also in terms of *higher stage*, because of the effect of more levee protection: the levees prevent the high discharges from spreading over the floodplain, so the rise in stage is confined to within the channel (Figure 3-49).

Figure 3-49. The stage of a flood flow for a given flood discharge. A) Before building of levees. B) After building of levees.

Is flood *discharge* increasing on average as well? There are two aspects to this question.

(1) In *urbanized* drainage basins, the percentage of area paved increases, thus decreasing the concentration time and tending to produce higher peaks in the hydrograph (stage and discharge) for a given rainfall. In rural drainage basins, bare-soil agriculture and clear cutting of forests have the same effect, although it seems not to be a really major effect. (Their effect on sediment yield is much greater, though.)

(2) Average rainfall in the drainage basin might be increasing because of *climate change* (although in much of the U.S., the prediction is that global warming will decrease precipitation rather than increase it). "Storminess" (increased incidence of major rainstorms) might increase as well.

Can Floods Be Controlled?

A better way of putting the question would be: how can flood damage be minimized, in the sense that a given measure affords more flood protection than it costs to institute?

(1) *Build higher and higher levees*. Yes, you can keep the floodwaters off the floodplain by building the levees higher. In a real sense, however, this is a self-defeating activity (Figure 5-50). Under natural conditions, the flood stage has built into it an inherently self-limiting effect: the floodplain area is so much greater than the channel area that, once the river goes beyond bank-full to inundate the floodplain, water-level rise is minor. (But remember that "minor" in a natural sense may be catastrophic for people living on the floodplain.) Building the levees higher restricts the river discharge to a much narrower width of the river valley, so the flood stage is much higher than would have been the case without the higher levees. Because of the effect of eliminating the floodplain-spreading effect, effective protection would necessitate impractically high levees. Moreover, if those high levees are overtopped, the flood damage to the surrounding floodplain area is likely to be much worse than if the levees had not been there.

(2) *Flood-control dams*. Clearly, the stage and discharge hydrographs can be reduced downstream of a given point on a river by building a water-storage dam there: catch the water so as to knock down the flood peak, and release the water slowly later in times of lower discharge. The problems, aside from the great cost of building dams, is that the effect lessens downstream, so a number of dams are needed. For decades, in the mid-1900s, there was a controversy over whether to build lots of small dams in upstream areas or a few big dams in downstream areas. It was finally concluded that *for effective flood control throughout the whole basin, both sets of dams would be needed.*

(3) *Land management.* Good agricultural practices (fallow planting; terrace agriculture; contour planting) should increase infiltration and thus lower runoff, and also make the concentration time longer. Careful consideration has shown, however, that the effect is measured in tenths of an inch of equivalent rainfall, so it would be effective against minor floods but not major floods. The great benefit is greatly decreased soil erosion—which is a different (but very important) matter.

(4) *Restrictions on human use of floodplains*. Clearly the best way to prevent flood damage is not to put anything susceptible to damage on floodplains! The expense of relocation is usually too great for this to be practical (although it's happening in some places; it's a policy issue rather than a scientific issue), but intermediate measures are effective: establish floodways and spreading areas that are used for such things as agriculture (of certain kinds) and recreation, but not permanent structures. This doesn't solve the problem for already urbanized areas, though.

Can Floods Be Predicted?

Simple answer is just "no". Well, in a sense you *can* predict floods, but only after the rain has stopped falling! There are two aspects to this:

(1) Given the rainstorm, predict the runoff in order to create a flood hydrograph

(2) Given the flood hydrograph, watch the evolution of the hydrograph as the flood wave moves downstream. This technique is called ***flood routing***.

Various approximate techniques have been developed, and continue to be refined, for predicting the nature of the flood wave downstream of a rainstorm in a drainage basin. The assumptions and approximations are many: distribution of rainfall; extent of runoff vs. infiltration; time of concentration to create the hydrograph. It works okay for very small catchments but not so well for large catchments. It's a standard engineering technique for very small catchments (much less than one square mile). Once a flood is underway, however, there are good flood-routing techniques for predicting its future course down the river. Flood routing is treated at length in textbooks on fluvial hydrology.

The other aspect of flood prediction has to do with the *statistics or probability of floods*. One assumes that, given the climate and the nature of the drainage basin upstream of a given point on a river, there is some underlying frequency distribution of floods, in terms of either stage or discharge. Any record of floods is then a sample population (in the parlance of statistics) from that underlying frequency distribution. So if you know the frequency distribution, you can give the probability of a given flood in a given time period (that's not a prediction, but it's very useful in making a cost–benefit analysis or decision). If you are interested in learning more about how fluvial hydrologists go about figuring out the probability of a given flood in a given time period in a given river system, go to the following advanced topic.

ADVANCED TOPIC: THE PROBABILITY APPROACH TO ERIVER FLOODS

1. The important question that arises is: *what can be done about establishing the probability of very large floods*? This is extremely important, but it’s fraught with difficulties. Here’s how it’s done, given a long record of discharge at some location along the river (see Figure 5-51):

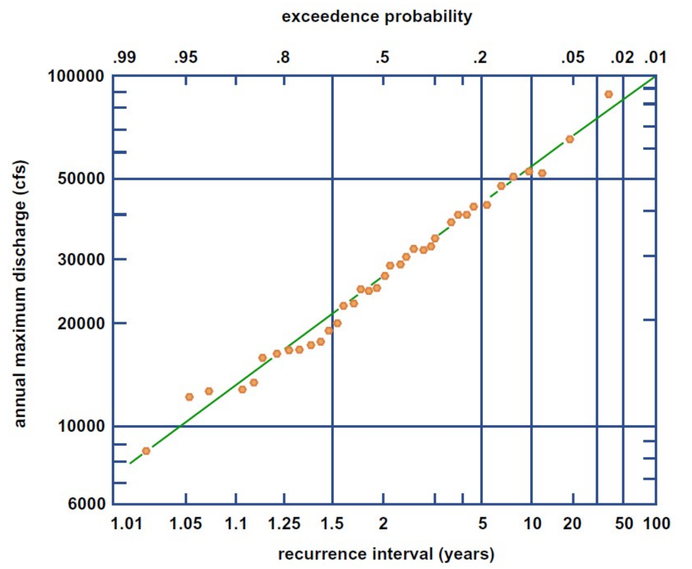
- Determine the largest momentary discharge in each year.

- Rank these in order of decreasing discharge (the highest being number one, the lowest being equal to the number of years of record).

- Compute the ***recurrence interval*** (also called the ***return period***), in years: the average number of years within which a given event (i.e., a flood of given magnitude) will be equaled or exceeded), by the formula $T = (n+1)/m$, where T is the recurrence interval, n is the number of years in the record, and m is the rank of the event (or, equivalently, the probability p of a given event being equaled or exceeded in any given year; that’s called the ***exceedence probability***).

- Plot the results in a graph with recurrence interval and/or exceedence probability on the horizontal axis and the annual maximum discharge on the vertical axis.

- Fit a smooth curve through the points in some way.



Flood-frequency curve plotted on logarithmic probability paper, Tana River at Garissa. Keny. 1934-1970. The scale at the top is the probability that the discharge is equaled or exceeded in any given year. The bottom scale, recurrence interval, is the average number of years in which the annual peak equals or exceeds the discharge given on the ordinate.

Figure 5-51. An example of a flood-frequency curve. (From Leopold, 1994.)

2. The results are reliable for the mid-range of the graph, but they become very uncertain near the high-discharge end because of the small numbers of very large events. Fitting such a tail of a distribution is always tricky business, but it is important if one is to try to extrapolate to much longer times, which is one of the important goals of such an exercise. Various techniques have been in use for fitting such a curve. Usually one uses a kind of graph paper with logarithmic vertical (discharge) axis and a horizontal axis rubber-sheeted so that certain theoretical frequency distributions plot as straight lines. (That makes it easier to fit the tails of the distribution.) Several such distributions have been in use. The trouble with blind curve fitting is that one or two unrepresentative or outlying points at the high-discharge end can bias the curve significantly. Some hydrologists prefer to fit the curve by eye, using judgment about whether to include such seemingly outlying points.

3. One valuable approach is to try to determine, from historical records, whether the greatest discharge in the period of record was exceeded in earlier times. If the greatest discharge in the period of record can be established to have been greater than any other for a number of years before the period of record, the period of record can be extended back in time and the ranking redone. This tends to shift the unrepresentatively large outlying point to the right and bring it more closely in line with the rest of the distribution.

Effects of Urbanization

Urbanization can have major effects of the magnitude and frequency of floods. Here’s an example (Figure 5-52): Seneca Creek, near Rockville, Maryland, drains an area of 100 square miles. Its watershed has been subject to much urbanization in recent decades. Take the record of discharge, 61 years long, from 1931 to 1991, and divide it into two equal parts, early and late. The average annual flood in the period 1931–1960 was 3000 cfs, in the period 1961–1991 it was 6000 cfs—twice as great! Likewise, bank-full discharge occurred 1.2 times per year in first period but 2.2 times per year in second period.

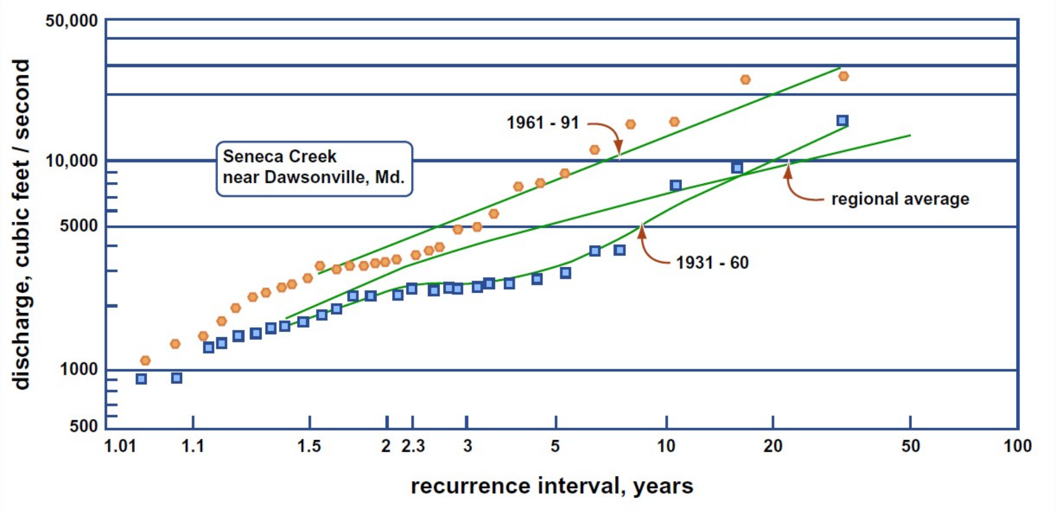


Figure 5-52. Flood-frequency curve for Seneca Creek, Maryland. (From Leopold, 1994)

Rainfall was almost exactly the same in the two periods, so the great differences have to be accounted for entirely by the effects of urbanization effects. The stream is not artificially leveed, so presumably the effect has been caused by a decrease in concentration time; the more pavement, the faster the runoff, and the shorter the concentration time.

Effects of Aggradation

There’s a serious long-term problem in areas, especially in the lower reaches of large rivers, where crustal subsidence is non-negligible (i.e., the substrate beneath the active channel is moving downward relative to a datum like sea level) and/or where sea level itself is rising. (The former effect is more important than the latter.) The subsidence can be a combination of two effects:

- deep crustal subsidence

- subsidence caused by compaction of the sedimentary substrate of the river

The river channel aggrades to keep pace with subsidence, and in consequence it finds itself higher and higher above its floodplain —because, under conditions of net aggradation, deposition can be much faster in the channel, where the river is transporting most of its sediment, than over the floodplain. At certain times the river breaks through some particularly vulnerable point along its natural levees and finds its way out onto the floodplain, which is now at a significantly lower elevation, and establishes a whole new course for long distances down the floodplain. The process is called ***avulsion***; the river is said to ***avulse***. The river then proceeds to build up its channel bed and its levees, again to be high above the floodplain, and avulse once more. Avulsions often develop over a period of time rather than happening all at once: more and more of the river discharge is diverted through the gap in the levee, flood after flood. The end result, however, is the same.

To prevent avulsion, humans in avulsion-prone areas tend to build higher and higher levees, in the hope of preventing such an unfortunate event. The channel bed gets higher and higher relative to the surrounding floodplain (Figure 5-53).

This is happening today in the lowermost reaches of the Mississippi River. Much of the city of New Orleans lies well below the river. The Mississippi has an unfortunate preference for flowing down the Atchafalaya River to Morgan City. The U.S. Army Corps of Engineers maintains a gigantic headworks on the left bank of the river upstream of New Orleans to prevent the entire discharge of the Mississippi from going down the Atchafalaya; at present, about a third of the discharge leaves the Mississippi at that point.

How long such a situation can be maintained is a matter of some controversy. If you would like to read an engaging account of the problem with the lower Mississippi, see the book by John McPhee (1989).

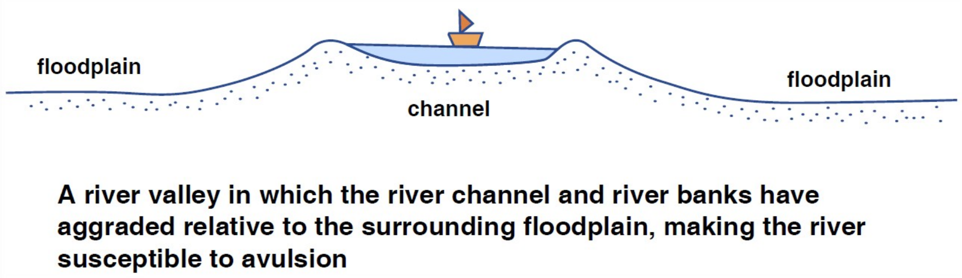


Figure 5-53. A river valley in which the river channel and river banks have aggraded relative to the surrounding floodplain, making the river susceptible to avulsion.

China has an even bigger problem with the Huang He (in English, the Yellow River), because they have been fighting the river for far longer than we have been fighting the Mississippi. I know from having been in China that the bed of the Huang He at the city of Kaifeng is almost fifteen meters above the streets of the city, only several kilometers from the river! The Chinese are working hard to come up with a way of solving this problem in the long term; for the near term (a few more decades) the levees are considered to be adequate.

5.13: Some Practical Aspects of Rivers

Water Supply

Many large and small cities on rivers use river water for municipal water supply. The advantage is that it's there: all you have to do is pump it out of the river rather than drill and maintain deep wells or build and maintain often distant reservoirs. One big disadvantage, though, is pollution. You can chlorinate for pathogenic organisms and filter for turbidity, but you can't easily extract dissolved chemicals introduced upstream. This is especially a problem for large cities located on the lower reaches of major rivers, like New Orleans.

Irrigation

In many areas, rivers are important sources of irrigation water as well as groundwater. You don't have to pump: just divert some of the river into irrigation canals and let it flow by gravity into agricultural areas adjacent to the river.

There are problems, however:

- As with municipal water supply, this may use up a large percentage of river discharge by the time the river empties into the ocean!
- That water has to go somewhere. Where does it go? (1) It's lost by evaporation and transpiration (loss by evapotranspiration is enormous in irrigated agriculture); (2) It infiltrates, then flows as groundwater back into river downstream. Chemical pesticides used on cropland partly remain in the soil and partly are leached with return groundwater flow into the river.

Dams

The issue of dams is a complicated one. Humankind has been building dams since way back in prehistory. There are several kinds of dams: earth fill, rock fill, and concrete. The technology of dam construction is well advanced (but, occasionally, there still are dam failures!).

What are the main uses of dams? (Note: these uses are often, perhaps usually, combined.)

- water supply
- flood control
- hydropower generation
- recreation

Below (in no particular order) are listed some of the aspects of dams that need to be taken into consideration by urban planners and other governmental authorities as well as by environmental scientists and engineers:

- cost–benefit analysis
- safety
- displacement of humans
- land loss
- ecosystem disturbance
- evaporation loss
- siltation
- downstream degradation

If rivers carried no sediment, problems with dams would be far less serious. All rivers, however, carry sediment, and all but the smallest are alluvial rivers. In such rivers, sediment is in one way or another the biggest problem in reservoirs behind dams. Unless special (and very difficult and costly) engineering measures are employed, the reservoir is a sink for sediment arriving at the upstream end. *All such reservoirs therefore have a finite lifetime*. These lifetimes are usually measured in decades, and often not many decades.

What happens when the reservoir is mostly filled with sediment?
(1) Dig it out—which is almost always impractical. (2) Let it go over the spillway with the water. The problem is most severe with

respect to flood control and irrigation: the capacity of the reservoir becomes negligible. The problem is not so severe for hydropower generation, because the difference in water level between the water surface behind the dam and the turbines at the base of the dam is still there, but there are serious engineering problems connected with how to use the water without passing the sediment through the generating facility as well.

Degradation of the river bed below dams is another big problem. If you store the sediment in the reservoir instead of letting it pass, sediment transport rate just downstream of the dam is zero, and the river tends to entrain bed sediment to establish its equilibrium sediment load. The river then cuts down into its bed.

It’s often not realized how much water is lost to evaporation from the water surface of the reservoir behind the dam, especially in summer and in arid climates. Various engineering schemes have been proposed for reducing such evaporation, but none has proved practical.

Cooling Water

Water temperature in a given river depends upon a number of natural effects: climate, season, residence time, and heating by energy dissipation. This thermal regime can be seriously disrupted by the need to supply cooling water to power plants, which are often located along rivers just for accessibility to cooling water. The main problem is the biological effect on river fauna and flora downstream.

Stabilization

Rivers, both meandering and braided, have an annoying tendency to shift their channels laterally. (Well, annoying for certain groups of people.) This creates obvious problems for residents along the river. In recent decades the engineering response has increasingly been to make rivers like artificial channels by lining the banks with various kinds of large heavy platy or interlocking objects, first at places of greatest vulnerability and then, increasingly, all along the river banks. Most large rivers in the United States are now partly stabilized in this way. The Mississippi is mostly stabilized. This was not always the case, as a glance at a map of the states adjacent to the Mississippi readily shows: little pieces of one state are now stranded on the other side of the river!

CHAPTER OVERVIEW

6: Lakes

6.1: The Origin of Lakes

Introduction

How do you define a lake? Perhaps the best way of defining a ***lake*** is that it’s *a body of water surrounded by land with no connection to the oceans*. Here are some comments on this definition:

- It excludes all the oceans.
- There’s a conventional size restriction that excludes little puddles and small ponds.
- Some bodies of water that are called seas are thus lakes; examples are the Caspian Sea and the Dead Sea.
- The salinity of the lake water is not part of the definition. But in this chapter, I’ll deal only with fresh-water lakes.

The Origin of Lakes

Although the origin of lakes is not really a part of this course, here are some comments that might be of interest to you. Lakes originate in a variety of ways:

- subsidence of land below the groundwater table (Figure 4-1A).
- isolation of a part of the ocean, either by local constructive processes of sediment deposition or by crustal uplift (Figure 4-1B).
- glacial erosion and deposition on the continents (Figure 4-1C).
- miscellaneous ways: volcanoes, damming by landslides, or meteorite impacts.

Here are some comments on the foregoing list:

- The first two items involve earth movements, and they’re very slow.
- Some large lakes caused by isolation of an arm of the ocean are called “seas”, although they really are lakes. If you could somehow close the Mediterranean Sea at the Strait of Gibraltar, it would be a lake—and in recent geologic times it’s widely believed that it once *was* closed to form a gigantic lake, which at one point might actually have evaporated to dryness!
- Lakes of glacial origin are very important in the northern part of the North American continent nowadays, because of the extensive glaciation in recent geological times.

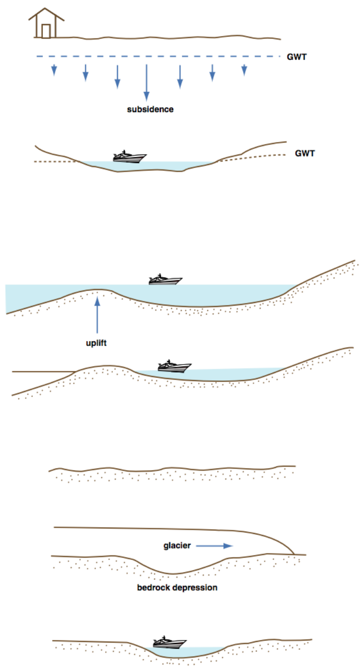


Figure by MIT OCW.

Figure 6-1. Ways lakes are formed.

The study of lakes is called ***limnology***. Limnology has several important aspects:

- water balance
- temperature structure • water circulation
- chemistry
- biology
- bottom sediments

Study of the water circulation, especially of large lakes, is based on the same principles as in the oceans, and tends to be practiced by the same set of specialists as in oceanography. The bottom sediments are studied by much the same techniques as the bottom sediment of the oceans. In this chapter I'll consider only the first two items on the above list.

6.2: The Geometry of Lakes

You can measure various geometrical things about lakes:

- area of water surface
- water volume
- maximum depth
- shoreline length
- lake length: this is equal to shortest distance through the water surface between the most distant points along the shore (Figure 6-2). This gets tricky when lakes have complicated surface geometry.
- lake breadth: this is equal to the length of a line perpendicular to the length defined above, at any point; the mean breath is equal to the area divided by the length (Figure 4-2).

In plan view, lakes range from equidimensional or irregular to very elongated. In certain special situations lakes are approximately circular—for example, kettle lakes (see Chapter 5).

In cross section, lakes are usually much wider than they are deep, especially large lakes. Small lakes are often deeper relative to their width than large lakes.

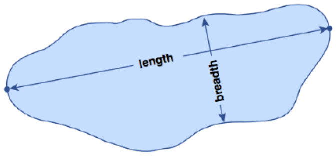
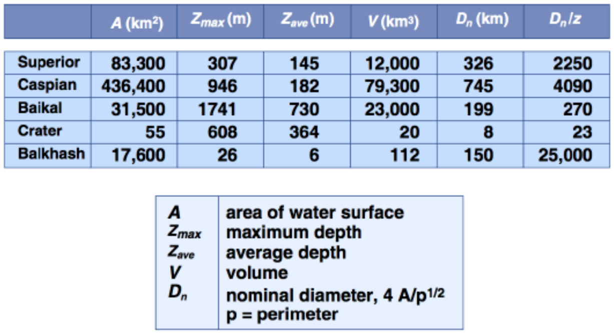


Figure by MIT OCW.

Figure 6-2. Length and breadth of a lake.

Table 6-1 gives some data about several well-known large lakes around the world.



	A (km²)	Z_{max} (m)	Z_{ave} (m)	V (km³)	D_n (km)	D_n/z
Superior	83,300	307	145	12,000	326	2250
Caspian	436,400	946	182	79,300	745	4090
Baikal	31,500	1741	730	23,000	199	270
Crater	55	608	364	20	8	23
Balkhash	17,600	26	6	112	150	25,000

A	area of water surface
Z_{max}	maximum depth
Z_{ave}	average depth
V	volume
D_n	nominal diameter, $4\ A/p^{1/2}$
	p = perimeter

Figure by MIT OCW.

Table 6-1. Data on large lakes of the world.

6.3: Residence Time

The concept of the residence time of water in a lake is a natural one, because typically the water stays in a lake a long time, and velocity is much smaller than the source of inflow and outflow. The concept is not particularly interesting for rivers, though, because the velocity is about the same everywhere. The ***residence time*** of a lake is defined as *the volume of the water in the lake divided by the discharge into or out of the lake*.

Letting the volume of water in the lake be *V* and the discharge to and from the lake be *Q*, The residence time *Tr* is

$$T_r = \frac{V}{Q} \qquad (1)$$

(A comment on the discharge here: this is easy to deal with only for stream discharge, not for groundwater discharge.) You can easily check that Equation 1 provides the right physical dimensions for *Tr*, namely time. In terms of the dimensions mass **M**, length **L**, and time **T**, the dimensions of *Tr* are by Equation 1

$$\frac{\mathbf{L}^3}{\mathbf{L}^3/\mathbf{T}}$$

or, canceling out the **L** dimensions, just **T**, which is what we should have expected.

The residence time is important because it tells you something about how long it might take to clear pollutant from a lake. The trouble with it is that it *deals only with the average properties of the lake*. A given drop of water or molecule of pollutant may actually reside in the lake for a very short time or a very long time.

Here's an example of residence time in a lake. Look at a lake that's one kilometer by one kilometer in surface area and ten meters deep, with a stream feeding the lake with an average discharge of 20 cubic meters per second. This is a rather small river, but it's larger than just a stream. You could have this kind of discharge in a river with a mean velocity of about half a meter per second, a width of 20 meters and a depth of 2 meters. By Equation 1 the residence time then is

$$T_r = \frac{V}{Q} = \frac{(1000\text{ m})(1000\text{ m})(10\text{ m})}{20\text{ m}^3/\text{s}} = 5 \times 10^5\text{ s} \qquad (2)$$

or a little less than one day.

On the other hand, very large lakes like Lake Superior or Lake Baikal can have residence times of many hundreds of years! Two important implications are that *it takes a lot longer to pollute large lakes but also a lot longer to clean them up*.

6.4: The Water Balance of Lakes

Lakes have many sources of both inflow and outflow. The main sources of inflow are:

- streams
- precipitation onto the lake surface
- ground water

The main sources of out flow from a lake are:

- streams
- evaporation from the lake surface
- groundwater

Various configurations of streams are possible:

- *Stream or streams flow in, stream flows out* (Figure 4-3A). Usually only one stream flows out. In this kind of configuration the stream can be viewed as simply a fat place along a river. This is the most common kind of lake.
- *Stream or streams flow in, no stream flows out* (Figure 4-3B). The loss is made up by groundwater outflow and/or net evaporation. This is common in arid and semiarid regions; the streams are usually ephemeral. Lakes like this tend to be playa lakes. The Great Salt Lake is a good (and large) example.
- *Stream flows out, no stream flows in* (Figure 4-3C). The lake is fed by groundwater, usually not just by precipitation.
- *No stream flows in, no stream flows out* (Figure 4-3D). The lake is both fed and drained by groundwater movement. Lakes like this are common in glaciated areas underlain by a thick mantle of highly permeable glacial sediment, as on Cape Cod.

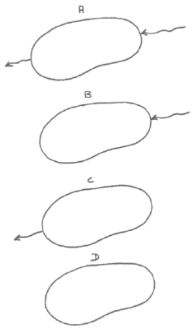


Figure 6-3. Configurations of streams flowing into and out of lakes.

6.5: The Vertical Temperature Structure of Bath Water

The vertical temperature structure of lakes is a surprisingly intricate matter, and one that's more important than would seem at first thought. Remember that the density of water varies significantly with temperature, so if there's exchange of heat between the lake and its environment, giving rise to temperature differences, buoyancy forces will cause motions of the lake water until a gravitationally stable arrangement is reached, with less dense water overlying more dense water. What makes these processes interestingly complex is the strange density maximum at 4°C, noted back in Chapter 1.

In one sense, the water in your bathtub is a good model of a small lake. Think about the temperature and density structure of your bath. First of all, here are the two obvious scenarios for how you draw your bath:

- First you run water that's too cold, and then, to make it warmer, you run water that's too hot (or vice versa). Then you swash it all around for a few seconds to wind up with a nice homogeneous well mixed bath at the desired temperature.
- You run it in all at the same temperature and then simply turn the tap off.

In both of these scenarios, if you measured the temperature with a little thermometer you'd find very close to the same temperature at all points. If you stuck in your little cream whipper coated with slowly soluble dye and beat the water, you'd generate a lot of turbulence and also set up a complicated circulation pattern through the entire bath, and the dye would rapidly become mixed uniformly throughout the bath.

You might think that the uniformly mixed situation described above is the normal situation in lakes, but it isn't. To illustrate, here's another bathtub scenario. First you put in hot water and then cold water, or vice versa, and then the phone rings and you don't get a chance to mix them. When you come back to the bathtub you measure the vertical temperature profile. You'd find that the temperature is somewhat lower near the bottom than near the water surface (Figure 4-4). This temperature difference from bottom to top depends partly on the original temperature difference between the cold water and the hot water you used, but just as importantly on how fast you filled the tub. The stronger the water jet hitting the water surface, the greater the mixing between cold water and hot water.

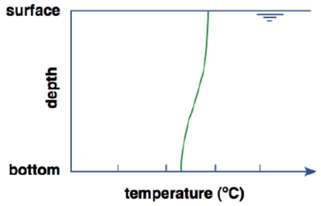


Figure by MIT OCW.

Figure 6-4. Temperature profile in your bathtub.

Why the temperature difference from bottom to top? Because *cold water is denser than the hot water* (provided that its temperature is above 4°C), so it tends to stay at the bottom (if you ran the cold water first) or find its way to the bottom (if you ran the hot water first). The only reason you don't end up with a perfectly stratified situation, with all the original cold water as a lower layer and all the original hot water as an upper layer, is the mixing occasioned by the jet from the faucet as the jet impinges on the water surface during filling.

With a little extra effort you can indeed set up an almost perfectly two-layer temperature-stratified bath. You get the best results by filling first with cold water and then with hot water, and by letting the water jet fall on a spreading plate that floats on the surface (Figure 6-5). The purpose of the spreading plate is to break the momentum of the downward jet, so that the warm water flows slowly off the plate horizontally and has no chance to mix with the cold layer below. That leads to two nearly homogeneous layers, cold below and hot above, separated by *a thin zone of very sharp temperature change* (effectively a temperature discontinuity), called a ***thermocline*** (Figure 6-6). In real lakes, the thermocline is also called the ***metalimnion***; the zone above is called the ***epilimnion***, and the zone below is called the ***hypolimnion***.

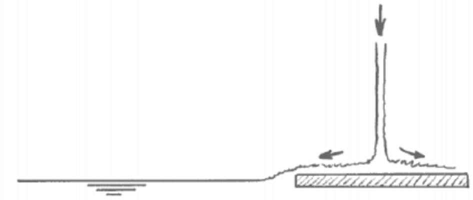


Figure 6-5. Drawing a two-layer bath.

Now if you put in your little dye-coated beater, what would you see? If you put the beater in either the upper (hot) or the lower (cold) layer, you'd produce a lot of turbulence in that layer, but the other layer would be little affected: each layer acts largely independently of the other (Figure 6-7). If, however, you put the beater at the interface, you'd mix some of the hot and cold together to produce water with intermediate density and temperature, and that intermediate water would drift out laterally to form a new and easily recognizable intermediate layer (Figure 6-8).

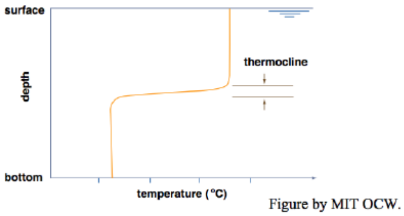


Figure 6-6. Your two-layer bath.

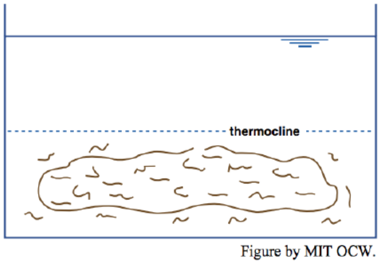


Figure 6-7. Near-independence of flow in the two layers.

What's going on here is that in arranging a cold layer below and a hot layer above you've produced a gravitationally stable stratification, and it takes *work* (literally, in the sense of physics—force times distance) to mix the waters of the two layers. The motions produced when the beater was in just one layer weren't enough to disrupt the stratification. You'd need a more powerful mixer to break down the stable density stratification.

Here are a few other instructive things you could do in your bathtub to gain some insights into the thermal regime of lakes:

• Line your bathtub the best you can with a coating of thermal insulation, including a floating insulating lid. In this way you eliminate all possibility of changing the temperature and therefore the density of the bath water by exchange of heat with the environment. Start with a thermally stratified bath, as above, and then come back hours later and re-measure the vertical temperature profile. You'd find that *the thermocline has thickened* (Figure 6-9). Why? By conduction of heat from the hot layer to the cold layer by molecular motions. This effect, called thermal diffusion, is slow but inexorable. In this situation of total isolation, *the water in the bath would come to have a uniform temperatur*e.

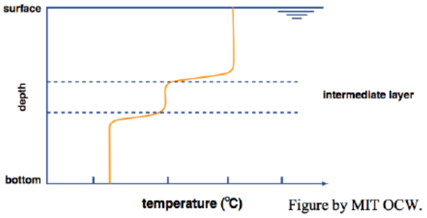


Figure 6-8. Beating the interfacial zone to make an intermediate layer.

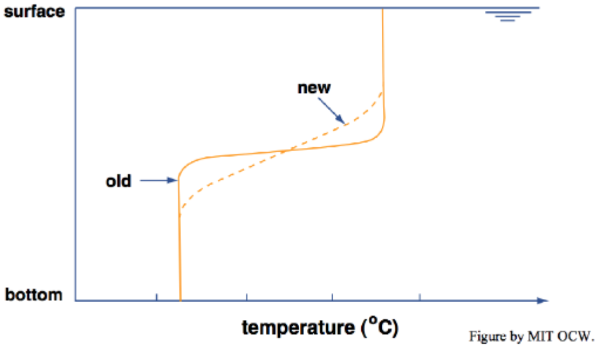


Figure 6-9. Thickening of the thermocline by molecular diffusion.

BACKGROUND: DIFFUSION

Diffusion is an important process in a great many natural environments. ***Diffusion*** is *a flow of a material or property in some medium as a consequence of the existence of a spatial gradient in the concentration of that material or property in the presence of random motions of the material of which the medium is composed.*

That's a rather long and abstract definition. Here are some examples of diffusion:

- The warming of the handle of a metal frying pan when it's set on the burner.
- The spread of a cooking odor throughout the house, even though there's no organized air circulation in the house.
- the widening of a plume of black smoke as it rises from the top of the chimney.

Here's a thought experiment (one you couldn't actually do) to make the concept of diffusion more concrete. Suppose that you could place a vertical airtight partition down the middle of a sealed room, to make a "left" side and a "right" side, and then color all of the air molecules on the right side of the partition green and all of the air molecules on the left side of the partition red. Keep in mind that the air molecules are zipping continuously this way and that, colliding now and then with the walls of the room and with each other, because of their thermal energy.

Now magically remove the partition, instantaneously, and think about what happens. There's a balanced exchange of molecules across the vertical plane where the partition was located: at any given time, just about equal numbers of molecules are passing across the plane in the two opposite directions. But at the beginning, all of the molecules passing from right to left are red molecules, and all of the molecules passing from left to right are green molecules. We say that there is a net flow of red molecules from left to right and a net flow of green molecules from right to left. That's the essence of diffusive transport. Eventually, of course, the concentrations of both red molecules and green molecules become evened out everywhere throughout the room. After that, even though molecules are still zipping around, there's no net transport, because there's no longer any spatial gradient in concentrations of red or green molecules.

The random motions of the medium can involve the individual atoms or molecules, or random motions of turbulent eddies in a fluid. The example of the pan on the stovetop involves the vibrations of the atoms of the metal handle. The temperature of the pan is higher than that of the handle, meaning that the speed of vibration of the atoms is greater. The flow of heat down the handle is just a manifestation of the tendency for the speed of vibration of the molecules to even out, by the interaction of adjacent atoms. The other two examples, however, involve the motions of turbulent eddies in a fluid—although in the case of the cooking odor the spread could be entirely by molecular diffusion, as in the case of the red and green molecules in your partitioned room, if the air in the house is very still. In a fluid medium, turbulent diffusion is much more effective than molecular diffusion, if the medium is turbulent.

In the example of heat transport by diffusion, what's being diffused is a property, the temperature of the metal. In the case of diffusive transport of the cooking odor, it's the concentration of the gas or colloid that we sense as the odor. In the case of the smoke coming out of the smokestack, it's the concentration of smoke particles.

The rate of diffusive flow of the property or material, called the ***diffusive flux***, is proportional to the spatial gradient of the concentration of that property or material. The proportionality constant is called the ***diffusivity***, or the ***diffusion coefficient***. In some cases it can be derived from theory; in many other situations, however, especially when fluid turbulence is involved, it just has to be a measured empirical value.

- Leave the insulation around the walls of the tub in place leave the top lid off. Draw a hot bath, and then keep track of the vertical temperature profile as a function of time. You'd find that the water temperature stays vertically uniform as it cools, and that the water cools more and more slowly as the temperature approaches the ambient temperature (Figure 6-10).
- Draw a cold bath, and do the same. The result would be spectacularly different! As the surface water is warmed, a cold layer is maintained below, and this stratification persists until all of the bath is at the same temperature (Figure 6-11). You'd also find that the time it takes to develop a uniform bath is longer than in the previous case.

6.6: The Thermal Structure of Real Lakes

Now you can apply your knowledge of the thermal structure of bathtubs to real lakes. Think about a lake in a temperate region, with good contrast between summer warmth and winter cold, as in New England. Also, assume that the lake isn't ridiculously shallow. Start with a hypothetical all-warm lake (but see the end of this section, after on full annual cycle).

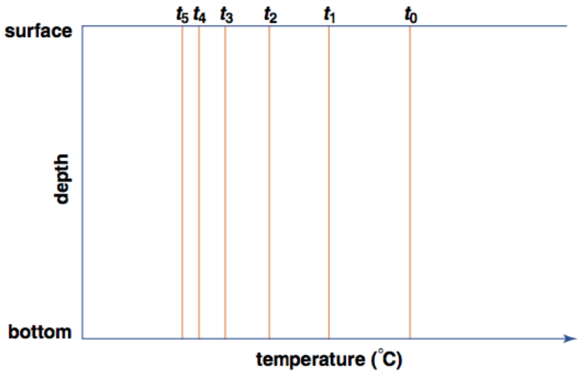


Figure 6-10. Time series of temperature profiles during cooling of your hot bath from the surface.

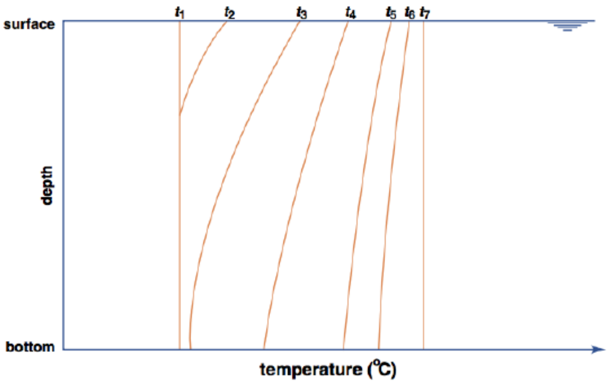


Figure 6-11. Time series of temperature profiles during warming of your cold bath from the surface.

Late fall: The surface waters are cool, with a uniform temperature profile and free vertical mixing (Figure 6-12A).

Early winter: The picture above holds until the surface water is about 4°C. Cooling past that point produces less dense water. Now stable stratification develops, because colder water remains at the surface. Vertical convection turns off (Figure 6-12B).

Middle winter: The lake reaches 0°C at the surface and becomes covered with ice. There's gradual cooling downward by conduction. There's no mixing, because of the stable stratification. The bottom water stays at nearly 4°C, unless the lake is shallow and/or the winter is very cold (Figure 6-12C).

Early Spring: The ice melts, the surface warms, and convective instability develops, because below 4°C the warmer water is more dense than the colder water (Figure 6- 12D).

Middle Spring: The surface water reaches 4°C, and the whole lake mixes, because 4°C water at the surface is denser than the water at any level and convection operates through the entire depth. This is called the ***spring overturning*** (Figure 6-12E).

Late Spring to Early Summer: The surface warm mixed layer lies above the thermocline. The cold water below the thermocline is gradually warmed by conduction (Figure 6-12F).

Early Fall: There's cooling at the surface, and mixing downward to the level at which the temperature equals the surface temperature (Figure 6-12G).

Late Fall: The surface develops the same temperature as the bottom, now greater than 4°C by conduction. Mixing is complete. This is called the ***fall overturning*** (Figure 6-12H).

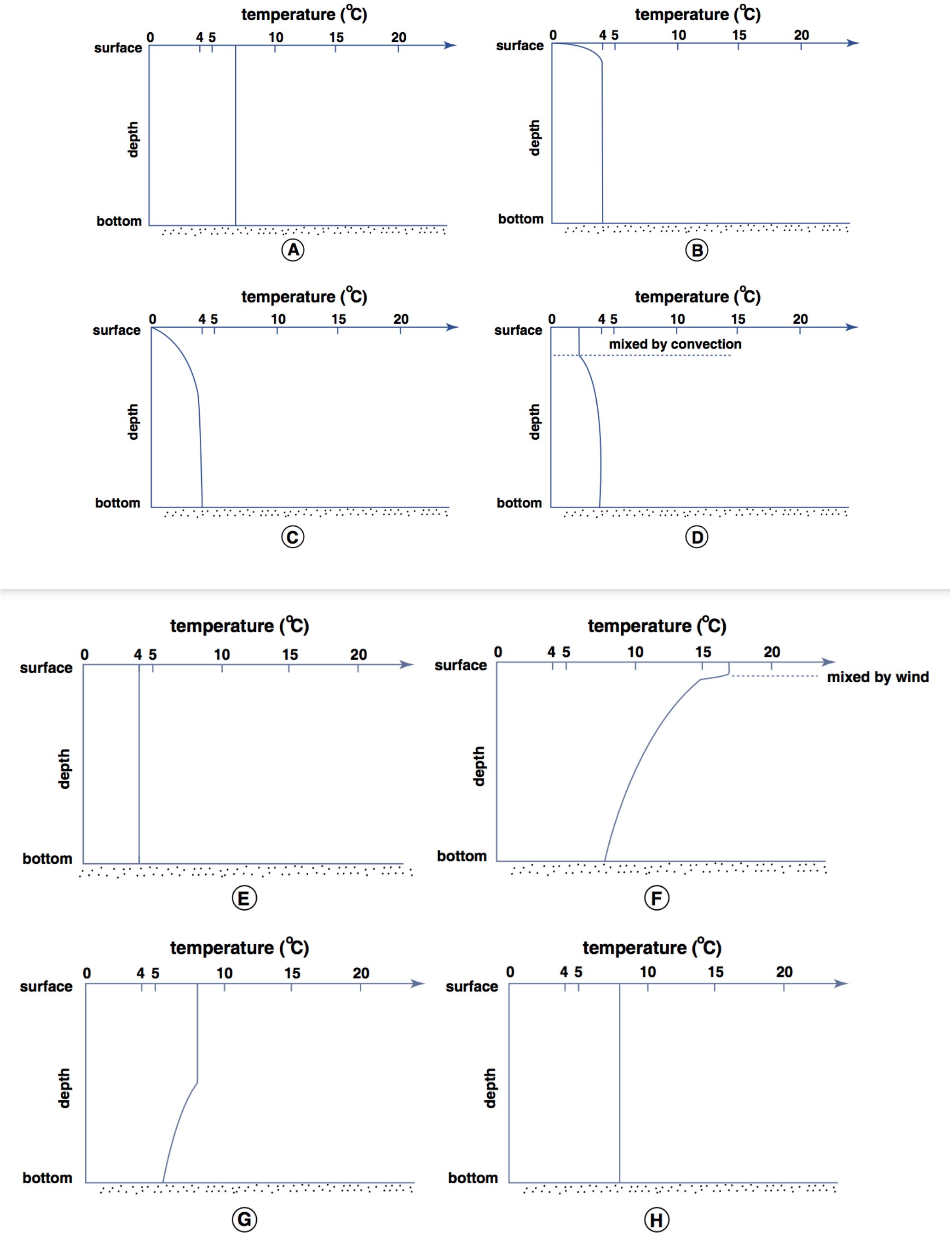


Figure by MIT OCW.

Figure 6-12. Temperature profiles in a temperate-region lake.

6.7: Classification of Lakes by Thermal Regime

Lakes can be classified on the basis of their thermal behavior in the course of a year. The hypothetical lake considered in the last section is just one possibility for the annual changes in thermal regime.

First we need to distinguish between holomictic lakes and meromictic lakes.

Holomictic lakes are of the same salinity (essentially zero) throughout, so that if the temperature becomes the same everywhere in the lake, the density also has to be the same everywhere in the lake, and the water of the lake can circulate freely. In such holomictic lakes, only the temperature, not the salinity, controls the density.

Meromictic lakes, on the other hand, can't circulate freely even if the temperature becomes the same everywhere, because of greater salinity in the bottom waters. Any higher-salinity water produced within the lake or introduced from outside the lake will find its density level and then spread out horizontally to form a distinctive layer, and the resulting density stratification prevents or at least inhibits vertical circulation.

Holomictic lakes can be further subdivided into dimictic lakes, warm monomictic lakes, and cold monomictic lakes, depending on the history of surface-water temperature in the course of the year, which in turn is a function of the climate.

Dimictic lakes are those in temperate regions where the annual fluctuations in air temperature are such that the surface water temperature of the lake is above 4°C for a part of the year and below 4°C for part of the year. Dimictic lakes *circulate freely by overturning twice a year*. The lake described in Section 7 is a dimictic lake.

Warm monomictic lakes are those in tropical regions where the air temperature never gets very low during the year. The surface water temperature of thelakestaysabove4°C,so*there'sfreeverticalcirculationonlyatthetimeof coldest surface water temperature*.

Cold monomictic lakes are those in polar regions where the air temperature never gets very high during the year. The surface water temperature of the lake staysbelow4°C,so*there'sfreecirculationonlyatthetimeofwarmestsurface water temperature*.

The basic reason why lakes show such diverse circulation behavior is that maximum in water density at 4°C.

6.8: The Lifetime of Lakes

How long do lakes last? There’s an inevitable tendency for lakes to disappear, by sedimentation to the point where the lake basin is occupied by sediment rather than water, so the water surface is replaced by land surface, perhaps with a stream or river flowing across that land surface.

One cause for the disappearance of lakes is the mechanical deposition of inorganic mineral sediment (gravel, sand, and mud) carried in mainly by *streams* and rivers feeding the lake, or to a much lesser extent by wind or by mass movements of soil down slopes to the edge of the lake. The sediment load of a stream feeding the lake tends to become segregated into coarse material deposited near the mouth of the stream to form a delta, and fine material deposited from suspension widely over the lake bottom (Figure 6-13).

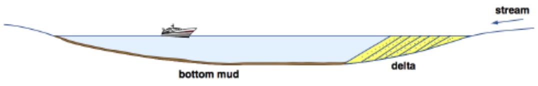


Figure by MIT OCW.

Figure 6-13. Sediment deposition in lakes.

The other important cause for the disappearance of lakes is biogenic sedimentation of non-decomposed organic matter, mainly plant material, on the lake bottom. This process is called ***eutrophication***. The important factor here is *availability of oxygen in the bottom waters*.

• If the biogenic productivity of the lake (mainly at the surface) is relatively low and the vertical mixing of water in the lake is relatively high, then the bottom water is well oxygenated, and most or all of the organic matter that settles to the bottom is decomposed before it has a chance to build up an organic bottom- sediment deposit.

• On the other hand, if the biogenic productivity is relatively high and the vertical mixing of water in the lake is relatively low, then the bottom water is deficient in dissolvedoxygen, and a deposit of organic sediment builds up on the lake bottom.

Because the vertical circulation is controlled almost entirely by physical effects, as discussed in earlier sections, the critical factor in eutrophication is biogenic productivity, which in turn is a function of supply of dissolved nutrients in the water entering the lake. In most lakes, phosphorus is the limiting nutrient. One of man’s strongest effects on lakes (in urban areas, suburban areas, and rural agricultural areas) is the introduction of much greater concentrations of nutrients, especially phosphorus, leading to greatly accelerated eutrophication.

CHAPTER OVERVIEW

7: Glaciers

- 7.1: Introduction to Glaciers
- 7.2: Classification of Glaciers
- 7.3: Distribution of Glaciers
- 7.4: Glacier Ice
- 7.5: The Budget of Glaciers
- 7.6: Movement of Glaciers
- 7.7: Deformation of Ice
- 7.8: The Thermal Structure of Glaciers
- 7.9: Glacial Meltwater
- 7.10: Glacial Erosion
- 7.11: Glacial Sediment Transport
- 7.12: Glacial Deposits

7.1: Introduction to Glaciers

Before the era of universal air travel, which commenced less than half a century ago, few of the world's population had seen a glacier. I suspect that majority of class members in this course have seen a glacier—if not close up, then out of a jetliner window. In the Canadian Rockies, you can drive to within almost a stone's throw of the terminus of the Athabasca Glacier, a classic active valley glacier. In many other parts of the world, valley glaciers are accessible to even casual day hikers. The great ice sheets of the world, in Antarctica and Greenland, remain much less accessible.

In the broad context of geologic history, the Earth is in an "icehouse" time, with recurrent major ice-sheet advances across the Northern Hemisphere continents. (There have been several other such icehouse periods in Earth history, separated by long intervals of ice-free times, called "greenhouse" periods, with no evidence of glaciation.) The Earth has only recently emerged from the latest episode of continental glaciation. Does it surprise you to learn that a mere twenty thousand years ago the Boston area was beneath a mile of glacier ice moving slowly southward toward its terminus south of what is now the south coast of New England?

The Earth is in many senses a glacial planet:

- 10% of the Earth is covered with glacier ice (about 15 million square kilometers).
- About 40% of the Northern Hemisphere in winter is covered with solid water at any given time (land and sea).
- 75% of the Earth's fresh water is in glaciers.
- Surficial deposits by glaciers cover a large percentage of the Earth's land surface.
- Glaciers have a profound effect on the Earth's climate (as well as being in turn controlled by climate).

And glaciers are not without practical importance:

- Role of ice sheets on climate
- Control on sea level
- Potential source of fresh water

Disciplines: in the US, people who deal with glaciers or their products are usually allied to *geology*; in other English-speaking countries, usually to *geography*. Disciplines relevant to this chapter:

- glaciology
- glacial hydrology
- glacial geology
- geomorphology
- sedimentology
- stratigraphy

What is a ***glacier***?:

a body of ice and recrystallized snow (plus refrozen meltwater), on land (or, if floating, then connected to land) and moving by deformation under its own weight.

Why are there glaciers?

A necessary and sufficient condition is an excess of snowfall over snowmelt at some locality for a long enough time to build up ice thick enough for it to flow under its own weight.

7.2: Classification of Glaciers

One good way of classifying glaciers is by the extent to which their shape and movement are affected by the underlying bedrock topography.

Glaciers not strongly constrained by underlying topography

Ice sheet: Superimposed on underlying topography; largely or entirely submerges that topography; ice flow reflects largely the size and shape of the glacier, less the shape of the ground. Ice sheets smaller than about 50,000 km^2 in area are called ***ice caps***.

Ice dome: About symmetrically over land area involved; the top of the dome may be over bedrock highs or bedrock lows. Highest: 4300 m, Antarctica. The convex profile is a reflection mainly of flow mechanics, including the nature of the bottom roughness.

Outlet glacier: A stream of ice that extends beyond an ice dome and drains it. Outlet glaciers are often very large: several hundred kilometers long and tens of kilometers wide. Commonly they are commonly associated both with large ice sheets and with smaller ice caps. Most of the Greenland ice sheet and about three-quarters of the Antarctic ice sheet is drained in this way.

Ice shelf: A floating ice sheet that deforms under its own weight. The slab of ice is squeezed between the atmosphere and the ocean. The ice shelf has to be anchored at several points. In Antarctica, ice shelves constitute about 7% of ice area. The ice shelf may be fed partly or not at all by land glaciers. Accumulation: snow on the surface; land glaciers; bottom freezing. Ablation: melting (at the top or the bottom); calving.

Glaciers strongly constrained by topography

Ice field*:* An approximately level area of ice, not an ice cap because not domelike; the flow reflects the underlying bedrock topography (ice fields are usually nestled in among mountains). Size: a few square kilometers to very large. Ice fields grade over into small ice caps. Requirement for existence of an ice field: high and overall gentle topography.

Valley glacier*:* A glacier flowing in a rock valley and surrounded by rock walls (therefore long and linear). Fed from an ice field or a cirque (see below). Usually 10–30 km long, but up to 100 km. The terminus can be on land or in the sea (by calving). A valley glacier can also disgorge from mountains and spread out as a flat mass in piedmont areas to form a what is called a ***piedmont glacier***. A good example of a piedmont glacier: the Malaspina Glacier, in Alaska, 40 km across. Valley glaciers are usually vigorous glaciers. And because they are typically present at low as well as high latitudes, they are among the most accessible of glaciers.

Cirque glacier*:* A small ice mass, fairly wide relative to its length, occupying a bedrock hollow or basin, usually on a mountain slope.

7.3: Distribution of Glaciers

The total area of the Earth's surface covered by glaciers is 14.9 million km^2

- Antarctic ice sheet: 12.5 million km^2
- Greenland ice sheet: 1.7 million km^2
- All the rest: 700,000 km^2 (many ice caps, mostly less than about 10,000 km2; many thousands of small glaciers, mostly valley glaciers)

Aside from the Greenland ice sheet, most of the larger glaciers in the Northern Hemisphere are mostly on Iceland and the Arctic Islands of Canada, because of the distribution of land and sea.

It's a lot more difficult to get the *volumes* of ice in glaciers than to get the *areas* covered by glaciers. Things have gotten better in this respect, though, because of use of echo sounding, similar to oceanographic depth sounding. But one still has to go out there to the glacier to do it.

- *Antarctica*: 21.5 million km^3 (sea-level equivalent about 60 m) *Greenland*: 2.4 million km^3 (sea-level equivalent about 5 m)
- *All other*: 180,000 km^3 (largest ice cap: Vatnajökull, Iceland, 3100 km^3, and fairly accessible).

7.4: Glacier Ice

The transition of new snow to glacier ice is similar to the deposition, diagenesis, and metamorphism of a sediment to form a metamorphic rock. (And in a real sense, glacier ice *is* a metamorphic rock.)

New dry snow has low bulk density and high porosity, and enormous internal surface area. Crushing, compaction, and exchange of water between flakes and air at low temperatures by sublimation and deposition, and at temperatures near melting by melting, refreezing, evaporation, and condensation, tend to round grains, producing the nearly spherical grains of granular snow. Except in the coldest environments, this takes a few days to a few weeks. In the process, the snow settles and gets more compact, and small grains disappear at the expense of larger ones. The recrystallized granular snow, still friable and porous, is called ***firn*** when it is more than a year old. The firn is then converted to glacier ice as compaction due to increasing weight of overlying younger snow continues. Air is forced out, decreasing the intergranular space. Recrystallization proceeds by solution at points of contact and deposition in the interspaces.

When the permeability to air has reached almost zero, the firn has become glacier ice. This change occurs in a relatively narrow density range around 0.82–0.84 g/cm^3, at depths of greater than 30 m. Crystal size has increased to about 1 cm. Firn is transformed into glacier ice in one year to a few hundred years, depending on temperature and rate of accumulation. Then, depending on both temperature and bottom slope, something like 50 m of ice is needed for the ice to flow plastically under its own weight (there's more detail on flow of glacier ice in a later section). The ultimate density of glacier ice approaches 0.90 g/cm^3, a bit short of the density of 0.917 g/cm^3 of pure ice because of the inevitable residual presence of little air bubbles.

7.5: The Budget of Glaciers

Economy vs. Regimen

First off, the term ***accumulation*** applies to all of the ways *glacier ice mass is added to a glacier*, and the term ***ablation*** applies to all of the ways *glacier ice is removed from a glacier*.

The term ***economy*** refers to the relative magnitude of accumulation and ablation. A glacier with accumulation greater than ablation over some period of time (much longer than just a single year) is said to have a *positive economy*, and gains ice volume with time. A glacier with accumulation less than ablation is said to have a *negative economy*, and loses ice volume with time. A glacier with a positive economy not only thickens but also extends its terminus in the downglacier direction. A glacier with a negative economy becomes thinner, and the terminus retreats upglacier.

The term ***regimen*** refers to the absolute values or magnitudes of accumulation and ablation, irrespective of their balance. A glacier with large values of accumulation and ablation is is said to have an *active regimen*, whereas a glacier with small values of both accumulation and ablation is said to have an *inactive regimen*. So there are four different combinations of economy and regimen.

Glacier Bookkeeping

Ice accumulates on a glacier by a variety of processes: snow (plus sleet and other solid precipitation), rain that gets frozen onto or into glacier, rime, frost, and avalanches. Snow is by far the most important of these.

Modes of ablation are melting (plus runoff and evaporation), sublimation, deflation by wind, and calving. The important ones are *melting* and, for glaciers that terminate in water, *calving*.

Occupy an observation point that's fixed relative to the bedrock base or walls of the glacier, and consider any point on a the surface of the glacier.
Look at the vertical thickness of water (i.e., the height equivalent to solid ice) added to or lost at a point.

In glacier bookkeeping, an important concept is the ***balance year***: *the time between two successive dates with minimum mass of glacier ice at a given location* on the glacier. Note: balance years are not necessarily 365 days, because the time of minimum ice mass depends upon the weather throughout the balance year, and the balance year can differ from point to point on the glacier.

Imagine plotting a graph showing the vertical ice-equivalent height added or subtracted as a function of time for a whole balance year (Figure 5-1). (Does the concept of "vertical ice-equivalent height" make sense to you? You have to convert accumulation and ablation into ice thickness.)

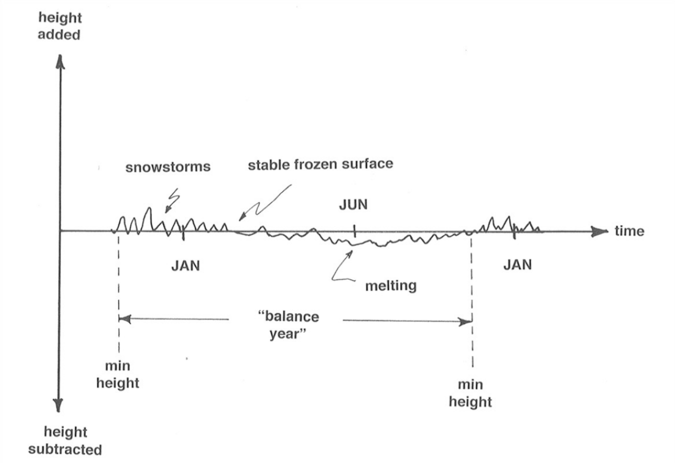


Figure 7-1. Keeping track of accumulation and ablation as a function of time on a glacier, for one balance year.

Now split the graph in Figure 1 into positive and negative parts (i.e., monotonically nonincreasing and monotonically nondecreasing), and then integrate (that is, keep track of a "running sum") the two parts separately to get an *accumulation curve* and an *ablation curve* (Figure 7-2). The difference between the accumulation curve and the ablation curve is called the ***mass-balance curve***. This curve is sort of the *net change as you go along* through the balance year. Don't worry that it's mostly above the horizontal axis; that's only because we chose to work on the basis of minimum-height times as the starting and ending points of the balance year. The difference in elevation of the mass-balance curve and the horizontal axis at the end of the balance year is the *net balance over the balance year*, positive at this particular place and in this particular year.

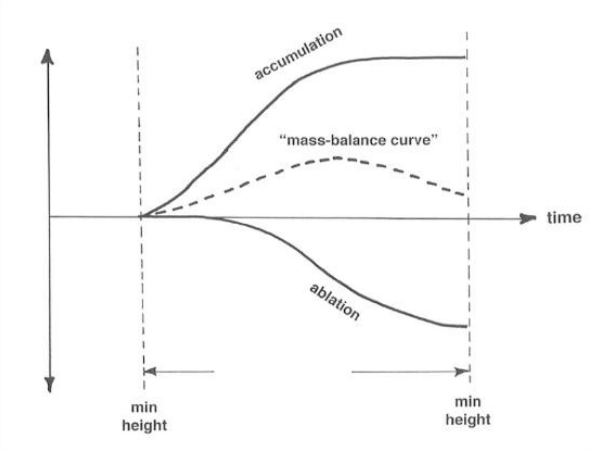


Figure 7-2. Vertical ice-equivalent height added to or subtracted from the surface of a glacier during a single balance year.

BACKGROUND: INTEGRATION AND INTEGRALS

This is not the place for a full-scale calculus lesson, but for those of you who have not been exposed to the beauties of differential and integral calculus, here's the concept of integration and integrals.

First, a few words about mathematical functions. You can think of a function as a magic box: you put a number into the box, and out comes a value that's associated with that number. You're probably more used to thinking of a function in strictly mathematical terms, in the form of an equation like $y = x^2$: plug in a value for x, and the function gives you the value of y.

In mathematics, the process of integration involves a summation of the values of some function, as the value of the function varies over its range. If the function is defined only over a range of discrete values of the input variable (1, 2, 3, ... , let's say), then the summation is simple and straightforward. All of you have done that. But how about when the function varies continuously over all values in the input variable, as, for example, the function $y = x^2$ above? There's a big conceptual leap involved here. The mathematical process by which the value of the function is summed continuously is called ***integration***, and the value of the sum thus obtained is called an ***integral***.

Obviously this is not the place either to develop the concept with mathematical rigor or to prescribe how to actually do integration. Here's a simple example to show the results. Figure 7-3 shows the results of integrating the function $y=x^2$ from $x=0$ to $x=2$. The curve in Figure 7-3 is that of the function $y=x^2$, and the shaded area under the curve between $x = 0$ and $x = 2$ represents the value of the integral, which turns out to be exactly 2/3. Your common sense tells you it has to be a bit less than 1, which is what it would be if the function were the straight line $y = x/2$.

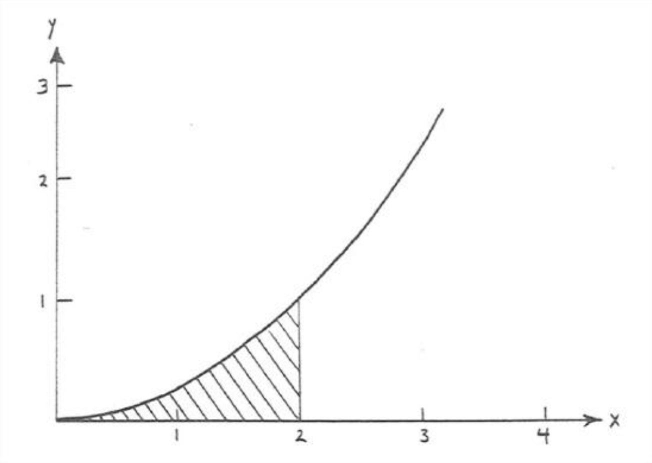


Figure 7-3. Integrating the function $y = x^2/4$ from $x = 0$ to $x = 2$.

If somebody challenged you to find such an integral, without actually knowing how to do it mathematically, your clever mind might hit on the strategy of dividing up the *x* axis into a great many little segments and drawing rectangles with the value of *y* at the their tops, then figuring out the areas of all of the litle rectangles, and adding up all of those areas (Figure 7-4). That would be a good approximation to the integral you want to find. If you're even sharper than that in such matters, you could imagine making the rectangles skinnier and skinnier, without limit. At that point you would be very close to the heart of the mathematical concept of integration!

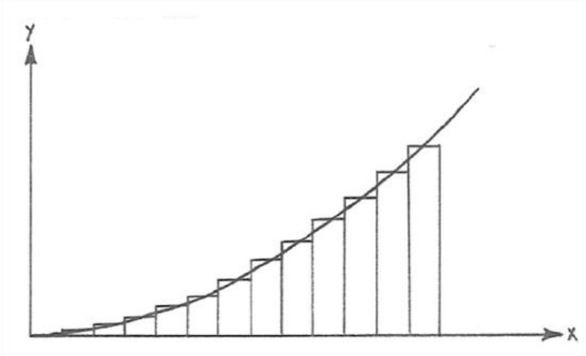


Figure 7-4. How you can evaluate the integral by approximating it with a lot of little rectangles that follow the curve.

The curves in Figures 7-1 and 7-2 differ from year to year at a point. Also, obviously they differ greatly from point to point: at points high on the glacier there's a strongly positive net balance, maybe with no ablation all year, whereas low on the glacier there's likely to be a strongly negative balance, with not much accumulation (presumably there's always some) but lots of ablation.

Figure 7-5 is a plot of ice-equivalent height gained by accumulation and lost by ablation as a function of elevation on a given glacier. In the right-hand part of the graph, high on the glacier, accumulation is greater than ablation, whereas in the left-hand part of the graph, low on the glacier, ablation is greater than accumulation. At a certain elevation the two curves intersect; this elevation is called the ***equilibrium line***. Upglacier there's still some of last year's solid water (firn and glacier ice) left when new accumulation begins at the start of the new budget year; downglacier, all of last year's snow melted before new accumulation starts. (Don't worry about the areas under the curves; they don't have to be equal, because they depend on the distribution of glacier area with elevation.)

Another thing you could do is integrate the accumulation and ablation curves in Figure 7-2 over the entire surface of the glacier to get the total accumulation and ablation. The result would look something like Figure 5-6. A little thought should convince you that glacier mass is unchanged through a given balance year if and only if Area 1 equals Area 2. If $A_1 < A_2$, the balance for that year is negative, and the glacier loses mass; if $A_1 > A_2$, the balance for that year is positive, and the glacier gains mass.

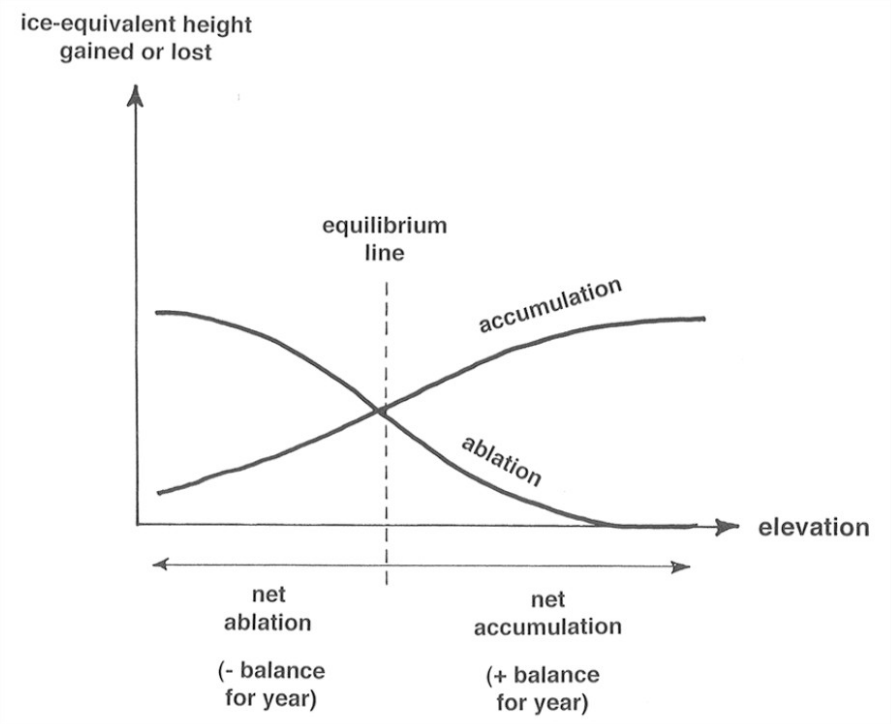


Figure 7-1 separately to obtain an accumulation curve and an ablation curve.

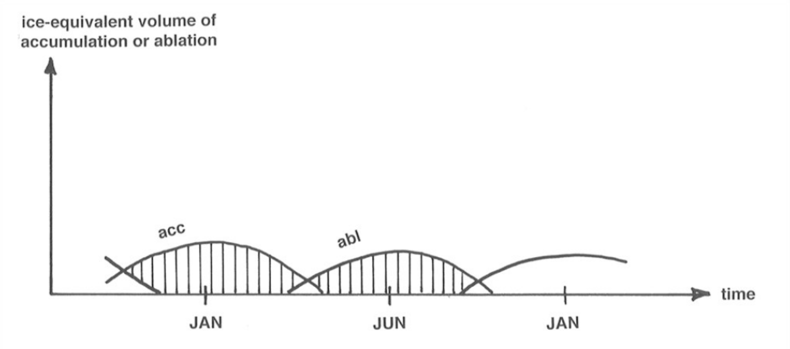


Figure 7-6. Ice-equivalent height gained by accumulation and lost by ablation as a function of elevation on a glacier.

It is very difficult to obtain accumulation and ablation curves at points on a glacier. It's easiest to measure (i.e., estimate) the net balance at many points in both the accumulation area and the ablation area and then integrate over the area of the glacier to find the net gain or loss of glacier mass for the given balance year. How? Dig pits down to last year's snow in the accumulation area and then measure ablation at fixed stakes in the ablation area. (There's some distortion involved, because of glacier movement.) Keeping track of accumulation and ablation as a function of time, as in Figure 7-4, is very time-consuming.

ADVANCED TOPIC: THE SITUATION AROUND THE EQUILIBRIUM LINE

Features and relationships near the glacier surface around the equilibrium line are more complicated than I've indicated so far. To see this, imagine a survey from head to foot of the glacier at the end of summer melting season, to look at near-surface materials (Figure 7-7). Here are some comments on the various zones labeled in Figure 7-7.

- **dry snow zone:** no melting even in summer.
- **percolation zone:** some surface melting in summer. Water percolates down into snow at 0°C and refreezes, thereby warming the surrounding snow. Two characteristic forms of ice are deposited: ***ice layers*** *or* ***ice lenses***, formed when water spreads out at some relatively impermeable horizon, and ***ice pipes*** or ***ice glands,*** formed by freezing of vertical water channels. The depth of saturation increases downglacier to the saturation line—the point where by the end of summer all of the snow deposited since the end of the previous summer has been saturated.

- **soaked zone:** by the end of the summer, all the current year's snow has been saturated and raised to 0°C. Some meltwater also percolates into deeper layers.
- **superimposed-ice zone:** in the lower part of the soaked zone, at lower elevations, there's so much meltwater that ice layers and patches among the firn merge to form a continuous mass of ice, called ***superimposed ice***. Summer melting exposes this ice. The exposed part that remains at the end of the melting season is called the ***superimposed-ice zone***. The ***firn line*** (the boundary between the current year's snow and the newer ice) is easy to find.
- **ablation zone:** the zone below which all of the firn and superimposed ice from the current year has been melted away, to expose older ice.

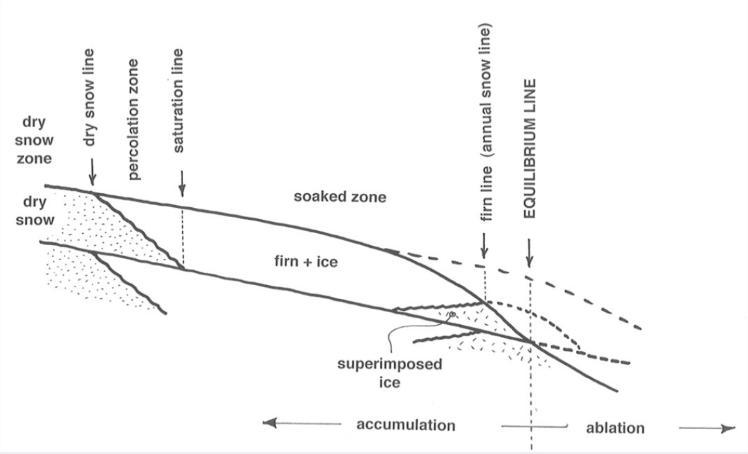


Figure 7-7. Schematic streamwise vertical cross section through a glacier in the vicinity of the equilibrium line.

This is all highly idealized. Most glaciers don't show all the zones, or at least not every season. The dry snow zone may be absent; melting eliminates all the superimposed ice if the net balance is negative; and the relative importance of evaporation and percolation varies. In a series of negative years, one can often see several "firn lines" or "firn edges" if development of superimposed ice is not important (Figure 7-8).

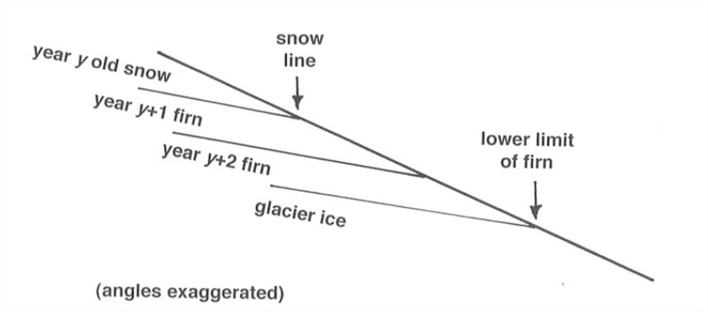


Figure 7-8. Detail of firn lines developed during a series of years, in vertical streamwise cross section.

A concept closely related to the equilibrium line is the ***annual snowline***: the lower limit of current year's snow on the glacier surface. What are the controls on the snowline? Mainly *winter precipitation* and *summer temperature*. The annual snowline is the local manifestation of what's called the ***regional snowline***: a band with a width up to a couple of hundred meters in elevation in which the local snowline lies on regional scale (This band is patchy locally but very consistent regionally.) Figure 7-9 shows the average position of the regional snowline as a function of latitude, as well as the average precipitation.

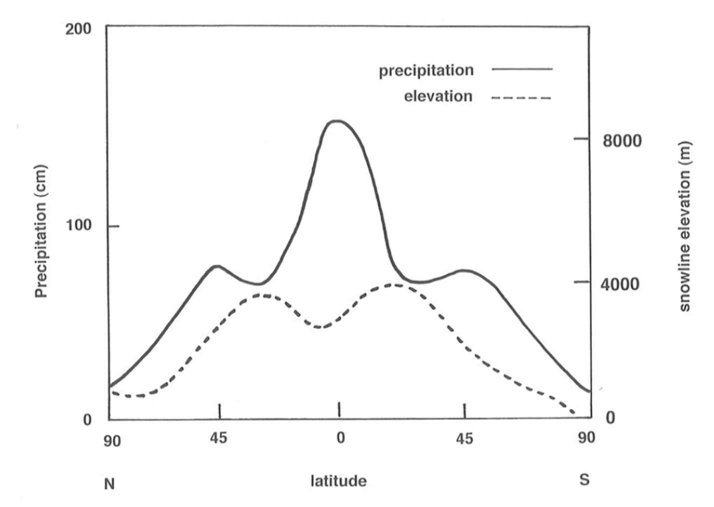


Figure 7-9. Combined plot of precipitation and snowline elevation as a function of latitude.

7.6: Movement of Glaciers

Movement of Glaciers

General Aspects of Glacier Movement

An important consequence of the balance considerations in the previous section is the glacier tends to thicken in its upper part and thin in its lower part, thus increasing its surface slope. The glacier flows under its own weight, passing ice across the equilibrium line to maintain an equilibrium slope. Figure 7-10 is a highly schematic view of how the glacier does this. Figure 7-11, a streamwise vertical cross section through a representative valley glacier, is a less idealized way of looking at this.

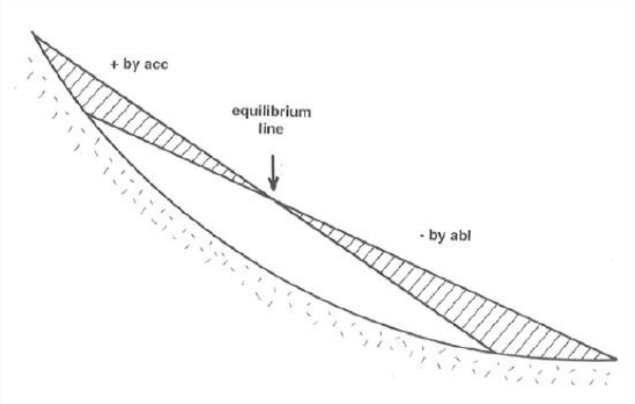


Figure 7-10. Schematic view of gain and loss of ice in a sloping glacier.

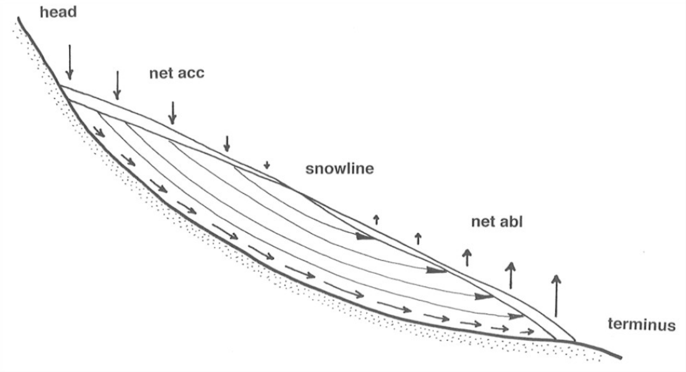


Figure 7-11. Streamwise vertical cross section through a representative valley glacier, showing accumulation, ablation, and flow lines.

The discharge (volume rate of flow) of a glacier is greatest at the snowline. Why? Because if the profile is to stay the same, the annual discharge through any cross section equals the integral of accumulation minus ablation over the whole area of the glacier above that cross section, and it’s at the equilibrium line that this integral is greatest, because upglacier of the equilibrium line, where the accumulation curve and the ablation curve intersect, the accumulation curve is everywhere above the ablation curve; see Figure 7-5. Does glacier flow fastest there too? Maybe or maybe not; it depends on how the thickness of the glacier varies in longitudinal profile.

Figure 7-12, similar to Figure 7-10, shows a vertical cross section through an ice sheet. In a typical ice sheet the equilibrium line is near the margins of the glacier, because accumulation is small over a broad area but ablation is concentrated near the margins. Note the increase in near-bottom velocity toward the margin and the very slow movement of ice deep in the center of the glacier. If you want to find the oldest ice, that’s where to look.

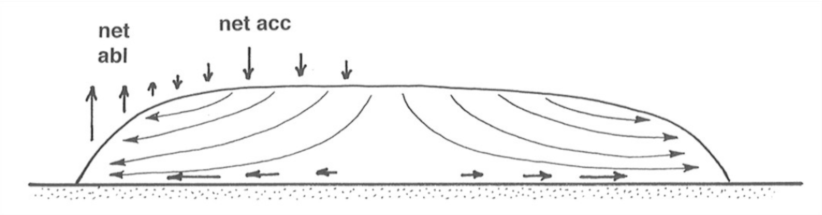


Figure 7-12. Vertical cross section through a representative ice sheet, showing accumulation, ablation, and flow lines.

Velocity of Glaciers

A glacier doesn't move fast enough for the motion to be perceptible directly to the eye, although in some cases speeds are up to several tens of meters per day, so you could almost see it moving, if you had just the right observation post where you could get a good view of the contact between the base the glacier and the underlying bedrock. But usually speeds are of the order of meters per week or meters per month.

Surface velocities on a valley glacier can be measured easily by planting a straight row of stakes across the glacier, surveying their positions relative to fixed points on the bedrock walls, and then coming back at a later time to resurvey the positions to see how far the glacier has moved downvalley. (This kind of thing is not nearly as easy for ice sheets and ice caps, though.) The first systematic work of this kind was in the French Alps in the 1840s.

If you looked at the resurveyed positions of the row of stakes, what kinds of profiles would you see? Figure 7-13, a map view, shows what's generally found: an approximately parabolic or U-shaped velocity distribution. Just as with the free-surface flows we looked at in Chapters 1 and 5, this variation in velocity across the profile is a manifestation of internal shearing deformation.

Figure 7-14 shows that you can resolve the total movement into two components: ***internal deformation,*** and ***basal slip***, or basal sliding. Basal slip varies in importance from *zero* in cold, slow glaciers to *large* in warm, fast glaciers.

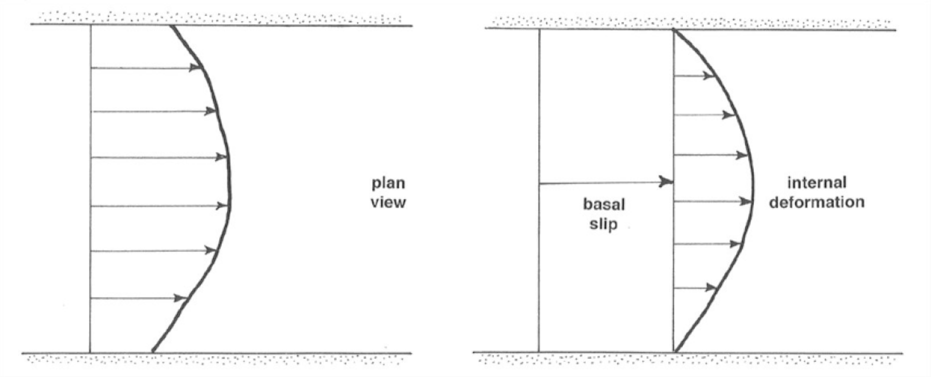


Figure 7-13 (left). Map view of a valley glacier, showing velocity distribution.

Figure 5-14 (right). Map view of a valley glacier, showing resolution of movement into basal slip and internal deformation.

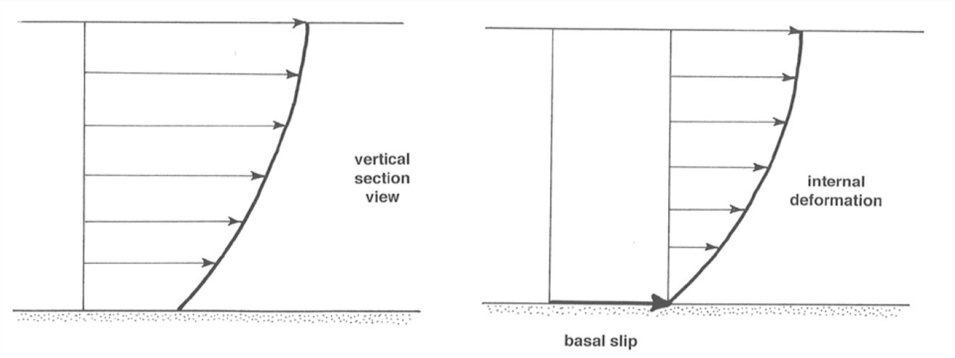


Figure 7-15 (left). Vertical streamwise cross section through a valley glacier, showing velocity distribution.

Figure 7-16 (right). Vertical streamwise cross section through a valley glacier, showing resolution of movement into basal slip and internal deformation.

Figures 7-15 and 7-16 show that the picture of velocity distribution is similar in vertical section. Again the motion can be resolved into a component of basal slip and a component of internal deformation.

Figure 7-17 is an example of an unusually complete result on the velocity of ice at the glacier surface, relative to bedrock, as a function of streamwise position on a valley glacier.

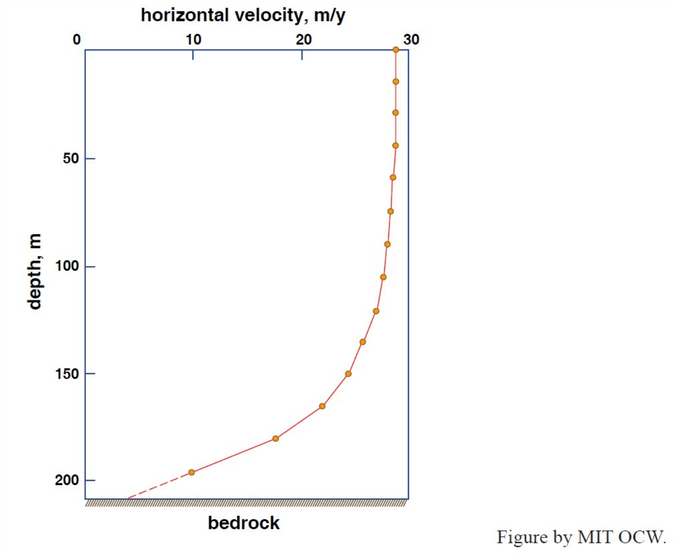


Figure by MIT OCW.

Figure 7-17. The vertical velocity profile in the Athabasca Glacier, Canada. (From Savage and Paterson, 1963.)

One other aspect of glacier velocity is important to understand. Because of accumulation in the upper part of the glacier, what is at the glacier surface at a given time is found at some depth below the glacier surface at a later time. The velocity at which some marker object moves downward from the surface is called the ***submergence velocity***. The corresponding velocity of upward movement of the marker toward the glacier surface in the zone of ablation is called the ***emergence velocity***. In glaciers in temperate regions, a typical value of submergence velocity in the accumulation zone is 1 m/yr, and a typical value of emergence velocity in the ablation zone is 3–4 m/yr.

ADVANCED TOPIC: HOW THE SURFACE VELOCITY OF A GLACIER IS MEASURED

1. How is the surface velocity of a glacier actually measured? Basically, by standard surveying techniques—but there's more to it than meets the eye. In the following discussion, refer to Figure 7-18. Suppose that you implant a vertical stake in the surface of the glacier somewhere in the ablation area. The slope angle of the ice surface in this part of the glacier is α, and the thickness of the glacier, the minimum distance from bedrock to the glacier surface, is *H*.

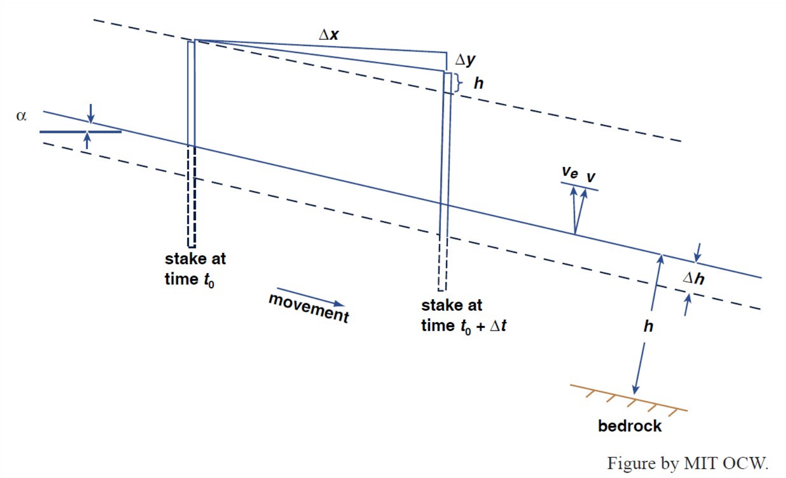


Figure by MIT OCW.

Figure 7-18. Definition sketch for analysis of the surface velocity of a glacier.

2. Wait a time Δ*t*, while the stake moves some distance down glacier. The horizontal component of movement of the stake, relative to the bedrock beneath the glacier, between time *to* and time $t_o + \Delta t$ is Δ*x*, and the vertical component, again relative to bedrock, is

Δy. Clearly the horizontal component of time-average velocity U of the glacier, relative to bedrock, is $\Delta x/\Delta t$. Likewise, the vertical component of velocity, again relative to bedrock, is $\Delta y/\Delta t$. And the resultant velocity of the glacier relative to bedrock is the vector sum of these two velocity components.

3. Note in Figure 7-18 that the ice surface is lowered (or raised) by a distance ΔH normal to the ice surface during time Δt. Think about the net rate of change in ice thickness, $\Delta H/\Delta t$. Part of $\Delta H/\Delta t$ is accounted for by downward melting of the ice surface, but part is caused by streamwise compression or extension of the whole glacier at this cross section, independently of ablation or accumulation at that point.

4. Let's deal first with the downward melting of the ice surface. This leads to another interesting vertical velocity: the vertical velocity of the ice relative to the local plane of the glacier ice surface. This is the rate at which the ice surface would rise or fall vertically if there were no ablation or accumulation. It's this vertical velocity that was called the emergence velocity (in the ablation zone) or the submergence velocity (in the accumulation zone) in Paragraph 6.2.7. It's so called because if the glacier is at equilibrium with a balanced economy the emergence velocity is how fast things embedded in the ice approach the surface, and the submergence velocity is how fast things put on the ice surface "move away" from the surface to become embedded in the glacier.

5. Shown at the top of the stake at time $t_o + \Delta t$ is the distance h, the change in elevation of the top of the stake relative to the plane of the ice surface (*not* relative to the bedrock beneath the glacier). By use of some trigonometry, the emergence velocity V_e can be written

$$V_e = \frac{h}{\Delta t} = \frac{\Delta x \tan \alpha - \Delta y}{\Delta t} \quad (1)$$

6. The component v_e of V_e perpendicular to the ice surface is just $V_e \sin \alpha$. This represents the part of the net rate of change of ice thickness H caused by melting; positive v_e corresponds to negative $\Delta H/\Delta t$. The other part of ΔH, remember, is caused by extension or compression; call it v_d. The sign of $\Delta H/\Delta t$ is determined by the relative values of v_e and v_d: in the ablation zone, if v_d is negative then $\Delta H/\Delta t$ has to be negative, because positive v_e contributes to negative $\Delta H/\Delta t$. Only if v_d is positive and greater in absolute value than *ve* does the ice surface rise relative to bedrock despite down melting of the ice surface.

6.2.8 So far we've talked only about surface velocities. How does one find the velocity of a glacier at depth? It's a big job. Drill a borehole from the surface to the base of the glacier with a thermal drill, at a speed of something like a meter per hour. This is fairly easy in glacier ice that's at its melting temperature, but it's very difficult in glacier ice that's below its melting temperature. After you finish the hole you have to install a casing, both to keep out melt water and to keep the hole from closing up by deformation!

6.2.8 Then put an inclinometer down the hole and read the inclination of the hole as a function of depth; that allows you to plot the vertical profile of the hole (relative to the glacier surface) as a function of time and therefrom find velocities (relative to the glacier surface). This relative-velocity profile is converted to an absolute-velocity profile by combining it with the already- measured absolute surface velocity. But what you cannot determine by this method is the vertical component of velocity at depth. Finally, if your hole reaches bedrock, you can indirectly measure the basal slip. And even if the hole ends a short distance from the base of the glacier, you can get a good idea of the basal slip just by extrapolation. Figure 5-19 shows an example of a vertical velocity profile measured in this way.

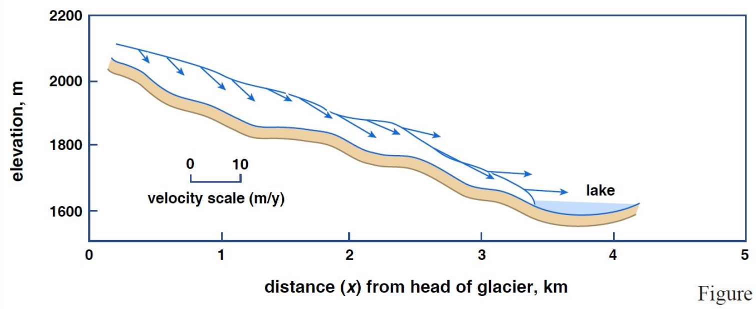


Figure by MIT OCW.

Figure 7-19. Surface velocity vectors from South Cascade Glacier, Washington, U.S., (From Meier and Tangborn, 1965.)

7.7: Deformation of Ice

As in many crystals, the way ice crystals are deformed or sheared when a stress is applied is by propagation of dislocations through the crystal. A ***dislocation*** is *a line defect in a crystal that disrupts the otherwise ideal and regular arrangement of atoms or molecules*.

Figure 7-20 shows a simple, idealized example of a dislocation. If we sheared this crystal, the dislocation could migrate simply by breakage of bonds and formation of a new bonds. The net effect is to move the “defected” plane to the right relative to the lower part of the crystal. The dislocation shown in Figure 21 is a line defect that extends indefinitely in one dimension; there are other kinds of dislocations as well.

Ice deforms mainly by propagation of dislocations along the *a* axis (that’s what the direction perpendicular to the hexagons in the ice structure are called; see the section on ice structure in Chapter 1), so slip is along the basal plane (that is, the plane perpendicular to the *a* axis). Here’s a loose but not misleading analogy: think of ice as a pack of cards oriented along the basal planes—they’re easy to deform by simple shear along these planes. It’s been shown experimentally that there’s no preferred direction of gliding within the basal plane itself, and it can be demonstrated theoretically that to within a few degrees there shouldn’t be any. There can be gliding by dislocations in nonbasal planes, but that’s much harder—stresses 10–20 times as great are needed—and apparently it’s unimportant.

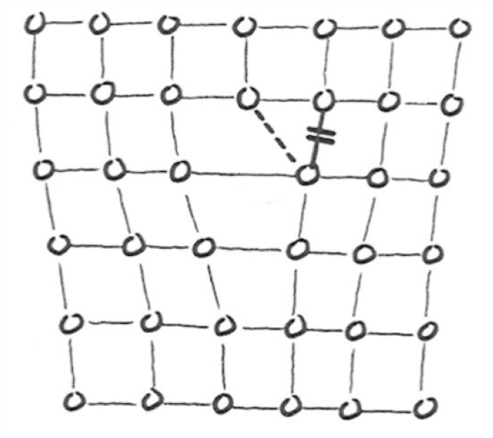


Figure 7-20. Idealized example of a dislocation.

Another way of looking at the nature of deformation of ice is to compare its behavior with other materials in a graph of deformation rate vs. applied shearing stress (Figure 7-21). I mentioned in Chapter 1 that certain fluids, air and water included, show a linear relationship between the applied shearing stress and the rate of shearing deformation. Such fluids are called ***Newtonian fluids***. Fluids that show some other kind of relationship between stress and rate of strain are called ***non-Newtonian fluids***. Ice is one of those. Materials like ice are harder to deform with increasing stress, so that the curve of deformation rate against shear stress is convex upward.

ADVANCED TOPIC: THE FLOW LAW FOR ICE

The relationship between applied stress and rate of deformation for a continuum is called a ***flow law***. What is the flow law for ice? The flow law for ice (both single-crystal and polycrystalline) is of the form

$$\text{rate of shearing deformation} = A\,(\text{shearing stress})^{n} \quad (2)$$

where A is a coefficient and n is an exponent. (In the case of water or air, the exponent n is just 1.)

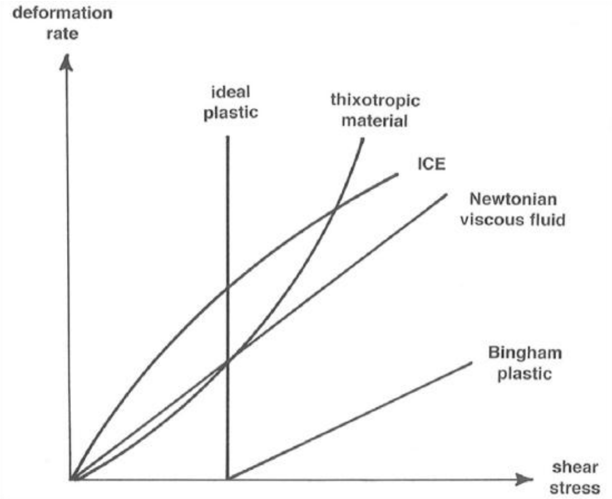


Figure 7-21. Graph of deformation rate against applied shear stress, showing behavior of various kinds of materials.

In simple shear (Figure 7-22),

du(3)

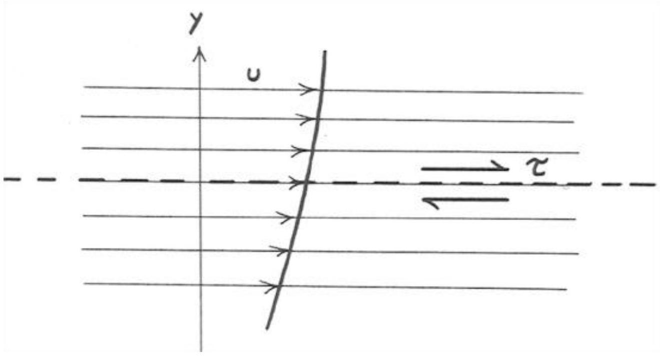


Figure 7-22. Definition sketch for analysis of simple shear in glacier ice.

The derivative *du*/*dy*, the rate of change of velocity with distance perpendicular to the planes of shearing, is what I referred to above as the rate of shearing deformation. (The coefficient 2 gets there by virtue of the way the rate of deformation is defined; don't worry about it.) In Equation 3 the coefficient *A* is like 1/viscosity. It depends strongly on temperature. As you might expect, it's much smaller (that is, the viscosity is much greater), by two orders of magnitude, for polycrystalline ice than for single-crystal ice oriented with the basal plane parallel to the shear planes. This is because deformation of ice is by internal gliding along the basal planes, and in polycrystalline ice, most crystals are not oriented favorably for this. The exponent *n* is hard to measure accurately, even in the laboratory! The value usually cited is 3 for polycrystalline ice—but there's nothing exact about this value. And *n* doesn't depend importantly on either temperature or pressure.

ADVANCED TOPIC: THE DOWNSLOPE FLOW OF GLACIERS

One thing we can do by way of theory on glacier movement is to consider the equilibrium of a given reach or streamwise segment of the glacier by writing a force-balance equation. This parallels the derivation of the fundamental resistance equation for rivers in Chapter 5.

Think about a somewhat idealized glacier of uniform thickness and infinitely wide lateral extent flowing down a planar bedrock surface (Figure 7-23). There must be a balance between the downslope component of gravity (the driving force) and the friction between the glacier and its bed (the resisting force), just as in steady uniform flow of water in an open channel. Accelerations of the glacier are very small and can be ignored for this purpose, although they are not unimportant in other respects.

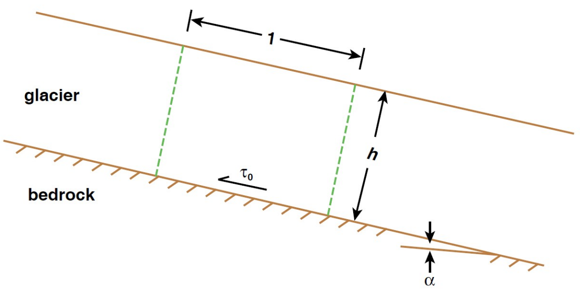


Figure by MIT OCW.

Figure 7-23. Definition sketch for force balance on a mass of ice in a valley glacier, to derive a resistance equation for glacier flow.

3. Writing the force balance for a block of the glacier with unit width and unit length, as shown in Figure 7-23 (just as for steady uniform channel flow of water; see Chapter 5 on rivers),

$$a_\tau = \rho g h sin\alpha \text{ (4)}$$

where τ_o is the shear stress, ρ is the density of ice, *g* is the acceleration of gravity, *h* is the thickness of ice, and α is the slope angle.

Here are two ways Equation 4 is important:

- It provides the best way of estimating the boundary shear stress of a glacier: typically 0.5–1.5 bars (1 bar is approximately 1 atmosphere). Note: this is far smaller than the hydrostatic pressure at the base of the glacier.
- This doesn’t tell us anything about velocities in the glacier. But it can, if we combine it with the flow law for ice. See the following development if you’re interested in the details.

First modify the resistance equation, Equation 4, to get τ as a function of position above the bed (Figure 7-24):

$$\tau = \rho g(h-y) sin\alpha \text{ (5)}$$

Now we have two equations for τ:

$$\tau^n = \frac{1}{2A}\frac{du}{dy} \text{ (6)}$$

$$\tau^n = [\rho g sin\alpha(h-y)]^n \text{ (7)}$$

Eliminating τ from Equations 6 and 7,

$$\frac{du}{dy} = 2A[\rho g(h-y) sin\alpha]^n \text{ (8)}$$

This simple differential equation is easy to solve:

$$u = 2A(\rho g sin\alpha) n \int_0^h (h-y)^n dy + c \text{ (9)}$$

To evaluate the constant of integration *c*, use the boundary condition that $u = u_s$, the surface velocity, at $y = h$. You find that $c = u_s$. So

$$u_s - u = \frac{1}{n+1} 2A(\rho g sin\alpha)^n (h-y)^{n+1} \text{ (10)}$$

If $u = u_b$ and $y = 0$, where u_b is the basal-slip part of the glacier velocity, Equation (10) becomes

$$u_s - u_b = \frac{1}{n+1} 2A(\rho g sin\alpha)^n h^{n+1} \text{ (11)}$$

Using the value of 3 for *n*, Equations (10) and (11) become

$$u_s - u = \frac{A}{2}(\rho g sin\alpha)^3 (h-y)^4 \text{ (12)}$$

$$u_s - u_b = \frac{A}{2}(\rho g sin\alpha)^3 h^4 \text{ (13)}$$

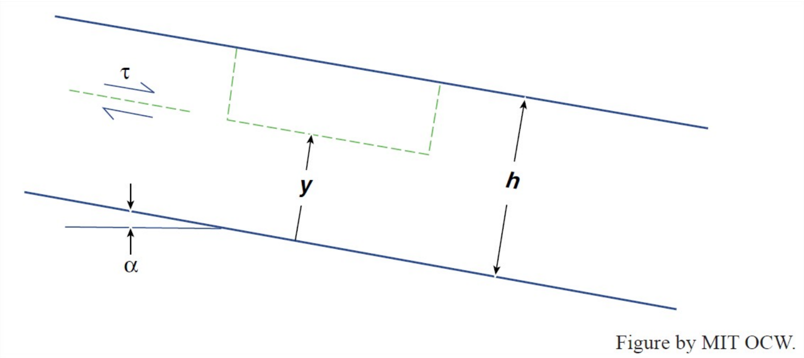


Figure 7-24. Definition sketch for force balance on a mass of ice in a valley glacier, to derive a velocity distribution along a line normal to the base of the glacier.

Note in Equations 12 and 13 that, other things being equal, glacier velocity varies as the third power of the slope and the fourth power of the thickness.

All of this is for steady uniform flow, but because glaciers have very small accelerations, and to a first approximation they are sheets, it's not bad. But temperature has a very important effect, through variation of the value of *A*.

How does a theoretical result like this compare with measured velocities? Not bad. So far, such comparisons have been made only for valley glaciers. (To do so, one has to extend the model slightly to deal with a non- infinitely-wide channel, but that's straightforward).

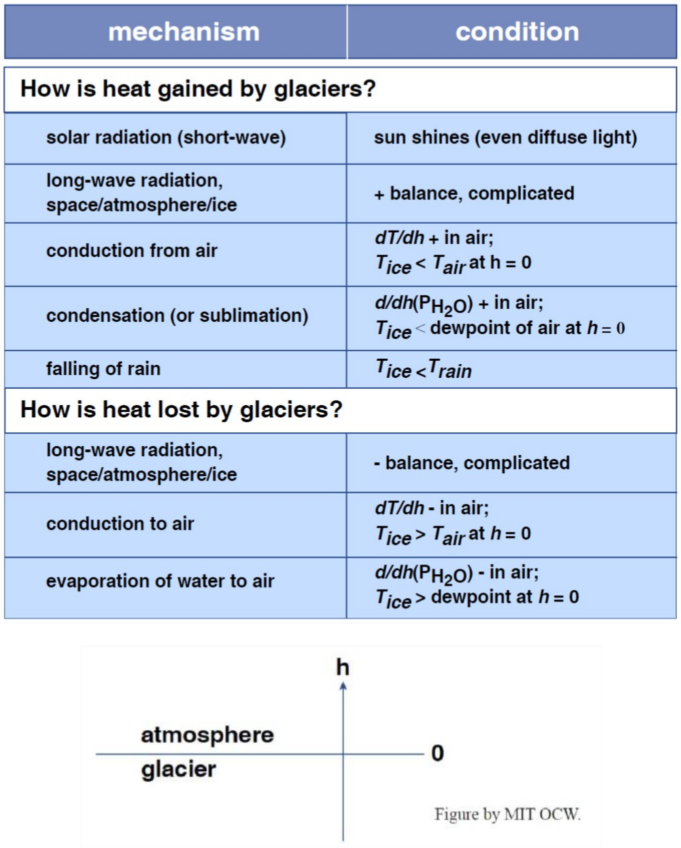
7.8: The Thermal Structure of Glaciers

The Thermal Structure of Glaciers

Processes of Energy Exchange at Glacier Boundaries

A glacier can gain or lose heat in several ways. It’s clear that these processes operate both at the surface of the glacier and at the base of the glacier. What are these processes, and what is their relative importance? Table 1 lists the ways heat can be gained or lost by a glacier at its surface.

First of all, keep in mind that the combined effect of all of these processes depends on whether the glacier is at its melting point or below its melting point (Figure 7-25). Obviously, if the temperature of the glacier is below the melting point, then the effect of addition of heat is to raise the temperature and the effect of *extraction* of heat is to *lower* the temperature. But if the temperature of the glacier is at the melting point, then addition of heat serves to melt glacier ice. Of course, extraction of heat when the glacier is at the melting point lowers the temperature below the melting point.



mechanism	condition
How is heat gained by glaciers?	
solar radiation (short-wave)	sun shines (even diffuse light)
long-wave radiation, space/atmosphere/ice	+ balance, complicated
conduction from air	dT/dh + in air; $T_{ice} < T_{air}$ at h = 0
condensation (or sublimation)	$d/dh(P_{H_2O})$ + in air; T_{ice} < dewpoint of air at $h = 0$
falling of rain	$T_{ice} < T_{rain}$
How is heat lost by glaciers?	
long-wave radiation, space/atmosphere/ice	- balance, complicated
conduction to air	dT/dh - in air; $T_{ice} > T_{air}$ at $h = 0$
evaporation of water to air	$d/dh(P_{H_2O})$ - in air; T_{ice} > dewpoint at $h = 0$

Table 7-1. Heat-exchange processes associated with a glacier.

An important indirect factor in all of this is heat conduction within the ice. This is the way heat is moved to or from the glacier surface, and it controls rates of radiation or conduction of heat at the glacier surface, because surface radiation and conduction is a function of the surface temperature of the ice.

$$\text{If } T < 0, \text{ change } T$$

$$\text{If } T = 0, \begin{cases} \Delta Q -, & \text{decrease } T \\ \Delta Q +, & \text{melt ice} \end{cases}$$

Figure 7-25. Effect of heat-exchange processes on the temperature of a glacier.

Here are some miscellaneous descriptive points about Table 7-1, keyed by number:

(1) The ***solar constant*** is *the rate at which the Sun delivers heat to the Earth*. It's conventionally taken to be the value that would be measured just outside the Earth's atmosphere when the distance between the Sun and the Earth is at its mean annual distance. The value of the solar constant is very close to 2 cal/cm^2-min.

(2) The ***albedo*** of the Earth is *the percentage of incoming solar radiation that is reflected directly back to space*, on the average. The albedo of a glacier varies considerably: snow surfaces have an albedo of 0.7–0.9, but glacier ice has an albedo of only 0.2–0.4.

(3) The surface energy exchange is greatly different under clear skies and under cloudy skies. Under clear skies, radiation directly to space is involved; typically the glacier loses heat, unless the air is very warm. Under cloudy skies, the direction of net flux of energy depends mostly on the relative temperature of clouds and ice.

(4) This would be minor without the effect of the wind. When the wind blows, there's turbulent diffusion, which you know from the section on fluid dynamics in Chapter 1 to be much greater than molecular diffusion; then conduction to or from the overlying air can be very important.

(5) This is a minor effect.

(6) This isn't important if the ice is at its melting point; 10 cm/day equals one day of long-wave radiation. But if the rain freezes when it falls, then this can be an important effect.

Energy exchange at the base of the glacier is a simpler matter, and the magnitudes of energy flux are not only much smaller but also far less variable. Geothermal heat is enough to melt about 5 mm of ice per year. This may not sound like much, but in a glacier whose basal ice is at the melting point it plays a significant role in glacier movement, by way of the lubricating effect of the thin film of water that's continuously produced and then slowly drained away. There's also heat from internal friction. This tends to be produced mainly in the lower part of the glacier, both because of basal slip and also because shearing is strongest in the lowermost part of the glacier. The rate of heat generation by internal deformation varies from considerably less to considerably more than the heat flux from the bedrock below, but it's about the same order of magnitude.

Thermal Characteristics of Glaciers

The temperature characteristics of glaciers are important because deformation of ice, and therefore flow of glaciers, is strongly dependent on the temperature of the ice. At first thought it may seem that all glaciers are cold. But an important distinction can be made between

cold ice: ice below its pressure melting point (there can be no liquid water), and

warm ice: ice at the pressure melting point (there's at least a little liquid water, under or between grains).

Recall from Chapter 1, in the section on water, that the melting point of ice falls slightly with increasing pressure. That translates to about a 2°C decrease beneath a thick ice sheet. Therefore there's actually a downward temperature gradient in a glacier, although it's small.

You can't necessarily classify an entire glacier as warm or cold, because commonly the upper part or the geographical interior of the glacier has cold ice and the lower part or the lower-latitude fringes has warm ice. But in a given geographical region of the glacier, if the glacier is such that in winter all of the ice is below the melting point, and only the surficial part is raised to the melting point in summer, the glacier is said to be a ***cold glacier***. On the other hand, if all of the ice is raised to the melting point in summer, and only a part is cooled below the melting point in winter, the glacier is said to be a ***warm glacier***.

One of the most important consequences of the thermal structure of glaciers has to do with basal phenomena. Therefore we can talk about ***warm- based glaciers*** and ***cold-based glaciers***. It's generally believed that cold-based glaciers show little or no basal slip; the ice is frozen fast to the bedrock, and all movement is by internal deformation. On the other hand, a warm glacier has a thin layer of water at its base, facilitating basal sliding.

Cold ice is formed in two different ways:

(1)*Accumulationisatatemperaturesolowthatthere'snosurface melting during the summer*. This is the case over most of the Antarctic ice sheet. The temperature of the firn and ice below the level of seasonal temperature change is approximately the same as the mean annual air temperature at the site. But temperature increases downward, because of geothermal heat. See Figure 7-26. Two opposing tendencies determine the course of the curve in Figure 7-26: (1) the value of geothermal heat flux from below, and (2) the rate of firn accumulation, which tends to "carry cold downward" into the ice sheet.

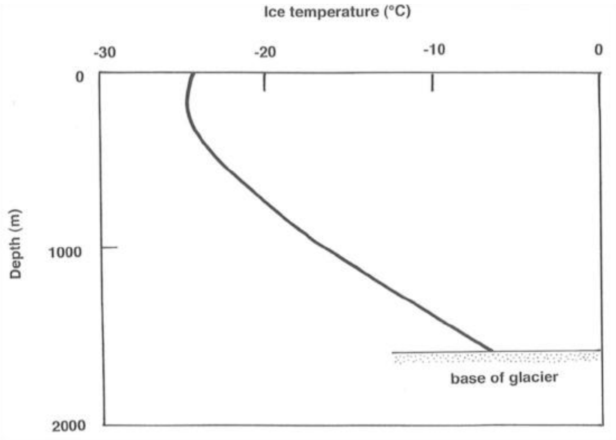


Figure 7-26. Graph of ice temperature vs. depth within a thick cold glacier.

(2) *Cooling the surface layer by winter cold*. This effect extends down as much as 20 m. This happens at the surfaces of all glaciers in winter.

How is warm ice formed? By heating to raise the ice to its melting point (anywhere). This happens at the surface of the glacier, by one or more of the heat-transfer processes listed above, to form a surficial warm layer, and then meltwater percolates down and warms the ice by refreezing. This is an important effect: when 1 g of meltwater freezes, enough latent heat is released to raise 160 g of ice 1°C. Note that this warms the ice but doesn't melt it. Note also that latent heat is the only source for warming at depth, because there's almost no temperature gradient and therefore almost no conduction.

Warm ice is produced at the base of the glacier by basal heat sources. This is favored by the following circumstances:

- thick ice
- high surface temperature
- low accumulation rate
- high ice velocity

Over large areas of both the Antarctic and Greenland ice sheets, the basal ice is at its pressure melting point!

An important point is that when even a thin layer of warm ice is produced at the base of the glacier, the temperature gradient is about zero, so all the basal heat (frictional and geothermal) is used for melting, because there's no conduction. This provides a continuous supply of meltwater at the base of the glacier.

7.9: Glacial Meltwater

Glacial Meltwater

General

Glacial meltwater is *the liquid water produced by ablation of glaciers*. Meltwater is in most glaciers by far the most important product of ablation; it's much more important than evaporation. Of course, in glaciers that terminate in the ocean, calving is more important.

The importance of glacial meltwater is twofold:

- It's intimately involved with the movement of glacier ice, by way of its influence on both creep and basal slip.
- Meltwater can carry enormous quantities of glacial sediment and deposit that material nearby or far away from the glacier.

Whereas the activity of the glacier ice itself is greatest near the equilibrium line, the activity of meltwater increases to a maximum at the terminus of the glacier.

Glacial meltwater is abundant on the surfaces of all temperate glaciers below the snowline. Even during the melting season, surface meltwater isn't common above the snowline, because the water readily sinks into the melting snow. Downglacier of the snowline, however, surface flows of meltwater during the melting season are common. Meltwater is present on polar glaciers only locally and temporarily, because it soon refreezes.

Usually the meltwater streams on the surface of a glacier plunge down into the body of the glacier before they reach the terminus; remember that liquid water is more dense than ice, so the meltwater tries to find a way down into the glacier. It's a little like pouring water into a pan of semi-connected ice cubes. You know what eventually happens: the ice cubes end up floating in the water. In a real sense, the meltwater of a glacier has a tendency to work its way underneath the glacier so as to cause the glacier to float in its own meltwater. The only thing that keeps a real glacier from becoming like the semi-connected ice cubes floating in your pan is that the balance between production and drainage of meltwater favors drainage over production.

Meltwater is classified on the basis of where it is in the glacier, as ***supraglacial meltwater*** (on top of the glacier), ***englacial meltwater*** (inside of the glacier), and ***subglacial meltwater*** (beneath the glacier). In the same way, the sources of meltwater can be viewed as surficial, internal, and basal. On most glaciers the surface sources are much greater than the internal or basal sources, by one or even two orders of magnitude. Surface sources are strongly seasonal, but internal and basal sources are largely unaffected by the seasons.

Surface sources:

- The main source is ablation during the summer melting season. This drops off sharply upglacier—just the reverse of normal fluvial watersheds.
- Rainfall on the ablation area, mainly in the warm season, is another source. In a narrow technical sense this isn't meltwater, but it's indistinguishable from true meltwater and is always considered in the same way.

Internal and basal sources:

- If the ice at the base of the glacier is at its pressure melting point, geothermal heat melts ice there rather than being conducted upward. Depending on the local value of heat flow, this accounts for one or two centimeters of ice thickness per year. Films of meltwater at the bases of glaciers have been observed directly.
- *Frictional heating by both internal deformation and basal sliding* cause melting if the ice is at its pressure melting point. This generally accounts for something like 0.5 to 5 cm of ice per year.
- Some meltwater is produced by melting by the heat generated by *the friction of the meltwater flow itself*. The quantities are not important, but this effect seems to be important in creation and maintenance of englacial and subglacial drainageways.
- *Groundwater flow* out of the regolith and bedrock beneath the glacier is locally important.

Drainage Routes

We observe surface runoff in the form of supraglacial channels, but we also see these surface meltwater streams plunging into crevasses and vertical tubular holes called ***moulins***. And we see important meltwater streams emerging from beneath the glacier at the terminus. So there must be important englacial and subglacial meltwater routes as well as supraglacial routes.

It’s easy to make casual studies of surface meltwater drainage, but there have been few detailed studies. And it’s difficult to study subglacial and especially englacial streams, for obvious reasons. But there are some fairly reliable theoretical approaches that are broadly consistent with what’s known about englacial and subglacial drainage and its depositional consequences.

Surface streams form wherever more meltwater is produced than can be absorbed locally into the glacier or held as pore water in firn or snow at the glacier surface. The stream channels range in size from tiny rills to large channels several meters wide and deep. The streams may or may not be incised into valleys on the glacier surface; that depends on the relative rates of channel downmelting and interchannel ablation.

Meltwater streams on glacier-ice surfaces have a strong tendency to form meanders, much like streams flowing on land. Meandering on ice is no better understood than meandering on sediment, but presumably there’s some kind of instability that involves preferential melting at certain points and less melting, or even refreezing, at other points. In many other cases, however, the courses of the channels are determined by lines of structural weakness in the ice.

Supraglacial streams also tend to form *dendritic stream networks*, as do stream systems on land. But there are several important differences:

- The network is *dense and rill-like*; it’s hard to cut major trunk streams when the ice is moving and deforming.
- The drainage pattern shows *a strongly subparallel pattern*, because of the relatively steep slope of the glacier.
- The *drainage density decreases upglacier*, because meltwater production decreases upglacier.
- The channel pattern is *highly changeable*: it looks different each year depending on the development of new englacial drainageways. Remember that the whole glacier surface moves downslope, but the zones of overall extension and compression of the glacier caused by subglacial bedrock topography, which tend to control englacial drainage routes, stay in one place. Moulins develop, are used for several seasons, and then are abandoned as new ones from upglacier take their place.

Dynamics of Meltwater Flow Beneath the Surface

Here are some basic observations on englacial and subglacial meltwater flow:

- In the uppermost zone of the glacier, water flows approximately vertically down crevasses and moulins in free-fall flow. These vertical drainageways have a definite water level in them that can be measured fairly easily. (Try not to fall into one, though.) This water level changes with time, on scales of hours to days, depending on air temperature, sunshine, and rainfall.
- Typically, the changes in water level from moulin to moulin are correlated throughout the entire glacier, suggesting the existence of a connected water table within the glacier.
- This water table can vary from right at the glacier surface, at times of maximum meltwater production in the early part of the melting season, when there aren’t a lot of easily exploitable passageways, to very deep (probably all the way to the base of the glacier) at the end of the melting season, when the passageways are fully developed but the supply of meltwater has dropped off.
- There have been few observations of the shapes of englacial drainageways below the moulins. These drainageways are thought to be nonvertical, with a large horizontal component to their orientation. One piece of evidence: circular englacial tunnels seen in the faces of freshly calved icebergs around the Greenland ice sheet. (A circular tunnel is the equilibrium shape if the water is flowing in a closed conduit in the ice, because approximate isotropy of the polycrystalline ice is usually a good assumption if the ice is far from the base of the glacier, where shearing by internal deformation is strongest.)
- *Travel times*: dye is injected at points of submergence and monitored where subglacial streams emerge. Speeds of movement are something like 1–2 km per hour, but there’s wide variability. This means total travel times of one to several days on ordinary valley glaciers. But speeds must increase downstream within a given drainage system: the flow in the large ice tunnels at the downstream end of the network is probably very high, many meters per second. The spectacular fountains formed where such subglacial tunnels discharge underwater in glacier-margin lakes is good evidence of this.
- What’s known about flow in large subglacial tunnels? Often the flow emerging from the tunnels has a free surface, but this free surface usually slopes upward relative to the roof in the upstream direction, so it’s reasonable to suppose that only a short distance upglacier the flow is closed-conduit flow, with no free surface. And the streams that terminate beneath glacial

lakes are clearly flowing full all the way to the end. At times of low discharge near the end of the melting season, however, most or all of the conduits have free surfaces in them.

- Any glacial tunnel below a few tens of meters, where ice can flow plastically, tends to close up completely by inflow of ice, if it’s not maintained open by some other means. There are two ways the tunnel can be held open: by water pressure equal to the hydrostatic pressure in the ice itself, or by melting of the walls of the tunnel by the flowing water. Figure 7-27 shows a graph of the half-life time for closing of a vertical ice tunnel as a function of overburden pressure for ice at its pressure melting point.

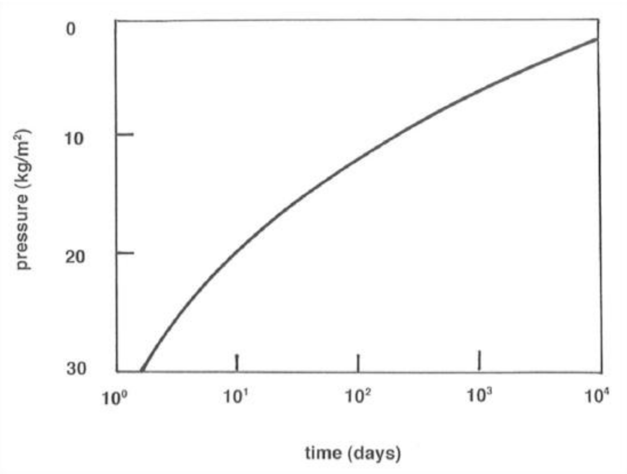
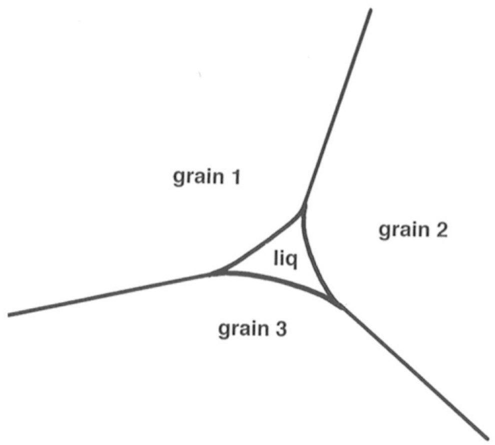


Figure 7-27. Half-life time for closing of a vertical ice tunnel as a function of overburden pressure for ice at its pressure melting point.

- Polycrystalline glacier ice at its pressure melting point is known both observationally and theoretically to be permeable to flow of water. There are veins or tiny passageways at three-grain linear boundaries (Figure 5-28), which meet in fours at four-grain point junctions. These junctions have generally tetrahedral shape. This can be observed in careful microscopic work with ice, and it also can be justified in terms of surface-energy arguments. The angle of junction between two ice grains and liquid water is about 20° (Figure 5-29). Water can therefore always percolate through a warm glacier, whatever the state of the large passageways.



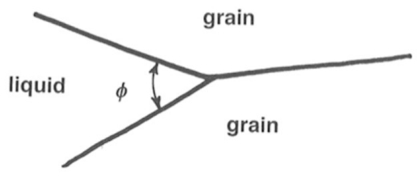


Figure 7-28 (left). Sketch of water passageways at three-grain linear boundaries in glacier ice.

Figure 7-29 (right). The angle of junction between two ice grains and liquid water is about 20°.

The basic idea about the hydraulics of meltwater flow within a glacieristhat*there’safeedbackbetweenwaterpressureandicepressurethat controls the size of the flow tunnels*. To understand the nature of these adjustments, look at a simplified tunnel (Figure 7-30) that extends vertically from the surface to the base of the glacier and then horizontally to the terminus. The water table—the free surface of the water in the vertical part of the tunnel—is shown near the surface of the glacier.

If the tunnel is too big, it's able to carry a greater discharge than is supplied from the glacier surface, and the water level in the vertical tunnel falls. That decreases the water pressure farther along in the tunnel to a value less than the ice pressure around the walls of the tunnel, so the tunnel closes up, thus constricting the flow through the tunnel and causing the water level to rise until the water pressure builds up to be equal to the ice pressure, stabilizing the diameter of the tunnel.

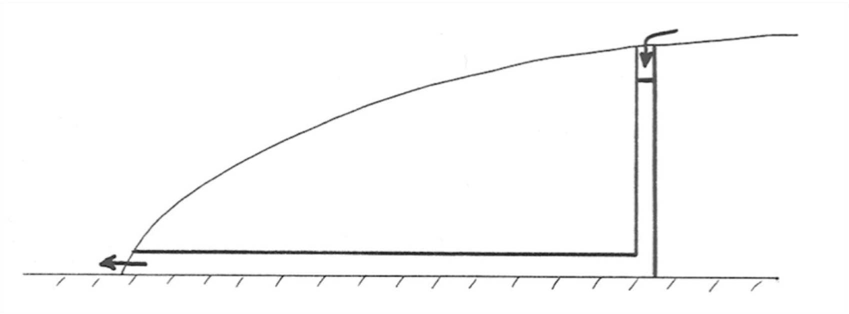


Figure 7-30. A simplified water-flow tunnel through a glacier.

Likewise, if tunnel is too small for a given meltwater discharge, the water table in the vertical part of the tunnel rises, thus increasing the water pressure to a value greater than the ice pressure at the tunnel wall, so the tunnel opens up by radial outflow of the ice wall, leading to increased discharge and a fall in the water table. The tunnel diameter thus again becomes stabilized at a value for which the water pressure is equal to the ice pressure on the tunnel walls.

In accordance with the foregoing argument, it's usually assumed that in a steady state (constant discharge, certain water level, certain tunnel diameter) the water pressure is equal to the ice pressure at every cross section in the tunnel. But there has to be one significant correction to this. Meltwater discharge through the tunnel tends to melt the tunnel walls, by two effects: (1) the heat generated by friction in the flow, and (2) heat carried from the surface by meltwater that's slightly above freezing. So at equilibrium the ice walls of the tunnel have to flow inward toward the center of the tunnel at a finite rate to balance this rate of wall melting. The adjustment described above is modified in such a way that the water pressure is a little greater than the ice pressure.

A good case can be made for another important consequence of this wall-melting effect: larger passageways grow at the expense of smaller passageways. The reason? (1) more heat relative to wall area is generated by viscous friction in the larger passageways than in the smaller passageways; and (2) more heat relative to wall area is carried by above-freezing water from the surface in the larger passageways than in the smaller passageways.

The consequence is that because of this differential growth of larger passages, the three-dimensional network of passageways in the glacier tends with time to become dendritic, with tributaries joining into ever-larger trunk passageways. This is broadly consistent with the few observations of the pattern of internal meltwater passageways in glaciers.

ADVANCED TOPIC: THE HYDRAULICS OF MELTWATER FLOW WITHIN GLACIERS

1. The vertical and then horizontal passageway in Figure 7-30 that was usedtoillustratetheforegoingpointsisclearlyunrealistic.*What'sthedirectionof meltwater flow in the ice?* That question leads in turn to what makes water flow inside a glacier in the first place. The answer to that latter question is: spatial gradients in the difference between the actual water pressure and the hydrostatic pressure (i.e., the pressure that would be measured at the given point if the water were not moving).

2. Figure 7-31 shows this effect in a simplified passageway. The hydrostatic pressure is constant all along the horizontal segment of the passageway, but the water flows from the higher tank to the lower tank because of the gradient in water pressure caused by the difference in water level between the two tanks. If that doesn't convince you, just consider that the state of flow in the horizontal segment would be exactly the same if you increase the water level in both tanks by the same vertical distance, thereby changing the hydrostatic pressure but not the gradient in pressure.

Figure 7-31. Flow of water in a horizontal passageway due to a gradient in water pressure.

3. Now look at the situation in a real glacier. Think about the water pressure and ice pressure at a point P within the glacier (Figure 7-32). H is the elevation of the ice surface above an arbitrary datum, and z is the elevation of Point P above that same datum. The ice pressure p_i at Point P is

$$p_i = \rho_i g(H - z) \quad (14)$$

and the water pressure *pw* at the same point is approximately

$$p_w = p_i \text{ (15)}$$

by the line of reasoning in Paragraphs 9.3.2 through 9.3.4. (Forget about the small effect of melting of wall ice.)

Figure 5-32. Definition sketch for analysis of the effect of water pressure and ice pressure at a point in a glacier.

4. Water tends to move through the network of passages in the direction of the gradient of a potential Φ that can be represented as a family of smoothly curving surfaces with the property that the direction of most rapid decrease in water pressure is everywhere normal to the surfaces. This is just a generalization of the idea that in a one-dimensional situation like a straight circular pipe the water moves in the direction of decreasing pressure. (It's just like the gravitational potential function that describes the direction of fall of bodies at all points near the surface of the Earth, which is the direction of most rapid increase or decrease of potential energy as you move the body up or down in the gravity field of the Earth.)

5. The potential can be represented by a equation like

$$\Phi = \Phi_o + p_w + \rho_w g z \text{ (16)}$$

where the first term on the right is just an arbitrary additive constant, the second term is the actual water pressure, and the third term is the hydrostatic pressure that would be produced by a column of motionless water above the given point.

6. Substituting Equations 14 and 15 into Equation 16 to get rid of *pw* and *pi*, and ignoring the arbitrary constant Φ_o,

$$\Phi = p_i + \rho_w g z = \rho_i g(H-z) + \rho_w g z = \rho_i g H + (\rho_w - \rho_i) g z \qquad (17)$$

(Remember that the direction of drainage will be normal to the equipotential surfaces.)

7. The dip angle α of these equipotential surfaces can be found with the aid of Figure 5-33:

$$tan\alpha = \frac{\frac{\partial\Phi}{\partial x}}{\frac{\partial\Phi}{\partial z}} \text{ (18)}$$

or, solving for the angle

$$\alpha = arctan\frac{\frac{\partial\Phi}{\Phi x}}{\frac{\partial\Phi}{\partial z}} \text{ (19)}$$

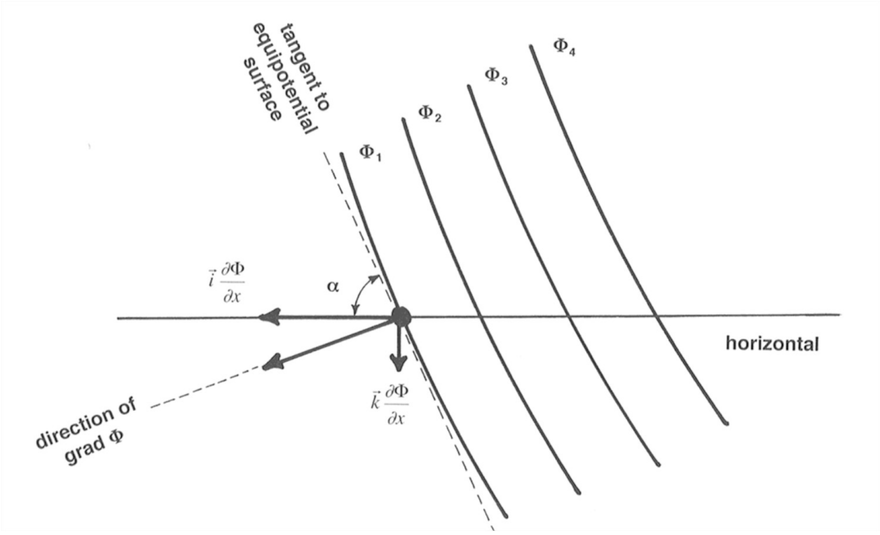


Figure 7-33. Flow of water in a glacier in response to the gradient of a potential.

Given that $\rho_i = 0.9\rho_w$, the result in Equation 19 tells us that englacial tunnels slope downglacier about 11 times as steep as the glacier surface!

8. Then what happens when the passageways reach the base of the glacier? Subglacial tunnels are constrained to follow the locus of steepest descent of the component of the potential function Φ parallel locally to the glacier bed. (Just think in terms of the curves lying on the glacier bed that are formed by the intersections of the equipotential surface with the bed, and then taking directions on the bed that are normal to those intersection curves.) On a horizontal glacier bed, this is in the same direction as the surface slope.

But if the glacier bed isn’t level, the tunnels can cross bedrock divides. There can even be subglacial lakes, where there are “hollows” in the equipotential surfaces (i.e., where the directions normal to the equipotential surfaces dip locally at a gentler angle than the bed of the glacier).

7.10: Glacial Erosion

Introduction

Glaciers are very effective in eroding, transporting, and depositing bedrock. How do we know that? There are three major lines of evidence:

- *We can see material in transport by modern glaciers*. In the most general way, where could such material have come from? Preexisting loose material (regolith; core stones; joint-bounded blocks); solid bedrock eroded by the glacier; and material that fell onto the glacier from bedrock weathering above. And we see glaciers making deposits.
- *We observe deposits that by excellent evidence were deposited by glaciers*.
- *We observe landforms that are the result of glacial erosion*.

The subject of glacial erosion is a difficult one. We know it happens, but it's hard to observe how it happens. Very few tunnels have been driven to the base of a glacier to watch erosion, and those haven't been representative anyway, in terms of depths and times involved. Also, very few experiments have been made. So there's a lot of deduction and speculation, and this can be very dangerous. And this is not a very graphic or photogenic topic.

It's generally agreed that there are two kinds of erosional activity of glaciers: ***abrasion*** and ***plucking*** (also called ***quarrying***). These affect bedrock on different scales (although there are intergradations). For each, I'll discuss evidence and possible mechanisms.

Glacial Abrasion

Tools (*rock and mineral particles, large and small, held in the base of the moving ice*) can abrade the underlying rock surface. Basically, this involves wearing away particle by particle.

What's the evidence that this happens? Mainly glacial striations and rock flour.

Glacial striations, or ***glacial striae***, are *subparallel striations or grooves cut on the bedrock base of the glacier by tools frozen into the basal ice*. They are a very common (although by no means ubiquitous) feature of glaciated areas. They are commonly are found on rounded undulating surfaces of glacially abraded bedrock. There's a wide and continuous range in size, from microscopic, at the small end, to meters deep, meters wide, and hundreds of meters long. (The biggest grooves were probably made not by single tools but by groups of tools.) Striations are also on the larger tools themselves. Often the bedrock surface shows two or more intersecting directions, indicating either that the direction of ice flow changed or that the tools were rotated relative to direction of movement. The finest and most delicate striations are cut on soft fine-grained but nonfractured rocks like carbonates. They are coarser in medium-grained to coarse-grained rocks like granite or sandstone. Striations are fairly readily weathered, so they are best seen soon after they're made or after they're freshly uncovered of overlying sediment. Striations are known from ancient glaciations as well as Pleistocene glaciations.

If abrasion happens, there must be a fine-grained product. What's the nature of this product, and what happens to it? Most of what's produced by abrasion is *mineral fragments*, evidently mostly less than 100 micrometers, and predominantly fresh. This material, expressively called ***rock flour***, is largely carried out of the glacier by meltwater; glacial streams have very high suspended-sediment concentrations of grams to tens of grams per liter, and even up to a few hundreds of grams per liter, which turn the streams a characteristic dilute-milk whitish color.

A case can be made for the necessity of continuous removal of the fine-grained abrasion products in order for abrasion to continue; otherwise the ice– rock interface would become clogged by this stuff, like overused sandpaper.
There are two ways for this to happen: it can be carried obliquely upward by flow of ice, under special circumstances of compressive flow, or more importantly, it can be washed out by meltwater. Remember that in a warm-based glacier, even if no meltwater is supplied from above, there will be a thin layer of flowing water, due to geothermal heat and friction, along the base from melting.

Direct observations on production of rock flour have been few and simple. One thing you might try to do is make marks or holes on the bedrock near the terminus and hope for readvance and then re-retreat within your lifetime. This points up the obvious difficulty in making systematic observations. Another more subtle difficulty is the possibility of weathering of the bedrock before readvance. The observation most often cited is some recent work that involved tunneling under a glacier and planting two rock slabs, one of marble and the other of basalt, at the base of an active glacier, retrieved after 9.5 m of passage of tool-studded ice over them. Both slabs became striated; the marble lost 3 mm of thickness, and the basalt lost 1 mm.

How about *the mechanism or mechanisms of abrasion*? Probably many of you have seen striations, but how are they made? By the ice dragging tools across the bedrock, you would say; yes, but under what conditions does this happen?

The first and most important condition for abrasion is that *the glacier ice has to be moving at the base*. That seems to mean that abrasion is important only under warm-based ice, for which there is basal slip. Cold-based ice is "glued" (frozen fast) to bedrock, and so doesn't drag tools across the bedrock. Only if there are very large tools that stick up into faster-moving ice and are rotated and pushed down against bedrock can there be abrasion in this case. Abrasion is generally conceded to be minimal in cold-based glaciers, especially when they're relatively clean.

But even in warm-based glaciers there's a considerable problem in keeping the tools in contact with the bedrock. Consider an isolated tool at the base of the glacier. As it rides along bedrock and exerts a normal force on the bedrock, the bedrock exerts an equal and opposite normal force on the tool (Figure 5-34). If there's basal slip in the first place, then the ice at the base is at its pressure melting point, and this means that the tool will retract slowly into the ice, and therefore stops abrading the bedrock!

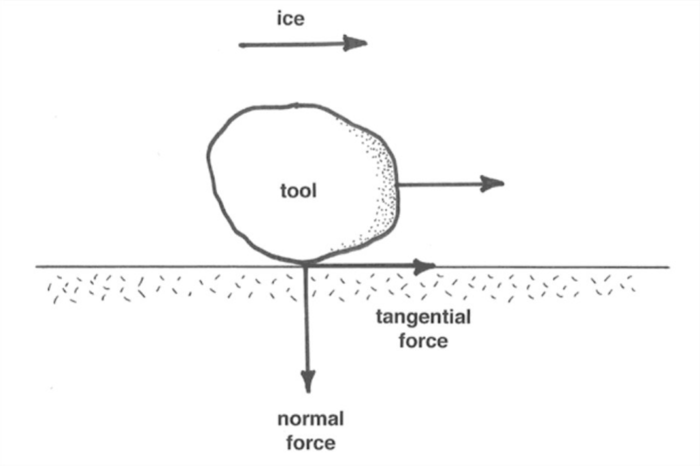


Figure 7-34. Forces and motions associated with the motion of a tool in contact with the sole of a glacier.

But there are invariably many tools, presumably of all sizes, packed in near the base of the glacier. Higher tools tend to hold lower tools to the base. Remember that the ice is also exerting tractive forces on the tools to keep them moving forward against the tangential resistance of the bedrock sole. Imagine a large tool, being pushed forward by the ice, and in turn pushing a smaller tool downward onto the bedrock (Figure 5-35): this produces a very large abrasive force on a small area for a long time, because the large tool retracts only very slowly, given its large area and relatively small ice–rock contact force.

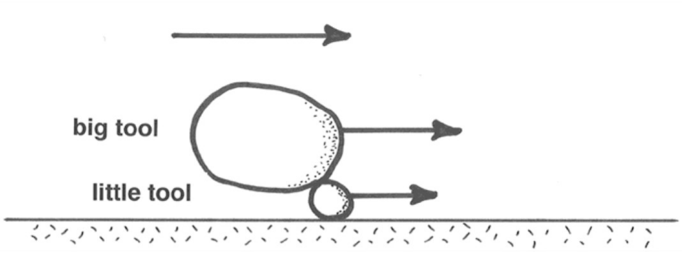


Figure 7-35. A big tool pressing upon a little tool in contact with the sole of a glacier.

But why wouldn't a condition eventually be reached in which all tools have attained a state of retraction? It seems as though this has to happen, eventually. So for continuing abrasion a mechanism is needed for replenishment of tools at the base of the glacier.

Something else to remember is that the tools themselves are abraded approximately as fast as the bedrock, if the two are of about the same resistance to abrasion. (If the tool is softer than the bedrock, the bedrock tends to abrade the tools, whereas if the bedrock is softer than the took, then the tool tends to abrade the bedrock.)

So tools are progressively used up, and new ones have to be supplied from somewhere. Probably the most efficacious abrasion happens in a situation where fragments are incorporated into the ice base by quarrying somewhere and then carried downglacier to abrade softer rocks.

The whole field of abrasion micromechanics is (as you can see) in a very speculative state. It's ripe for further observation, experiment, and theoretical work. (But there are difficulties with all three).

Striations tell you the orientation of the ice movement at the time they were cut. Unfortunately they don’t give you the direction as well as the orientation of the movement: some are asymmetrical one way, others asymmetrical the other way. An obvious thing to do is map the orientations of striations over large areas to get an idea of the patterns of glacier movement. This has been done many times (Figure 5-36 is a good example), and the results are valuable.

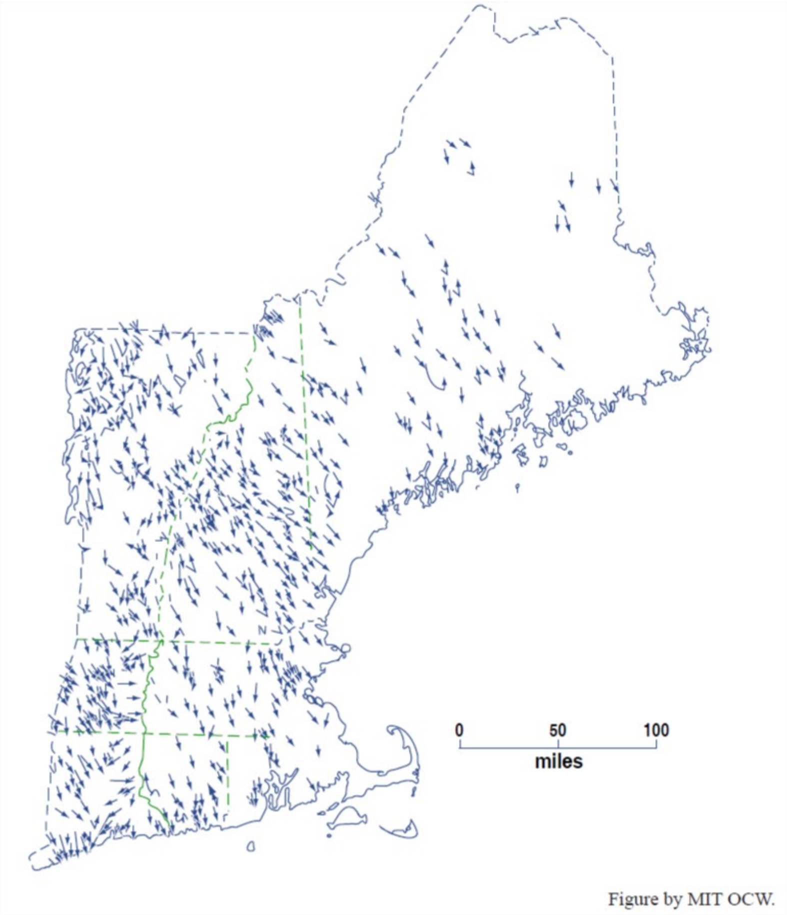


Figure 7-36. Map of glacial striae in New England, compiled from many sources. (From Flint, 1971.)

But there are some serious problems and limitations:

- Commonly there are very strong local variations in orientation because of local topography. So you either have to make very detailed maps or do some averaging. Striations indicate that flow can be locally at 90° to the main flow in small troughs oriented across the ice movement, or even “eddies” in cavities.
- Directions of ice movement can change with time as the geometry of ice sheets and ice caps changes as they wax and wane. In particular, there can be almost complete reversal if an outlying ice cap develops on a locally higher area during general retreat of an ice sheet (Figure 7-37). This seems to have happened in North America just south of the St. Lawrence depression.

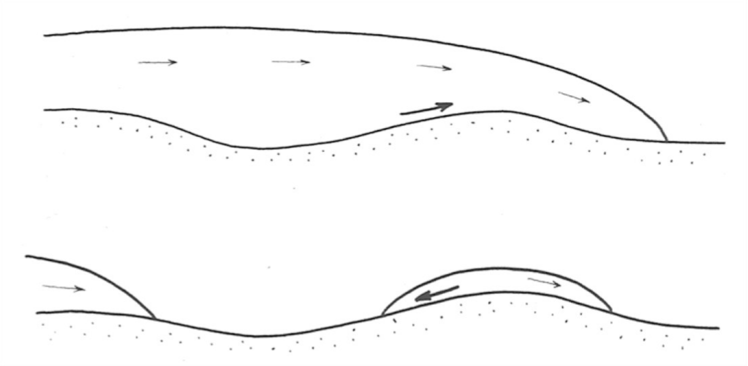


Figure 7-37. Reversal of ice-flow directions as an ice sheet retreats past a high land area.

A related point to remember is that striations record only the very latest abrasion, and this could be radically different in orientation from earlier, and perhaps more important, abrasion. It's probably typical that the last ice movement is unrepresentative of the main ice movement.

Finally, another important point is that striations on bedrock surfaces have been observed to be produced by several other mechanisms:

- drifting ice (icebergs or ice floes)
- debris flows
- snowslides and avalanches

And unfortunately you can't tell which mechanism just by looking at striations. Other evidence is usually available, though.

Another minor abrasional feature often discussed in the literature is ***friction cracks***. These are of several characteristic kinds, all apparently produced by the normal and tangential forces exerted cyclically by tools on bedrock at regularly spaced points in the downstream direction. These cracks occur in trains analogous to the chatter marks produced in certain machining operations, and they are probably a manifestation of stick-slip friction.

Some of these friction cracks show removal of material from the bedrock surface; if they are concave upglacier they are called ***crescentic gouges***, and if they are convex upglacier they are called ***lunate fractures***. Others, called ***crescentic fractures***, show no removal of material from the bedrock surface; these are almost always (or even always?) convex upglacier.

Controls on the geometry and spacing of friction cracks are unclear. Some experiments on scoring of optical glass with steel balls have been made to simulate, crudely, the conditions under which friction cracks are made. From these experiments it's known in a general way that crescentic fractures are produced when there's no rolling, and crescentic gouges are produced when there's rolling. In any case, it's clear that friction cracks are not reliable indicators of the direction of ice movement.

Glacial Quarrying

Glacial quarrying, ***glacial plucking***, and ***joint-block removal*** are approximately equivalent terms for a process involving *incorporation of relatively large discrete fracture-bounded blocks into the moving ice at the base of the glacier*. Typically this process is viewed as operating on fragments that are decimeters to several meters across.

Glacial quarrying must be a very important mode of glacial erosion, because how else can you get all the large fragments observed to be carried in glaciers and in glacial deposits (assuming that there is a definite limit to the material initially available before the glacier covered the area)?

The evidence for glacial quarrying is perhaps best summarized by discussing what is usually considered to be the most characteristic form of glacial erosion: ***stoss-and-lee topography***.

Just as common as simply abraded surfaces is a landform that involves hillocks or knobs of bedrock on scales of meters to tens of meters to even a few hundreds of meters in plan view. These hills have a strong tendency to be asymmetrical, with gentle and streamline-molded upstream sides and steep downstream sides. The upstream sides are smoothly rounded in the large and striated in the small. The downstream sides are rough, blocky, and craggy.

The usual interpretation of stoss-and-lee topography is that it developed by erosion by moving ice, involving (minor) abrasion on upstream side and (major) quarrying on downstream side. Jointing patterns in the bedrock are thus important in controlling the shape of the stoss-and-lee topography.

Stoss-and-lee topography is less good in giving the orientation of ice movement (to maybe 10–20°) but is definitive evidence of direction and is less likely to be produced only by the last phase of glaciation.

There is disagreement in the literature over whether stoss-and-lee topography is a transient or steady-state aspect of glacial erosion.

The mechanism for glacial plucking in general is very poorly understood. An obvious requirement is the existence of fractures and joints in the bedrock. (Bedrock with tight and widely spaced joints is just not likely to be quarried.)

Joints and fractures can obviously predate the glacier, either because the rock was jointed long before or was fractured just before. One popular European school of thought holds that most glacial quarrying is the result of fracturing that takes place just before the arrival of the glacier; active frost wedging in a periglacial (that is, near the glacier) climate in advance of the expanding glacier prepares the way.

Another interesting possibility is that the production of joints is occasioned by the presence of the glacier itself. Dilatation joints are produced along surfaces congruent to the bedrock surface by unloading of overlying rock, by erosion, or by artificial means in quarries. In quarries, sometimes a rock face bursts outward, producing such a dilatation joint, because the rock has tendency to expand upon unloading. Some workers think that such dilatation joints can be produced beneath a glacier: as a glacier erodes rock and fills a deeper and deeper depression, hydrostatic pressure at the glacier base, although greater than if the glacier were not there, is much less than it was before the rock was eroded, because the density of ice is much less than that of rock.

What could be interpreted as good evidence of this is the existence of prominent dilatational jointing congruent to the walls of major deep glaciated valleys. (Remember that the valleys must have been produced by the glacier, and not just there from before.) Also, one can make a comparison of the stresses needed: given an ordinarily deep glaciated valley, effective unloading is more than a few hundred meters, and rock bursts resulting in dilatational joints in quarries are known to occur for unloading of less than 10 m.

One still needs to have joints perpendicular to rock surface. There's some suggestion from recent studies that moving glacier ice can produce its own steeply dipping joints: The data that support such an idea are from measurement of joint orientations at a great number of glaciated localities, together with the direction of ice flow from striations. Sets of shear joints and extension joints oriented symmetrically with respect to direction of movement are found. This is highly suggestive, because how could it be coincidental?

A related matter is that of the processes of mobilization of material in the lee of stoss-and-lee topography. What causes the quarrying there? For one thing, there's the possibility of freeze–thaw processes at base of glacier itself. Look at a typical stoss-and-lee hillock in a warm-based glacier (Figure 5-38). Local pressure at base is considerably higher over the stoss side than over the lee side. (This pressure difference has actually been measured.) So, because the pressure melting point of the ice decreases with increasing pressure, ice tends to melt on upstream side, and the meltwater produced flows around to lee side, where it refreezes in the region of lower pressure. If this happens on a larger scale, it could wedge out blocks on the steep lee side of the hillock, given the preexistence of fractures.

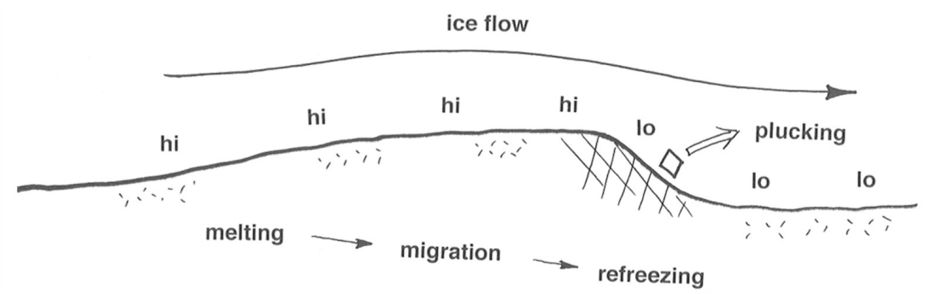


Figure 7-38. A possible mechanisms for plucking, based on melting and refreezing in response to change in pressure over a bedrock knob at the base of a warm-based glacier.

Which is the more important quantitatively, abrasion or plucking? It's generally agreed that plucking is more important than abrasion. But the basis for this view is not extensive.

The classic study was done close to home by Jahns in 1943 on glaciated granite hills of eastern Massachusetts. Well developed sheeting joints presumably predate glaciation and show the original form of the hills. The present form of hills relative to original form shows that there has been much more glacial erosion on the lee sides, by plucking, than on the stoss sides, by abrasion (Figure 7-39).

Another important line of evidence lies in the long profiles of glaciated valleys. Characteristically such profiles show large-scale steps, and even excavation of deep rock basins with up to several hundred meters of closure, now filled with glacial debris or lakes (Figure 7-40). On the assumption of regular preglacial river-valley profiles (a very good assumption), this shows that erosion rate varies greatly from place to place. Also, places where rock basins are excavated tend to have well jointed and therefore easily quarried material, whereas the bedrock highs don't, and are therefore subject only to abrasion.

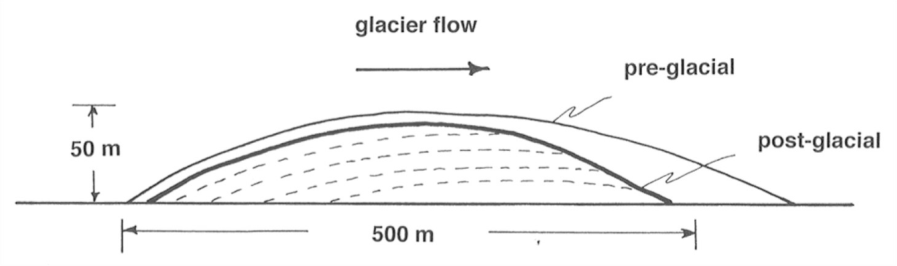


Figure 7-39. Streamwise cross section through a glaciated bedrock hill, showing the relationship between sheeting joints and land-surface topography.

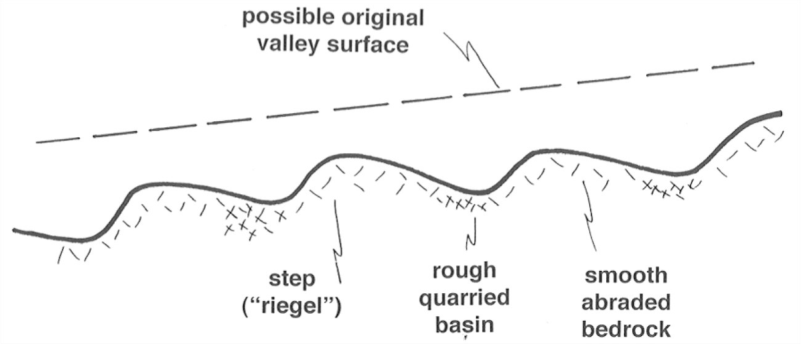


Figure 7-40. Steps and basins along the longitudinal profile of a typical glaciated valley.

Rates of Glacial Erosion

It would be nice to be able to make an estimate of the rates of glacial erosion, and then to compare these with rates of fluvial erosion. Much effort has been expended in this direction.

There have been four major approaches to this problem: 335

- Erosion of artificial markers placed beneath a glacier (already mentioned). Problems with this approach:

— Time scales are too short.

— What about pre-weathering? (But this is a problem only if the marker is placed in an ice tunnel.)

— What about spatial non-uniformity? (The sampling grid would have to cover a large area.)

— This measures abrasion only, and by the nature of abrasion this involves too small a measurement area.

- Measure sediment transport in subglacial streams emerging from beneath a warm-based glacier. Problems with this approach:

— One needs to accumulate very long-term records and make closely spaced observations, because such streams are notoriously highly variable.

— It's difficult (or impossible) to measure bed load.

— One has to look at moraines too; not all debris goes out in streams (but probably usually most, except in cold glaciers).

- Reconstruct the preglacial surface. Problems with this approach:

— How do you know what the preglacial surface was? There's lots of disagreement about the criteria to use. (This approach works best in glaciated valleys, but there are problems even there.)

— Effects of glacial erosion are hard to factor out from concurrent fluvioglacial erosion and fluvial erosion during interglacial periods.

- Compute volumes of glacial drift. Problems with this approach:

— One needs to work in a very large area.

— Glacial drift grades out into fluvial and marine deposits; how do you separate out the glacial component in those deposits?

— It's not always easy to identify the form and depth of the base of the drift.

— Later erosion of drift deposits must be taken into account.

Method 2 is probably the best, but it works only for temperate (warm) glaciers.

Some very general conclusions:

- Active temperate valley glaciers erode at much greater rates than rivers in but nonglaciated areas.
- Ice sheets (warm or cold) moving slowly over low-relief areas produced relatively low erosion even over all of the Pleistocene.

Example: the Antarctic ice sheet seems now to be eroding very slowly. Erosion probably reached a peak early, well before the Pleistocene, when the glaciers were more active and there was more material available, and it's now slow, because of lack of easily erodible material, inactive regimen, and cold base.

Here's some discussion on two additional specific topics in glacial erosion: *glaciated valleys* and *cirques*.

Glaciated valleys. In mountains that are not mainly the work of glaciers, preexisting valleys are modified during glaciation. In low-relief regions covered by ice sheets, often prominent valleys (now partly filled by sediments or lakes) have less obvious relation to preexisting drainage; their position is determined by preexisting drainage, but valleys become more differentiated as the glacier exploits weak rock. The Finger Lakes of New York State are an excellent example of the latter effect.

The transverse profile of glaciated valleys is commonly U-shaped: very steep or nearly vertical side walls, and gently rounded bottom. In some cases there's even a good parabolic fit. Nonglaciated (stream) valleys in mountainous regions, on the other hand, are commonly V-shaped. Glaciated valleys have much steeper sidewalls. See Figure 7-41.

Figure 7-41. Transverse cross sections through A) a nonglaciated valley (V shaped) and B) a glaciated valley (U shaped).

The change from a V-shaped stream valley to a U-shaped stream valley could simply be by widening, or by both widening and deepening. There seems usually to be both widening and deepening. Evidence:

- Hanging tributary valleys (Figure 7-42), which even if glaciated would have far less glacial erosion. Very common.
- The long profile shows steps and rock basins, already illustrated.

Figure 7-42. Cartoon of a hanging alley.

Why does a glacier develop a U-shaped valley? This is difficult to answer. Two possibilities are commonly mentioned in the literature:

- Concentration of nivation (see below) in the lower parts of valleys, as the approaching glacier weakens rocks for removal.
- Development of dilatation joints, especially in lower parts of the valley, where the ice is thicker.

Both of these effects would tend to act more in lower parts of the valley profile.

Cirques. A ***cirque*** is *a rounded basin, partly enclosed by steep cliffs, cut into a mountain slope*. Cirques are about circular, or at least fairly equidimensional, in plan. Their size ranges from tens of meters to several kilometers across. They can be located either at the heads of glacialed valleys or independently as indentations in smooth slopes. Cirques may or may not have (or have had) a glacier in them.

Small cirques can be formed without the presence of a glacier (Figure 7-43). Think about a firn bank that occupies a slight preexisting depression or a shady spot. Freeze–thaw cycles in the warm season result in frost wedging, and then this material moves downslope by a combination of mass wasting and surface runoff (This *combination of freeze–thaw and mass wasting* is called ***nivation***.) This results in a small cirque. In such a small cirque, a firn bank acts as a passive water source.

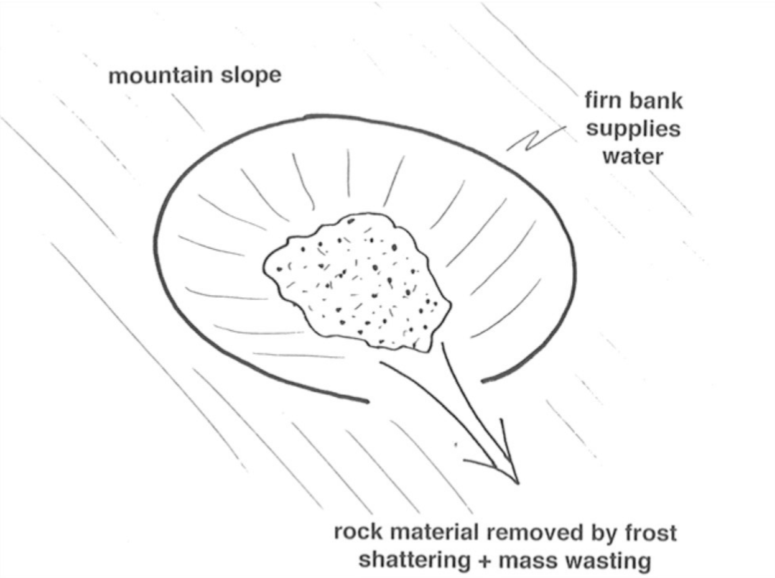


Figure 7-43. Early stage in the development of a cirque.

If the firn bank gets big enough, a small glacier (called a ***cirque glacier***) is formed. Then the moving ice can enlarge the cirque to much greater size. Larger cirques tend to have a bedrock basin (often occupied by a lake after the ice melts, called a ***tarn***, and a sill at the downslope edge (partly rock, partly moraine). Of course, large cirques grade over into the steep amphitheater-like heads of major glaciated valleys. Cirque glaciers are typically half-moon-shaped, and they move in a characteristically rotational way (see Figure 7-44, which shows the results of detailed study of a conveniently small and fairly regular cirque glacier in Norway). Note the almost circular-arc base, the almost planar surface, and the strongly rotational movement. The glacier consists of a series of ice layers separated by discontinuities (ablation textures and mineral-organic dust coatings). Most of the movement is accounted for by rotational sliding around a horizontal axis, but there is some deformation of annual layers as well.

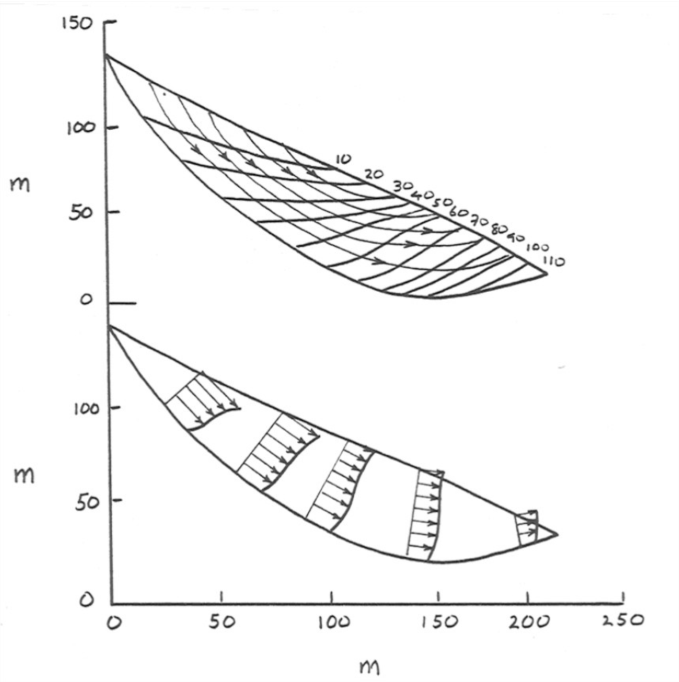


Figure 7-44. Streamwise vertical cross section showing flow lines and velocity profiles in a small cirque glacier.

Alpine Sculpture

Glaciation in mountainous areas tends to produce characteristic glacially sculptured landforms. The assemblage of such landforms is usually termed ***alpine sculpture***.

As the climate in a rugged mountainous area gradually changes to become more favorable to the development of glaciers, small cirque glaciers form first and then expand to become valley glaciers. The heads of the valley glaciers may expand and merge to form extensive snowfields. Eventually an ice cap covers most or all of the mountainous area, extending out into the lowlands beyond. Then, as the climate ameliorates again, the process operates in reverse, and the glaciers shrink back to valley glaciers, then to cirque glaciers, and finally disappear altogether.

Figure 5-45 shows the sequence of glacially sculptured landforms associated with the cycle of glaciation outlined above. During the expansion of the glaciers, valleys are deepened and widened, leaving U-shaped valleys often separated by sharp ***cols*** and ***arêtes***. Three or more large cirques at the heads of glaciated valleys may meet to for pyramidal faceted peaks known has ***horns***.

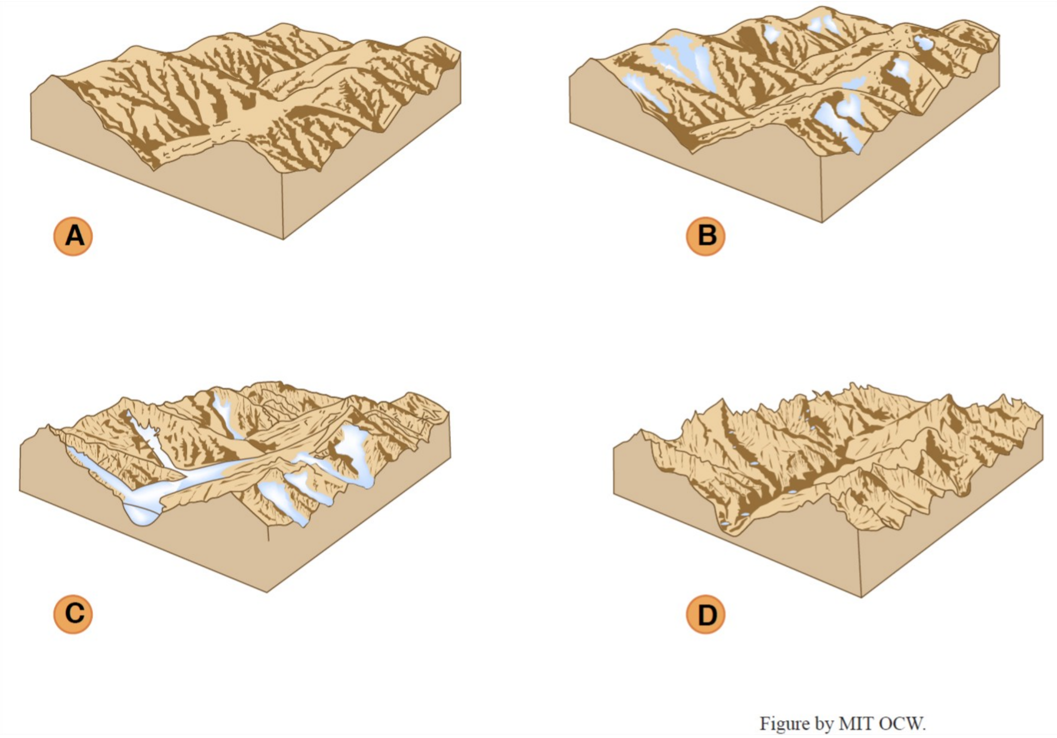


Figure by MIT OCW.

Figure 7-45. Sequence of glacially sculpted landforms associated with the cycle of glaciation and deglaciation of a mountainous area.

At the time of maximum glaciation, much or all of the former sharp glacially sculptured topography is smoothed as the entire area is worn by the moving ice of the ice cap. Then, as the ice caps shrink back to valley glaciers and then cirque glaciers, the mountains are sculpted once again into the landforms noted above.

7.11: Glacial Sediment Transport

The term ***load*** is used for *all drift that's in transport by a glacier at a given time*. As with fluvial sediment transport, keep in mind the distinction between the load and the transport rate. The ***transport rate*** of a glacier is *the time rate of passage of sediment past some cross section through the glacier that's stationary relative to the underlying bedrock*.

The load is generally classified on the basis of where it's transport in the glacier:

- **supraglacial load:** load transported on the surface of the glacier. It gets there by falling onto the glacier. It's thus restricted mainly to valley glaciers. But near the terminus of the glacier, drift can reach the surface by ablation of the ice, and also by upthrusting.
- **englacial load:** load transported within the glacier. The quantities of englacial drift are always much smaller than of supraglacial drift and subglacial drift, because the load is obtained by the glacier at the bottom and the top of the glacier. Solids deposited on the surface of the glacier in the area of accumulation are buried and thus become englacial drift.
- **subglacial load:** load transported at the base of the glacier. This constitutes most of the load, just because most material is entrained at the base. Cold-based glaciers have little subglacial load. The dirtiest warm- based glaciers might have concentrations of up to tens of percent by volume, for thicknesses of a few meters above the base. The way this is known is by examination of now-dead ice at a stagnant terminus.

7.12: Glacial Deposits

Introduction

Here are some comments on the nature of sedimentary materials deposited by glaciers:

- It *tends to be fresh*. After a glacier picks up the loose preexisting material, it wears away fresh bedrock. And there's no weathering while the material is in transport. So the material deposited beneath the glacier, or dumped at the terminus, tends to be fresh. (But the deposit itself can undergo subsequent weathering.)
- It faithfully reflects the composition of the upglacier source rocks.
- It *varies widely in composition*, because it reflects the composition of the bedrock directly upglacier, which of course can be anything. It consists of both mineral grains and rock fragments.
- It *tends to be poorly sorted*. A glacier is indiscriminate in terms of the particle sizes it carries, so deposits directly from the glacier are likely to be very poorly sorted. Glacial deposits are among the least well sorted of all sediments. But if the material is reworked by water or wind, it can end up being fairly well sorted. (But it's still called glacial sediment if it's recognizably related to glacial action.)
- Particle shape is sometimes characteristic: multifaceted "flatiron" shapes are common among the larger, gravel-size clasts. This characteristic shape is caused by abrasion while in successive orientations as a tool at the base of the glacier. Only some, not all, of the large clasts show this characteristic shape.
- The larger clasts are often striated, just like the underlying bedrock. Beware, however, that other agents of transport (like debris flows) can produce striations on gravel-size clasts.

Notes on Classification of Glacial Deposits

It's important to understand that classification of glacial deposits is two-fold:

- texture of material in general
- form or geometry of deposit

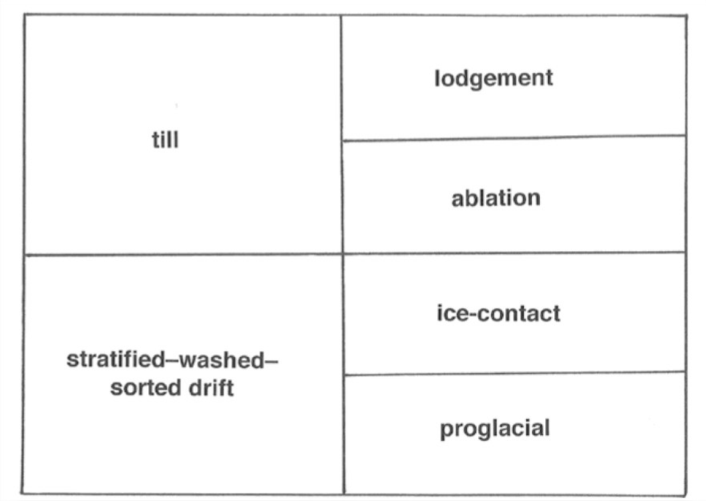
Composition is typically not taken into account in classification.

There's a partial but not nearly total correlation between these two bases for classification. It's common to have the same textural kind of deposit involved in more than one morphological kind of deposit, and the same morphological kind of deposit showing more than one textural type—thus the need for separate ways of looking at glacial deposits. (One could argue that this could be the case for sediments in general, but usually we can't study them together as suitably as in glacial deposits.)

Glacial Drift

Glacial sediments have long been called ***drift***. That word dates from before the glacial theory, when it was thought that this characteristic material was deposited by flowing water ("drifted" in). Glacial drift is *all material in transport by glacier ice, all deposits made by glacier ice, and all deposits predominantly of glacial origin even though not deposited directly by a glacier*. Glacial drift is thus highly varied in texture, composition, deposit morphology, and origin. In the following I'll discuss the nature of the material first, and then the deposit geometry later.

Glacial drift is usually subdivided into two major categories*:* ***till***, on the one hand, and material variously called ***stratified drift***, ***washed drift***, or ***sorted drift***, on the other hand (Figure 7-46). (A really good term for this second kind of drift has not yet been invented.)



till	lodgement
	ablation
stratified–washed–sorted drift	ice-contact
	proglacial

Figure 7-46. Classification of glacial drift.

The distinction between till and s/w/s drift is basically *descriptive*: till is poorly sorted and largely nonstratified, and s/w/s drift is much better sorted and characteristically well stratified. But the distinction is universally viewed as having a clear *genetic* basis: till is deposited directly from glacier ice, without the effect of flowing meltwater, whereas s/w/s drift is material that has been picked up by flowing meltwater and redeposited somewhere else.

Till

General

Till is a genetic term applied to *all unstratified and unsorted deposits made directly by or from glacier ice*. The lithified equivalent of till, as seen in the ancient sedimentary record, is called ***tillite***.

Till is subdivided in turn into lodgement till and ablation till (Figure 7-47). ***Lodgement till*** is *deposited directly from the moving ice beneath the moving glacier*, whereas ***ablation till*** is *deposited at the glacier terminus as the ice melts and drops its load*. The distinction between lodgement till and ablation till is thus a genetic one.

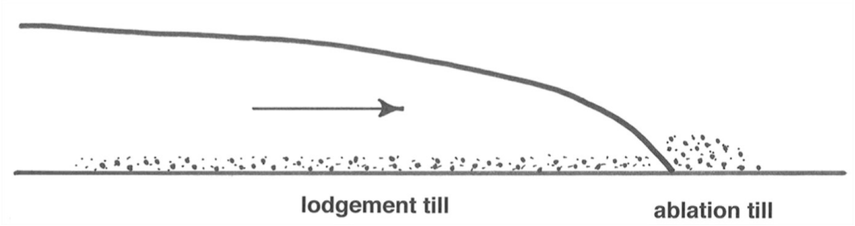


Figure 7-47. Lodgement till and ablation till.

Lodgement till

Classic lodgement till is a very characteristic sediment:

- The particle size ranges from large boulders continuously down to clay- size material (mainly ***rock flour***: *ground-up very-fine-grained mineral material*) Tills are often described as *boulder clays*.
- Typically the larger clasts "float" in the matrix, so the sediment could be called *matrix-supported*.
- There's usually *no stratification*: till is typically a structureless deposit.
- Often there's a subtle statistical *preferred orientation*, or fabric, especially of the larger clasts; the long axes tend to be oriented parallel to glacier flow.
- The sediment tends to be *cohesive* even if it's young; the degree of compaction is such that often the stuff has to be dug with a pick instead of a shovel. Tills, even young ones, are often lithified enough to be jointed.
- It has low porosity and very low permeability.

In terms of genesis, it seems clear that till like that described above is deposited particle by particle from the base of an active glacier. Till known or thought to have been deposited in this way is called ***lodgement till***; remember that this is a genetic term.

Apparently it’s common in the peripheral parts of an ice sheet for the glacier to be overloaded at its base, so instead of the glacier picking up material, it’s depositing material. But the relative importance of the various deductively reasonable mechanisms of lodgement are largely unclear.

Here are a few comments on the mechanics of subglacial deposition. Three mechanisms of subglacial deposition can be envisioned:

- Pressure melting at the base of the sediment-laden ice releases particles at a rate faster than the rate at which comminution by abrasion plus removal by meltwater can operate, and the excess particles left beneath the ice are plastered onto the depositional surface.
- Plastering happens because the frictional drag on sediment particles in movement over the sediment becomes equal to the tractive force of the glacier ice. This might happen where the ice is decelerating and the friction is decreasing.
- Gradual reduction of basal ice velocity leads to shearing over successive layers of ice–debris mix at the base, and then the water is eventually removed by melting.

There have been just a few direct observations of lodgement processes at the base of a glacier. These indicate that at least the first two processes actually happen. And there’s no reason to believe that the third can’t happen as well.

Another important thing about lodgement till is that under some conditions it can undergo flow by shearing after it’s deposited. The moving glacier above the sediment exerts a shearing force on the upper surface of the deposit, and if the shearing force exceeds the shear strength of the material, the till flows. Here’s a qualitative treatment of the forces involved in such flow.

The shear strength of a continuous granular medium like till is a reflection of the frictional resistance the material affords to shearing. Think of the resistance to shearing as a matter of friction: the frictional stress, which tends to make the material shear along shearing planes, can be viewed in the classical way as the product of the normal stress exerted on the plane of shearing and the coefficient of sliding friction. The high water content together with the low permeability of typical lodgement till beneath the glacier means that the effective normal stress on the shearing planes is greatly reduced by the water pressure in the pores, because the water pressure lessens the particle-to-particle contact forces. If that doesn’t make sense to you, just think in terms of *buoyancy* or *submerged weight*: fill a drinking glass with loose sand and think about how much the grain- to-grain contact forces are lessened when you fill the glass with water interstitial to the sand grains.

The slow flow of subglacial lodgement till has actually been observed directly in a few cases. All you have to do is drive a tunnel under the base of the ice (and keep it open!), and then drive a segmented rod down into the till. Come back sometime later, excavate the rod, and see how its little segments have moved in the downglacier direction. Figure 5-48 shows an actual measurement of this kind.

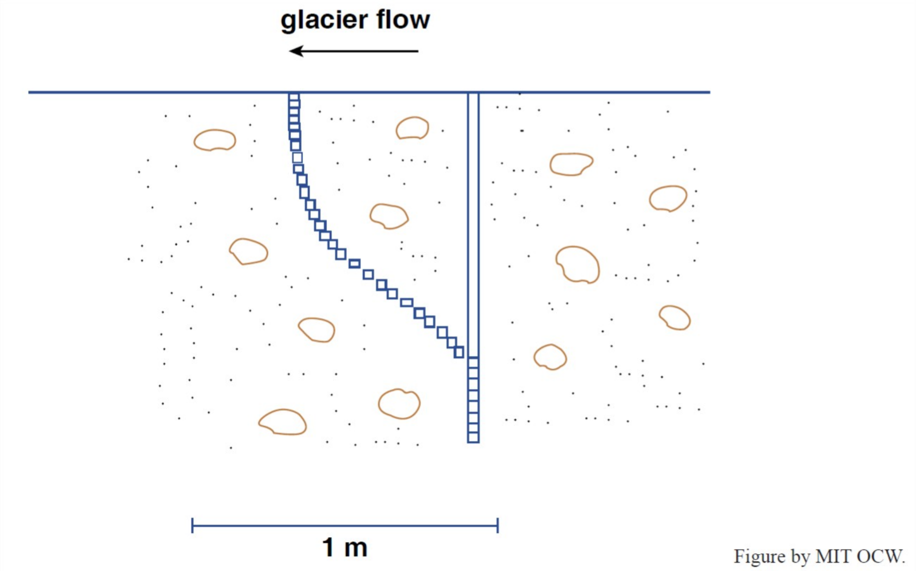


Figure 7-48. Slow flow of subglacial lodgement till.

Ablation till

The other kind of till, called ***ablation till*** (also a genetic term, remember) is easier to understand. It’s *deposited right at the terminus of the glacier, in the process of ablation there*, not underneath the glacier. This is one of the things that can happen to the load of the glacier that's delivered all the way to the terminus rather than being lodged beneath the glacier.

Although I've never seen any definitive pronouncements, my guess is that more sediment is delivered to the terminus than is extracted from the glacier from below, upglacier of the terminus. This material is simply released from the glacier ice as the ice melts—hence the term ablation till.

The important thing about ablation till is that it's not as rich in very fine material as lodgement till, because the liquid water melted out of the glacier tends to carry that material away in suspension, leaving behind the coarser material, of gravel, sand, and silt size (along with some fraction of even the finest material as well). So ablation till is more friable and easier to dig than lodgement till. It's not difficult to tell the two apart in this way.

You may be thinking that what I've said about ablation till makes it seem closely related to, or indistinguishable from, what I called washed drift. The important point is that the material that was deposited to form the ablation till—what's there for you to see—was deposited directly from glacier ice; only the finest fraction is winnowed by the meltwater and deposited somewhere presumably far away. Of course, there are likely to be gradations between ablation till and washed drift, depending on how much meltwater happens to be around and flowing in the given locality, and therefore on how much of the coarser material is transported and redeposited.

Two-Layer Till Sheets

From what was said above about the deposition of till, it should make sense to you that till typically appears as extensive sheets and blankets that mantle the landscape continuously or semicontinuously. (The mantle of till tends to be thicker in valleys than on hills, and thicker on the lee sides of hills than the stoss sides.) The thickness of till sheets ranges from less than a meter to many meters, and in some cases even tens of meters. Till sheets take on a wide and interesting variety of geometrical forms; I'll elaborate in a later section.

Commonly, though not nearly always, one sees a ***two-layer till sheet***: *a lower layer of lodgement till, and an upper layer of ablation till*. The upper layer of ablation till tends to be spottier and less continuous than the lower layer of lodgment till, but the two-layer structure is nonetheless characteristic.

It's easy to understand how the two-layer arrangement of till comes about (Figure 5-49). As the glacier advances, it's likely to form an underlying sheet of lodgement till. Then, during retreat, ablation till deposited at the retreating terminus of the glacier is spread backward as a sheet across the underlying sheet of lodgement till.

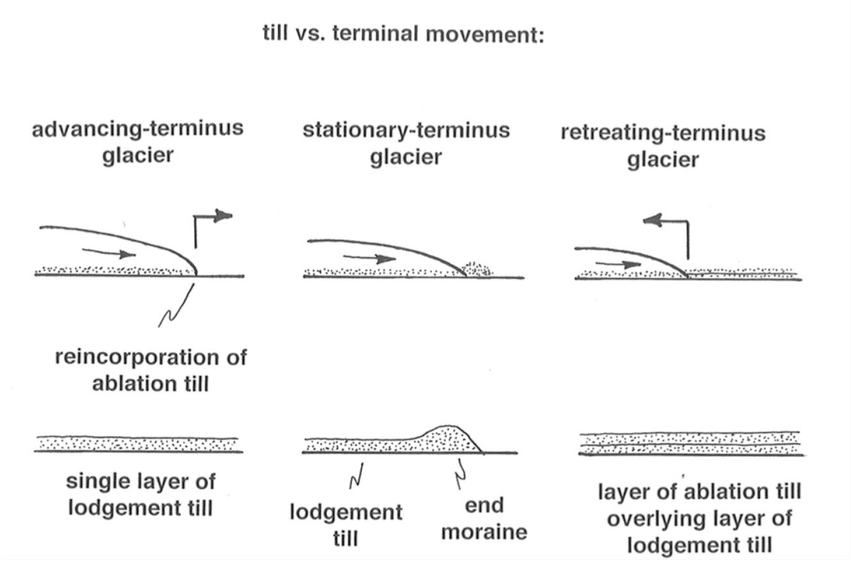


Figure 7-49. Relationship between till deposition and terminus movement.

The Morphology of Till Sheets

When till sheets are viewed morphologically or topographically, they are usually termed ***ground moraine*** (note the without-the-article usage).

The topography of ground moraine varies widely (Figure 7-50). Often one sees a fairly regular series of ridges and furrows oriented parallel to the direction of ice flow. The relief of these ridges and furrows varies from less than a meter high and only a few hundreds of meters long, to giant features, up to a few tens of meters high and a few tens of kilometers long. The ridges are strikingly straight and parallel. This kind of topography can best be appreciated from the air or on air photos. Ground moraine showing such ridges and furrows is called ***fluted moraine*** or ***fluted ground moraine***.

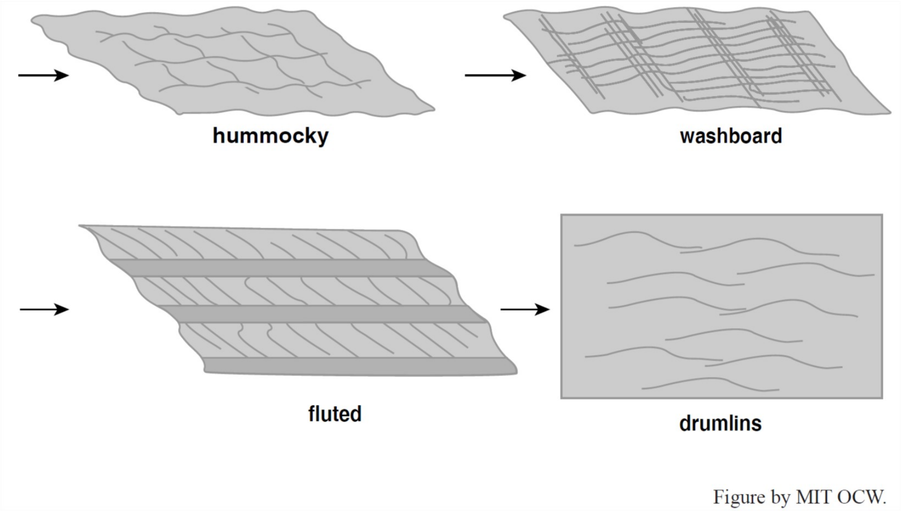


Figure 7-50. Varieties of surface topography of ground moraine.

How is the ridge-and-furrow topography produced? Most theories rely on lateral flowage of till beneath the glacier, somehow. The idea here is that till moves laterally in a semiliquid or plastic state subglacially in response to subglacial hydrostatic pressure field. For example, low pressure in the lee of boulders or rock knobs should produce flow into that area, and these low-pressure areas might tend to extend themselves downstream. All such explanations are highly speculative.

Drumlins are another very characteristic kind of ground-moraine topography. You've all seen drumlins: *glacially streamlined hills*, ideally almost half-ellipsoidal in shape. Drumlins are found in large numbers in certain areas, called ***drumlin fields***. Drumlins can be cored by either till or bedrock hills. Often the drumlin consists of a streamlined tail of till extending downstream from a bedrock knob. There's a whole spectrum of features intermediate between discrete and rather blunt drumlins to long and regular longitudinal ridges and furrows. Often what are called drumlins are very elongated hills that might better be called longitudinal ridges.

A great many theories on the origin of drumlins has been proposed, but none of them has become really widely accepted. The problem is the classic one that we can't observe them in the process of formation. Actually drumlins may be a polygenetic phenomenon. The usual approach is to invoke some kind of depositional response to subglacial pressure differences, either by flowage of till or preferential lodgment. Clearly a dynamical instability of some kind must be involved, or else the till surface would remain planar. But this kind of approach doesn't seem to be appropriate or necessary to explain rock-cored drumlins with just a veneer of till, especially if the drumlins are more or less isolated.

Just to confuse you further, let me point out that some kinds of ground moraine show transverse topographic elements. Such ground moraine is usually called ***ribbed moraine*** or ***washboard moraine***. It's clear that such topography is produced subglacially, not at the glacier terminus. But again, understanding is in a poor state: there are lots of theories, all of them speculative.

Finally, it's only fair to point out that much ground moraine shows no strongly organized topographic elements, just a seemingly random collection of high and low areas, often rather subdued. Apparently in many cases none of the various mechanisms leading to flow-parallel or flow-transverse features are at work.

Stratified Drift

Introduction

The term ***stratified drift*** refers to glacial drift that has been reworked by glacial meltwater and then deposited either in direct contact with glacier ice or at some point more or less far away from the glacier, in a wide variety of depositional environments. In a kind of order-of-magnitude sense, stratified drift is just as common and important as glacial till.

The term stratified drift is not an especially good one. What's called stratified drift is indeed usually prominently stratified, but the term doesn't do well at capturing the essence of the stuff. Alternative terms, perhaps somewhat better but not a lot better, are ***sorted drift*** and ***washed drift***.

Here are some miscellaneous important points about stratified drift:

• It's mostly sand or sand plus gravel; the fines are carried away in suspension.

• It occurs in large to small but basically *isolated deposits*. Sometimes it's in two-dimensional belts, but it's never in extensive sheets like till sheets. The main reason is that streams are by their nature localized.

• Stratified drift commonly forms deposits with considerable *relief*, so it's often seen as hills.

• It almost always shows excellent and striking *stratification*, because it's laid down by flowing water stratum by stratum, and conditions of deposition usually vary with time.

• It commonly shows *cross-stratificatio*n on various scales, because of the existence of ripple and dune bed configurations under a wide variety of conditions of flow and sediment.

• It commonly shows *deformation* features, because it's often deposited against steeply sloping surface of ice, and, when the ice inevitably melts, the sediment slumps and slides.

• It's an important *source of sand and gravel* for making concrete (in glaciated regions that is; in nonglaciated regions, sand and gravel is usually a lot harder to come by).

Watch for sand and gravel pits and quarries for examination of cuts through stratified drift. The trouble is, the good viewing doesn't last for long.

In this section I'll deal mostly with the mechanics of transport and deposition of stratified drift, and the characteristic form or geometry of the drift deposits.

Classification

It's not easy to classify stratified drift, because there's basically a continuous gradation among the various types. But there are some characteristic types, and we'll deal with these.

There are two major categories of stratified drift:

• **ice-contact stratified drift:** deposited in direct contact with, or at least in the immediate proximity, of, glacier ice, active or inactive.

• **proglacial drift:** drift that is carried out well beyond the glacier terminus by flowing water. This passes over into sediment that is not recognizably of glacial origin.

Ice-Contact Stratified Drift

First off, I want you to have two rather different mental pictures of a retreating glacier, which is ultimately going to leave the stratified-drift deposits we see (Figure 7-51). I'll give call two contrasting modes ***active-ice retreat*** and ***stagnant-ice retreat***. (These terms are expressive but not really official.)

• **active-ice retreat:** there's a fairly steep ice slope in the downglacier area; the ice flows actively all the way to the terminus; there's a well defined terminus even though the glacier is retreating; meltwater discharges directly off or from the glacier and forms a well defined stream, except when the terminus lies in a glacial lake.

• **stagnant-ice retreat:** the glacier surface has a fairly gentle slope near the terminus; excessive ablation reduces the thickness of the glacier over a substantial distance, thereby giving rise to a broad belt at the downstream end of the glacier where the ice can no longer flow; the dead melting ice then sits around over a large area, and the active terminus is now located far upglacier; meltwater streams flow between, around, under, and over this dead ice, depositing stratified drift, and the dead ice sheds sediment as it melts as well.

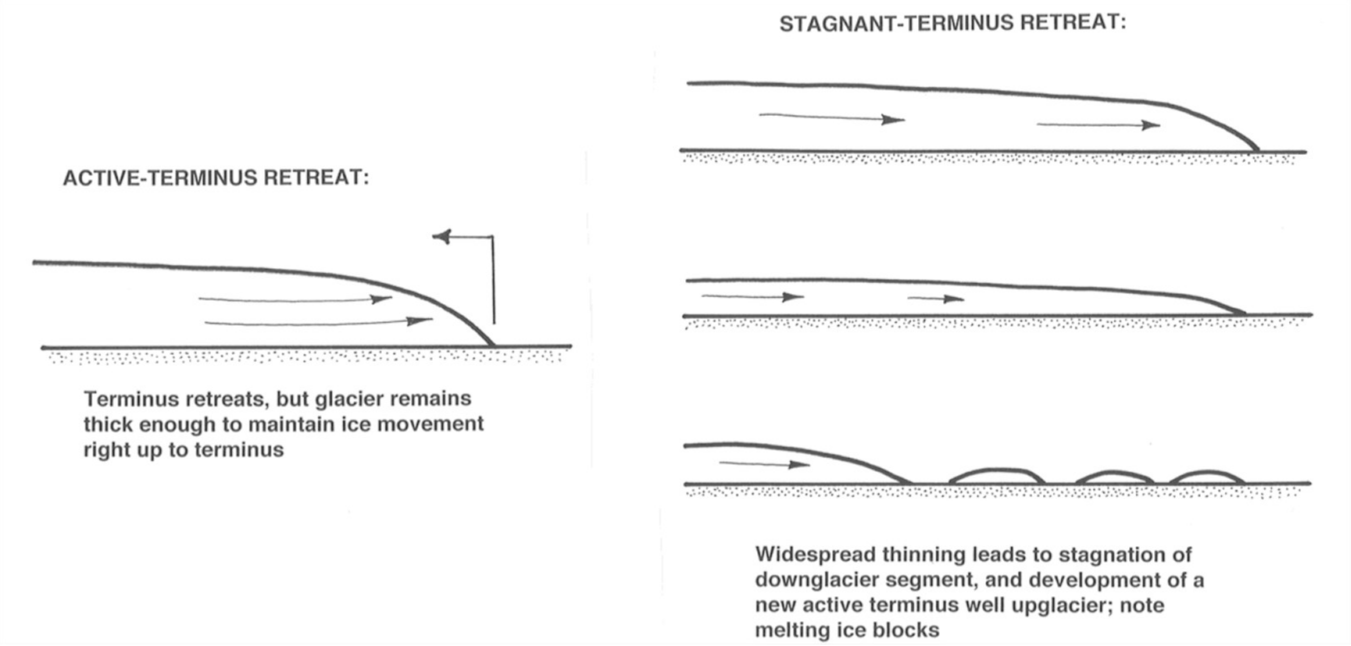


Figure 7-51. Active-terminus retreat vs. stagnant-terminus retreat.

An important point: most retreating glaciers in modern times show active-ice retreat rather than stagnant-ice retreat, either because they are valley glaciers or outlet glaciers, which usually remain well integrated even during retreat, or, if they are ice sheets with broad terminus fronts, they calve into the ocean. But stagnant-ice margins seem to have been very common at times of rapid ice-sheet retreat at the end of the Pleistocene, because there were widespread ice sheets on flat, low land in temperate regions. So we suffer somewhat from a lack of highly realistic modern models for deposition of the stratified drift we see resting on the landscape today.

Think first about deposition of sediment in meltwater streams before those streams leave the glacier.

• *supraglacial streams* could be carrying sediment that’s deposited as a bed and then abandoned when the stream disappears; the sediment would then eventually let down onto the land surface by melting. Apparently this is not common.

• *englacial streams*: same thing.

• *subglacial streams*: these are by far the most important in carrying sediment, both because streams tend to migrate to the base of the glacier and because that’s where most of the sediment is carried by the glacier.

Subglacial streams certainly deposit a lot of sediment as they leave the glacier. But how about deposition upstream of the glacier terminus? There may or may not be deposition there, depending upon the particular relationship between sediment load and carrying capacity. Figure 7-52 shows two possibilities. Sketch A shows a subglacial stream with uniform transport capacity; all the sediment is carried through over bare rock, and no sedimentary record of the tunnel remains after melting of the glacier. Sketch B, however, shows a situation in which sediment transport rate decreases downstream for some reason, to leave a deposit on the floor of the tunnel. In this case, upmelting of the tunnel roof is likely, to make room for the deposit as well as the flow. A sediment ridge that follows the course of the stream is left after the glacier melts.

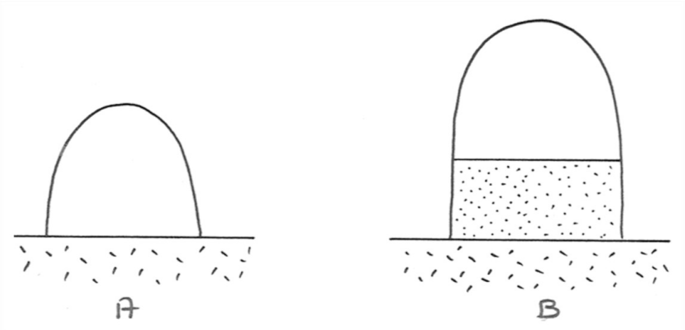


Figure 7-52. Subglacial stream tunnels A) without and B) with a sediment bed.

Remember that flow in subglacial tunnels is mechanically like pipe flow, in that there's likely to be no free surface, except very near the glacier terminus or at times of unusually low meltwater discharge. The floor of the tunnel is either bedrock or a sediment bed, and the walls and roof are of glacier ice.

What does the cross section of the subglacial flow tunnel look like? A natural first guess would be semicircular, with a horizontal planar floor. But if there's a full bed of sediment, it's possible for the floor to be either concave upward or convex upward instead of planar (Figure 7-53). The meager evidence from the exhumed record of ice-tunnel deposits suggests that the sediment bed is planar or convex upward, because the stratification overall tends to be as shown in Figure 7-54.

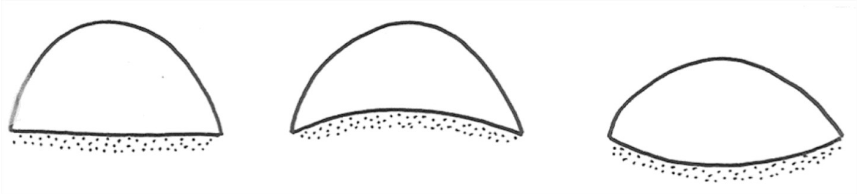


Figure 7-53. Possible cross-sectional shapes of subglacial ice tunnels.

Figure 7-54. Planar and convex-up deposits of subglacial ice tunnels.

There's a problem with this interpretation, however, because we never have direct independent evidence that the flow was really in a closed conduit. What about the possibility that the stream was flowing open to the sky in a canyon between two ice walls? That would always lead to a transversely horizontal bed rather than a convex-up bed.

One can account hydrodynamically for a transversely arched bed by appealing to *secondary circulations* in the flow (Figure 5-55). Such secondary circulations are known to be present in closed conduits of noncircular cross section. There are various theories on the origin of such secondary circulations, but they are not yet well understood. The circulation pattern shown in Figure 7-55 would lead to a component of sediment movement on the bed from the walls to the center, until an equilibrium transverse profile is reached in which there's a balance between this tendency toward upslope transport and the pull of gravity back down the slope toward the corners of the cross section.

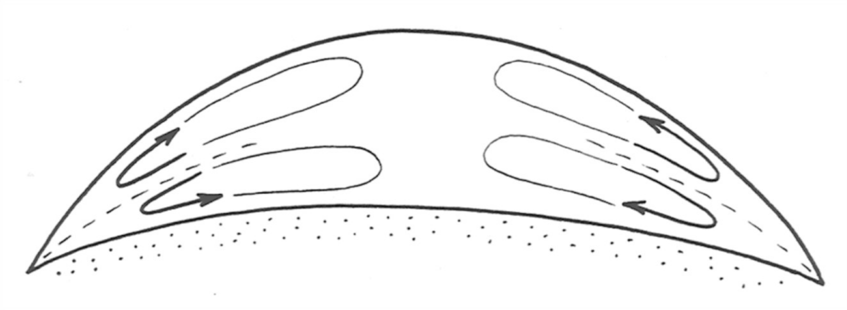


Figure 7-55. Secondary circulations sin flow in subglacial ice tunnels.

Deposits of the kind formed in ice tunnels are called ***eskers***, a nongenetic term for *any long and more or less isolated ridge of stratified drift, oriented at a small angle to the overall direction of ice flow, and presumably deposited by meltwater beneath or at the terminus of the glacier*. Most eskers seem to have been deposited by meltwater streams either within the glacier, as discussed above, or as they emerge from beneath the glacier.

Eskers have rather variable features, probably reflecting a polygenetic origin (Figure 5-56). Here are some common characteristics of eskers:

- kilometers to well over 100 km long
- up to a few tens of meters high, and up to a few hundreds of meters wide
- either flat-topped or arched
- almost always all sand and gravel, well stratified; no till

- the sediment is often slumped at the margins
- often sinuous, like streams
- sometimes show tributaries and branching
- sometimes pass over bedrock divides (by as much as 200 m!) at the lowest saddle point
- sometimes end in what look to be fans or deltas
- show evidence of high-velocity flow

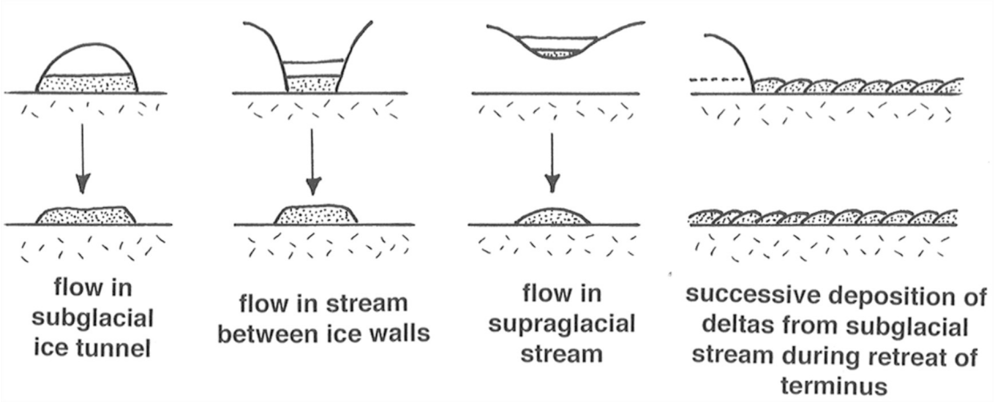


Figure 7-56. Possible mechanisms for development of eskers.

Many, maybe most, eskers seem to have been deposited by subglacial streams in closed flow; this accounts well for their passing over bedrock divides, for the arched bedding, and for the existence of tributaries. But of course an open-to-the-air esker formed on glacier ice could have been let down onto the bedrock divide by melting of the ice.

Many, but certainly not most, eskers must have been deposited at the downstream end of a meltwater stream where the stream reached the glacier terminus. The long length of the esker is then accounted for by year-by-year shifting of the site of deposition upglacier, as the terminus retreats. Some eskers are *beaded* (that is, they consist of a chain of closely spaced mounds of glacial deposits), and this mode of origin accounts well for the beading. Keep in mind that whatever the mode of origin the long length of eskers is accentuated by the retreat of the glacier.

Aside from eskers, many kinds of deposits are left by melting of ice at the terminus of a retreating glacier. These might generally be called ***ice- disintegration deposits***. These are suites of interrelated but distinctive depositional features produced by wasting of the thin, stagnant, marginal part of a glacier, mostly by separation of the ice into isolated masses. The marginal zone of the glacier in which such features are produced might be as much as ten kilometers wide. Instead of being carried to an active terminus, the drift accumulates by various processes associated with running water, in channels or in lakes, and in various positions relative to the melting ice: in channels or openings between or beneath ice blocks, and over the surface of wasting ice.

The *surface morphology* of these deposits varies widely. There’s a mostly continuous spectrum or gradation of named features: *kames*, *kame terraces*, *kame deltas*, *collapsed masses*, and *eskers* (already discussed). All of these can be observed forming today near the termini of glaciers that are retreating in an inactive-terminus mode. One of the best examples is the Malaspina piedmont glacier in Alaska.

Kame terraces. In areas of active meltwater production in valley glaciers, meltwater streams often flow down the valley along the sides of the glacier. This often leaves linear deposits of stratified sand and gravel on the bedrock slopes of the valley after the glacier melts (Figure 7-57). Often these deposits are at more than one level, as a function of different stillstands of the terminus. It’s easy to confuse kame terraces with stream terraces in a nonglaciated valley.

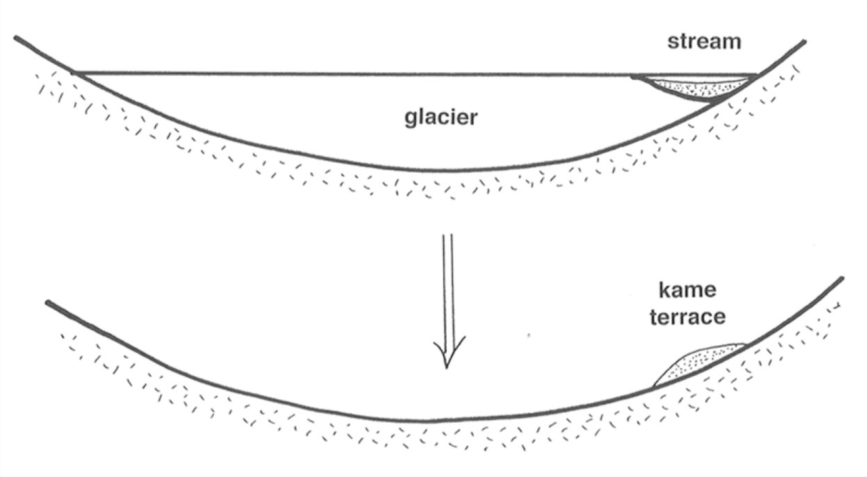


Figure 7-57. Development of a kame terrace.

Kames. The term ***kame*** is used for a whole variety of hills, knobs, and ridges (positive relief forms), more or less isolated from one another, of stratified sand and gravel *deposited against or adjacent to wasting stagnant ice* (Figure 5-58). The sediment may be derived in part from the adjacent ice itself, but the larger kames are probably formed by deposition of sediment carried from upstream by meltwater streams. Kame deposits are characteristically in part deformed, because of the collapse of sediments when the supporting ice melts away.

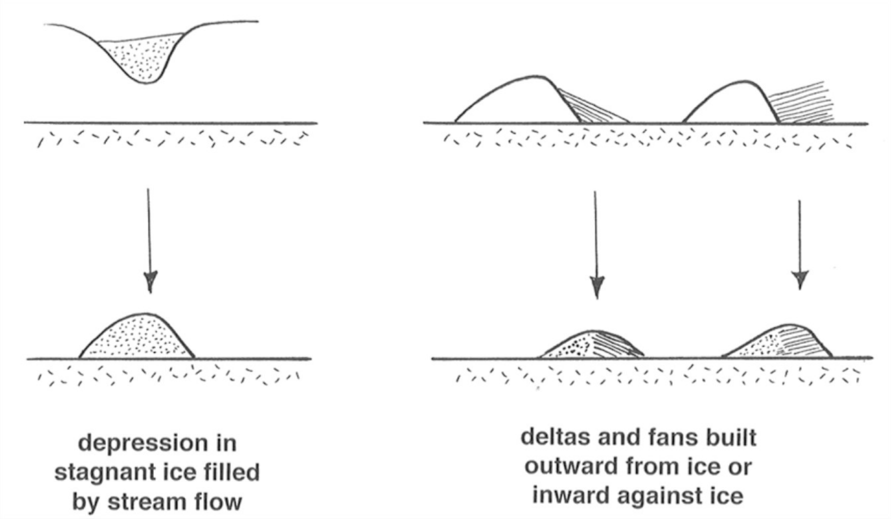


Figure 7-58. Development of kames.

Kame Deltas. Small glacial lakes are common right next to melting glacier ice, because of local damming of meltwater drainage by either the ice itself or sediment deposited by the ice. Streams flowing between, or from the surfaces of, the melting ice masses tend to deposit deltas in the lakes. After the ice melts, the delta forms an isolated and often flat-topped mass of sediment (Figure 7-59).

Collapsed Masses. Another common ice-disintegration feature, for which there is no very felicitous term, consists of irregular masses of sediment deposited on top of wasting ice and then let down onto the land surface as the ice melts. The so- called collapsed masses tend to be blanketlike, but they are highly varied, with sharp variations in thickness. This is sort of a “wastebasket” category.

Kettles. ***Kettles*** are *depressions formed when a thick layer of drift is deposited around or over an isolated ice mass and then the ice later melts* (Figure 7-60). If the groundwater table later lies above the floor of the depression, the kettle has a lake within it, called a ***kettle lake***. Kettles often are present in large numbers in areas of outwash plains; such an outwash plain is called a ***pitted outwash plain***.

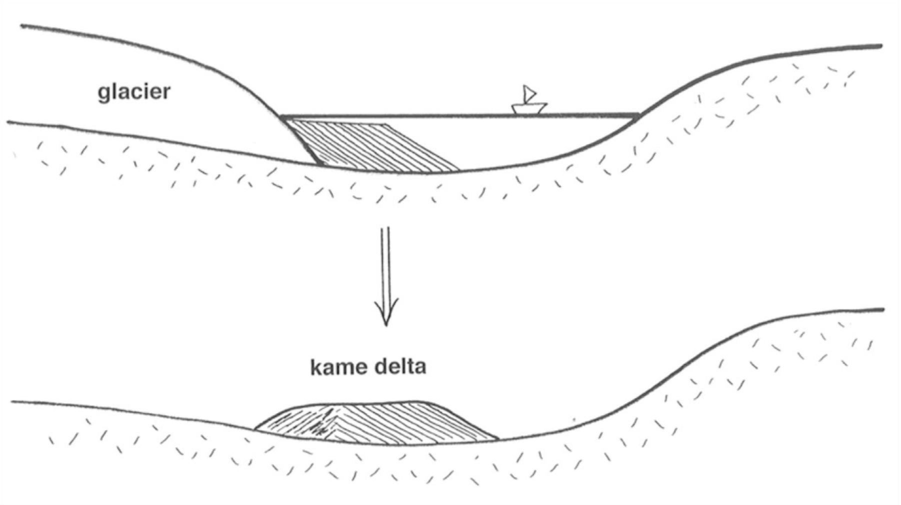


Figure 7-59. Development of a kame delta.

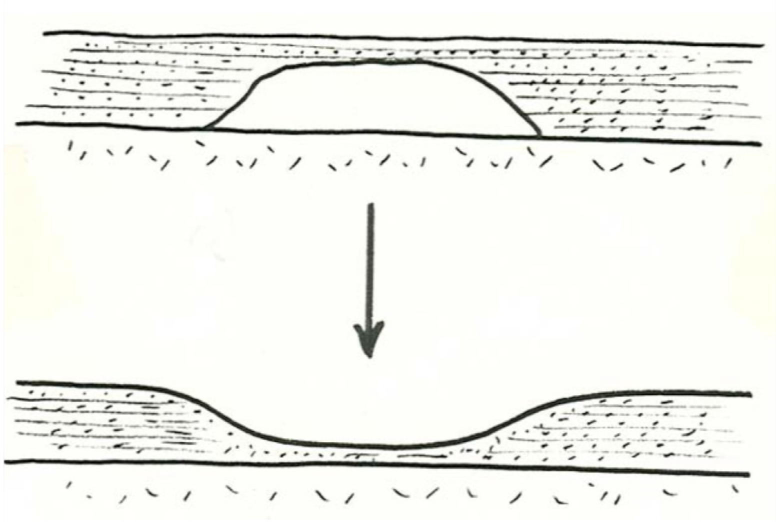


Figure 7-60. Development of a kettle.

Proglacial Drift

Proglacial drift is drift deposited in various ways, by water or even by wind, more or less far removed from the glacier. The idea that proglacial drift is abundant makes sense. Look at a typical active warm glacier. There's abundant meltwater, so large meltwater streams originating well upglacier from the terminus are important during the melting season. These streams characteristically carry high sediment loads. In such situations, most of the drift deposited by the glacier is deposited downstream from the glacier terminus. This happens whether the glacier is advancing, retreating, or stationary; if the glacier is advancing, then the glacier later overrides some of its own proglacial drift. Here are the important kinds of proglacial drift:

- outwash (glaciofluvial)
- lake deposits (glaciolacustrine)
- marine deposits (glaciomarine)
- loess (wind-blown silt)
- sand dunes (usually reworked outwash)

Glaciofluvial deposits (these are also called glacial ***outwash***)

- Mainly *sand and gravel*; fine material (silt and clay) is either carried farther downstream into areas not recognizably glacial, or blown away by the wind

- Essentially *fluvial* in nature; channel pattern and sedimentary structures are not directly or recognizably glacial

- Typically shows *rapid downstream decrease in grain siz*e, because of overall aggradation

- Outwash streams are typically *braided* (factors: high sediment load; no bank stability)

The typical form taken by outwash is a fan or cone of fluvial sediment, deposited with the apex or head at the ice margin. The reasons for fan development are basically the same as for alluvial fans in nonglacial situations: abrupt decrease in gradient as the stream leaves the glacier and abrupt relaxation of channel constraints leads to decrease in capacity and therefore deposition. Outwash deposits tend toward *two distinctive forms*:

- **Outwash plains:** broad linear front, many outwash streams, series of coalescing fans. This leads to deposition of a broad wedge or blanket, thinning and fining away from the glacier (Figure 7-61). Also called a ***sandur***, from the Icelandic (plural: ***sandar***). One of the best sandar is at the margin of Vatnayökull, the largest ice cap in Iceland.

- **Valley trains:** braided outwash filling a glaciated valley downstream of an active valley glacier (Figure 7-62).

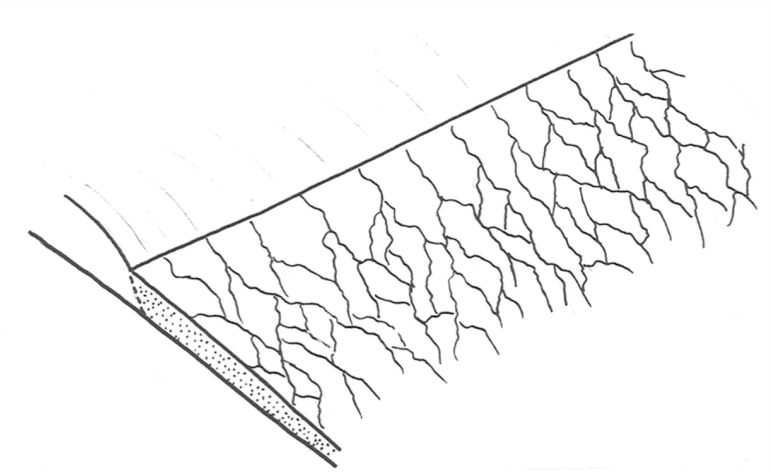


Figure 7-61. A glacial outwash plain (a sandur)

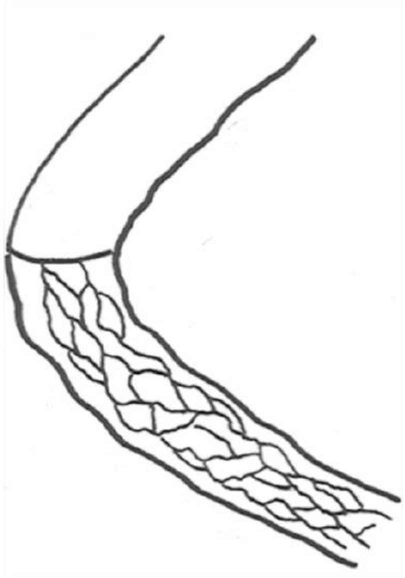


Figure 7-62. A valley train.

Commonly where both meltwater and drift are very abundant (as in wasting of a warm ice sheet in the Pleistocene), the outwash plain or fan is built up to be so thick that it is built headward over the terminus of the glacier itself and merges with ablation moraine. Often it buries stagnant ice near the terminus, leading to a pitted outwash plain as the buried ice melts. Earlier end moraines are often buried as well. This is well displayed on Cape Cod and Long Island.

Glaciolacustrine Deposits

Glaciers create three kinds of lakes:

- glacier-margin lakes dammed by the ice itself (short-lived)

- glacier-margin lakes dammed downstream by earlier-deposited glacial sediment (longer-lived but still temporary)
- bedrock depressions made by glacial erosion, later filled with water (long-lived)

Sediments deposited in glacier-margin lakes are very common, because valleys are dammed by ice or by ice-disintegration deposits.

- Probably *more common during retreat that advance*
- Invariably *temporary*, because ice dams melt soon, and glacial-sediment spillways are degraded by erosion
- Range in *size* from very small and lasting only years, to enormous, lasting thousands of years.

How does one explain the relatively long lives (some thousands of years) of some glacier-margin lakes? If the retreating glacier forms a long-term but moving dam in a valley, and drainage is over a rock divide, the lake remains either until the ice melts or until a sediment-formed divide is cut to the level of the bedrock spillway. Example: Glacial Lake Hitchcock in central Connecticut and Massachusetts.

Some points about sedimentation in glacial lakes (Figure 7-63):

- *Deltas* of coarse sediment are common
- Varves (*annual coarse–fine couplets deposited on the lake bottom*) are formed by differences in summer and winter suspended-sediment deposition (Figure 5-64).
- Dropstones are possible, from floating ice derived from the glacier.

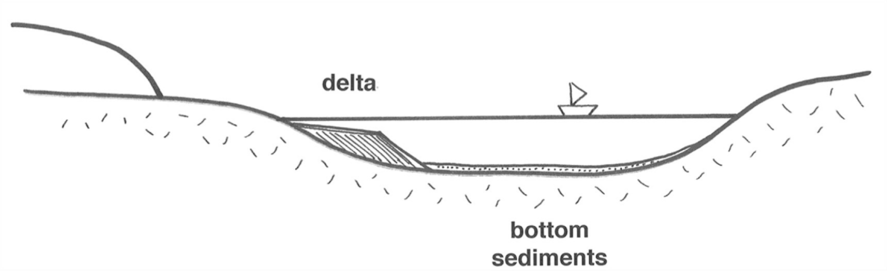


Figure 7-63. Sedimentation in glacial lakes.

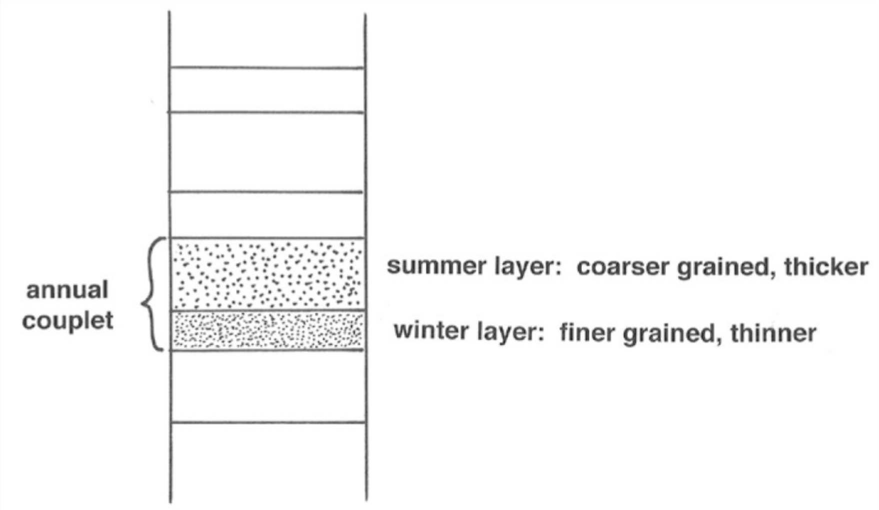


Figure 7-64. Glacial varves.

Loess

The term ***loess*** is used for *blankets of wind-deposited silt on the land surface*. Much but by no means all loess is derived from wind erosion of broad outwash plains marginal to Pleistocene ice sheets. Some is derived also from extensive wind erosion in large deserts, not associated with glacial deposits.

Loess is unconsolidated to semiconsolidated (by slight to moderate simple cementation), and usually buff to yellow to tan in color, reflecting an oxidized state. It is unstratified to only vaguely stratified. It has the interesting property of standing in vertical slopes even though it's easily dug with a shovel (interlocking angularity of grains, plus slight cementation?), and sometimes it even shows columnar jointing. It consists of relatively well sorted and angular grains usually in the fine silt to coarse silt range (average grain size 0.01–0.05 mm). Quartz is usually the dominant mineral. The lack of stratification is probably due to bioturbation by plants and animals, together with the relative uniformity of supply.

Loess forms blankets from less than a meter to many tens of meters thick (over 200 m in the central parts of China). Thickness is well correlated with grain size. In North America, loess is widespread in east-central to west-central US (Kansas, Nebraska, Iowa, Missouri, Wisconsin, Illinois, Indiana, Kentucky, Ohio) and also the Pacific Northwest. Coverage in North America: 1.6 x 10^6 km^2.

Glaciomarine Deposits

Icebergs produced by calving of large glaciers into the ocean often contain abundant drift, if the glacier is an active warm-based glacier. As the icebergs drift in the ocean and melt, they release this load, which settles to the sea floor along with fine sediment derived from elsewhere.

Glaciomarine deposits are characteristically well stratified but poorly sorted at the same time. The good stratification presumably comes about by annual and longer- term fluctuations in sediment supply from the icebergs. A distinctive feature of glaciomarine deposits is the presence of ***dropstones***: unusually large iceberg- derived clasts which bow down the sediment upon impact and which are then buried by later strata that arch over the dropstone.

READINGS

Benn, D.I., 1998, Glaciers and Glaciation. Oxford University Press (one of the best recent books on glaciers and glacial geology)

Flint, R.F., 1971, Glacial and Quaternary Geology. Wiley, 892 p (*the* classic book, outdated but never surpassed in its excellence)

Patterson, W.S.B., 1994, The Physics of Glaciers, Third Edition. Pergamon, 480 p. (the source on the physics of glaciers most accessible to the nonspecialist)

Strahler, A.N., 1966, A Geologist's View of Cape Cod. The Natural History Press, 115 p. (a small, readable book, old but still good, that deals in part in a very engaging way with the glacial geology of Cape Cod; long out of print)

Sugden, D.E., and John, B.S., 1976, Glaciers and Landscape. Arnold, 376 p. (a good but aging source, not highly technical)

CHAPTER OVERVIEW

8: Coasts

8.1: Introduction to Coasts

Introduction

The term *coastal studies*, which is in common use for a great variety of approaches to coastlines, covers a large area of endeavor. What do I mean by the term *coastal*? There are various definitions or interpretations, depending on how much or how little is included seaward and landward of the shoreline. The ***shoreline*** is fairly generally taken to be *the line or trace where the sea meets the land*, and this is fairly well defined except in areas with estuaries, tidal flats, etc., unless you want to quibble about where the high-water line is.

In this course I'm going to interpret the term ***coast*** in a fairly broad sense to mean *a marine area extending from the shoreline out onto the continental shelf for some distance, together with some land area immediately landward of the shoreline*. This admittedly sloppy definition succeeds in generally including areas both seaward and landward of the shoreline in which a great variety of processes operate that most people would categorize as coastal processes. I intend the definition to include only the innermost part of the ***continental shelf***, the part that's most strongly affected by the adjacent shoreline. (The continental shelf is the broad belt of shallow water adjacent to the coastline. In many areas of the world it is as much as a few hundred kilometers wide, and water depths at the shelf edge are no more than about two hundred meters.) In many coastal areas there's a well defined ***coastal plain*** that may be well over a hundred kilometers wide; only the part nearest the shoreline, directly affected by modern coastal processes, is included in my definition.

Coastal environments are unusually varied. If you let your mind go blank and I say the word "coast" to you, what image do you first conjure up? Rugged rocky sea cliffs? Wide sandy beaches? Shining coral reefs bathed in transparently blue water? These are only some of the many important coastal environments in the world today. How can I possibly deal with all these environments in one small part of this course, you might ask? Obviously I can't, so I'll concentrate on just a few kinds of coastlines and the most important processes that operate around them.

There's a lot of coastline in the world today: by one estimate, the coastlines of the world are almost 450,000 km long. One difficulty with an estimate like that is that the closer you look at a given stretch of coastline, the longer it comes out to be. On a large scale, that's because maps always involve a certain degree of generalization, depending upon the scale of the map. But the effect is still there even when you're looking at a small segment of the coastline right at your feet, because how do you take account of the outlines of the individual little sand grains?

Here's a list of the important factors that govern the nature of a coast, together with a few initial comments about each. Keep in mind that there's a rather strong interdependence among these factors.

- **sediment supply from land.** Clearly this is important in determining whether there is any sediment for the ocean to shape into beaches, deltas, barrier islands, etc. And the *size* of the sediment is important too: if only mud is supplied to the shoreline (as in certain low-latitude areas with hot and humid climate), you shouldn't expect to have sandy beaches!
- **climate.** The climate of the land area inland from the coast has various indirect effects on the nature of the coast: *size and supply of sediment* (see above), *river runoff*, and in some cases *glaciation*.
- **sea-level history.** Sea level hasn't stayed the same relative to the land: it rises and falls at rates ranging from something of the order of a millimeter per century, as a representative value throughout much of geologic time, to as much as a meter per century during rapid continental deglaciation, as happened in very recently in geologic time (within the last 20,000 years!). The nature of the coastline is not much affected by very slow changes in sea level, because coastal processes have plenty of time to equilibrate, but rapid changes in sea level are known to have strong effects on the nature of the coastline.
- **hydrodynamic setting.** By this vague term I mean the picture of *waves, tides, and currents* that affect the coastline. The relative importance of these three kinds of water motions varies greatly, and the nature of the coastline depends strongly on their relative importance.
- **tectonic setting.** In many areas the Earth's crust is unstable, and undergoes both vertical and lateral movements that can be very rapid on geologic time scales and can be substantial even on the time scales of long-term coastal processes. You have to think about the effect of tectonism on the climate and sediment supply landward of the coastline and on the submarine topography offshore, as well as the more direct effects on sea level at the coastline.

8.2: Classification

What are coasts like generally, and how can they be classified? I won't try to develop any formal or exhaustive classification here, but there are several important varieties I want you to be aware of and thinking about.

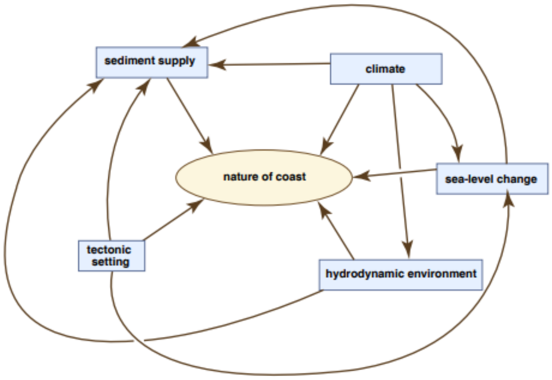


Figure by MIT OCW.

Figure 8-1 An attempt to show, in a simple diagram, the important relationships among the factors that govern the nature of coasts.

Look at the present shorelines of the world. What do you see?

Beaches. Beaches are perhaps the most common kind of coastline. It's easy to define a ***beach***: *an accumulation of loose sediment at the shoreline, shaped by the action of shoaling and breaking waves*. Sediment size ranges from fine sand to coarse gravel. Beaches range in extent from little pockets only meters to tens of meters long, to many hundreds of kilometers without a break.

Rocky or cliffy coasts. Along such coasts there's little or no loose sediment above mean low tide, so the bedrock of the continent is exposed to slow marine erosion. (The bedrock is usually resistant, or else the forces of the sea would produce enough sediment to form a sedimentary coastal environment like a beach.) There are all gradations from rocky coasts to beachy coasts.

Tidal flats. These are usually found in protected embayments or along coasts where waves are not important. They typically show an interlacing network of tidal distributary channels, with coarse sediment in the channels, passing into broad areas of fine sediment away from the channels.

Estuaries. An ***estuary*** is *a body of water with salinity lower than the open ocean, with restricted access to the ocean, where fresh water mixes with salt water*. There are several varieties, the main kinds these days being drowned-river estuaries and barrier-system estuaries. They seem to be unrepresentatively common today, presumably because of the effects of the most recent Pleistocene glaciation and deglaciation. Estuaries are often closely associated with tidal flats.

Deltas. A ***delta*** is *a body of sediment delivered to the coastline by a river, and built out into a body of water*, a lake or the ocean. Deltas show a great variety of geometries, and they can exist in combination with many of the other kinds of coasts.

Reefs. A ***reef*** is *a marine structure, built by organisms, that withstands the erosive action of waves*. Reefs are common in low-latitude areas with perennially warm water and low supply of siliciclastic sediment from land. They are often found in combination with carbonate-sand beaches.

Muddy open shorelines. These are not as common, and lie well outside the coastal experiences of most of us (including me). They are found mainly in the tropics in areas where substantial fine siliciclastic sediment is supplied from land but no carbonate sediment is produced near the shoreline, by reason of either too-cold water or swamping by the siliciclastics. (Why are they more common in the tropics? Presumably because of the nature of terrestrial weathering.) *Mangrove swamps*, important along some low-latitude shorelines, should be included here as well.

Coasts can also be classified into three kinds on the basis of *the effect of sea-level change*:

Emergent. Sea level *falling* relative to the land. These exist today, but they are not common. They are found only in areas where tectonic uplift has more than offset the postglacial sea-level rise.

Submergent. Sea level *rising* relative to the land. These are very common now, because of the recent rise in sea level.

Stable: These are not common now, but they must have been throughout much of geologic time.

Keep in mind that this is an atypical time, because of the large changes in sea level worldwide since the disappearance of the last great Pleistocene continental ice sheets. Sea level rose to nearly its present level from minus 120–130 m between 20,000 yr BP (years before the present) and now, and most of that was between 18,000 yr BP and 6,000 yr BP!

Coasts can be usefully classified on the basis of their *plate-tectonic setting* into leading-edge coasts and tailing-edge coasts:

Trailing-edge coasts. These are tectonically undifferentiated, once sedimentation has become established well enough to mask the complexities of initial rifting; slow subsidence, relatively low sediment supply, wide shelves; the east coast of North America is a good example.

Leading-edge coasts. These show varied uplift and subsidence; topographically differentiated (basins and ridges both offshore and onshore); relatively high sediment supply, narrow shelves; the west coast of North America and the west coast of South America are two different examples.

Finally, coasts are naturally classified on the basis of *hydrodynamic setting* into four basic groups, obviously with intergradations among them:

Tide-dominated coasts. Large tidal range; day-to-day sediment movement by strong tidal currents overshadows the effects of important but infrequent storms.

Wave-dominated coasts. Large breaking waves from distant storms overshadow the effects of important but infrequent nearby storms.

Storm-dominated coasts. Waves and currents produced by large coastal storms overshadow other effects.

Current-dominated coasts. Throughgoing strong ocean currents impinging from the adjacent deep ocean dominate sediment movement and shaping of the coast. (These are the least common of the four kinds.)

8.3: Tides and Tidal Currents

Introduction

Go to any seacoast and build a tide gauge to obtain a record of sea level as a function of time (over a period of weeks or months) by somehow damping out or averaging over or filtering out the effects of waves on time scales of seconds and storms on time scales of days to weeks. What would you observe? In most places, you would find regular and systematic fluctuations in water level with dominant "periods" of about half a day or about one day, together with more subtle longer-term patterns on time scales of months and even years.

And if you then set up a similar tide gauge in a different place, you would find the same general effect (and the same frequency) but with differences in details of the shape of the time record, and also differences in timing of highs and lows.

As you all know, these effects you're observing are the tides. It doesn't take much brilliance on anybody's part to relate the tides to the moon, because of identity of tidal periods and lunar periods. This was known in ancient times, and probably in even more distant prehistoric times. (It might trouble you, though, that the times of high and low tides are not coincident with the times when the moon is overhead; more on that later.)

Here's the plan of action for this section on the tides:

- Forces involved in generating the tides, and the origin of the tidal bulge
- How the Earth's rotation under the tidal bulges accounts for the observed tidal variations
- Forced-wave nature of the tides, which accounts for the phase lags observed
- Complicating effects needed to account for real tides: standing waves and Coriolis effects
- Tidal currents: periodic motions of the water in conjunction with the rise and fall of the tides

How Forces in the Earth–Moon System Cause the Tidal Bulge

First, here are some comments on the Earth–Moon system, as an astronomical entity governed by the laws of mechanics. Usually one thinks of the Moon revolving about the Earth once in about each month. But it's better to think of the Earth and the Moon as revolving around a single point, the center of mass of the Earth-Moon system. The Earth–Moon system together forms a mass unit, and it's a well known principle in dynamics that the motions of this unit with respect to external forces can be analyzed in terms of an equal mass concentrated at the center of mass of the system. This is not the place for details, but here's a homey example: throw a skinny dumbbell up in the air, and watch the way the center of mass follows a parabolic trajectory, just like a concentrated mass point, even though the dumbbell is twirling wildly. Because the Earth is much more massive than the Moon, the center of mass of the Earth–Moon system is actually *inside the Earth*, about 4600 km from the center (Figure 6-2)!

How do the Earth and the Moon stay in equilibrium? The rate of revolution of the two bodies about each other is such that the overall gravitational force of attraction is just in balance with the overall centrifugal force that tends to make the bodies fly apart from one another.

Given this *overall* balance, however, the *local* balance between centrifugal forces and gravitational forces differs from point to point within the Earth, depending upon the position the point relative to Earth–Moon axis. It's this differing balance that causes the tides, by virtue of creating *a bidirectional tidal bulge* that draws the waters of the ocean out both in the direction of the Moon and in the direction away from the Moon, relative to directions normal to the Earth– Moon axis (Figure 6-3). (The solid Earth undergoes the same effect, as what are called earth tides, but because the solid Earth is a lot more rigid than the waters of the oceans the Earth tide is far smaller, only of the order of ten centimeters, than the ocean tide, which is a few meters, more or less.)

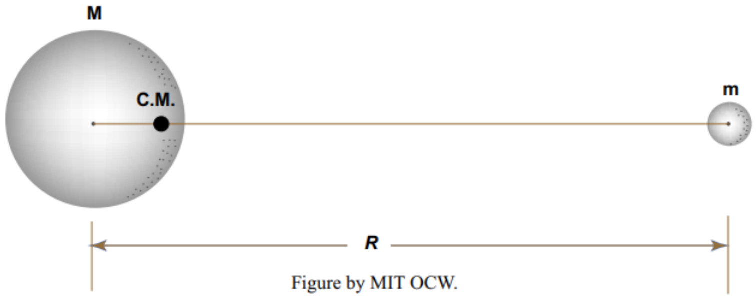


Figure by MIT OCW.

Figure 8-2. The Earth–Moon system.

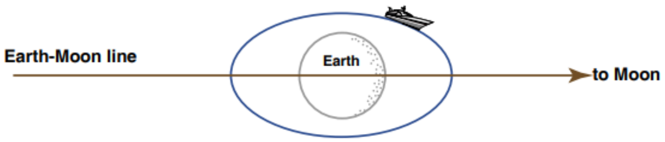


Figure by MIT OCW.

Figure 8-3. The double tidal bulge.

I imagine that the existence of this double-sided tidal bulge seems counterintuitive to you: *why not just in the direction of the Moon?* The following paragraphs attempt to provide a qualitative explanation, which itself will probably not seem immediately obvious to you; if you would like to see a more rigorous explanation, see the “advanced topic” section below.

The basic idea has to do with the relative magnitudes of two forces: the gravitational attraction of the Moon, and the centrifugal force on the Earth. As noted above, overall these two forces have to be in balance, or else the Earth– Moon system would not be in stable equilibrium with each other. The fundamental point is that, despite this overall balance, the relative magnitudes of the two forces vary from point to point within the Earth. The force of gravitational attraction is straightforward, but the centrifugal force needs some careful explanation. To help you understand it, try simulating the Earth–Moon system with your fists, as follows.

Make two fists. Hold the left fist pointing upward, and the right fist about a foot away, pointing downward. Your left fist represents the Earth, and your right fist represents the Moon. You can easily make your right fist revolve around your left fist in a circular orbit. But to simulate the motion of the Earth– Moon system, you need to move your left fist at the same time in a much smaller circular motion in the same sense of revolution, in such a way that the center of mass of the Earth–Moon system, which lies just inside your left fist, stays in the same place, relative to the floor below you. (It’s easier to do than to describe or to illustrate; I’ll demonstrate in class.)

Important: ignore the Earth’s rotation, because that’s not relevant to the origin of the tides. Concentrate on the circular revolution of the Earth (your left fist). Every point in your left fist is experiencing a centrifugal force, because of the circular motion, and that centrifugal force is the same at every point in your fist. Moreover, that centrifugal forces is always directed opposite to your right fist. (To understand that, you have to be doing the movements right.)

At the same time, every point in your left fist is experiencing a gravitational attraction from your right fist. That gravitational attraction is always directed toward your right fist. So at every point in your left fist there are two opposing forces, which are almost in balance. In contrast to the constant centrifugal force, the gravitational attraction of your right fist is slightly greater on the side of your left fist nearest your right fist and slightly smaller on the side of your left fist farthest from your right fist.

Now think about the relative magnitudes of the centrifugal force and the gravitational force. On the side nearest your right fist, the gravitational force is slightly larger than the centrifugal force, resulting in a net force directed outward from the surface of your left fist. On the side farthest from your right fist, the gravitational force is lightly smaller than the centrifugal force, also resulting in a net force outward from the surface of your left fist. These upward forces on opposite sides of your left fist are the forces that raise the tidal bulge on the two sides of the Earth!

Advanced Topic: Quantitative Treatment of the Double Tidal Bulge

I mentioned above that the Earth–Moon system is characterized by a balance between gravitational force and centrifugal force. By Newton’s law of universal gravitation, the gravitational force is

$$F=\frac{G m_l m_2}{r^2} \tag{8.3.1}$$

where, given two bodies, F is the gravitational force of attraction between the two bodies, M_1 is the mass of body 1, M_2 is the mass of body 2, r is the distance between the centers of mass of the bodies, and G is the gravitational constant, equal to 6.667 x 10-8 cgs units. (Important note: this G is ***not*** the same as the acceleration of gravity g on Earth!). So, for the Earth–Moon system,

$$\text{centrifugal force }=\frac{G M m}{R^2} \tag{8.3.2}$$

where M is the mass of the Earth, m is the mass of the moon, and R is the distance between the centers of mass of the Earth and the Moon.

But at any point on or in the Earth the local gravitational force and the local centrifugal force are not necessarily in balance. In terms of a unit mass located anywhere in the Earth, the centrifugal force (per unit mass) is the same everywhere, *Gm*/*R*2, found by dividing the overall centrifugal force by the mass of the Earth, but the gravitational force per unit mass varies from point to point because the distance from that point to the center of the Moon varies.

Why is the centrifugal force the same everywhere? (This is tricky to think about; at first thought, it seems as though it should vary from place to place.) I think the key here is to neglect the rotation of the Earth around its own axis; that's a *different* centrifugal-force effect, not related in any way to analysis of the tides in terms of the revolution of the Earth and the Moon around their common center of mass.

If we assume that the orbits of the Earth and the Moon about their common center of mass are circles (actually they're ellipses with very small eccentricity), a little thought should convince you that every point on the (nonrotating) Moon describes a circular path in space, and (important) the radii of all the circles are equal—although the circles are nonconcentric. And the same holds true for all points of the Earth, even though a nonrotating Earth revolves around a point actually located inside the Earth.

But how about the effect of the Earth's rotation? That simply changes the equilibrium shape of the Earth to be a prolate spheroid instead of a sphere; it's a matter of a balance between the "pancake" effect of the spin and the "glob" effect of the Earth's own gravity field. The picture of tidal forces, based on considerations of gravitational versus centrifugal forces for a nonrotating Earth, is *superimposed* on this equilibrium shape.

Now think about the local balance of gravitational and centrifugal forces at two special points on the Earth: the point on the Earth's surface that lies along the Earth–Moon axis on the side *nearest* the Moon (called the ***sublunar point***), and the point on the Earth's surface that lies along the Earth–Moon axis on the side *farthest* from the Moon (called the ***nadir point***). Set up a coordinate system as shown in Figure 8-4. What I'm going to do is write the sum of the local gravitational force per unit mass and the local centrifugal force per unit mass at each of these special points.

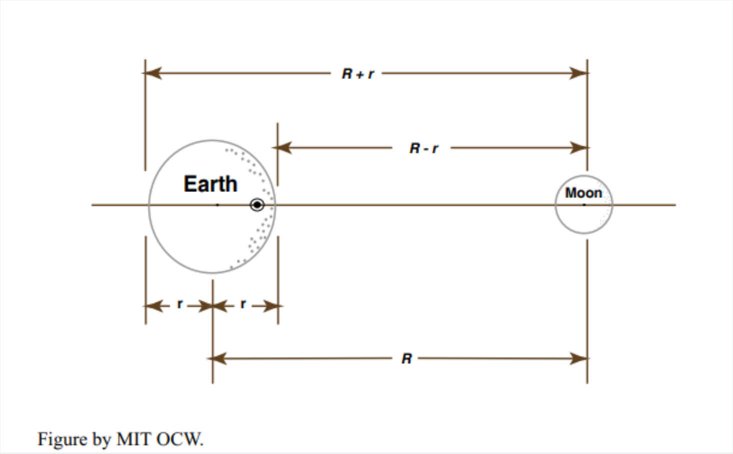


Figure by MIT OCW.

Figure 8-4. Definition sketch for developing the equilibrium theory of the tides.

At the sublunar point the local gravitational force per unit mass is $+Gm/(R-r)^2$, where $R-r$ is the distance from the sublunar point to the center of mass of the Moon. The local centrifugal force is - Gm/R^2, as discussed above; the minus sign is there because the centrifugal force acts in the direction opposite the gravitational force. So the *sum* of these two forces, which I'll call F_T, the net force per unit mass on Earth material located at the sublunar point, is

$$\begin{aligned}
F_T &= -\frac{Gm}{R^2} + \frac{Gm}{(R-r)^2} \\
&= \frac{-Gm(R-r)^2 + GmR^2}{R^2(R-r)^2} \\
&= Gm\left[\frac{R^2-(R-r)^2}{R^2(R-r)^2}\right] \\
&= Gm\left[\frac{2rR-r^2}{R^2\left(R^2-2rR+r^2\right)}\right]
\end{aligned}$$

This looks formidable, but since $r \ll R$, then $r_2 \ll rR$ and $rR \ll R_2$, so to a very close approximation

$$\begin{aligned} F_T &= Gm\left[\frac{2Rr}{R^2\left(R^2\right)}\right] \\ &= \frac{2Gmr}{R^3} \end{aligned}$$

On the opposite side of the Earth, at the nadir point, the same considerations on the net force per unit mass lead to

$$\begin{aligned} F_T &= -\frac{Gm}{R^2} + \frac{Gm}{(R+r)^2} \\ &= Gm\left[\frac{-(R+r)^2 + R^2}{R^2(R+r)^2}\right] \\ &= Gm\left[\frac{-2rR - r^2}{R^2\left(R^2 + 2rR + r^2\right)}\right] \\ &= Gm\left[\frac{-2rR}{R^2\left(R^2\right)}\right] \\ &= -\frac{2Gmr}{R^3} \end{aligned}$$

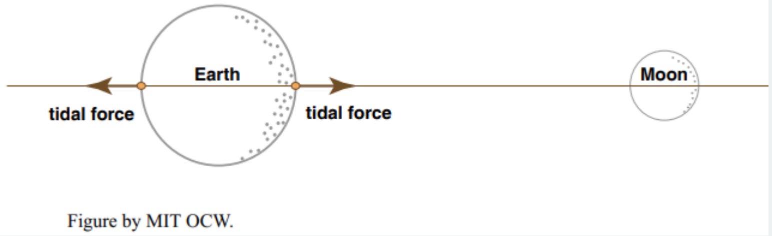


Figure by MIT OCW.

Figure 8-5. Tidal forces at the sublunar point and the nadir point.

At this point we should expand our consideration from one dimension to two dimensions and think about all the points on the intersection of the Earth's surface with any plane that passes through the Earth–Moon axis.

That's more complicated mathematically, because we have an angle to worry about, but it's straightforward. I won't do it here. The result is shown qualitatively in Figure 8-6, which shows for a number of representative points the gravitational force, the centrifugal force, and the resultant of the two, and in Figure 8-7, which shows just the resultant.

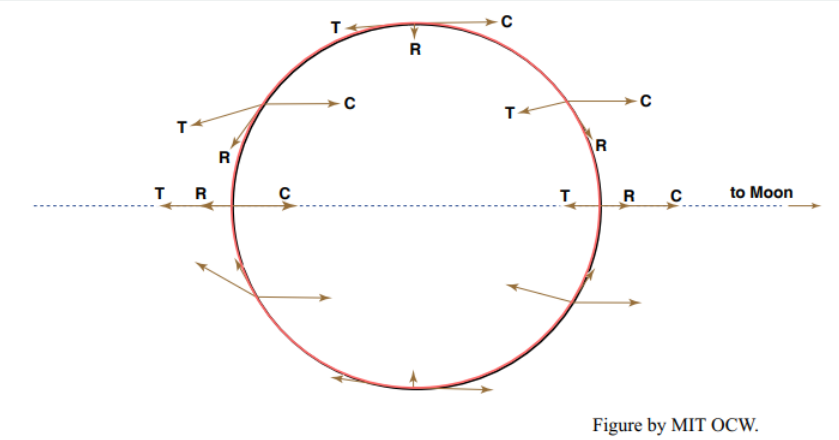


Figure by MIT OCW.

Figure 8-6. Tidal forces. T = gravitational force, C = centrifugal force, R = resultant force.

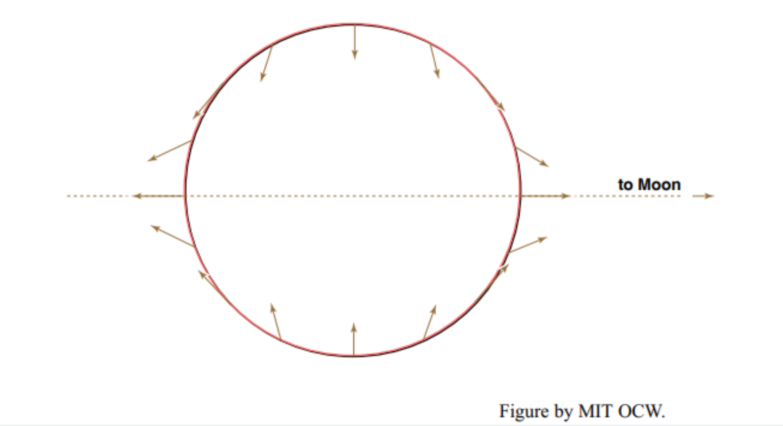


Figure 8-7. Resultant tidal forces.

The forces shown in Figure 8-7 can themselves be resolved into components along the local vertical and the local horizontal. If you plug in some numbers, you would find that the local vertical component of *FT* is only about 10-7 that of the Earth's gravitational pull on an object. So the vertical component of the net force is has a negligible effect on things like ocean water at the Earth's surface. But the *horizontal* component of this net force is of about the same magnitude as the horizontal pressure-gradient forces that drive the water motions in the ocean. So the horizontal forces are important, and they can't be ignored. They are what cause the tides.

Figure 8-8 shows just the horizontal components of the net force *FT* at various points around the Earth. They're zero at the sublunar and nadir points and at all points around the great circle that's normal to the Earth–Moon axis, and they're at a maximum around two small circles normal to the Earth–Moon axis and at 45° to that axis.

It's these horizontal forces that cause the tides. They pile up ocean water in two symmetrical bulges at the sublunar point and the nadir point and lower the water level in the vicinity of the great circle normal to the Earth–Moon axis.

The Sun produces a tidal effect of the same kind as that of the Moon, but the effect is smaller. Although the Sun has far greater mass than the Moon, it's also much farther away from the Earth. The tidal effect of the Sun is about 46% of that of the Moon. You can show this by forming the ratio

$$\frac{2Gm_{\text{moon}}\, r / R^3_{\text{moon}}}{2Gm_{\text{sun}}\, r / R^3_{\text{sun}}} \tag{8.3.3}$$

using the term on the right-hand side of Equation 4 or Equation 5 and substituting the appropriate values.

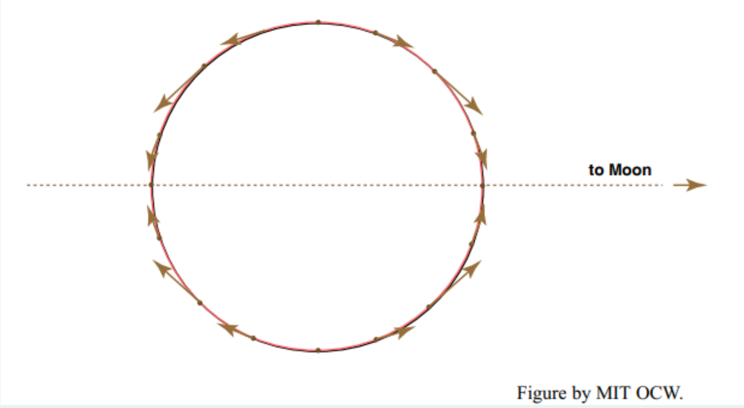


Figure 8-8. The horizontal component of the resultant tidal forces.

At this point you have to visualize that the Earth rotates on its axis under the tidal bulge, which, remember, is locked into position relative to the Moon. To you, the observer on the Earth, a tidal bulge seems to speed around the Earth twice a day, causing the rise and fall of the tides at your station. The period is the same as half the time it takes for the moon the reoccupy the meridian on successive days (24 hr 50.47 min, to be exact).

Some Complexities of the Equilibrium Tide

What I've presented above is basically what's called ***the equilibrium theory of the tides.*** Newton laid the foundation by developing this concept, and the details were well worked out by over a hundred years ago. But this is really only the beginning of the story of the equilibrium tide, because we would have to get more deeply involved in analysis of the relative motions of the Earth, the Moon, and the Sun to take account of all the variations. Below are just two examples of such matters—but they are the two most important ones.

In general the Earth's rotation axis is not perpendicular to the Earth– Moon axis. So when a point not on the Equator rotates under the two-ended tidal bulge the high tide at that point alternates between a higher high tide and a lower high tide (Figure 8-9). In Figure 8-9, a station at x_1 has a higher high tide now, but about 12 hours later, when that same station experiences another high tide, that high tide is a lower high tide, because now the station is at x_2, off the center of the tidal bulge. This is one cause of what's called ***diurnal inequality***, whereby *one of the two semidiurnal tides is larger than the other at a given station*.

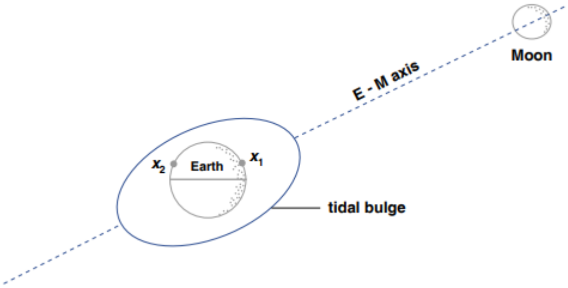


Figure by MIT OCW.

Figure 8-9. The orientation of the tidal bulge relative to the Earth's axis of rotation.

A striking worldwide effect of the tides is an alternation between greater and smaller tidal ranges on a time period of one lunar month. Tides with the *greater* tidal ranges are called ***spring tides*** (no relation to the spring season of the year!), and tides with the *smaller* tidal ranges are called ***neap tides***. The explanation of the spring–neap cycle is simple, and it has the side benefit of explaining also the phases of the Moon—something humankind in modern society has largely lost touch with, except for a swarm of boaters, smaller numbers of campers and fishermen, a rapidly shrinking group of old-fashioned gardeners and farmers who still insist on planting by the Moon, and a tiny band of scientists keeping knowledge alive.

In the course of one lunar month, the Moon makes one full revolution around the Earth. The Earth is at the same time revolving about the Sun, but it travels only a small fraction of a revolution in the time it takes for the Moon to go once around the Earth. Figure 8-10 shows the Moon at various positions around the Earth during one of its revolutions. The sense of rotation of the Earth around its own axis is shown also. The view in Figure 6-10 is perpendicular to the plane of revolution of the Earth around the Sun (called ***the plane of the ecliptic***), and looking down onto the Northern Hemisphere of the Earth.

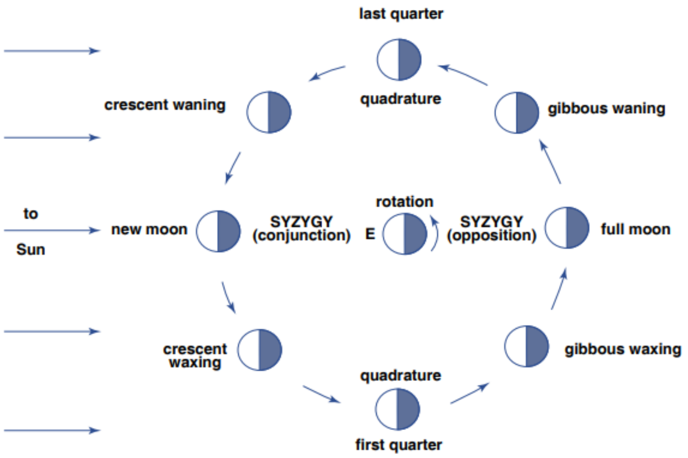


Figure by MIT OCW.

Figure 8-10. The phases of the moon

Twice during the month the Sun, the Earth, and the Moon are approximately in alignment. These are called ***syzygies*** (singular: ***syzygy***). The syzygy for which the Moon is toward the Sun is called ***conjunction***, and that for which the Moon is away from the Sun is called ***opposition***. When the line between the Moon and the Earth makes a right angle with the line between the Sun and the Earth, the Moon is said to be in ***quadrature***.

Obviously, the Sun and the Moon reinforce each other in their tidal effect at the time of syzygy, giving rise to spring tides, and they tend to cancel each other's effects at the time of quadrature, giving rise to neap tides.

You can also see in Figure 8-10 the explanation for the phases of the Moon. At the time of conjunction, you can't see the Moon; that's called the ***new moon***. At the time of opposition, you see the ***full moon***. At the times of quadrature you see a ***half moon***, officially called the ***first quarter*** and the ***last quarter***. The full set of terms is given in Figure 8-10. If you think about it just a little, you can also see from Figure 8-10 why the Moon rises a little later each day throughout the lunar month; it's because the sense of rotation of the Earth on its axis is the same as the sense of revolution of the Moon around the Earth, both being as shown in the figure.

There are many other such astronomical effects on the tides, having to do with

- the relative positions of the Earth, the Moon, and the Sun,
- the relative distance of the Moon from the Earth as a function of time during one lunar month, and
- the relative distance of the Sun from the Earth as a function of time during one year.

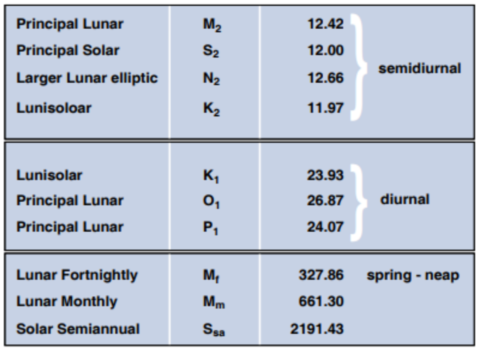
Accordingly, the tide varies complexly with time. These variations can be sorted out mathematically into what are called ***tidal constituents***: *regular periodic variations, with different amplitudes and periods, which are related directly or indirectly to the astronomical variations*. These constituents are given names and code designations. Table 8-1 shows the most important ones. The principal lunar tide, M2, is the most important; it's the one derived in the "advanced topic" section above.

Tides on the Real Earth

That something is disastrously wrong with the equilibrium theory of the tide should be apparent to anybody who makes intelligent observations of the tides: the time of high tide is not the same as the time when the Moon is at its highest point in the sky. There's *a time lag between the time the Moon is highest in the sky and the time of high tide*. That time lag is called the ***lunitidal interval***. It's different at different points on the Earth, but it's always the same at any given point.

The reason for this effect is tied up with the fact that *the tide is a wave*: it has two crests and two troughs around the world, and it moves as a progressive wave from east to west. As the tide wave passes, the water oscillates, just as when a much smaller wind-generated waves passes, as described in Chapter 1, but the water undergoes no net translation over one tidal cycle. (Actually that's often not true in coastal environments with complex shoreline geometry and sea- bed topography, but it's certainly true as a broad spatial average.)

In the terminology of Chapter 1, the tide wave is a grossly *shallow- water wave*: the ratio of water depth (a few kilometers) to wavelength (basically halfway around the Earth) is extremely small.



Principal Lunar	M_2	12.42	semidiurnal
Principal Solar	S_2	12.00	
Larger Lunar elliptic	N_2	12.66	
Lunisoloar	K_2	11.97	
Lunisolar	K_1	23.93	diurnal
Principal Lunar	O_1	26.87	
Principal Lunar	P_1	24.07	
Lunar Fortnightly	M_f	327.86	spring - neap
Lunar Monthly	M_m	661.30	
Solar Semiannual	S_{sa}	2191.43	

Figure by MIT OCW.

At this point you have to make a clear distinction between *free waves* and *forced waves*. A ***free wave*** propagates at a speed determined by its own wavelength and by the depth of water in which it's propagating. A ***forced wave***, on the other hand, is forced by external constraints to travel at some speed, slower or faster than its free-wave speed. The tide wave is clearly a forced wave, because it's constrained, whether it likes to or not (and it *doesn't*; see below), to travel around the Earth once a day. Whether such a forced wave lags behind its constraint or rides out in front of it depends on whether the ratio of the free-wave speed to the forced-wave speed is less than one or greater than one.

It's well known from the theory of water waves that the speed of propagation of a shallow-water wave is equal to gh , where g is the acceleration of gravity and h is the water depth. You could easily compute from this that the free-wave speed of the tide wave is much less than the speed needed to get it around the Earth in one day. (Use 9.8 m/s2 for g, and assume a representative water depth of 4000 m in the oceans. The number of seconds in the lunar day of 24 hr 50.47 min is about 89,428.) So *the tide wave lags behind the apparent motion of the Moon around the Earth* (Figure 8-11).

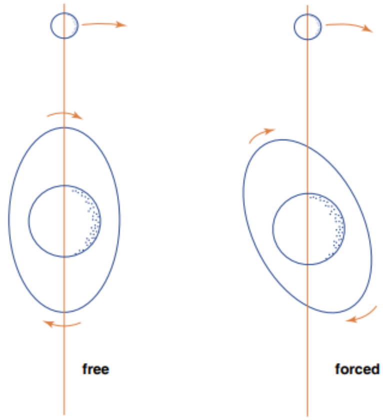


Figure by MIT OCW.

Figure 8-11. A forced tide wave vs. a free tide wave.

We're still not close to describing the true state of affairs in the real oceans. Only around the Southern Ocean, encircling a broad area in the Southern Hemisphere around Antarctica in the present configuration of the continents, can the tides do their thing as a wave that travels uninterruptedly all around the globe. All other areas of the oceans today are almost-closed basins, large or small, open only at their northern and/or southern ends. The critical question is: *How do the tides behave in such basins?*

The North Atlantic ocean basin is a good close-to-home example. If you ignore the minor effect of the narrow opening on the north into the Arctic Ocean, the North Atlantic is open only at its southern end. The basin feels a regularly rising and falling water level at its southern end. The effect of this rise and fall on the water levels and water motions in the basin is something that's not quite within your everyday experience, but it's something you could almost observe in your own bathtub. Here's a very simple home experiment you can make to study the effects:

Cut a sheet of plywood into long strips and nail the strips together to form a simple long trough closed at both ends (Figure 8-12). On the bottom of one end of the trough, place a separate short board of wood that stretches almost from one side of the trough to the other. Nail a vertical stick to this little board so that you can move it up and down regularly, to oscillate the water level at that end of the tank. That simulates the periodic rise and fall of the tide wave at the southern end of the ocean basin.

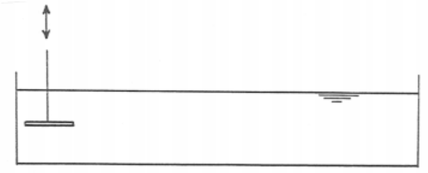


Figure 8-12. Making waves in a wave tank.

The up-and-down oscillation of the board at the end of your trough makes waves, and if your trough were endless you would just see a nice propagating train of waves, as described in Chapter 1. But the waves you make are reflected from the far wall of your trough. Those reflected waves travel back along the trough at the same speed as the oncoming waves, and they are of about the same amplitude (if you neglect the slight loss in energy when they're reflected at the vertical wall). The result is a beautiful pattern of ***standing waves,*** with ***nodes*** where the water level stays the same and ***antinodes*** where the oscillation in water level is at its maximum (Figure 8-13). The number of nodes depends on the length of your tank and the period of the oscillation you impose on it. So although waves are actually moving in both directions, all you see is their sum, a standing wave.

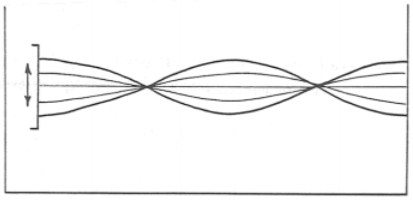


Figure 8-13. Standing waves in a wave tank. Each curve represents the water surface at a different time.

I used to have a wall phone in my kitchen, with one of those long spiraling lines to the handset. During particularly boring phone calls, I played at jiggling my end of the wire up and down to make standing waves on the wire. The faster I jiggled the wire, the more nodes I got. This is closely analogous to the behavior of the water in your trough.

Figure 8-14 shows the standing wave in your trough as a time- sequence cartoon for the simple case of only one node, in the middle of the tank. Given the size of real ocean basins, and the speed and period of the tide wave, this is what you're likely to find in the oceans. Note that the horizontal movement of the water is at its maximum at the nodes and is zero at the ends of the trough. This is just the kind of water motion observed in relatively long and narrow ocean basins.

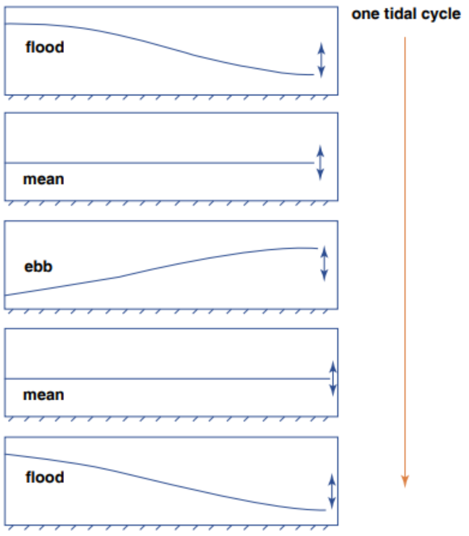


Figure by MIT OCW.

Figure 8-14. Time-sequence cartoon of a first-mode standing wave.

The Bay of Fundy is a good example of tidal standing waves. You probably know that the tides at the head of the Bay of Fundy are the highest in the world. Why? Because the size and depth of the Bay of Fundy are just about such that its natural frequency is close to the natural period of oscillation of seiching. (A ***seiche*** is *a periodic back-and-forth sloshing motion of water in an elongated lake or basin*. Once a difference in water level between the two ends of the water body is set up—by the wind, for example—the sloshing continues until it dies out by friction.) The amplitude of the forced oscillation is well known to be greatest when the forcing period is close to the natural period; this condition is called ***resonance***. You've all had lots of everyday experiences with resonance, whether you know it or not. Pumping on a swing is a good childhood example. Pushing your car in a rocking motion to get it unstuck on an icy road is a more grown-up example, as is sloshing the water in the roasting pan when you're trying to scrub the pan in the sink.

But this still isn't the end of the story! Unless the basin is very narrow, so that the water is constrained to move only back and forth, the Coriolis force has an important effect, unless you're near the Equator. (See the following background section to learn something about the extremely important, but also extremely counterintuitive, Coriolis effect.) Figure 6-15 shows how the Coriolis force affects the water motions in the basin. Keep in mind that in the Northern Hemisphere the Coriolis force acts to the right of the direction of motion. So when the water is moving from the back end of the basin toward the front end, the Coriolis force banks it up against the left wall, until the downslope component of gravity, acting away from the left wall, balances the Coriolis force. Then, when the water is moving toward the back of the basin, the Coriolis force banks the water against the right wall. The result is *a kind of circular sloshing motion* that's not easy to describe in a few words. This is called ***amphidromic motion***.

The Coriolis Effect

In a large unobstructed indoor area (a gymnasium or a warehouse would be best, but a big room in your house would suffice), build a giant flat horizontal turntable—just a disk mounted at its center point on a vertical rotating shaft (Figure 8-16). You can rotate the whole disk at any desired constant rotation rate. It would be best if you painted the surface of the disk a flat black, the better to observe the motions of the brilliantly white marker spheres you're going to roll around on the surface. To make things really exciting, be sure to coat the white marker spheres with a thick chalky coating of some sort that tracks off evenly onto the surface of the turntable as they roll about.

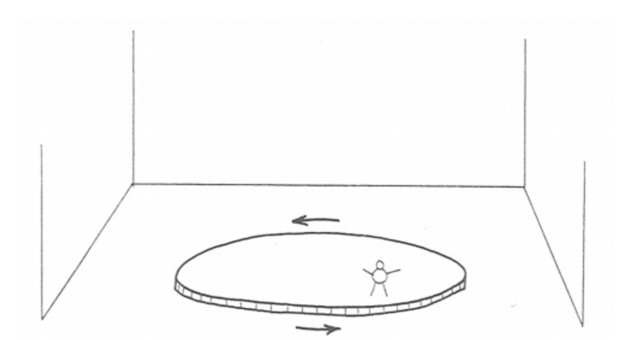


Figure 8-16. A turntable for investigating the effect of the earth's rotation.

To do things right, you're also going to need an observation perch above the turntable. This perch should be easily movable from place to place above the surface, but (and this is important) stationary relative to the floor of the room while in use. One of those mechanized cherry-picker seats, extending horizontally from the margin of the disk, would do nicely, if you can afford it.

Set the rotation rate, get into your perch, occupy a point just over the turntable, and roll one of your marked spheres onto the table, just as if you were at a bowling alley. Here's the big question: *What would the track of the sphere look like* on the turntable? (You're going to have to assume that the turntable exerts no substantial force on the rolling ball. That's not really true, but the effects are small enough that you can safely ignore them for the purposes of this demonstration. If you don't like that assumption, you can always imagine using a magic air-hockey puck that scoots frictionlessly over the turntable, leaving a powdery white trail behind it.)

The big jump that your powers of deduction or imagination have to take here is to see that *the track left by the ball on the table would be curved* (Figure 8- 17). (And, once you're comfortable with that idea, it might naturally occur to you to think about whether that curved track is a circular arc. The answer turns out to be NO, although the reasons are a little too intricate to deal with at the moment.)

If you can't afford the time and money to build the turntable, but you still want to get some useful results, here's a much simpler and cheaper way of demonstrating the phenomenon (Figure 8-18). Pin a big piece of posterboard to the wall so that it can be rotated about its center point, and have an assistant stand to one side and rotate the posterboard in a hand-over-hand motion as steadily as possible. Stand in front of the posterboard with a marker pen, and draw a line on the posterboard in such a way that *the tip of the pen is moving in uniform rectilinear motion relative to the underlying wall* (that is, a straight line at constant speed). That's tough to do, because you have to try to ignore the surface of the posterboard and the mark that's coming out onto it and instead concentrate on the imaginary path of the pen point on the motionless wall behind. You would find (Figure 8-19) that no matter where you start on the posterboard, and no matter which direction you choose for your line, *the mark on the posterboard is a curving arc!*

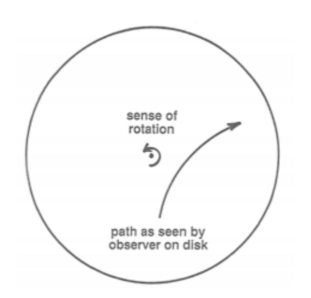


Figure 8-17. The curved track left by a ball rolling on a rotating turntable.

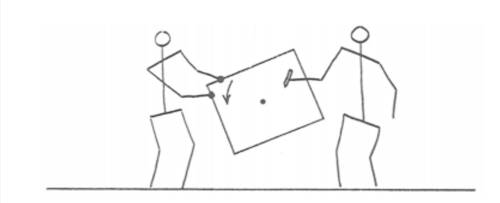


Figure 8-18. A simple way of demonstrating the Coriolis effect.

The next thing to do is have your assistant roll a marker sphere onto the turntable while you're riding on the turntable. Watch the ball as it rolls and leaves its circular-arc track. It will look to you as though some mysterious sideways force is continuously acting on the ball normal to its path to push it off its course. Something seems to be wrong with Newton's first law, which tells you that the ball should be moving in a straight line at constant speed. You know what the problem is, of course: the fictitious side force is an artifact of your observing the ball from the standpoint of the rotating turntable. If you reoccupied your perch and rolled a clean, chalkless ball onto the dimly lit black surface of the turntable, you'd see the ball roll in a nice straight line! The fictitious side force that seems to act on moving bodies in a rotating environment is called the ***Coriolis force***, after the nineteenth-century French engineer–scientist Gaspard Gustave de Coriolis (1792–1843), who first studied the effect. And the apparent acceleration of the sphere (it's a radial acceleration, not a tangential acceleration, in that only the direction changes, not the speed) is called the ***Coriolis acceleration***. The entire effect is called the ***Coriolis effect***.

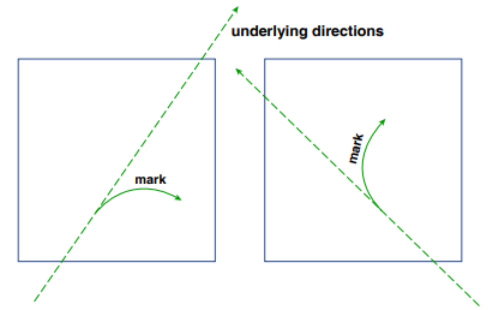


Figure by MIT OCW.

Figure 8-19. The results of the posterboard experiment.

In basins like this (Figure 8-20), there's some *point in the center of the basin where the tidal change in water elevation is zero*. That's called an ***amphidromic point***. Around the amphidromic point you can draw *closed contours that give the loci of equal tidal range*. These lines are called ***co-range lines***. Finally, you can also draw spoke-like *lines radiating outward from the amphidromic point toward the margins of the basin that show the locus of points at which the tide reaches a maximum at the same given time*. These lines are called ***cotidal lines***. A basin that has one amphidromic point with its associated cotidal lines and co-range lines is called an ***amphidromic system***. Figure 8-21 shows the North Atlantic as an amphidromic system. The complex geometry of the bottom topography and shoreline trace makes the details of the amphidromic system irregular, but the essential features are clearly there.

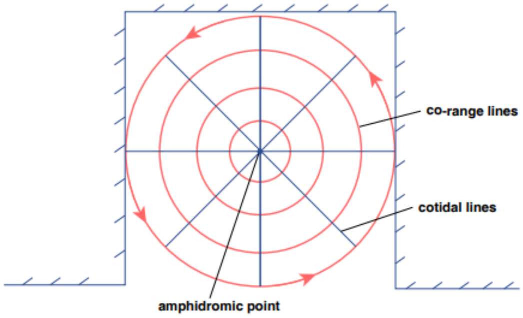


Figure by MIT OCW.

Figure 8-20. Elements of an amphidromic system.

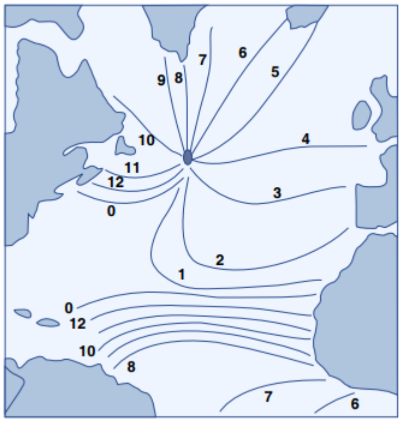


Figure by MIT OCW.

Figure 8-21. The amphidromic tidal system of the Atlantic Ocean.

Large amphidromic systems can act as drivers or forcers for smaller amphidromic systems. Smaller basins like large gulfs or bays around the margins of a large ocean basin have their own amphidromic systems, which are forced by the periodic tidal rise and fall of water level at their mouths as the tide sweeps amphidromically around the larger basin. But whatever the size of the basin, the foregoing exposition serves to explain nicely the progressive difference in times of low tide and high tide along a given stretch of coastline. Figure 8-21 shows why the times of low and high tides get later and later from north to south along the east coast of North America.

Tidal Currents

Tidal currents are the horizontal water movements associated with the tidal rise and fall of the sea surface. If you occupy a station in the ocean, you observe not only a systematic change in water level but also a systematic change in current velocity.

The connection between tidal water-surface elevation and tidal- current velocity is an indirect one, though. Because of the complexity of the actual tidal wave, depending on where you are there can be different relationships between the tide and the tidal current, and there can even be places with *tides but no tidal currents*, or *tidal currents but no tides*. But this shouldn't be too surprising, in light of our examination of standing tide waves in the preceding section. For the same reasons, there's no direct relationship between the times of high and low tides and the times of slack water, when no tidal currents are running.

Another thing: you have to average over a time long enough compared to the tidal cycle to factor out currents caused by random effects like winds and storms—but with the added complication that there may be nonzero average currents at a locality, caused by currents of other kinds in the oceans. Also, don't assume that the tidal currents themselves at a given locality have to average out to zero.

Why are there tidal currents? Tidal currents are simply the horizontal water motions associated with the passage of the tide wave, whatever the nature of that tide wave. Remember that tidal waves are very long compared to the water depth, so they are shallow-water waves. The water motion associated with such waves is a back-and-forth oscillation, actually an extremely flat ellipse, because the amplitude is so small relative to the wavelength (Figure 6-22). Even the sketch in Figure 6-22 has great vertical exaggeration, and also exaggeration of the length of the orbit.

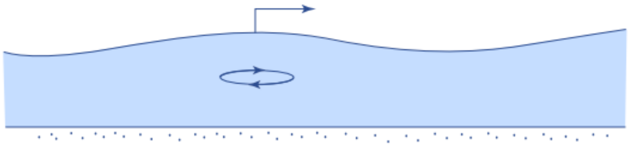


Figure by MIT OCW.

Figure 8-22. Water motions associated with the passage of the tide wave. (Vertically exaggerated.)

Some definitions:

flood current: the current that runs while the tide is rising
ebb current: the current that runs while the tide is falling
slack water: the period of no tidal current while the tide reverses

(Note: the concept of slack water is relevant only to areas with nearly bidirectional tidal currents.)

Complications in tidal currents arise because there’s usually more than one progressive tidal wave, because of reflection, and commonly there are standing-wave systems as discussed above. So there’s no general relationship between high and low tide, on the one hand, and the time of slack water and the speed of the tidal currents, on the other hand. That relationship varies from area to area. The time of slack water may or may not be close to the times of high and low tides.

Another complication is that there’s no necessary relationship between the speed of the tidal current and the tidal range at a given place. In a very general way, of course, the speed of the tidal currents increases with the tidal range. But the speed of the tidal current is a matter of hydraulics, given the tidal range, and it depends on the volume of water moved versus the size of the passage through which the water has to move. For example, in the Gulf of Maine there’s a large tidal range but weak tidal currents, whereas in Nantucket Sound there’s a small tidal range but strong tidal currents. At a given station, however, the speed of the tidal currents is proportional to the tidal range, so currents are strongest during spring tides and weakest during neap tides.

How does one measure tidal currents? It’s more difficult than measuring the tidal range. Like any other measurement of current, it requires a station that’s fixed relative to the bottom, and a current meter of some kind. The picture of tidal currents is well known only for populated coastal areas.

What are the typical magnitudes of tidal currents? They’re highly variable. Strong tidal currents are 2–3 m/s at the surface (or 75–80% of this when depth-averaged). And 1–2 m/s tidal currents are very common. Currents like that can move a lot of sediment—if the sediment is there to move.

How does one represent tidal currents? The best way is to plot the tidal-current velocity vector as a function of time over a complete tidal cycle, with all the vectors having a common origin. That kind of plot is called a ***tidal-current rose***. Figure 8-23 is an example from the Nantucket Shoals lightship. The following are some notes on Figure 8-23:

- The angular spacing of the vectors is not uniform, meaning that the turning of the current is not uniform in time.
- Low tide and high tide are not spaced equally in time.
- The maximum currents are not symmetrical in time.
- A floating object would make a closed traverse with the same shape (although not the same size!) as the curve connecting the tips of the vectors (and because the velocity vectors are turning clockwise, the object would move clockwise also).
- A current rose like that shown in Figure 8-23 is often approximately elliptical. If that’s the case, then it’s called a ***tidal ellipse***.
- Current roses can be open, as in Figure 8-23, or more closed, as in Figure 8-24. As the current rose becomes flatter, the tidal currents become bidirectional tidal currents.

8.4: Beaches

Shoaling Waves

First go back to Chapter 1 and review the material on surface gravity waves. Now it's time to think about what happens to a train of waves propagating from the deep ocean to the shoreline. The waves start as deep-water waves, but as they shoal the begin to "feel the bottom", in the sense that they make the water move at the bottom. This bottom friction causes some loss of wave energy. Also, if the orbital velocity of the water at the bottom becomes great enough to move sand, ***oscillation ripples*** (a kind of oscillatory-flow bed configuration) develop (Figure 8-25).

One of the most important effects connected with shoaling waves is that they *slow down*. It turns out that for shallow-water waves the wave speed c is proportional to the square root of the water depth. Because the wave period itself doesn't change (the waves keep on being supplied from offshore at the same period), by the relation $L = cT$ *the wavelength decreases*. Wave height H also increases, for reasons connected with conservation of wave energy that are too complicated to discuss here, and the combined effect leads to *increase in wave steepness*. This can't go on indefinitely (the limiting steepness H/L turns out to be one-seventh), so eventually the waves fall over on themselves, or in other words they *break*.

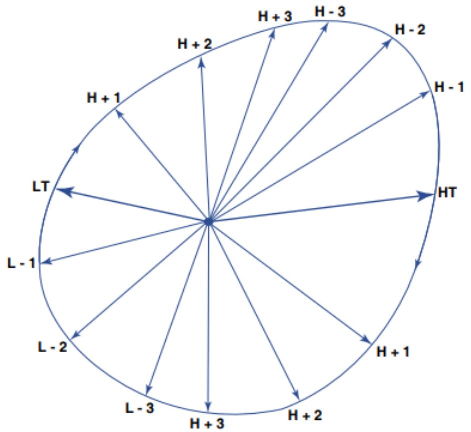


Figure by MIT OCW.

Figure 8-23. The tidal-current rose measured at the Nantucket Lightship.

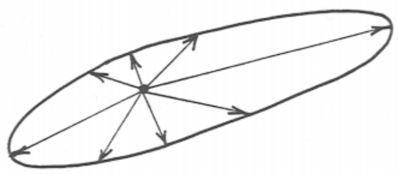


Figure 8-24. An example of a more elongated tidal-current rose.

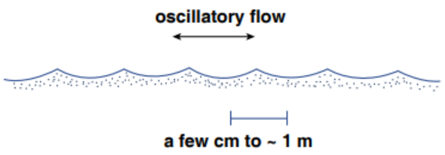


Figure by MIT OCW.

Figure 8-25. Oscillation ripples.

The form taken by the breaking waves depends on the rate of change in the steepness as the waves shoal. Three types of breakers are generally recognized (but keep in mind that there's a gradation among them):

> **spilling breakers**: the wave peaks up slowly, the crest becomes unstable and spills down the front surface of the wave, and the wave height slowly decreases (Figure 6-26A). These tend to occur with gentle bottom slope and originally steep waves.

plunging breakers: the wave peaks up rapidly, and the crest becomes a thin vertical wall that curls over forward and then plunges forward and downward, causing a catastrophic decrease in wave height (Figure 6-26B). These tend to occur with steep bottom slope and originally intermediate-steepness waves.

surging breakers: the wave peaks as if to plunge, then the base of the wave surges up the beach face, and the crest collapses and disappears (Figure 8-26C). These tend to occur with very steep beaches and originally low-steepness waves.

After the wave has broken, the water rushes up the beach as a moving mass, which can be thought of as a *wave of translation* (in contrast to former nature of the wave, wherein the water had no net translation). This called the ***swash***. The water carried up the beach as swash returns under the downslope pull of gravity as ***backwash***.

Another effect of the slowing of waves as they shoal is ***refraction***. You may have noticed that waves moving toward shore tend to curve around so as to be more nearly parallel to the shore (Figure 8-27). This is an inevitable consequence of the slowing of the waves as they move into shallower water. Here's an easily understandable analogy. The drill sergeant is putting a squad of marching persons through their exercises. At one point he/she tells the people on the left to take baby steps and the people on the right to take giant steps. You can easily envision what will happen: the direction of movement of the squad will curve around to the left. This is exactly the same refractive effect that makes the shoaling waves swing around to be more nearly parallel to the shoreline.

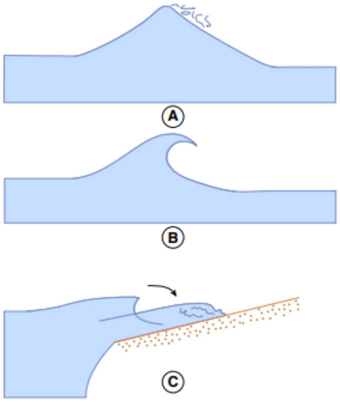


Figure by MIT OCW.

Figure 8-26. Varieties of breaking waves. A) Spilling breakers. B) Plunging breakers. C) Surging breakers.

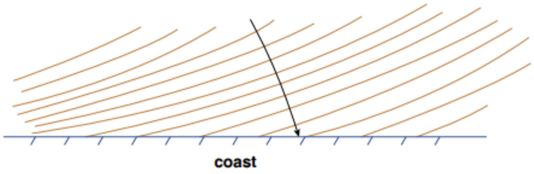


Figure 8-27. Refraction of waves as they shoal toward a shoreline.

Beach Profiles

It's easy to define a beach. A ***beach*** is *a mass of non-cohesive sediment along a shoreline that's molded by wave action*. The sediment size of beaches ranges from the finest sand to coarse gravel. The best way to approach the dynamics of beaches is to look at a vertical cross section normal to the shoreline; such a cross section is called the ***beach profile***.

To establish some terminology, Figure 8-28 is a sketch of a representative beach profile. I want to emphasize that beach profiles vary a lot in their geometry and in the kinds of features they show—both from beach to beach and at the same beach as a function of time. The sketch in Figure 8-28 is a typical snapshot, but it's not representative of all beaches.

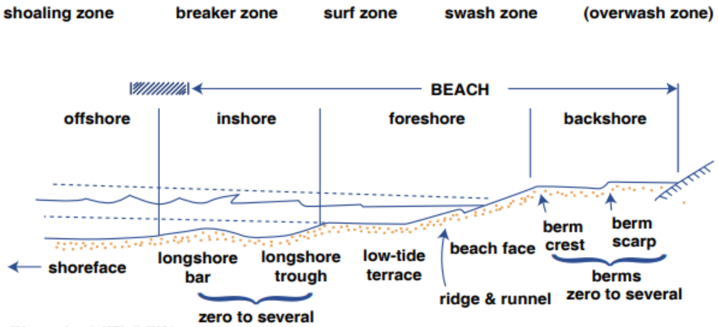


Figure by MIT OCW.

Figure 8-28. Shore-normal vertical profile through a beach.

(The *seaward limit* of the beach is a matter of definition; depending upon who's doing the defining, it can range from mean low tide to the seaward limit of breaking waves, which can be much farther offshore.)

Some general points about this profile:

- As discussed in the section on shoaling waves above, the strength and nature of bottom-water movements associated with breaking waves vary greatly and systematically from deep water to land.
- The beach profile is always changing with time, in an effort to keep up with the ever-changing wave state. We naturally assume that for each given wave state and beach sediment there's an equilibrium profile, in the sense that, whatever the initial geometry of the beach, the final equilibrium geometry attained by the beach after prolonged exposure to that wave state is the same. But because in the real world the wave state changes all the time, the beach is always adjusting toward a new equilibrium.
- The nature of sediment movement on the beach varies not only with the waves but also with the sediment size, which itself depends in a complex way on the long-term average wave state and on the nature of sediment supply to the beach.

Onshore–Offshore Transport

The main feature of sediment movement on the beach is the back- and-forth movement of the sand in response to the swash and backwash of the waves. Sand is carried up the beach as both bed load and suspended load by the relatively deep and strongly turbulent swash, and it's carried back down the beach mostly as bed load by the thinner and less turbulent backwash.

For a given beach sand, the vertical profile of the beach is very sensitive to wave conditions. At any given time there's nearly a balance between sand moved up the beach by the swash and sand moved down the beach by the backwash. The volume of backwash is always smaller than the volume of swash, because water percolates into the porous beach sand. To maintain the balance in upslope and downslope sand movement, the beach takes on a particular slope so that the force of gravity hinders upslope transport by the swash but aids downslope transport by the backwash.

What's therefore important in governing the slope of the upper part of the beach is *the ratio of backwash volume to swash volume*. As wave height increases, the volume of water moving up and down the beach increases, but because the volume lost by percolation doesn't increase very greatly the percentage of backwash compared to swash is greater than for waves with low steepness. So more material is moved down the beach, and the slope on the upper part of the beach decreases. The sand thus eroded from the upper part of the beach is parked farther offshore, although still within the beach system. During periods of good weather, with smaller waves, the ratio of backwash volume to swash volume decreases, and the slope of the upper part of the beach increases again, as sand is supplied from offshore. Figure 6-29 shows beach two profiles, one for small waves and one for large waves.

Ordinarily the sand that's parked offshore during periods of strong waves isn't lost to the beach system, and it's moved back onto the upper part of the beach—the part we see and lie on and play on—during periods of weaker waves. Only during unusually large storms can there sometimes be mobilization of great masses of sand and transport farther out to sea.

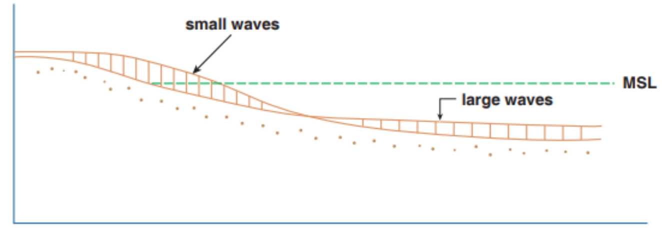


Figure by MIT OCW.

Figure 8-29. Beach profiles for times with small waves and times with large waves.

By much the same line of reasoning, *the slope of the beach varies directly with sand size*: coarser beaches admit of greater percolation, so for a given wave state the beach has to be steeper to strike the balance between upslope and downslope sediment movement. Beaches with fine sand have very gentle slopes, sometimes of only a few degrees. Such beaches can be hundreds of meters wide. On the other hand, beaches formed of coarse gravel in areas exposed to very strong waves can be as steep as the angle of repose of the gravel, well over 30°. (The ***angle of repose*** is *the slope angle that's assumed by apile of loose sediment that's formed by pouring the sediment from a point above the pile*.) On such beaches there's no backwash: the water of the swash percolates so rapidly into the beach that there's none left to flow back down the beach as backwash! Then the downslope pull of gravity is the only thing that counteracts the upward transport by the swash.

Longshore Transport

So far I've talked only about sediment movement in the onshore– offshore direction. Sand is also moved *along* the beach, in various ways. Such movement is called ***longshore transport***, or ***littoral drift***, or ***longshore drift***.

How is longshore transport observed? It's notoriously difficult to make direct measurements of sediment transport rate, especially in complicated situations like beaches, where various kinds of water movements are superimposed upon one another, but two lines of indirect evidence should be readily understandable:

- piling up of sand in front of groins and jetties, and depletion behind them (Figure 8-30);
- building of sand spits at headlands (Figure 8-31).

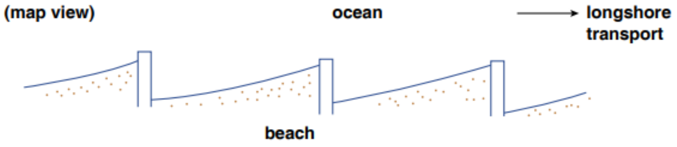


Figure by MIT OCW.

Figure 8-30. How sand piles up a the updrift sides of groins.

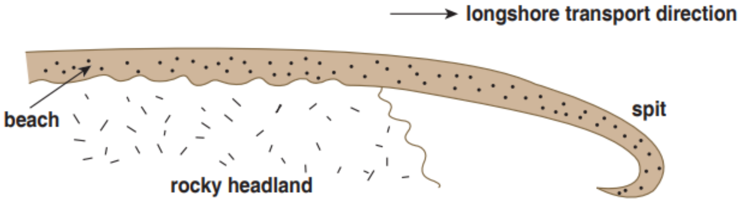


Figure by MIT OCW.

Figure 8-31. A sand spit at a headland.

Another good line of evidence, less obvious to the casual observer, is the presence of distinctive sediment along the beach downdrift from the mouth of a river (Figure 8-32).

Finally, you can imagine marking or tagging a small volume of sediment, with paint or even with a radioactive coating, and then looking for it in the downdrift direction after some period of time. That's an excellent way of getting qualitative information on longshore transport. But getting actual rates of movement, in terms of mass of sand moved per unit time, is much more difficult, largely because of problems connected with mixing of the marked sand downward into the underlying sand.

Longshore transport is complicated, because several different mechanisms contribute to it. One kind of longshore transport, called ***beach drifting***, is easy to understand: it comes about because the waves approach the beach obliquely rather than straight on. (It's generally the case that waves don't approach the beach straight on.)

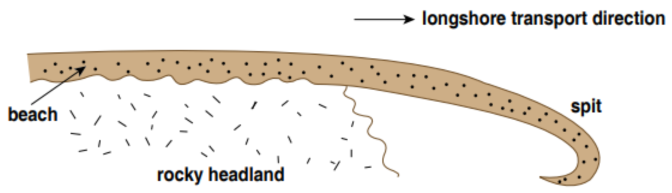


Figure by MIT OCW.

Figure 8-32. Downdrift movement of a tracer introduced to the coastline by a stream.

Despite the tendency for wave refraction to make the waves more nearly parallel to shore, waves still tend to be slightly oblique to shore when they break (Figure 8-27). So the swash (the wave of translation produced by breaking) runs not straight up the beach but diagonally, moving sand diagonally with it. As the swash loses its momentum and is pulled back down the beach by gravity, it flows more nearly straight down the beach, again carrying some sand with it. The net effect on the sand is to be displaced slightly down the beach in sawtooth fashion, in the direction of wave approach to shore (Figure 8-33).

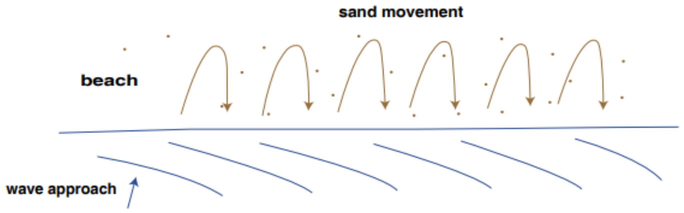


Figure by MIT OCW.

Figure 8-33. Beach drifting.

The other important cause of longshore transport is the presence, immediately adjacent to the shoreline, of a steady current flowing parallel to the beach. Such a current is called a ***longshore current***. The causes of longshore currents are varied. Here are three possibilities:

- the sawtooth movement of swash and backwash by obliquely approaching waves, described above;
- currents caused by complex processes associated with the oblique approach of the waves themselves;
- currents present in the nearshore zone but not caused by wave shoaling.

Whatever their origin, longshore currents can transport sand just like currents in a river. In fact, the ability of longshore currents to transport sand is enhanced by the suspension of sand by breaking waves in the breaker zone: there can be appreciable longshore transport by longshore currents even though the currents are not strong enough to move the sand by themselves.

Beach erosion is much in the news, especially in southern New England and on Long Island. The best way to think about beach erosion is in terms of the budget of sand involved in longshore transport, because we've seen already that little sand is lost to the beach by direct offshore movement. In this view, beach erosion in one locality is offset by beach deposition somewhere else. In areas where less sand is supplied from upstream than is transported downstream there's beach *erosion*, and the position of the beach moves *landward*; in areas where more sand is supplied from upstream than is transported downstream, there's beach *deposition*, and the position of the beach moves *seaward*.

In the context of Cape Cod or Long Island, you have to keep in mind that they are basically just gigantic piles of coarse sediment that were parked there while the terminus of the last continental ice sheet stayed in about the same place for a long time at the height of advance of the ice sheet, and after sea level rose to about its present position those piles of sediment have been exposed to wave action and longshore transport without any new source of sand, as from a major river system. So the only source of sand for the inevitable longshore transport comes from the beach itself, and the sediment cliffs behind the beach. On the other hand, lots of new beach is being built by longshore transport to the “downstream ends” of the beaches—Provincetown and Monomoy, in the case of Cape Cod.

Figure 8-34 is a sketch of outer Cape Cod, showing the direction of approach of the dominant sand-moving waves and the net longshore transport directions. South of a point about in the middle of the face of the outer Cape, net longshore transport is to the south, and north of that point, net longshore transport is to the north. The dividing point is called a ***null point***. For a long distance north and south of the null point there’s net erosion of the beach. But to the north and south of that zone of net erosion, there’s net deposition, as the spits extend themselves ever farther.

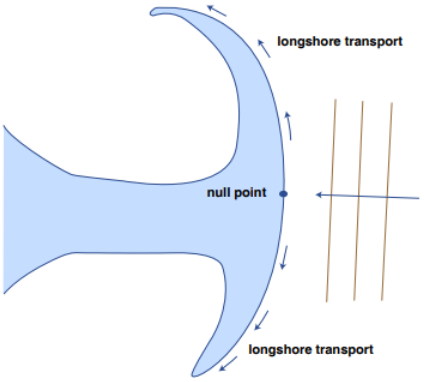


Figure by MIT OCW.

Figure 8-34. Longshore transport along the outer face of Cape Cod.

8.5: Deltas

Introduction

A **delta** is a body of sediment deposited at a point along a body of water where a sediment-transporting channelized flow of water enters the water body. Deltas range in size from those little decimeter-scale bodies that you can see forming when a rivulet of rainwater enters a puddle during and right after a rain, to giant bodies at the mouths of major rivers like the Mississippi. The reason deltas form—and I suppose that this is obvious—is that as the channel flow enters the water body it spreads out and thereby decelerates, dropping much of its sediment load at or not far from its mouth.

The concept of a delta is simple, but the geometry of large deltas in the real world is rather complicated and highly varied, owing to a variety of factors. More on that later; first, some material on the hydrodynamics of deltas.

The Hydrodynamics of Deltas

To understand deltas, you need to know something about jets. In fluid dynamics, a jet is a fluid motion created where a high-speed flow in a pipe or other conduit enters a large body of relatively still fluid. It's easy to visualize the classic jet (Figure 8-35). A pipe enters a large tank of still water. When the water flowing in the pipe enters the tank, it no longer feels the force (the downstream pressure gradient) that was driving the flow in the pipe. Its momentum carries it out into the water of the tank, but friction with the surrounding fluid robs it of its momentum, and it eventually slows to a stop. Most such jets are turbulent—unless they are very small, very slow, and/or consist of very viscous fluid—and the mixing of eddies at the margins of the jet entrain ambient water into the jet, causing it to widen.

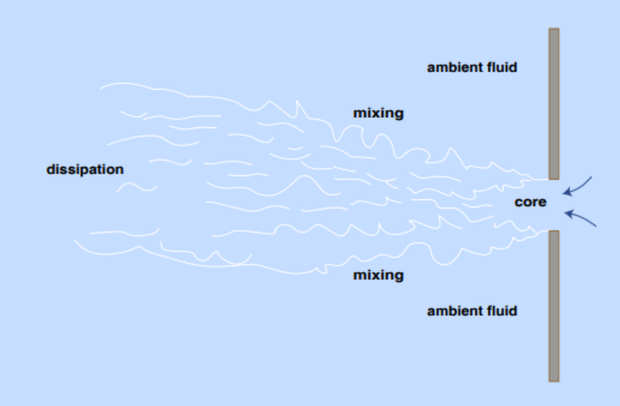


Figure by MIT OCW.

Figure 8-35. A jet entering a body of still fluid from a pipe. This is a cross-section view, cut through the axis of the pipe flow and the jet.

As you can imagine, the mixing at the margin of the jet tends to even out the properties of both the jet and the ambient fluid—which might be temperature, salinity, or concentration of suspended solids. Note in Figure 8-35 that there is a residual core in the jet, which has not yet been affected by the marginal mixing. This core shrinks to nothing in the downstream direction.

You're probably thinking: What does all of this hydrodynamics have to do with deltas? To make the connection more direct, think in terms of a jet with ambient fluid ambient fluid core mixing mixing dissipation a somewhat different geometry: a free-surface channel flow entering a body of still water, as might a river flowing into a lake (Figure 8-36). The essential nature of the jet is no different; what is different is the geometry. The jet still mixes with the ambient fluid, but it can't mix upward, only sideways and downward. If the channel flow is muddy water, you can imagine how the jet would appear from the air: a spreading mass of muddy water, surrounded by clear water, with a core of muddiest water extending outward from the entrance and the muddiness gradually decreasing outward in all directions.

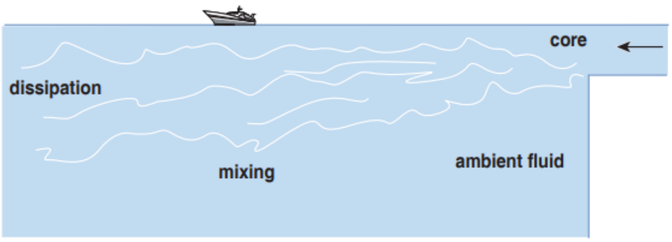


Figure by MIT OCW.

Figure 8-36, except that the shape of the jet is not axially symmetric.)

Now I'll throw in another complicating factor. The jet described in the preceding paragraph is a good model of a river entering a fresh-water lake or reservoir. But, as we all know, the ocean is salty, and, because it's salty its density is significantly greater, by an important few percent, than that of fresh water. That impedes vertical mixing of the jet with its surroundings, while not affecting its horizontal mixing greatly. The reason is that the density stratification that develops between the overlying fresh-water jet and the underlying salt-water medium is gravitationally stable, and it takes work to disrupt or break down that stratification by mixing. The jet may be able to do some vertical mixing, but the extent is much reduced. The jet ends up being largely in the form of a kind of horizontally oriented fan, spreading laterally but not downward (Figure 8-37).

Sediment Deposition in Deltas

Now that you have a good mental picture of the hydrodynamics of the delta environment, think about sediment deposition in the delta environment. Think in terms of the jet shown in Figure 8-36 or Figure 8-37, with the reservation that the depth of water in the water body is not infinitely deep—perhaps only several times the depth of flow in the approach channel (that is, the stream or river).

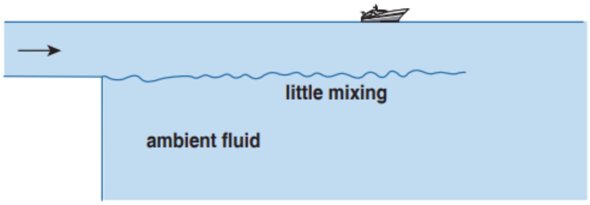


Figure by MIT OCW.

Figure 8-37: except that the shape of the jet is not axially symmetric.)

Suppose that the stream is carrying both bed load and suspended load. You know what's going to happen: the bed load tends to be dropped out rapidly, at and near the entrance of the jet, because the sediment particles have relatively large settling velocities, but the finer part of the load is carried for greater distances away from the entrance, participating in the lateral mixing while it settles slowly to the bottom of the water body.

Think now specifically about the deposition of the bed load. At first, of course, it just makes a pile of sediment at base of the water body below the jet entrance. Eventually the pile builds up to reach the jet entrance. From then on, as the sediment comes to rest as the jet emerges into the water body, a wedge of coarse sediment builds forward into the water body, as sediment is deposited at the brink of a sediment body and slides down an angle-of-repose slope. Figure 8-38 shows three stages in the process, and Figure 8-39 shows a plan view.

Note, in particular, from Figure 8-38 that in the later stages the delta body consists of three rather distinct parts: ***topsets, foresets, and bottomsets.*** The foresets we just dealt with. The bottomsets are readily understandable as well: as the foresets build forward, they bury earlier-deposited finer sediment, which fell from suspension at some distance from the jet entrance.

The topsets require a bit of further explanation. Keep in mind that the river flowing into the water body has some nonzero slope. (That's what keeps the water flowing, remember.) I purposely showed that slope in Figure 8-38. The outbuilding of the foresets has the effect of shifting the point of entrance of the jet in the direction out into the water body. That causes the downstream-most part of the river profile to rise at all points. That happens by sediment deposition. More specifically, a very small part of the passing bed load is extracted from the flow to come to permanent rest on the river bed, thereby building the profile of the river upward. These are the topsets of the delta

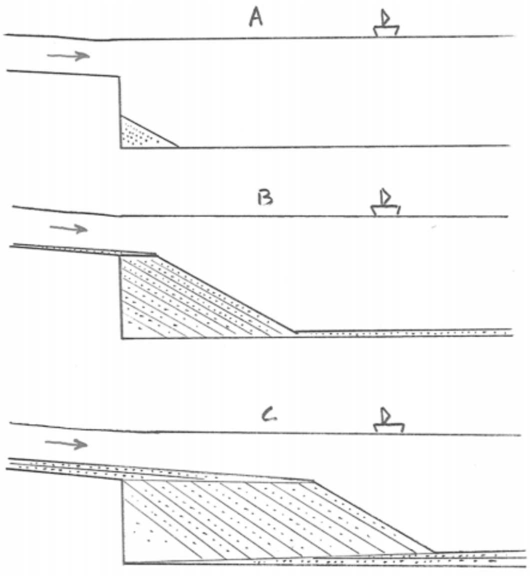


Figure 8-38. Building of a delta as a river enters a body of standing water. A) The earliest stage. B) A somewhat later stage. C) An even later stage, showing topsets, foresets, and bottomsets.

Finally, Figure 8-39 shows a plan view of the delta in Figure 8-38 after the delta has built out into the water body for an appreciable distance. The margin of the delta forms an arc, because sediment builds not just forward but also with a lateral component. Why? Basically because the tendency for deposition along the axis of the flow leads to a slight axial ridge, and then the flow tends to flow laterally off that ridge, down the slight sideways slopes of the delta body.

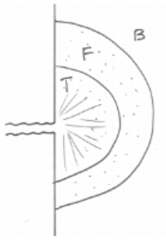


Figure 8-38. T, topsets; F, foresets; B, bottomsets.

Large Deltas in Nature

The delta described in the preceding section, and shown in Figures 8-38 and 8-39, is representative of the way deltas develop when a relatively small stream or river enters a water body in which currents and waves are minor. Such deltas are called **Gilbert deltas**, in honor of an early geomorphologist, G.K. Gilbert, who made the first systematic studies of such small deltas in the western U.S. in the late 1800s.

At this point we need to address the various complicating factors that set in when the delta-building river is larger and the water body into which the delta is building is not as placid as was assumed in the preceding section. Figure 8-40 shows one of several classifications of deltas.

Let’s deal first with the behavior of the river. The topset surface, which we’ll call the **delta plain**, is not a smooth surface. One reason for its nonsmoothness, even if only sand and gravel bed load is being carried by the river, is that the delta plain is braided. (Go back to the chapter on rivers to review the nature of braiding in streams and rivers.)

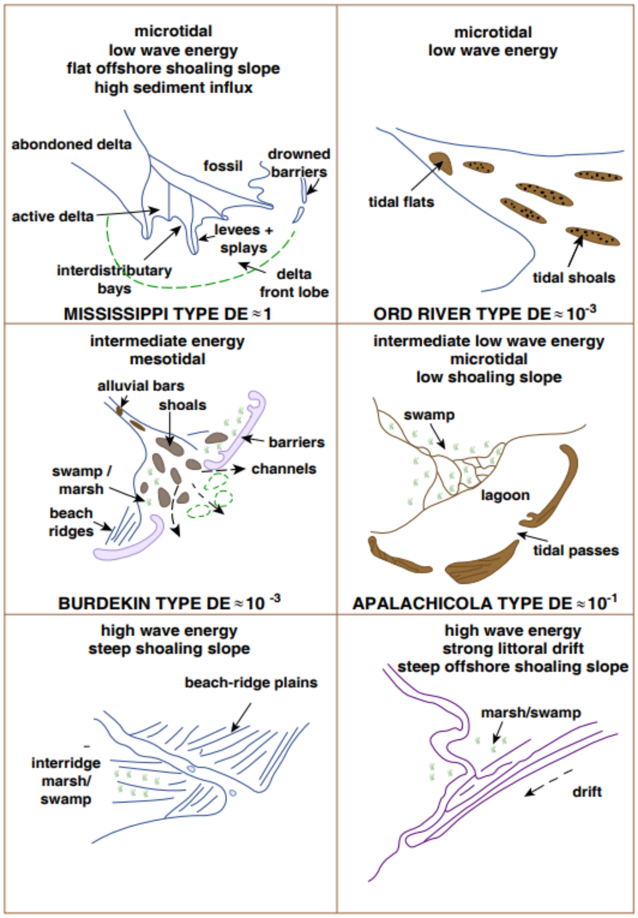


Figure by MIT OCW.

Figure 8-40. Classification of deltas. (From Carter, 1988.)

A more important reason, applicable to large mixed-load rivers, is that there is a strong tendency for upbuilding of the flow-carrying channel, deposition of natural levees, and occasional avulsion to relocate the channel to slightly lower areas on the aggrading delta plain. Typically there are two or more active channels, carrying water and sediment, operating on the delta plain at the same time. All the time, these shift around, avulse, and become abandoned, and new ones form. In that way, over time all of the delta plain is built up uniformly but with very complex internal structure in detail. The individual flow-carrying channels are called **distributary channels**. The next time you look at a map of Louisiana, note how the most recent active area of the Mississippi delta is building out into the Gulf of Mexico in the form of a hand-like body with several active distributaries (locally called “passes”).

Now we have to contend with the various effects that the water body itself has on the growing delta. We need to be concerned particularly with the action of waves, tides, and currents. Each of these factors is able to mold the morphology of the delta, in complex ways.

currents: The effects of throughgoing coast-parallel ocean currents are the easiest to understand. They tend to sweep the delivered sediment in one direction, causing slight to extreme asymmetry of the delta. Currents offshore of the mouth of the Amazon are so strong that even given the enormous sediment load of the river the delta does not protrude greatly into the ocean, the way the Mississippi delta does.

waves: ocean waves, approaching at a large angle to the coastline, tend to blunt the margin of the delta, giving it a fairly regular, arcuate shape. The Nile delta is the classic example.

tides: In regions with a large tidal range and strong tidal currents, deltas are shaped by strong back-and-forth flow in the various distributary channels, making for a strongly dissected delta body. The classic example is the Rhine delta, in the Netherlands.

The Mississippi delta is an excellent example of a delta that is dominated by the river itself rather than the water movements in the water body. That’s why the delta is able to build outward in long salients. Such a delta is picturesquely called a **bird’s-foot delta**. (Look again at a map of the Mississippi delta to get the concept.)

CHAPTER OVERVIEW

9: Deserts

9.1: Introduction to Deserts

I might define a desert as a land area characterized by sparse and infrequent rainfall. I wonder what images pop into your mind when I mention deserts: vast expanses of dry, rocky, rugged land? enormous shifting sand dunes? Both of these images are good representations of many deserts, but I think it's important to point out that

- most deserts have at least some vegetation, and
- sand dunes constitute only a small percentages of the area of most deserts.

It was once generally believed that the land surface of deserts is shaped mainly by the wind. But it's clear now *that except for areas of sand dunes, the desert landscape is shaped almost entirely by running water rather than by wind*. It rains, albeit infrequently, even in the driest of deserts. And in very dry deserts the vegetation is so sparse as not to provide much stability to the surficial layer of regolith, so the running water can readily mobilize the regolith and shape the landscape.

In the Atacama Desert in Chile the average annual rainfall is between one and two millimeters, and no rain at all falls for years at a time. But when it does rain in such a desert, a lot of rain—as much as several centimeters—can fall in a short time; that's when almost all of the sediment movement takes place.

Another misconception we Americans tend to have about deserts is that they are mountainous areas with characteristic relief made up of mountain ranges with intervening basins. In fact *most of the desert areas of the world are not mountainous*; they are broad plains of relatively low relief, except for large sand dunes in certain places.

9.2: Classification of Deserts

Deserts can conveniently be classified into three kinds:

- **High-latitude deserts**. At high latitudes, temperatures are so low most of the year, and sources of atmospheric moisture are so far away, that rainfall is very low. Water present near the surface is frozen for most or all of the year. In a sense, the surface of the northern part of the Greenland Ice Sheet and the interior part of the Antarctic Ice Sheet are deserts!
- **Mid-latitude deserts**. In temperate latitudes, deserts form only where for one (or both) of two reasons rainfall is scanty over large areas. Rainfall can be low because an area is located in the interior of a large continent, far away from marine sources of atmospheric moisture, like the interior of Asia. Or rainfall can be low because mountain ranges located upwind extract most of the moisture from the air. Such deserts might be called rain-shadow deserts. The deserts of the southwestern United States are of this kind.
- **Low-latitude deserts.** In the subtropical high-pressure belts of both hemispheres (the so-called horse latitudes), the atmosphere undergoes widespread subsidence as a consequence of broad patterns of the general circulation of the atmosphere, so cloudiness and precipitation is uncommon. Most of the great deserts of the world, in North Africa and the Middle East, are of this kind.

9.3: A Brief Summary of Desert Features

Introduction

Because eolian transport of sand, and the construction of eolian dunes large and small, is perhaps the most striking characteristic of deserts (even though, as noted above, fluvial sediment transport is much more effective in shaping the landscapes of deserts), I'll concentrate in the following sections on eolian saltation and eolian dunes. But in this section I'll mention just briefly several other aspects of deserts.

Alluvial fans

An **alluvial fan** is a fan-shaped or cone-shaped body of loose sediment deposited at the foot of a steep stream valley. Alluvial fans usually consist mostly of sand and gravel. Fans are deposited because the slope of the stream, and therefore the sediment-transporting ability of the stream, decreases abruptly where the stream leaves it mountain channel and spreads into a valley. At the head of the fan, the stream passes from being laterally confined to laterally unconfined, so on average it becomes wider and shallower, and therefore of gentler slope—although the transition is not really as abrupt as I indicated above.

The contrast in slope between the confined mountain stream and the unconfined fan surface is greatest in the very early stages of fan development, when there's only a small fan at the base of a steep mountain front, and much less in the later stages, when the fan has built to greater size and the stream has had a chance to cut a deep canyon and extend its headwaters far into the mountain mass.

Once on the fan, the stream progrades the fan locally, and in doing so the elevation of the stream bed is raised somewhat above the level of the adjacent fan surface. (By progradation I mean the tendency for the depositional surface to be built forward, in the direction of sediment transport.) At some point the stream 414 shifts its position to a lower area of the fan. In this way, over a long time period the stream sweeps irregularly across the entire fan, building it in a symmetrical, conical shape.

Alluvial fans are characteristic of deserts with high relief, especially in the southwestern US, but they are not common in many other deserts. You should also keep in mind that alluvial fans are not restricted to arid or semiarid regions: they form even in humid regions where there is high relief and rapid weathering. In such areas, however, alluvial fans tend to be heavily vegetated and not as noticeable.

Ventifacts

Rigid solid surfaces tend to be abraded by the impact of flying sand grains. The upwind-facing surfaces of outcropping bedrock or pieces of gravel are abraded to smooth and sometimes fluted surfaces. Such rocks are called **ventifacts,** or wind-worn stones. Loose pieces of gravel sometimes show a faceted shape formed by abrasion from two or three different directions—either by variability of wind direction or shifting of the position of the clast relative to the prevailing sand-transporting wind direction.

Desert Pavement

On old alluvial fans it's common to see a ground surface with a strikingly densely packed veneer of gravel. The substrate beneath consists of much finer sediment as well as gravel. The explanation is that the surface of an alluvial fan can become stranded, or cut off from fresh supply of sediment. This happens whenever there is tectonic uplift of the fan, causing the streams to be incised into the fan. Then slow **deflation** (removal of sediment by wind action) winnows the finer part of the surficial sediment, leaving only the coarsest fraction as a surficial residue. If you dig just below the surficial coarse layer you find a mixture of coarse material with much finer sediment. Only if the coarse surface layer is somehow disturbed can the wind act upon that underlying fine material.

Desert Varnish

Rock surfaces in the desert commonly show a dark coating, dark brown to almost black and sometimes even with a suggestion of darkest violet. The coatings, a small fraction of a millimeter thick (0.05–0.1 mm), are known from actual chemical analyses to be rich in iron and especially manganese, along with aluminum, silicon, and oxygen, but these elements seem not to be well crystallized into minerals.

Rocks most susceptible to desert varnish are relatively resistant to weathering: mafic and felsic volcanics, metamorphics, and well cemented sandstones. Rocks like felsic intrusives, which tend in deserts to weather fairly readily by granular disintegration, don't show good desert varnish, because the surfaces are degraded too fast by mechanical weathering. The same is true of limestones and even dolostones, because they weather by solution too fast.

The source of the Fe and Mn, as well as the mode of precipitation, are not entirely clear. The general consensus seems to be that Fe and Mn can be derived either from the interior of the rock itself or from soil in which rock fragments rest. (Other things—rock type and locality— being equal, rock fragments in regolith show better desert varnish than bedrock.)

Films of moisture must be invoked for transport of the Fe and Mn in solution; rainwater is an obvious possibility, but coatings of dew are likely to be even more important, because most desert rocks are wet for a much greater part of the time from dew than from rain. The Fe and Mn in solution could be transported by actual flow of the film, or by diffusion across it. The details of chemistry are not yet clear.

By various lines of evidence it's known that the time scales for development of a good coating of desert varnish vary greatly, from just a few decades in the most favorable circumstances to many centuries or even millennia.

Playas

In geomorphological usage, a **playa** is the barren, flat, and usually dry area at the lowest part of a basin of interior drainage. (In Spanish, playa means just beach or shore.) The typical playa is dry most of the time, except for occasional inundation by infrequent heavy rain in the watershed area. Playas vary in size from those you can walk across in a few minutes to those many kilometers across. With increasing frequency and magnitude of rainfall in the drainage basin, playas grade over into what might better be termed ephemeral lakes, and then into permanent interior-drainage lakes.

Most playa surfaces are virtually nonvegetated, except possibly at the very periphery, where relatively salt-free water emerges from drainage from the distal slopes of alluvial fans, which typically terminate at the playa edge. The lack of vegetation is clearly related to the elevated content of salts in the sediments of the playa.

The nature of playa surfaces varies widely, depending partly upon the ratio of dissolved load (salts in solution) to fine particulate sediment (mainly suspended load) in the flows reaching the playa, and partly upon the groundwater regime of the playa. Some playas, especially those that are topographically closed but have throughgoing flow of groundwater, are dry except soon after rains, because the water can drain out of the basin through an aquifer. Such playas have little or no salt in their sediments. Other playas, which are closed with respect to groundwater flow as well as to topography, tend to be moist long after rains, especially where the groundwater table, or at least the capillary fringe thereof, is always fairly near the playa surface. Such playas are characterized by much salt in their sediments, and some consist mostly or entirely of evaporites. (Evaporites are sediment deposits that originate from partial or complete evaporation of water that contains salts in solution.) The salt content of such playas is derived partly from weathering of rocks in the watershed (that salt may be from pore waters originally included in sedimentary rock or it may be from constituents of the minerals themselves) or from washing out of atmospheric salts, ultimately derived from the ocean. That not all evaporitic playas show the same content is good evidence that the salt content in the watershed is usually the more important of the two sources. Halite, gypsum, and borates are some of the common compositions of playa evaporites. The floor of Death valley is probably the most-visited playa, still in its natural state, in the United States.

9.4: Saltation

Earlier in the course I tried to give you a qualitative picture of the nature of sediment movement by flows of water. You learned that sediment is moved by water as bed load and as suspended load. The picture of sediment transport by air is somewhat different. This is basically because the ratio of sediment density to fluid density is almost three orders of magnitude greater in air than in water. This means that once a sediment particle is set in motion by a flow of air, the particle tends to follow a trajectory that is largely independent of the details of turbulence in the air, because of the great relative inertia of the particle. In water, on the other hand, sediment particles in transport tend to follow rather closely the patterns of motion in the water itself.

The forces the wind exerts on sediment particles resting on the land surface are not greatly different from those exerted by flowing water, and particles can be set into motion as bed load or lifted into suspension, just as in flowing water. But particles of sand size or greater, which are set into motion as bed load, are able to make much stronger impacts with the sediment surface, because of the relatively slight cushioning effect of the fluid. The consequence is far greater importance of ballistic or collisional effects in eolian transport. This is manifested in a characteristic mode of transport called saltation.

The term saltation is used for a mode of particle movement in which trajectories of particles show take-off from the bed at moderate to steep angles and descent to the bed at small angles. Figure 9.4.1 shows a typical saltation trajectory taken by a saltating grain of sand in air. Saltation trajectories are rather regular, and then they show little of the sinuosity that might be expected from passage through turbulent eddies. This is because of the typically large ratio of particle density to fluid density: the particles have great relative inertia, and they can cruise almost unaffected through eddies with rather different local fluid velocities.

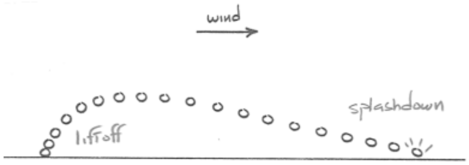


Figure 9.4.1: A typical saltation trajectory of a sand particle moved by the wind.

Saltation is the dominant mode of particle movement when a strong wind blows over a sand surface. Except in the very strongest winds, the saltation heights attained by the saltating grains seldom exceeds a meter, and the saltation lengths are mostly less than a few meters. There is, of course, a continuous distribution of jump heights and jump lengths, from zero to the maximum. There's also wide variability in take-off angles: they range from just a few tens of degrees to vertical. After colliding with some particularly immovable bed grains, some saltating grains even take off with an upstream component to their motion!

If you are lucky enough to be out on a dry sand surface during a strong wind, take the risk of getting some sand in your eyes and nose and mouth, and get down for a horizontal view at an eye level of a few tens of centimeters above the surface. You would see a hazy layer of saltating grains, which tails off imperceptibly upward. This well defined layer of saltating grains is called the **saltation carpet.** If you then looked downward at the sand surface, you woul see an abundance of surface grains being pushed and jogged along for short distances, just one or a few grain diameters at a time, presumably by being struck by saltating grains. That mode of movement is called **surface creep**—but there is actually no sharp break between surface creep and saltation.

To see individual saltation trajectories nicely you would have to be out there at night with a strobe light. The best thing to do is shine the strobe light straight down at the bed through a horizontal slit oriented parallel to flow; that way you see only a thin slice of the saltation carpet. The effect is striking. You could easily do the same thing in a homemade wind tunnel, which doesn't have to 418 be any bigger than tabletop size. The most common kind of wind tunnel for studies of sand movement (Figure 9-2) consists of a duct with a flared entrance, which passes into a large collection box with an exhaust fan on the other wall. You can start with the sand in the duct, or you can shower it into the upstream end through a slit in the roof of the duct.

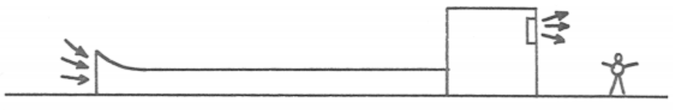


Figure 9.4.2: A wind tunnel of suction type.

Many aspects of saltation are not well understood. Among these is the mechanism that causes the initial rise of the grains. Two different effects might be important:

Aerodynamic lift forces

This is not the place to discuss the intricacies of the Bernoulli equation, which expresses the relationship between fluid pressure and fluid velocity in any flowing fluid. Suffice it to say that along any flow line in the flow there is an inverse relationship between the pressure and the velocity: where the velocity is high, the pressure tends to be low, and where the velocity is low, the pressure tends to be high. This is what makes airplanes fly (Figure $9.4.3A$): the shape of the wing section is such that the air travels a shorter distance around the lower surface of the wing than around the upper surface, so the velocity is lower and the pressure is higher. This pressure difference between the lower surface and the upper surface of the wing is called lift. The same is true of a sand particle resting on a sediment surface (Figure $9.4.3B$): the velocity is relatively low around the base of the particle and relatively high at its top, so there is a net upward pressure force in addition to the downwind drag of the wind. **Ballistic collisions between the moving particle and the bed particles**. When a flying particle makes impact with the bed, it imparts part of its momentum to one or more bed particles. One or more of the bed particles is likely thereby to be launched into motion above the bed. Take-off velocities and angles depend in some very complicated way upon the local geometry of arrangement of the bed grains and upon the velocity and location of impact.

Despite much research over the years, the relative importance of these two effects—aerodynamic lift and particle collisions—is still controversial.

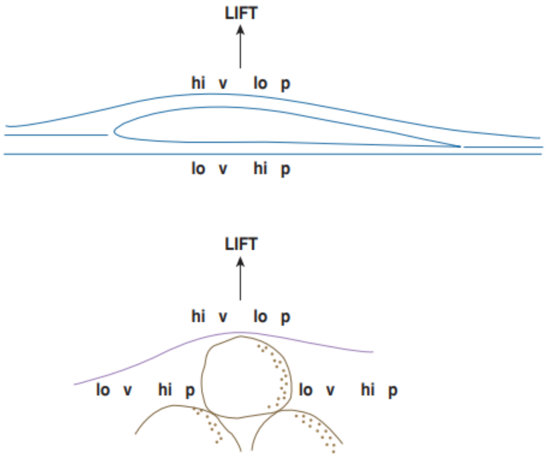


Figure by MIT OCW

Figure 9.4.3: Aerodynamic lift forces. A) On an airplane wing. B) On a sediment particle.

An interesting aerodynamic aspect to saltation has to do with the rotation rates of saltating particles. Ultra-high-speed motion pictures have revealed strikingly high spin rates of hundreds of revolutions per second. The sense of spin is such that that the top of the particle moves faster in the downstream direction than the bottom of the particle (Figure 9.4.4). The origin of such high rates of spin is still a mystery. But it's clear that the spinning motion gives rise to a non-negligible additional lift force on the particle, by what's known as the **Magnus effect**: the rotation of the grain alters the streamline pattern around it in such a way that instead of the pattern being symmetrical about the grain, the streamlines are closer together above the grain than below it (Figure 9.4.5). By the Bernoulli equation, this means that the fluid pressure at the grain surface is smaller above the grain than below, causing a net lift force. It's the same effect that makes for curve balls, and for those annoying slices and hooks.

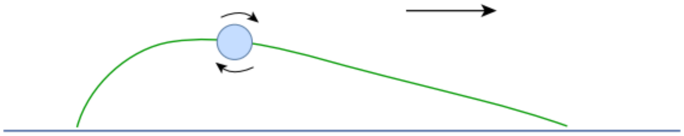


Figure by MIT OCW

Figure 9.4.4: The sense of rotation of a saltating particle.

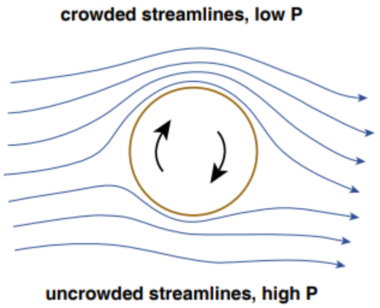


Figure by MIT OCW

Figure 9.4.1: The Magnus effect

Except in the very strongest of winds, saltating sand grains seldom attain heights greater than somewhere between one and two meters. So the Hollywood conception of a sand storm, in which the protagonist steps out of his tent into a blinding, choking cloud, is all wrong: in a real sand storm the air at one's eye level is clear, but one's legs are being bombarded by zillions of saltating sand grains. Hollywood sand storms are really dust storms, which can happen only where there's not only a strong wind but also a freshly available surficial layer of fine sediment, as from a recent flow of water after a heavy rain. Dust storms aren't very common in deserts, because these fresh supplies of sediment are made available to the wind so infrequently.

9.5: Eolian Ripplers and Eolian Dunes

You learned earlier that when a current of water flows over a bed of loose sand, if the current is strong enough to move some of the sediment several kinds of geometrical features, called bed forms, are molded by the flow. The same is true for flows of air over loose sediment. And in fact, just as in water flows, two different kinds of bed forms, ripples and dunes, are formed at widely different scales. A dynamical instability of some kind, whereby a planar sand surface is unstable to small disturbances, is involved in both cases, water and wind. The dynamics behind these instabilities is just as mysterious in the case of wind as in the case of water. It's not even known for sure whether subaqueous and subaerial ripples are the same dynamical kind of feature, and whether subaqueous and subaerial dunes are the same dynamical kind of feature.

Here's another home experiment for you to try. (I promise you that this will be the last home experiment I propose this semester!) It would be easy to build a home wind tunnel. (See Figure 9-2 above.) Just nail four wide pine boards together to make a duct several feet long, and arrange the downwind end to pass into a hole in a large box. In the other wall of the box, mount a window fan. Put a layer of sand in the duct, turn on the fan, and you have eolian transport. (It would help to improve the design in three ways: put a couple of Plexiglas windows in the duct so you can watch the sand move; make the upper surface of the duct removable so you can get at the sand and photograph it easily; and arrange a flap board against the downwind side of the fan to control the wind speed.)

Gradually increase the wind speed over a planar sand surface in your duct until the sand grains just begin to move. Very soon after that, fully developed saltation sets in. If then you gradually decreased the wind speed again, you would find that the minimum wind speed needed to maintain already existing saltation is much less than the wind speed needed to initiate saltation. A physicist would call this a hysteresis effect. (Can you recall an earlier example of hysteresis in this course?) These two wind speeds represent what are called the **saltation threshold** and the **fluid threshold**, respectively.

Crank up the wind again until there's saltation. You would have to wait only a matter of minutes until the originally planar sand surface becomes molded into a strikingly regular series of wind-transverse rounded ridges and troughs called **wind ripples or eolian ripples**. The typical spacing of wind ripples is of the order of several centimeters, although they get bigger in poorly sorted and/or coarser sediment. They shift very slowly in the downwind direction, by movement of sediment up the upwind flanks and deposition of that sediment on the downwind flanks.

There's a tendency for the particle size to be slightly coarser on the crests of the ripples and slightly finer in the troughs. Long ago R.A. Bagnold, a pioneer in the study of eolian sediment transport, devised a spectacularly visual 422 demonstration of this effect by making wind ripples in a thoroughly mixed sand of two different colors in a wind tunnel: as the ripples develop, the bed surface resolves itself into alternating stripes of the two different colors right before one's eyes!

Eolian dunes are not as easy to make in experimental wind tunnels, because their minimum scale is too big. Almost everything we know about dunes is from studies in the field. One of the big problems in the studying of eolian dunes is that there aren't many places in the world characterized by really steady winds: obviously the wind speed varies with time, but in most places the wind direction changes substantially too. Nobody really knows what the reference case of dunes formed by a steady wind over a full sand surface of effectively infinite extent looks like.

Eolian dunes take many different shapes, and there is no generally accepted classification. One reads about linear dunes, longitudinal dunes, crescentic dunes, seif dunes, oblique dunes, barchan dunes, star dunes, and many others. I think a fairly satisfactory but simplified way of classifying dunes is to recognize three kinds of dunes formed in full sand beds depending on orientation relative to a dominant sand-moving wind:

- **transverse dunes**, oriented normal to the wind
- **longitudinal dunes** (also called **seif dunes**), oriented parallel to the wind
- **oblique dunes**, oriented oblique to the wind.

It's not even clear whether these three basic kinds are dynamically different from one another or not. Dunes formed by winds blowing from various directions without a dominant sand-movement direction are called **star dunes.** Star dunes are found where sand-transport paths converge to form a sand "sink" or sand-storage area. Finally, dunes formed in areas where there is not enough sand to keep troughs from being floored by immobile substrate take a characteristic crescent shape are called **barchan dunes**. There's a continuous transition in geometry from barchan dunes to transverse dunes as the thickness of the sand increases relative to dune dimensions, so that less and less of the troughs expose immobile substrate.

Of course, dunes are found not just in deserts. **Coastal dunes** are common along many shorelines where onshore winds pick up beach sand and carry it inland to form belts of dunes, narrow or wide. The prerequisite for any area of sand dunes is a source of sand. In deserts that source might be fluvial sediment freshly deposited where a river finally loses all of its water by infiltration into the river bed, or just an area where sand from sandstone bedrock is freed by slow weathering.

If the sand supply is abundant and the transport is consistently in one direction away from the source, enormous areas of deserts can be covered with mobile sand. Such areas are usually picturesquely called **sand seas**. Sand seas are especially prominent in North Africa and the Arabian Peninsula.

Readings

Allen, P.A., 1997, Earth Surface Processes. Blackwell Science, 404 p. (Chapter 10)

Bagnold, R.A., 1941, The Physics of Blown Sand and Desert Dunes. Chapman & Hall, 265 p.

Bloom, A.L., 1998, Geomorphology; A Systematic Analysis of Late Cenozoic Landforms, Third Edition. Prentice Hall, 482 p. (Chapter 13)

Easterbrook, D.J., 1999, Surface Processes and Landforms, Second Edition. Prentice Hall, 546 p. (Chapter 17)

Greeley, R., 1985, Wind As a Geological Process, on Earth, Mars, Venus, and Titan. Cambridge University Press, 333 p.

Pye, K., and Tsoar, H, 1990, Aeolian Sand and Sand Dunes. Unwin Hyman,

---

CHAPTER OVERVIEW

10: Mass Wasting

10.1: Introduction to Mass Wasting

Everything at or near the surface of the Earth is pulled toward the Earth by the force of gravity. ("Gravity: it's a law you can live with.") That includes all Earth materials, rock and regolith. This is one of the central concepts of this course. Its importance lies in the fact that under certain circumstances those materials are moved downslope by the pull of gravity. Processes of downslope movement of surficial Earth materials under the pull of gravity are collectively termed **mass wasting**. (I've always thought that term to be rather infelicitous, but there it is.) This is a good time to look at the two background sections on gravity at the end of this section.

You might view the problem of mass wasting in terms of an atrocious paraphrase of a famous quotation: "To move downslope, or not to move downslope?" The criteria for initiation of downslope movement of Earth materials by gravity are not straightforward, but they are of great practical as well as theoretical importance: mass-wasting events are responsible for enormous loss of life.

Processes of mass wasting are highly varied in three important respects:

- mass of material involved
- speed of movement
- nature of movement

Masses of material vary from tiny mineral grains tumbling downslope to enormous masses with volumes of rock and mineral material as great as thousands of cubic kilometers. (Yes, thousands; that's not a typo.) Speeds range from imperceptibly slow, less than a meter per year, to hundreds of meters per second. The nature of movement ranges from intact masses, which retain their original structure, to those that become thoroughly mixed and homogenized during the movement. Given this wide range of characteristics, it should not come as a surprise to you that classification is difficult and terminology can be confusing.

BACKGROUND: GRAVITY

Everybody knows about gravity. For the sake of clarity, though, here's some background material on it.

First of all, there's a distinction between gravitation and gravity. **Gravitation** is the more general concept: it's an attraction between any two bodies of matter in the universe. As with many important things in physics, it began with Newton: he formulated what is now called Newton's law of universal gravitation, that any two bodies of matter exert an attractive force upon each other that's proportional to the product of their masses and inversely proportional to the square of the distance between them. (It's one of a number of manifestations of what physicists call "inverse-square laws".) Does it surprise you to learn that gravitation is an extremely weak force? It gets big only when one or both of the bodies is very massive, like stars or planets.

The Earth exerts a force of gravitational attraction upon us, and upon everything else on Earth as well. That's what's called the **force of gravity**. (We exert an equal and opposite force on the Earth, but we usually don't bother to think about that. It probably seems ridiculous to you to think that each of us exerts that big force on the whole Earth, but it's true.)

What we call **weight** is just the force of gravity the Earth exerts on us or on other bodies of matter. It's important to be clear on the distinction between mass and weight. The mass of a body is a measure of the amount of matter of which that body is composed; the weight of a body is the force of gravity the Earth (or any other solar-system body, for that matter) exerts on the body. That's why the astronauts weighed so much less when they were on the Moon, and why astronauts in deep space are weightless, even though they are not massless.

BACKGROUND: THE DOWNSLOPE COMPONENT OF GRAVITY

The force of gravity that the Earth exerts on us is directed toward the Earth's center of mass—which is almost identical to the Earth's center of volume. (It's just generally called "the center of the Earth".) The way "horizontal" is defined, anywhere on Earth, is just the plane that's perpendicular to the direction of the force of gravity, at that locality. Any still water surface is coincident with that horizontal plane—because if the surface were not horizontal, it would be sloping, and gravity would pull the water down the slope until no slope remains.

That leads us directly to the important idea that the force of gravity can be viewed as having a part that acts in all directions that are not coincident with the horizontal plane. That force is called the **component** of the force of gravity in that particular direction. To deal with that quantitatively, you need to know something about the mathematics of vectors. (A vector is a quantity that has both magnitude and direction, and behaves according to a particular set of rules; no details here.) You can resolve a vector into components along any set of axes, by certain mathematical rules.

Figure 8-1 shows how the force of gravity of a body resting on a sloping surface can be resolved into two components, one parallel to the surface and one perpendicular to it. What I want you to see in this is that the downslope component of gravity gets smaller as the slope gets gentler: it varies from the full force of gravity, in the vertically downward direction, to zero, on a horizontal, surface.

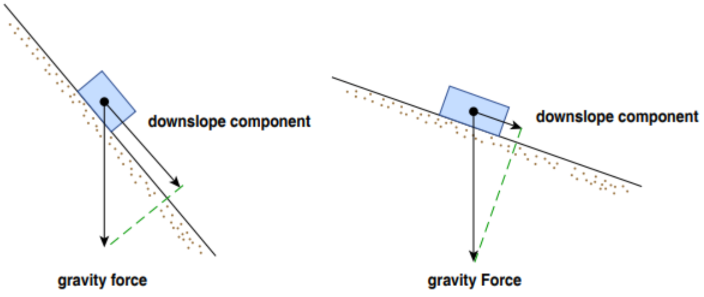


Figure by MIT OCW

Figure 10-1. Resolving the force of gravity into a component in the downslope direction. The lengths of the arrows represent the magnitudes of the components of the gravity force. Left: a steep slope. Right: a gentle slope.

All of this should make good qualitative sense to you, because it's part of the human experience: we are pulled down a steep slope more than down a gentle slope. The same is true of all rock and regolith on the Earth's surface as well, and that's what's important for this chapter.

10.2: The Controls on Downslope Movement

Envisioning the Problem

At this point we need to take a closer look at what controls whether a mass of material will slide down a slope. Just to get your thinking started, here are two seemingly dissimilar but actually closely related situations:

- Making a pile of dry sand by pouring the sand slowly and continuously down toward the same point (Figure 8-2A)
- Placing a thick layer of soil on a horizontal sheet of plywood and then slowly tilting the sheet until some or all of the soil slides off (Figure 8- 2B)

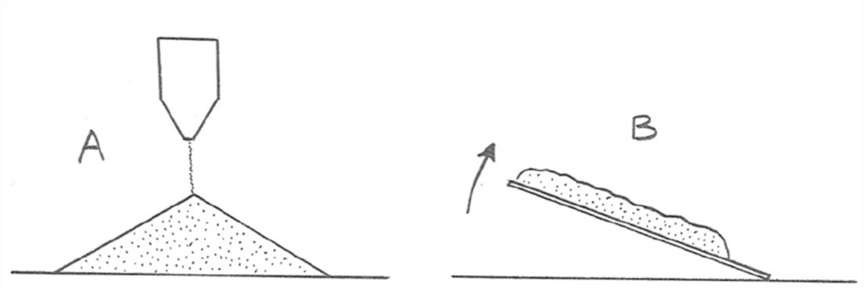


Figure 10-2. A) Making a pile of dry sand. B) Sliding a layer of soil off a silted plywood sheet.

You must know, from your own experience, that when you try to build the sand pile in Figure 8-2A to have steeper slopes, a surface layer of sand suddenly slides down the slope, thereby decreasing the slope angle. As you build the pile to a larger and larger volume, the alternation of slope-increasing intervals and slope-decreasing intervals balances out, resulting in a constant average slope angle. That angle is called the ***angle of repose***. For dry granular materials that are not extremely angular or jagged, the angle of repose lies in the range 30–35°.

Very angular blocks, of the kind that are common on talus slopes in mountainous areas, have greater angles of repose, but seldom more than 45°. (To me, such slopes always seem greater than they really are when I'm walking on them.)

As you gradually increase the slope of the plywood sheet in Figure 8-2B, at some point the whole mass slide off the sheet. ("Duh", you're probably thinking.) As a variant on the experiment, however, you might make the upper surface of the sheet ribbed or corrugated, with the ribs running in the transverse (horizontal) direction. How would that affect the value of the slope angle at which failure takes place?

Analyzing the Problem

In both of the situations described above, we need to think in terms of the force of gravity on a layer of material in the shape of a slab with its upper surface at the ground surface and its lower surface some depth below the surface (Figure 8-3). If the slab is stationary (not moving downslope), then there must be a friction force that's exerted by the material underneath the slab on the material of the slab. We can place this imaginary plane anywhere in the material we want. Whenever and wherever the downslope gravity force exceeds the friction force, the slab above the plane slides downslope. We say that the material has *failed* along that plane. (The friction force we are dealing with here is the same as what I called the shear strength earlier in the course.)

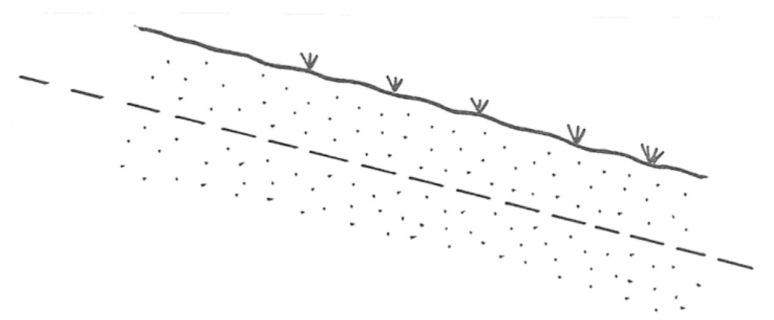


Figure 10-3. The force of gravity on a layer of deformable material on a slope.

BACKGROUND: FRICTION

Everybody knows about friction. Life without friction would be unbearable. Friction is essential for locomotion. We would all end up clumped together at the bottoms of local depressions in the land surface, unable to shift our positions in the least.

One instructive but technically incomplete definition of ***friction*** is that it's *the force that is generated when one mass of material slides past another mass of material with which it is in contact across a surface*. That definition is a bit too specialized, because it doesn't adequately cover the situation when a continuous medium, like water, undergoes a shearing deformation, as shown in Figure 10-4.

To get a handle on shearing deformation, think in terms of taking a thick telephone book and "racking" it so that the pages slide past one another, and then supposing that the shearing motion is continuous throughout, rather than page by page. In such shearing of a continuous medium, there is friction across any imaginary plane through the medium (Figure 10-4B).

The phenomena of friction are messy. They involve the details of the surface along which the friction force acts, on scale ranging from macroscopic bumps and corrugations down to the atomic scale. Only recently have physicists begun to inquire deeply into the nature of friction.

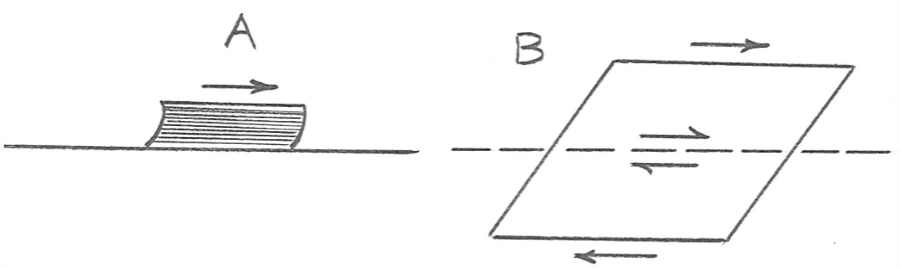


Figure 10-4. A) Shearing a thick telephone book. B) Shearing a continuous medium. The dashed line through the medium represents an imaginary plane of shearing. This single-barb arrows represent the friction force exerted by the material on one side of the plane on the material on the other side.

Friction can be dealt with by gathering or subsuming all of the messy phenomena into a single coefficient. (The fancy general term for that kind of thing is called *parameterizing*.) Figure 8-5 shows a slab being made to slide on a horizontal surface by exerting a horizontal force on it. When the slab is moving at constant speed, the force that needs to be exerted on the slab to keep it moving is equal to the friction force between the bottom of the slab and the underlying surface.

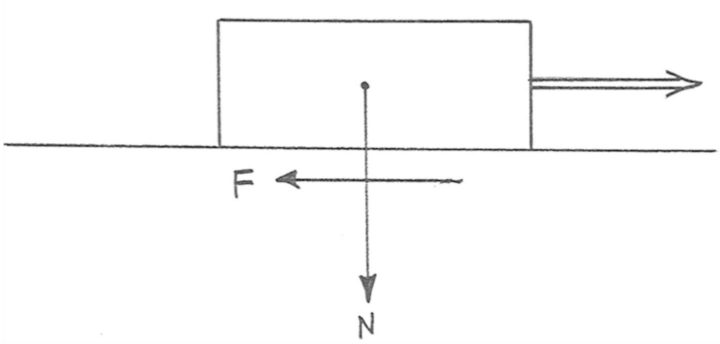


Figure 10-5. Friction on a slab sliding on a horizontal surface.

Let the weight of the slab, per unit horizontal area, be N, and the friction force, again per unit area of contact between the slab and the underlying surface, be F (Figure 10-5). These two quantities, N and F, are related by the equation $F = \mu N$, where μ is called the coefficient of sliding friction. It depends only on the nature of the surfaces in contact. Its value is usually less than one. There's an analogous coefficient for ***static friction***, in the situation where the force that tends to move the slab is not yet big enough to cause the slab to move. The coefficient of static friction is generally larger than the coefficient of ***sliding friction***.

If you run the experiment on sliding of a layer of material down your sheet of plywood, you would find that the failure plane (the plane where sliding is located) *is always at the bottom of the layer of material*, whether or not the surface of the plywood is smooth or corrugated. It's easy to see why. The downslope driving force (the weight of the material above a given plane parallel to the plywood sheet) increases with the thickness of the material above that plane, but the coefficient of friction is the same throughout the material. The shear strength of the material (or, what is the same, the static friction) is therefore overcome first at the deepest plane in the layer.

The situation with your sand pile is a bit different. What's happening there is that the slope is oversteepened at the tippity-top of the pile, as a lobe of new material is built. Eventually the lobe fails, and its material slides down the slope.

The Real World

Now, with all of the foregoing material on the downslope pull of gravity, the nature of friction, and the shear strength of materials as background let's think about *slope failure in nature*. The problem is easy to state. In most places, the land surface has some slope angle. That leads to the possibility of failure and sliding of material downslope. That's the essence of mass wasting. Most of the time, of course, the stuff just sits there on the slope, not moving. (But I will be amending that seemingly obvious statement, in a very important respect, in a later section.) At certain times and places, however, failure occurs, and there is sudden downslope movement, often fast, of some volume of material, often large. What are the factors that determine when and where that will happen? There are two aspects that are important to think about:

- What might cause previously stationary material on a slope to be mobilized?
- How deep below the surface will the failure plane be located?

Let's address the second question first. In nature, with its layer of regolith and weathered rock at the surface and solid bedrock, with its greater strength below, the depth at which failure first occurs depends on the competition between two effects: the tendency, discussed above, for failure to occur as deep as possible; and the general tendency for the shear strength of the material to increase, usually greatly, downward in the transition from regolith to bedrock.

The first question is more difficult to address. Several things, all of them important, might cause failure:

- The material on the slope might be loaded from above, by addition of new layers of material. If there is no concomitant increase in the shear strength of the underlying material, that might lead to failure.
- The slope might be steepened by differential deposition, whereby a layer of newly deposited material is thicker high up on the slope than low down on the slope. That increases the downslope component of the gravity force on the material.
- The base of the slope might be undercut, by stream erosion, for example.
- The shear strength along some plane in the material might be lowered by some effect.

The last of these calls for further comment. The shear strength of a granular material like regolith, as well as a solid but porous material like sedimentary rock, arises from the particle-to-particle forces. If that force can be decreased somehow, the shear strength is lessened. A good way to do that is to increase the pore pressure in the layer. (By ***pore pressure***, I mean *the water pressure of the water in the connected pore spaces in the material*.) Figure 10-6 shows this effect, in a very schematic way. Suppose that there is a layer of relatively permeable material that lies between an overlying layer and an underlying layer of effectively impermeable material. If in some way the fluid pressure in the permeable middle layer is pumped up, that extra pressure tends to lift up the overlying material, thus relieving some of the particle-to-particle forces and lowering the shear strength. That's conducive to failure, and sliding of the overlying layer downslope.

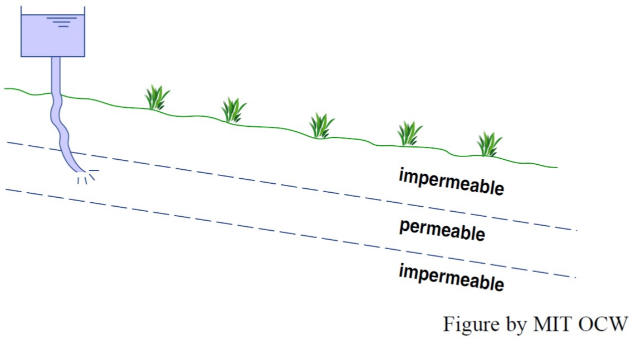


Figure by MIT OCW

Figure 10-6. Increase in pore pressure in a subsurface layer.

10.3: Kinds of Mass Wasting

Enough of theories and generalities. *What kinds of mass-wasting processes are important?* Mass downslope movement of bedrock and/or regolith is a complexly (almost hopelessly) multidimensional phenomenon. Here is a list of some of the factors in this complexity (I'm not pretending that it's a complete list):

slope of the land surface
composition of the surficial material
nature and degree of fracturing of bedrock
degree of weathering of bedrock
rate of downward transition from regolith to bedrock
size, shape, and sorting of regolith particles
porosity
permeability
degree of water saturation of medium
extent of excess pore pressure
fauna and flora
undercutting at base of slope
earthquake vibrations

Accordingly, *the mode, speed, and volume of downslope mass movements varies enormously*. Below is a list of some types of commonly recognized modes of movement (in alphabetical order). I think that an appropriate adjective in this situation is "bewildering". Many, if not most, of these grade into one another, without clear boundaries.

creep
debris avalanche
debris flow
debris slide
earthflow
gelifluction
lahar
landslide
mudflow
rock fall
rock slide
slump
solifluction
sturzstrom

The most general term in the foregoing list is ***landslide***. That term applies to *all perceptible movement of rock or regolith down a slope*. Two further qualifications are still needed, though:

(1) The adjective "perceptible" leaves out a class of very important mass-wasting processes—*creep* and *solifluction*—that are of great importance but are too slow to stand around and watch.

(2) There's a phenomenon called *debris flows*, whose nature lies somewhere in between what's generally considered to be landslides, on the one hand, and sediment-transporting streamflow (the topic of an earlier chapter) on the other hand.

Many classifications of types of mass wasting have been proposed. It's generally agreed that the definitive criteria for such classifications are as follows:

- *type of material in motion* (particle size, degree of coherence)
- *mode of motion* (falling, toppling, sliding, flowing)
- *speed of motion*

Rather than trying to present to you various technical classifications, I'm going to deal in the following sections with several of the most important kinds of mass wasting. Given what's said above, it seems natural to do this in the form of three sections: creep and solifluction; landslides; and debris flows.

10.4: Creep

The expressive term ***creep*** is used for *all slow downslope movements of regolith under the pull of gravity that are so slow as to be imperceptible except to observations with long duration* (days to weeks in the case of solifluction; years to decades to centuries in the case of slower creep). It is unspectacular in comparison to the sudden, large-volume, high-speed landslides that make the news reports, and only negligibly destructive (nobody dies from creep!). Its great geological importance arises from its ubiquity.

We perceive creep to be a continuous process, as over the years we observe the slow downslope movement it engenders, but in fact it's the sum of innumerable small and discrete movements of the slope-mantling regolith. These slow movements are brought about by a number of processes. Classic creep is brought about not by bodily movement of the surface layer above a plane of failure, as discussed at length in earlier sections, but by individual cyclic movements of the material.

The basic idea is this: particles or small masses of material at or near the surface are lifted upward perpendicular to the surface by any of a number of cyclic processes that involve lifting of material perpendicular to the surface and then lowering of that material in a much more nearly vertical direction. Here's a list of some of the important processes that are thought to contribute to creep:

- wetting and drying
- heating and cooling
- freezing and thawing
- transfer of material to the surface by burrowing organisms, then vertically downward collapse of material overlying the resulting cavities

All such processes affect only the very near-surface layer of the regolith, generally down to depths of no more than a few meters. The result is a kind of "sawtooth", step-by-step movement of material downslope. Figure 8-7 shows how this effect works in the case of soil particles lifted up by expansion and dropped down again by contraction. Similar figures could be drawn to show the sawtooth effect of other raising-and-lowering processes.

The sawtooth movement involves a variety of kinds of material:

- individual particles resting directly on the surface
- more extensive masses of the surface layer of regolith
- material brought up from deeper in the profile and deposited on the surface

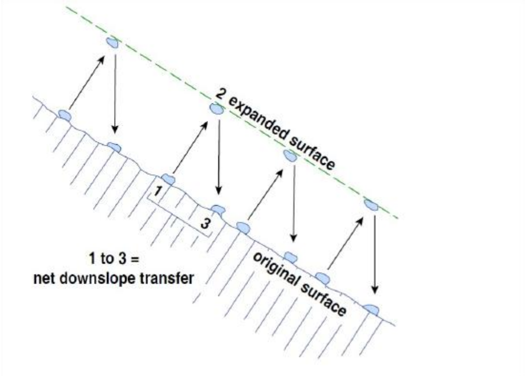


Figure by MIT OCW

Figure 10-7. Sawtooth path of a surface particle caused by expansion and contraction of the soil. (From Bloom, 1998)

The role of the lowly earthworm is believed to be of special importance in soil creep. (I seem to remember reading once that Darwin himself was aware of the importance of earthworms in soil movement.) The volume of soil processed by earthworms in an ordinary soil in humid temperate regions is staggering. As you must know from your own observations, earthworms work their way to the soil surface at night, leaving those telltale little piles of soil around their holes. In that way there is a net upward movement of soil material normal to the surface. Eventually the tube left behind by the burrowing earthworm collapses, and the tendency is for the collapse to be vertically downward under the pull of gravity. The net result is a sawtooth movement of soil material downslope.

Figure 10-8 shows common effects of creep. You are not likely to see all of them in one place, but all are clear manifestations of creep.

Solifluction is a special kind of creep, by which *a surface layer of water- saturated regolith flows imperceptibly slowly downslope over an impermeable lower layer of some sort*. The impermeability of the lower layer prevents drainage of the overlying soil, causing it to remain for long periods of time in a thickly soupy condition, which predisposes it to downslope flow. It is common in, but not restricted to, high-latitude regions of permafrost, where summer thawing affects only the surface layer, leaving frozen and impermeable material beneath. Solifluction happens also where there is an impermeable layer of clay-rich “hard pan” beneath a permeable surface layer, and even where there is an impermeable layer of bedrock just below a surface soil zone. In any case, solifluction is promoted by a high percentage of clay in the surface layer. Downslope speeds of solifluction, although still not something you can detect by standing there watching, are much higher than in other kinds of creep: up to some centimeters per day.

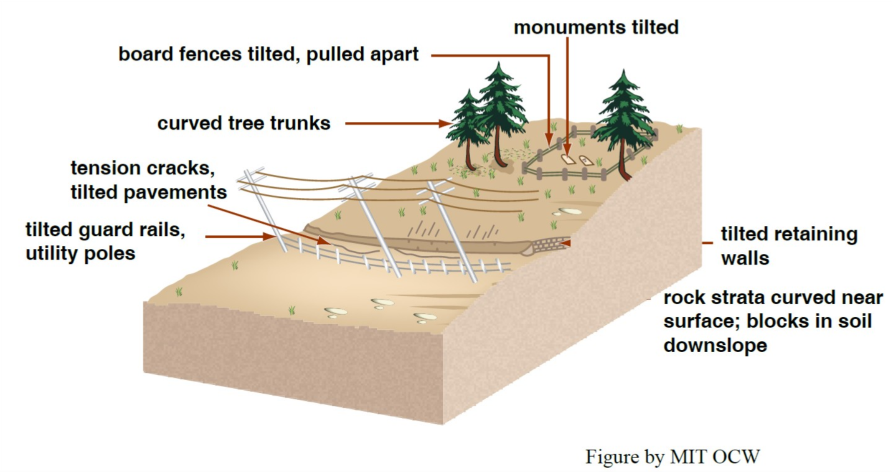


Figure 10-8. Common effects of creep. (From Bloom, 1998)

10.5: Landslides

As noted above, the term *landslide* is used for a variety of downslope movements of rock and regolith. These movements vary greatly in size, speed, and mode of movement. The simplest to understand (see Figure 10-9), and which are dealt with only briefly here, are ***rock falls*** (rapid fall of rock material down a cliff face), ***rock topples*** (downslope collapse of an originally vertical slab of rock that becomes detached along vertical fractures from a main rock mass), and ***rock slides*** (whereby a tabular mass of rock glides down a slope, which is usually underlain by more of the same rock, with planes of weakness parallel to the slope)

Another characteristic kind of landslide, in which *a large mass of earth or rock material moves downslope along a discrete shear surface of failure*, is called a ***slump*** (see Figure 10-10). The underlying surface is concave upward, and the mass rotates, in the sense of becoming more and more tilted upslope as it moves downslope. As the head of the slump evacuates the original space, a well- defined steep headscarp develops. The downslope segment of the slump, near the toe, commonly flows out onto the preexisting land surface as an earthflow, often with transverse ridges caused by compression as the earthflow feels the frictional resistance of the underlying surface. Slumps like this typically move slowly—not so fast that you can't get out of the way. They are especially common in regions with steep slopes, thick layers of regolith rich in fines, and occasional heavy rains. Many have occurred in coastal central and southern California in historic times. In many cases they have been triggered by careless undercutting of the base of a slope during road construction or other excavations.

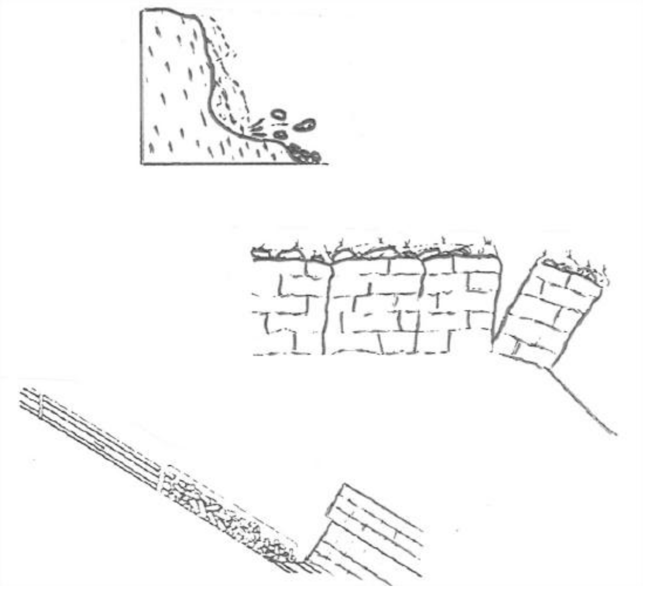


Figure 10-9. Rock falls, rock topples, and rock slides. (After Bloom, 1998)

The largest and most destructive of landslides originate by failure of enormous masses of partially weathered and strongly fractured rock perched high on a mountain slope. Failure, along a master underlying slip plane, causes the mass of material to accelerate downslope at speeds in excess of a hundred meters per second. Most of the famous landslides in history, and in prehistory as well, have been of this kind. Failure is sometimes triggered by heavy rains, especially in springtime after a thaw, or by undercutting at the base of the slope. Others happen just at random—when the time has come for failure. There may be premonitory slow movement before failure, but prediction is difficult. Volumes of rock material involved can be many cubic kilometers.

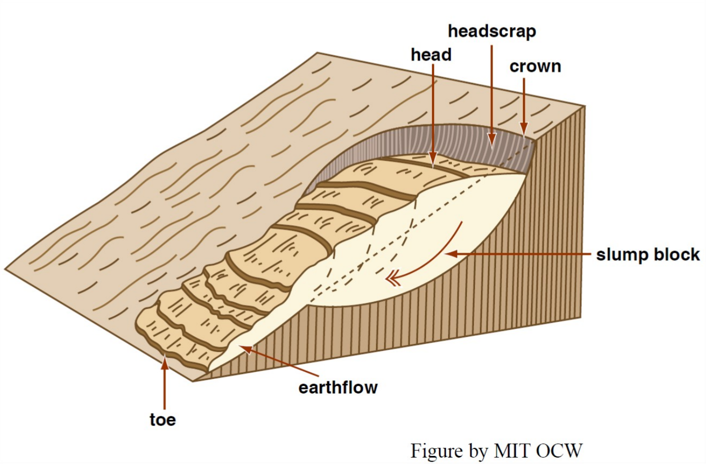


Figure 10-10. A slump. (From Bloom, 1998.)

In some cases, these large and rapid landslides have run out for long distances beyond the base of the slope down which they traveled, for several kilometers. The mechanism(s) that allow such long runouts over gentle slopes are not entirely clear. For some such landslides, it's been proposed, on good theoretical and observational grounds, that the slide became launched into the air, perhaps when moving over a "lip" of bedrock, and then rode on a cushion of compressed air. Friction at the lower surface would then be far less than in movement over solid materials. (Think in terms of air hockey.) When the cushion of air is finally depleted, by upward flow through the porous and permeable moving material, the slide grinds to a halt.

As you can imagine, some such landslides have resulted in great loss of life. The worst in recorded history was in Peru in 1970, which involved an enormous volume of rock and ice that moved at speeds of over 300 kilometers per hour and killed as many as 40,000 people. Another famous landslide occurred in northern Italy in 1963, when the entire mountainside, a slab 2 km long, 1.6 km wide, and 250 m thick, flanking the Vaiont Reservoir failed and slid into the reservoir, causing a catastrophic flood that killed more than two thousand people living in the valley below the dam.

This course is oriented toward what happens on the land surface of the continents, but it's important to point out here that the very largest landslides known have occurred undersea. It's only been in recent years, with the advent of increasingly sophisticated surveying and imaging techniques, that their extent and importance have become clear. The two locales where the largest of such landslides have been observed are on the flanks of the great Hawaiian shield volcanoes (much the greater part of whose volumes lie below sea level) and along the margins of the continents, below the break in slope at the edge of the continental shelf. Some of these are truly enormous, with volumes of thousands of cubic kilometers.

10.6: Debris Flows

Debris flows are *concentrated mixtures of water and loose rock and mineral material that flow downslope, usually in a preexisting channel, under the pull of gravity*. Speeds range from a slow walk to a speeding automobile. They differ from ordinary sediment-transporting streamflow in that the motivating force for the downslope movement comes directly from the pull of gravity on the sediment–water mixture. In sediment-transporting streamflow, in contrast, the water flows because of the pull of gravity, and the sediment is moved by the flowing water. (See the earlier chapter on rivers for a lot more material on sediment-transporting stream flow.)

The classic debris flow has sediment concentrations by volume of over fifty percent—almost to the point at which the sediment particles lock together and prevent movement. It also is characterized by extremely poor sorting of the solid material, from clay-size particles to house-size boulders. Up until recently, it was generally believed that the proportion of mud in a debris flow had to be high. It's know known, especially from laboratory experiments on debris flows, that such flows can happen even at concentrations of muddy material as low as five percent. But below that concentration of mud, debris flows can't happen.

The classic debris flow is nonturbulent to only weakly turbulent: the mixture moves smoothly and without the vigorous turbulent mixing that's so characteristic of sediment-transporting streamflow. Many debris flows, however—the ones that are relatively "thin" and somewhat more like ordinary streamflow—are clearly turbulent.

One of the big sticking points in accounting for the existence of debris flows has been *the mechanism (or mechanisms) by which large clasts can remain suspended in the flow in the absence of turbulence*. A number of effects seen to contribute.

- First of all, there's the **buoyancy effect**: the large clasts feel themselves to be immersed in a fluid medium with density much greater than that of clear water, because of the high concentration of finer sediment. The large clasts are therefore only slightly negatively buoyant.
- In flows with relatively high concentrations of clay-mineral particles, what is called **matrix strength**, caused by the electrostatic cohesive forces between clay particles, probably plays a role.
- Finally, the upward **dispersive effect** of strong collisions among clasts as the mixture is sheared is thought to be an important effect.

Debris flows are of much more than scientific interest. That's because deposition from debris flows is typically by rapid halt of the entire mass rather than incrementally along the path of the flow. Large, fast-moving debris flows can bury entire valleys, demolishing villages and all of their inhabitants, when the moving mass finally comes to rest. Stream valleys on the flanks of active explosive volcanoes in humid regions are especially susceptible, because volcanic ash weathers readily to fine-grained, clay-rich material, and heavy rains can mobilize such material into a massive debris flow, called a ***lahar*** (an Indonesian word).

READINGS

Bloom, A.L., Geomorphology; A Systematic Analysis of Late Cenozoic Landforms, Third Edition. Prentice Hall, 482 p. (Chapter 9)

Easterbrook, D.J., 1999, Surface Processes and Landforms, Second Edition. Prentice hall, 546 p. (Chapter 4)

CHAPTER OVERVIEW

11: Landscapes

11.1: Introduction to Landscapes

A course on the environment of the Earth's surface would be seriously incomplete without a chapter on the Earth's landscapes. An early chapter dealt with practical matters of describing the lay of the land by means of topographic maps. Now is the time to give attention to nature and origin of the Earth's landforms and landscapes.

Everybody knows about landscapes. Dictionaries define a ***landscape*** as *the aggregate of surface topographic features in some region as produced or modified by geologic processes*, or as *a region of the Earth's surface that the eye can see in a single view*. A ***land form*** is *some topographic feature of the Earth's surface that originated by natural processes*. You can think of a landscape as consisting of a number of individual landforms, of various kinds, in some definite relationship one to another. Hills, mountains, and valleys—of which there are many kinds—are examples of landforms. You have already learned much about specific landforms, in the earlier chapters on rivers and glaciers. Now is the time to have a more systematic look at the Earth's major landforms.

The study of landforms is a matter dealt with in the science of ***geomorphology*** (the study of the Earth's landscapes and the processes that shape them), a branch of Earth science that has been around since early in the modern age of Earth science, in the nineteenth century. Geomorphologists view themselves in part as geologists and in part as geographers. In the US, geomorphologists have tended to be geologists, whereas in the British sphere they have tended to be geographers. (Turf distinctions, although to a large extent artificial, have always been with science, as is true of many other fields as well.)

There is a longstanding dichotomy in geomorphology between the study of the history of development of landscapes, on the one hand, and the study of natural processes that shape the Earth's surface. The latter is usually referred to as "process geomorphology", and has been alive and well in recent decades. We have done quite a lot of it in this course. The former, however, which was first systematically developed by the preeminent early geomorphologist William Morris Davis (1850–1934) has had a checkered history.

The plan of this chapter is first to deal with classical Davisian geomorphology, along with some alternatives, and then to look at some of the prominent landscape elements that develop in areas of the Earth's surface—and they constitute much the greater part of the Earth's surface—where fluvial processes dominate. First, however, the following section presents some initial material on mountains and valleys.

11.2: Mountains and Valleys

Just to get you started thinking about the Earth's prominent landforms (there will be more detail later in the chapter), here are some comments about the nature of mountains and valleys.

As you will see soon, mountains are diverse not only in their scale and geometry but also in their processes of origin. I suppose that if you were to ask the proverbial "person in the street" about how mountains originate, he or she would tell you that they are somehow "built" to stand above their surroundings. There are indeed such ***"constructional" mountains*** (an only semi-official term), as the next few paragraphs describe.

Volcanic mountains, especially, come readily to mind as constructional mountains. Volcanoes emit lava (liquid, melted rock) and solid particles of a great variety of compositions and sizes. Much of this material is deposited in the vicinity of the fissure or vent, to build a hill or a mountain.

Volcanoes that emit basaltic lava, of relatively low viscosity, build mountains that are very broad, with gentle slopes. Such volcanoes are called ***shield volcanoes***. The great volcanoes of Hawaii and the Galápagos Islands are shield volcanoes. In a certain sense, Kilauea, on the big island of Hawaii, is the world's tallest mountain—if you measure the vertical distance from its base, in the deep ocean, to its summit!

Many volcanoes emit solid particles rather than liquid lava. Eruptions from such volcanoes tend to be explosive—often catastrophically so. The volcanic ash erupted from the volcano is thrown high into the atmosphere, and much of it settles to the ground over large areas, but some is deposited in the immediate vicinity of the volcano, to build a classic cone-shaped mountain. Such volcanoes are called ***cone volcanoes***. Mount Fuji, in Japan, is a classic cone volcano. Many such volcanoes consist of alternations of layers of lava and ash. The slopes of cone volcanoes are generally much steeper than those of shield volcanoes.

Mountains are also constructed by movements on faults. (A ***fault*** is *a widespread fracture in bedrock along which the material on the two sides of the fracture have moved relative to one another*.) Fault movements that have a vertical component to the movement create topography. Movements of major faults or fault zones can create great mountain ranges. Figure 11-1 shows a common example. The fault surface, usually approximately a plane, is inclined at a steep angle. The movement is such that the mass of rock overlying the fault surface moves downward relative to the mass of rock underlying the fault surface. The result is a mountain range with an adjacent valley. There are many such mountain ranges in the large area in the southwestern U.S. called the ***basin-and- range province***. Death Valley is a classic example: the main fault (called the "range-bounding fault") is on the east side of the valley; it is inclined to the west (or, in the parlance of geology, it *dips* to the west) at a moderate angle. The range of mountains east of Death Valley is going up, and the valley is going down, as a result of episodic movement on the fault. Of course, as the mountain goes up it is at the same time being worn down by erosion. The sedimentary products of the erosion are being deposited in the valley. The fill of the valley has attained a thickness of thousands of meters.

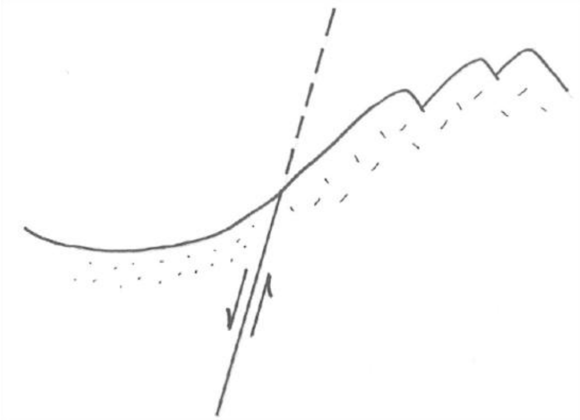


Figure 11-1. A mountain range and a valley formed by movement on a fault.

It may seem counterintuitive to you at this point, but you will learn in this chapter that most of the Earth's mountains are ***"erosional" mountains*** rather than constructional mountains. Such mountains come about by broad uplift, over a large area of the Earth's crust, and then erosion by streams and rivers, to leave elevated remnants of the originally uplifted area as mountains, surrounded by lowlands and valleys.

The central Appalachians of the eastern U.S., in Pennsylvania for example, are an excellent place to study erosional mountains. In much earlier geologic times the entire area was broadly uplifted by several hundred meters. Since then, erosion has excavated valleys, leaving higher ground, underlain by the rocks most resistant to erosion, as mountain ridges. The process has been most advanced in the eastern part of the area; westward, into northwestern Pennsylvania, the progress of erosion has been less advanced, and there is an extensive plateau, called the Allegheny Plateau, that is cut here and there by deep stream valleys.

You can see now that some valleys are produced by fault movements, and are usually occupied by streams and rivers, except in arid regions, whereas other valleys are the result of downcutting into broadly uplifted regions by streams and rivers.

11.3: Davisian Geomorphology

Early on, toward the end of the nineteenth century and into the early part of the twentieth century, an explanatory account of the development of landscapes was developed by application of a deductive approach: observe a large number of existing landscapes and then try to arrange them into a “movie” that represents the development of landscapes through time. Essential to this endeavor, of course, is a good understanding of the processes that shape landscapes in the first place. The problem is that just because such movies can be made does not mean that that’s how things work!

As the decades of the twentieth century passed, to a considerable extent this approach fell out of favor. In more recent times, however, with the advent of modern ideas about tectonism, brought about by the emergence of the plate- tectonics paradigm, along with development of better techniques of dating rocks and even surfaces, there has been a resurgence of interest in the historical aspect of landscapes.

In Davis’s time it was generally believed that orogeny (the building of mountainous terrain by tectonic activity) tended to occur as brief pulses, worldwide, and then, in the long time periods in between, there was tectonic quiescence while the land was gradually worn down. It was natural, then, for Davis to assume rapid broad uplift of a large region of the Earth’s continental crust and then slow reduction of elevations over very long periods of time. Davis concentrated on what is actually just a subset of the Earth’s land areas: those with substantial rainfall and well-developed river systems that mobilize and transport the products of weathering to the oceans. That leaves out vast arid and semiarid regions of the Earth’s surface, where differing geomorphic processes tend to leave rather different landscapes.

Davis perceived a cycle: rapid uplift, long-term quiescence while the landscape was worn down, then renewed rapid uplift. That cycle has been called the *humid geomorphic cycle*, or the *humid cycle of erosion*. It has been a central aspect in geomorphic thinking, but it has been controversial from early on, and its problems caused it to fall into disfavor by the mid-twentieth century. The elements of reality that it carries, however, has kept it alive in geomorphologists’ thinking, albeit with reservations. An account of it here seems therefore to be warranted.

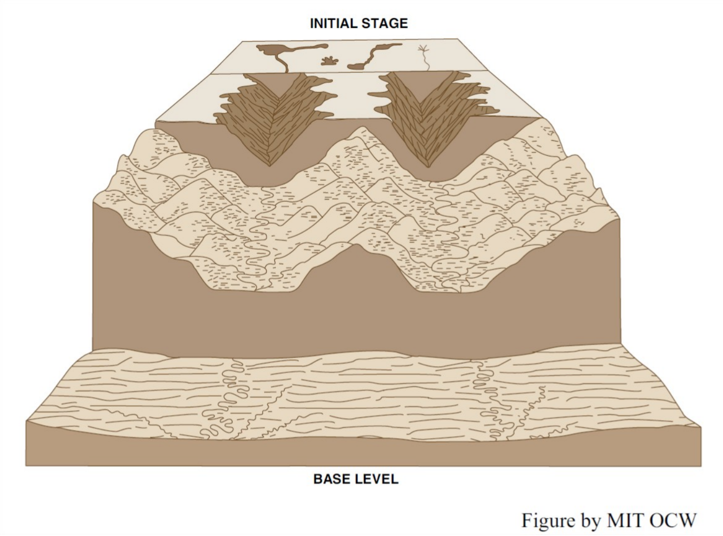


Figure 11-2, Part 1. Stages in the humid geomorphic cycle. A) Initial stage; B) youth; C) maturity; D) old age. (From Easterbrook, 1999).

A key element in Davis’s thinking was the ***peneplain***: a broad low-lying region, hundreds if not thousands of kilometers across, that is the ultimate stage in the cycle. A peneplain (“almost a plain”) is not nearly a planar surface: it slopes gently upward from the ocean shoreline, nowhere attaining elevations greater than some hundreds of meters, and it has subdued hill-and-valley topography. One big problem with concept of a peneplain is that, even though it makes perfectly good deductive sense, it’s difficult to identify any peneplains in today’s world! (Such a problem has a natural tendency to turn scientists off from an otherwise attractive concept.) You can view the Davisian cycle of erosion as uplift of a preexisting peneplain, progressive dissection of that peneplain, and eventual development of a new peneplain. But the big question is: does it ever really work that way?

Davis viewed his humid erosion cycle in terms of broad, qualitative stages, with no well-defined boundaries and with no specification of actual ages in years: **youth**, **maturity**, and **old age**, by loose analogy with the life histories of organisms.

After initial uplift (Figure 11-2A), in the stage of ***youth*** (Figure 11-2B), streams are rapidly incising into the newly uplifted surface. Most of the area consists of low-relief uplands, called ***interfluves***, between the active stream channels. Valley slopes are steep, and valley profiles are V-shaped. In ***maturity*** (Figure 11-2C), headward development of stream channels has mostly eliminated non-dissected interfluves. Stream valleys have widened, valley slopes have become gentler, and overall relief of the region is declining. In ***old age*** (Figure 11-2D), the region has been worn down to an even lower elevation, relief is much reduced, streams are of low gradient, and divides are broad and low. The region is close to being a peneplain.

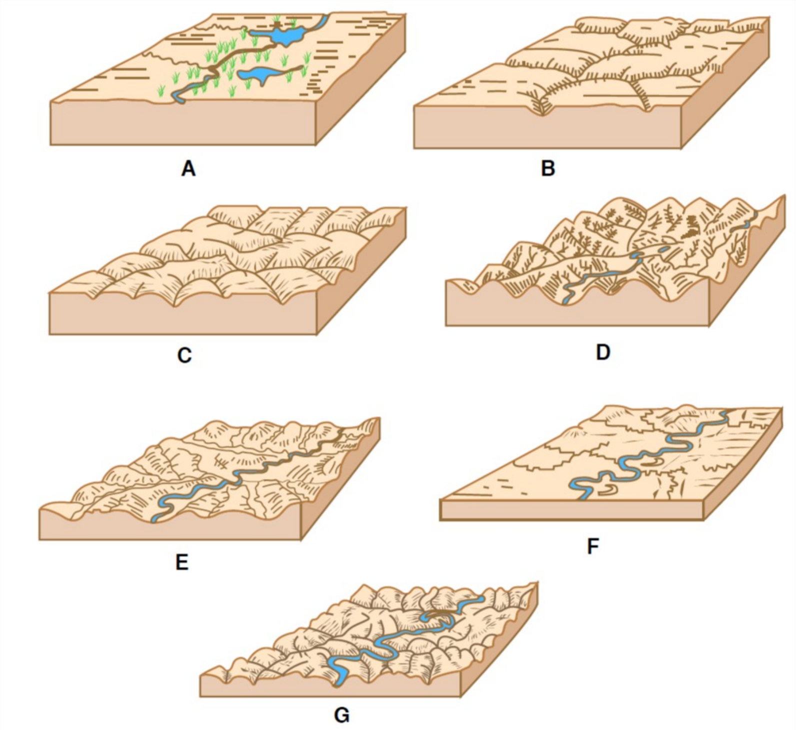


Figure 11-2, Part 2. Stages in the humid geomorphic cycle. (From Strahler, 1975.)

But does it ever really happen this way? You could make it happen, on a very small scale, on a large stream table or a backyard erosion plot. The trouble is that scale models of that kind, although often very revealing, are only qualitative: they are not dynamically similar (to use a term from fluid dynamics), in the sense that all forces, motions, and processes are in the same proportion in the model as in the natural case. Scientific common sense, tells us, however, that something akin to the humid geomorphic cycle must be applicable to large continental regions, provided that there is rapid uplift and then long-term tectonic quiescence.

It’s clear these days that tectonism (movements of the Earth’s crust) is not nearly as well-behavedly episodic as the early geomorphologists believed. It’s clear that in regions of plate convergence (ocean–continent subduction, as in the Andes, and continent–continent collision, as in the Himalayas–Tibetan Plateau region of southern Asia), strong regional uplift can last for tens of millions of years, and, all the while, streams are denuding the landscape—a far cry from the Davisian concept of rapid uplift followed by long quiescence. Even in such regions, however, now-uplifted remnants of the old, preexisting low-relief and low-elevation surface can persist for geologically long times. In fairness to Davis, I might point out here that he himself emphasized that interruptions to his cycle of erosion must be common and important.

A key concept in the humid erosion cycle is ***accordance of summits***. Someday, when you are traveling southwest from New England on Interstate Route 84 through southern New York State toward northeastern Pennsylvania, take a brief detour south into northwestern New Jersey. (There’s a convenient exit labeled “Mountain Road”, the last before the road reaches Port Jervis, where the three states of New York, New Jersey, and Pennsylvania come together.) Along a ridge of the Appalachians in northwestern New Jersey, at the highest point in the state, is a Washington Monument–like structure upon an almost treeless and spectacularly three-hundred-and-sixty-degree observation site. There you can let your eye sweep around the compass at the surrounding landscape, near and far. The striking thing about that vista is that to an excellent approximation one can fit to all of the distant mountain summits on the horizon a single plane, one that rises gently from east to west. The summits are in the form of ridges consisting of resistant rock units, all trending approximately parallel in a northeast–southeast orientation governed by the underlying bedrock structure of the Appalachian orogen. The intervening valleys are underlain by weaker rock, where streams have eroded the landscape to much lower levels. The inference is attractive to the point of being overwhelming that such topography is the result of uplift of a low- lying plain and then fluvial incision to exploit zones of weak rock while leaving the more resistant rock almost unaffected as divides between the various streams (Figure 11-3). Does that mean that the land was uplifted in a single episode, and then remained high while erosion proceeded, in the Davisian way? Not necessarily: perhaps uplift has been ongoing, concurrently with denudation. In either case, it’s difficult to avoid the deduction that there once was a low-lying plain that got uplifted and eroded!

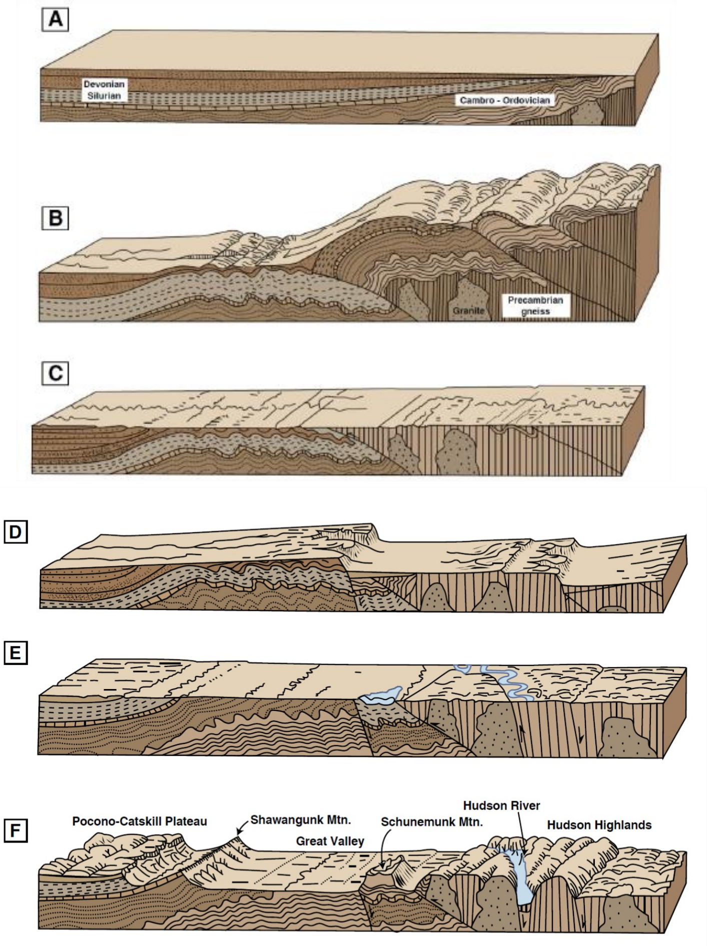


Figure by MIT OCW

Figure 11-3 (previous page). Uplift and dissection of a preexisting peneplain, to form the present topography of the central Appalachians. (From Strahler, 1975.)

11.4: How Fast Are the Continents Worn Down

Geomorphologists use the term ***denudation*** for *the overall, regional lowering of a continental land surface by processes of weathering, erosion, and transportation of bedrock material to the oceans*. Rivers, of course, are by far the most important agents of denudation, although ice sheets have been very effective at certain times and places in geologic history.

For a long time, geoscientists have been trying to develop estimates of rates of denudation. There are several ways of doing this—none of them perfect.

Measuring the sediment load of rivers. Because almost all of the products of denudation are carried by rivers, it should seem natural to you to measure the sediment loads of all of the word's rivers to get a yearly total of the sediment delivered to the oceans, and then imagine that volume to be spread over the area of the continents to get an annualized denudation rate. Many such estimates have been made. There are some serious problems, however. One is that it's notoriously difficult to measure the sediment load of rivers, especially bed load. Moreover, systematic measurements of the total load (suspended load, bed load, and dissolved load) are common on relatively few rivers around the world. Abroaderproblemis:*howrepresentativeislastyear'ssedimentload,say,of denudation on time scales of millions of years?* It's generally agreed that humankind has greatly accelerated denudation, owing largely to agriculture. And then there's the matter of natural fluctuations in climate, on scales ranging from centuries to millions of years. Despite all of that pessimistic commentary, however, here are some figures. Worldwide estimates are in the range of several centimeters per millennium, but values range widely, from just a few centimeters per millennium, for tectonically stable regions without high relief, to far higher values, even up to a few meters per millennium, in tectonically active areas with high relief and abundant rainfall (a widely cited example is the eastern mountains of Taiwan).

Deposition rates in the ocean. The obvious alternative to measuring the sediment loads of rivers is to estimate the volume of sediment that has been deposited in the oceans over some period of geologic time. Most of the solids load of rivers is deposited along the continental margins (but dissolved load is not easily taken into account in this way). Geophysical methods of measuring sediment thickness, together with coring to define the stratigraphy, lead to regional estimates of denudation. The advantage of this method is that it should provide estimates of denudation over geologically long times and also show changes in rates through time. Estimates for the eastern seaboard of North America are a few centimeters per millennium, in the same ballpark as estimates made on the basis of sediment loads of rivers.

Rates of erosion of land areas. If the thickness of continental rocks removed in some interval of geologic time can be measured or estimated, that gives the most direct indication of denudation rates. In some places it has been possible to make a direct measurement—for example, where a basalt flow of known age covered a preexisting land surface and then has been removed along with some thickness of the underlying material. In many regions, study of metamorphic facies and thermal history has led to estimates of the thickness of rock unroofed owing to long-continued uplift, especially in orogenic regions. Again, rates vary widely but are generally in the range of a few centimeters per millennium to a few tens of centimeters per millennium.

A striking conclusion from data of the kind noted above is that, given denudation rates even of just a few centimeters per millennium, which seems to be a conservative estimate, a continent like North America would be worn down to low-lying terrain in geologically short times, of the order of ten million years!

Just divide the average elevation of North America, of the order of five hundred meters ($5 \times 10^5\ cm$), by a denudation rate of, say, five centimeters per millennium ($5 \times 10^{-3}\ cm/yr$). The result: ten million years. Why, then, are all of the continents not just low-lying plains? (One factor is that rates slow way down as elevation and relief decrease.) We are forced to appeal to *long-continued uplift*: orogenic, in the case of uplift of mountain ranges, which is relatively rapid and relative restricted in area, or epeirogenic, which is relatively slow but covering relatively large areas. (The adjective *orogenic* is used to describe building of mountains in a local region; the adjective *epeirogenic*, less easy to grasp, is used to describe broad vertical movements that affect large parts of a continent.)

That leads to the issue of the competition between uplift rates and denudation rates. You might think in terms of the two end-member cases:

- **Uplift rates are much greater than denudation rates**. This would be the case in an tectonically active but arid region with rapid uplift of the crust: the land is coming up fast, but rainfall to weather the rock and remove the weathering products is scarce. The maximum elevation in such an area would then be controlled by rock strength. There seems to be a self-limiting process at work here: when the uplifted mass becomes sufficiently high, it literally flows out sideways, at depth, on a regional scale of many hundreds of kilometers, by a complex of tectonic processes we need not deal with here, thus limiting the maximum elevation. That seems to be happening in the Altiplano, in the Andes, and in the Tibetan plateau, in southern Asia.

- **Denudation rates are as great as uplift rates.** Owing to high rainfall in a humid climate conducive to rapid rock weathering, denudation rates can keep up with even very high rates of uplift, limiting mountain elevations to just a few thousand meters despite the very rapid uplift. One of the classic areas of that kind is the eastern mountains of Taiwan.

11.5: Drainage Development in Newly Emergent Regions

Now for some more concrete and closer-to-home stuff for you. First, here's some background on the recent geologic history of the Eastern Seaboard of North America (see Figure 9-4). Like other passive continental margins (those facing an expanding ocean with a mid-ocean spreading ridge in the middle, and no strong tectonism caused by plate interactions along the continental margin), the east coast of North America underwent long-continued slow subsidence, caused by thermal contraction of the cooling lithosphere as it moved away from the ridge crest. A seaward-thickening wedge of siliciclastic sediment was deposited along the continental margin, dating back to not long after the breakup of Pangea, about two hundred million years ago. It's also known, however, that in more recent times, in the past few tens of millions of years, there has been general, broad uplift of the continental margin of North America, for unclear reasons. (That's known from study of the stratigraphy of the offshore sediment wedge.) The present coastal lands of the Eastern Seaboard have a landward-thinning veneer of almost flat-lying marine deposits, called the ***coastal plain***. The coastal-plain wedge laps up onto much older rock with generally complex structure generated during the Appalachian Orogeny, which led up to the assembly of Pangea. The Eastern Seaboard has thus been undergoing erosion for many millions of years, and in the process the thickness of coastal-plain sediments has been reduced and their western edge has retreated eastward, exposing more and more of the older rocks.

Got the picture? Now think about the development of river drainage during the uplift of the past few tens of millions of years. As the continental-shelf sediments are gradually exposed to form a widening coastal plain, a new pattern of drainage becomes established on the gently sloping sedimentary cover. It's natural to assume that the streams and rivers flow down the slope, generally toward the southeast in most areas. Such streams are called ***consequent streams***—because they are consequent upon emergence of the region above sea level. These streams, and their tributaries, gradually denude the region as they extend headward and lower their profiles. If the sedimentary cover of the coastal plain is uniform, the drainage pattern is likely to be dendritic (see the chapter on rivers).

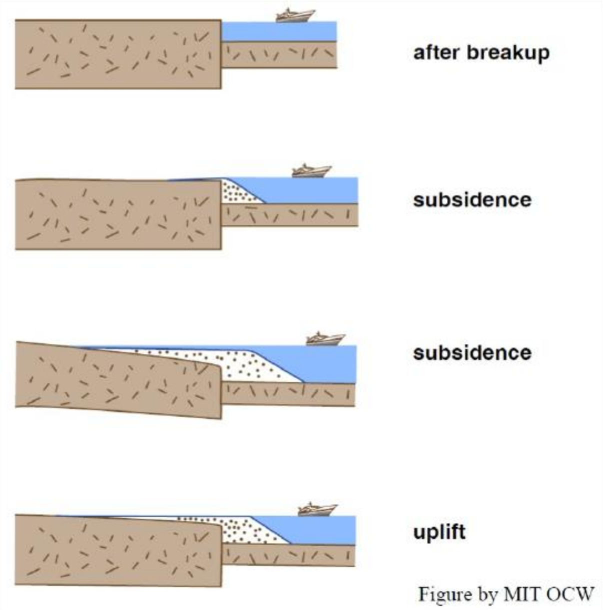


Figure 11-4. Recent geologic history of the Eastern Seaboard of North America.

Eventually, the rivers and streams lower their courses to the point where they encounter the "basement" rock underlying the coastal-plain sediments. At that point, certain tributaries to the larger consequent streams encounter belts of weaker rock, and their courses become adjusted to follow those belts of weaker rock. Such streams are called ***subsequent streams***. The orientations of the subsequent streams can vary widely, but in a region like the Eastern Seaboard of the US, where the structural trend of the older rocks underlying the coastal-plain strata is generally parallel to the shoreline, the subsequent streams tend to be perpendicular to the consequent streams (Figure 11-5).

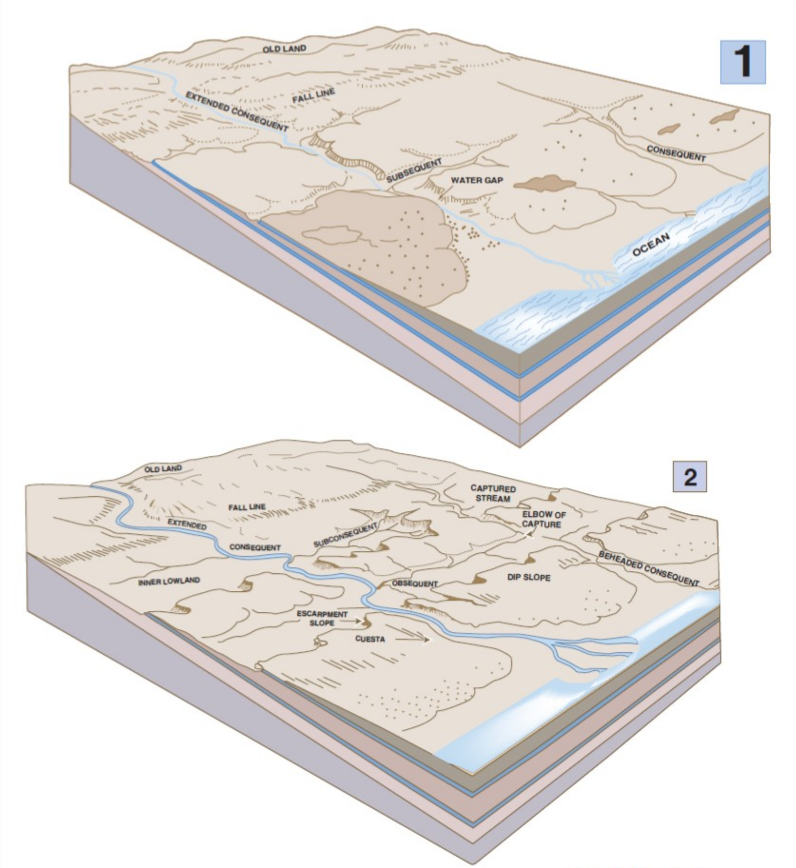


Figure by MIT OCW

Figure 11-5. A) Early stage in the development of drainage on an uplifted coastal plain. B) Later stage of development. (From von Engeln, 1942.)

In some places, certain of the consequent streams are larger than others, just because of the accidents of the original topography of the coastal-plain surface. These larger streams are likely to lower their profiles faster than adjacent consequent streams, especially if they are flowing on a weaker substrate. Then the subsequent tributaries to such consequent streams are able to extend themselves headward at lower elevations than the adjacent consequent streams. The lower subsequent stream then occupies a greater and greater drainage area at the expense of the drainage areas of adjacent subsequent streams, and eventually the divide might be eliminated altogether. The result is a diversion of the adjacent consequent stream into the drainage system of the larger consequent stream. This process is called ***stream piracy***, or ***stream capture*** (Figure 11-6). The captured consequent stream is said to be a ***beheaded stream***. Its valley, downstream of the point of capture, is then disproportionately large for the new, lesser discharge carried by the stream. The stream in such a valley is called an ***underfit stream***. One tip-off that capture has happened is an abrupt change of course, often at right angles (Figure 11-6B).

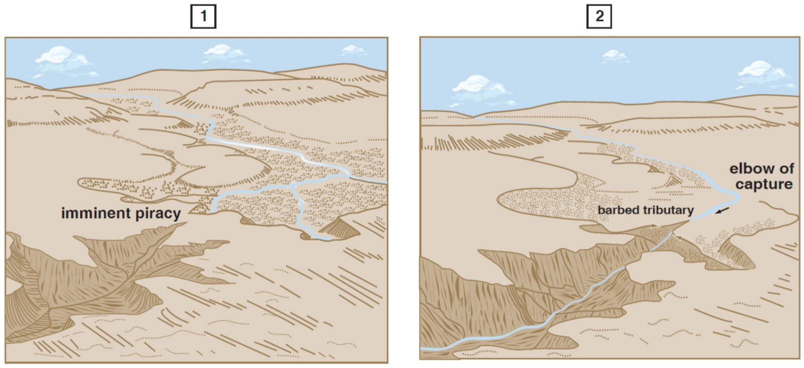


Figure by MIT OCW

Figure 11-6. Development of stream capture. A) Just before capture. B) Well after capture. (From von Engeln, 1942)

Consequent streams tend to encounter belts of rock that are strongly resistant to erosion. Ordinarily, however, they have sufficient erosive power to cut through such a belt of erosion-resistant rock as they lower their profiles, although the site may be a zone of rapids. As the adjacent belts of weaker rock are lowered by subsequent streams that are tributary to the consequent stream, a topographic ridge develops at the site of the belt of stronger rock, and a notch in the ridge, called a ***water gap***, develops (Figure 11-7). The Delaware water gap, on the Delaware River between northwestern New Jersey and northeastern Pennsylvania, and the Hudson Highlands water gap, along the lower course of the Hudson near West Point, where the Hudson cuts through the Precambrian highlands on its way south, are good examples of water gaps in the eastern US. Another good example is the Potomac River just downstream of its confluence with the Shenandoah River near Harper’s Ferry. The rapids there figured prominently in George Washington’s dream of making the Potomac River the principal gateway to the rapidly developing west after the War for American Independence.

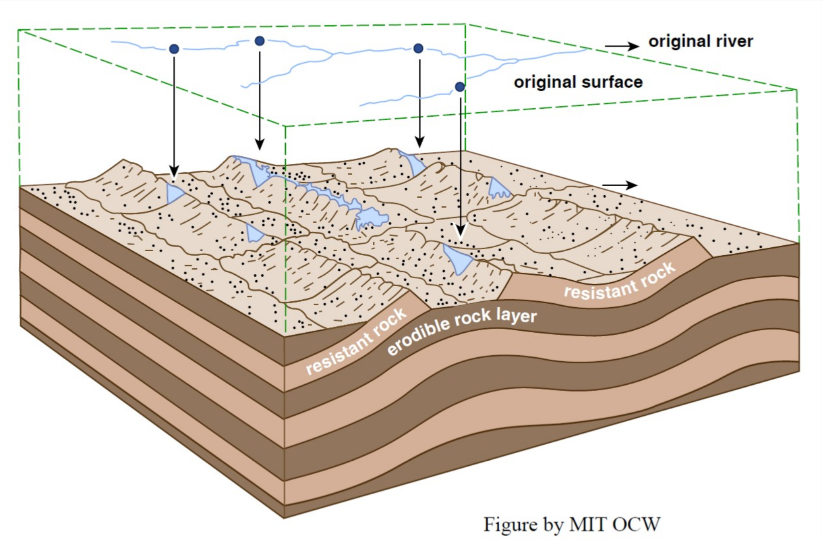


Figure 11-7. Development of water gaps and wind gaps. (From Bloom, 1998.)

Sometimes, when an aggressive subsequent stream captures an adjacent less aggressive consequent stream, a water gap through which the downstream reach of the captured stream flowed becomes abandoned. As the belts of weaker rock on either side of the ridge become lowered by further denudation, the former water gap appears as a ***wind gap***: a notch in the ridge high above the level of the adjacent valleys (Figure 11-7). You can see several wind gaps in the southeasternmost ridge of the valley and ridge province on Pennsylvania, as you drive southwest on Interstate Route 78 from the New Jersey border toward Harrisburg. (Also, if you drive across the Susquehanna River on Pennsylvania Route 581 just south of Harrisburg, look north up the Susquehanna River to see a classic water gap.)

11.6: Some Common Geomorphic Features Produced by Fluvial Erosion

Much of the landforms and landscapes we observe and admire are the result of fluvial erosion of regions underlain by sedimentary rock units, whether still flat-lying or now deformed. In this section we look at some of the common landscape features in such regions.

mesas and buttes. Many plateaus in arid or semiarid regions are surfaces capped by a particularly resistant horizontal sedimentary layer or sedimentary unit, underlain and overlain by softer, more easily erodible units. If the overlying weaker material is stripped away down to the level of the resistant unit, a plateau results. When fluvial erosion breaches the plateau and removes some of the weak rock underlying the resistant unit, broad areas of the plateau remain in partial or complete isolation from one another. Such a broad area is called a ***mesa*** (Figure 9-8). As the area of the mesa shrinks by wasting of the edges of the resistant layer and removal of more of the soft underlying material, the mesa becomes a ***butte*** (Figure 9-8).

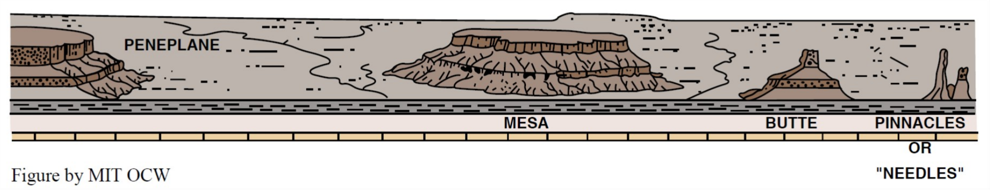


Figure by MIT OCW

Figure 11-8. Buttes and mesas. (From Lobeck, 1939)

cuestas and hogbacks. When a layer or unit of erosion-resistant rock, underlain and overlain by weaker rock, is gently dipping, fluvial erosion results in the resistant layer forming an asymmetrical ridge, called a ***cuesta***, with a gentle slope parallel to the layer and a steep slope perpendicular to the layer (Figure 11-9). The gentle slope is called a ***dip slope***. If the dip of the resistant unit is greater, the ridge is more nearly symmetrical; in that case, the ridge is called a ***hogback*** (Figure 11-9). When the dip of the strata is even steeper, the dip slopes sometimes

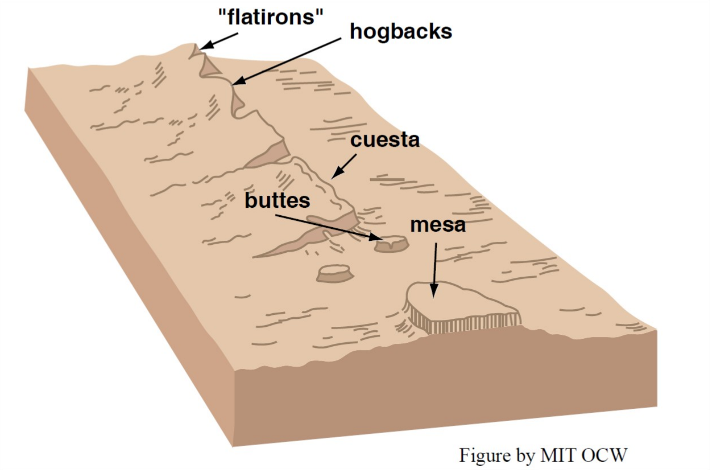


Figure by MIT OCW

Figure 11-9. Cuestas, hogbacks, and flatirons. (From Bloom, 1998)

READINGS

Three of the most modern comprehensive treatments of geomorphology:

Allen, P.A., 1997, Earth Surface Processes. Blackwell Science, 404 p.

Bloom, A.L., 1998, Geomorphology; A Systematic Analysis of Late Cenozoic Landforms, Third Edition. Prentice Hall, 482 p. (especially Chapter 6)

Easterbrook, D.J., 1999, Surface Processes and Landforms, Second Edition. Prentice Hall, 546 p. (especially Chapter 15)

Two old books, with some outdated concepts but with superb illustrations (both photos and line drawings) of geomorphic features:

Lobeck, A.K., 1939, Geomorphology. McGraw-Hill, 731 p.

von Engeln, O.D., 1942, Geomorphology; Systematic and Regional. Macmillan, 655 p.

A classic textbook that covers a variety of geomorphic features and processes. Later editions are dumbed down.

Strahler, A.N., 1975, Physical Geography, Fourth Edition. Wiley, 643 p. + index and plates.

The classic book on fluvial geomorphology:

Leopold, L.B., Wolman, M.G., and Miller, J.P., 1964, Fluvial Processes in Geomorphology. Freeman, 522 p.

CHAPTER OVERVIEW

12: The Hydrologic Cycle, the Sediment Cycle, and the Carbon Cycle

---

12.1: Introduction - Systems, Cycles, Reservoirs, and Fluxes

Thinking about the Earth as a system has become, in recent years, much in vogue. The basic idea of a ***system*** is that it is *something that consists of a number of distinctive and diverse parts that function together by a variety of interactions and exchanges of energy and matter*. Your motor vehicle is a good example of a fairly small and simple system: it is a machine with a number of moving or nonmoving parts, which consumes motor fuel and emits various exhaust substances as it travels from place to place. The Earth is, of course, a far larger and also far more complicated system, whose component parts operate on time scales from seconds to many millions of years. This is the basis for what is nowadays called ***Earth system science***. Here I'm only touching upon some important aspects of Earth systems science that are relevant to this course: the hydrologic cycle and the carbon cycle. A whole course could be devoted to those and certain other cycles that are important in the context of the Earth's surface environment.

The concept of cycles plays a fundamental role in systems thinking— although it's not easy to frame a suitable definition of such cycles. A ***cycle*** might best be described as a characteristic succession of events and processes , involving certain kinds of Earth materials, by which the materials reside in certain kinds of places and move among such places in certain ways. Cycles operate through an indefinitely long span of time.

Much of the material that has been covered so far in this course involve, in one way or another, water and sediment. Earlier chapters have dealt in much detail with the Earth's water and sediment: where they are, and how the move. Now it's time for us to be more integrative in our thinking about the Earth's water and sediment. To that end, the next two sections deal with what are called the *hydrologic cycle* and the *sediment cycle*.

Geochemists, in particular, are fond of thinking about certain substances from the standpoint of cycles. Such cycles are called ***geochemical cycles***. The most important, in the context of the Earth's surface environment, is the *carbon cycle*. The carbon cycle is especially important because carbon is the fundamental basis for life. It is also of great importance for the Earth's climate, because carbon dioxide (CO2) is the second important of the so-called greenhouse gases, which play a key role in climate. (Water vapor is the most important, but we humans have virtually no control over the content of water vapor in the atmosphere, whereas we have great influence on the content of carbon dioxide in the atmosphere, owing to the burning of fossil fuels.) The concluding section of this chapter is devoted to some aspects of the Earth's carbon cycle that are most relevant to the Earth's surface environment. Geochemical cycling of certain other elements, most notably nitrogen and phosphorus, have received study because of their role as nutrients.

What pops into your mind when you hear the word *reservoir*? Probably a body of water, small or large, that is impounded behind a dam. In Earth systems science, the term ***reservoir*** is used for *a distinctive kind of place where a certain kind of material is stored, or resides, for some period of time*. That definition probably doesn't enlighten you much, so here are some examples of near-surface reservoirs you have encountered earlier in the course: the atmosphere; glaciers; the soil layer; the aggregate of bodies of fresh water on the continents (rivers and lakes).

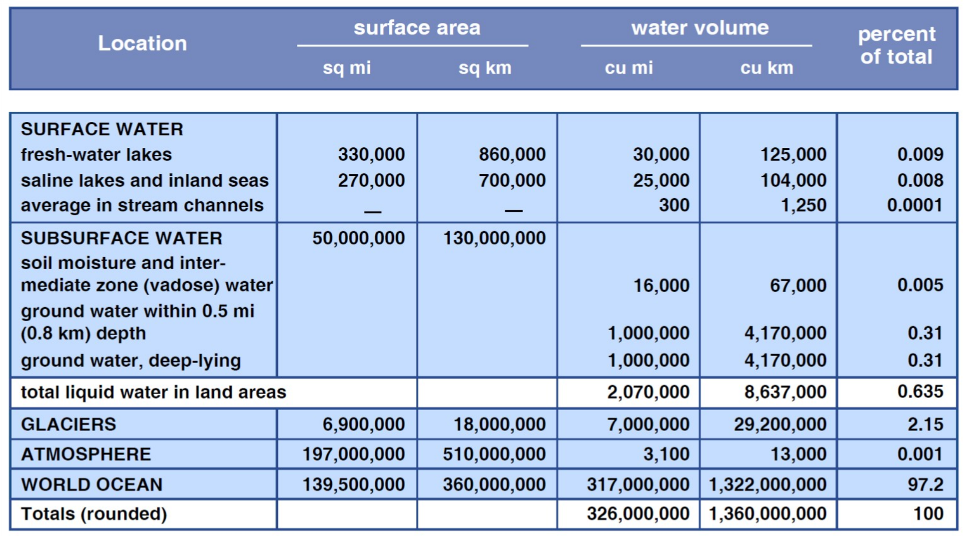
Material moves into and out of reservoirs. *The rate at which a given material moves between reservoirs* is called a ***flux***. If the flux of material into and out of a given reservoir is the same for some period of time, that reservoir is said to be in a ***steady state***. Commonly, however, the flux in and the flux out are not equal.

Another concept that is useful to know about when thinking about systems is ***residence time***. In an earlier chapter you learned about the residence time of water in lakes. The same concept can be applied to any substance that resides in some reservoir for some period of time.

12.2: The Earth’s Water, and the Hydrologic Cycle

The earth would be a boring place without water. Fortunately, there’s a lot of water on Earth. Table 12-1 gives an approximate accounting of where the earth’s water resides. By far the most water is in the oceans. Of the Earth’s fresh water, most is tied up in glacier ice, with groundwater a distant second; the fresh water we are most familiar with—streams, rivers, lakes, and what’s in the atmosphere in the form of water vapor, clouds, and precipitation—is a tiny part of the Earth’s stock of near-surface water.

The trouble with an accounting technique like that of Table 12-1 is that it doesn’t tell anything about *the movements and transformation of the earth’s water*. You all know that water on, below, and above the earth’s surface is in a continuous state of movements and transformations. The picture of these movements and transformations, viewed in their entirety, is called the ***water cycle***, or the ***hydrologic cycle***. Figure 12-1 gives one version of the hydrologic cycle. Most of the pathways and processes shown in Figure 12-1 should be easily understandable to you. They vary a lot in their magnitude (how much water, per unit time on average, is involved in each of them), but that’s not shown in the figure. For the purposes of this course, what’s most important is evaporation, condensation, and precipitation.



Location	surface area		water volume		percent of total
	sq mi	sq km	cu mi	cu km	
SURFACE WATER					
fresh-water lakes	330,000	860,000	30,000	125,000	0.009
saline lakes and inland seas	270,000	700,000	25,000	104,000	0.008
average in stream channels	—	—	300	1,250	0.0001
SUBSURFACE WATER	50,000,000	130,000,000			
soil moisture and intermediate zone (vadose) water			16,000	67,000	0.005
ground water within 0.5 mi (0.8 km) depth			1,000,000	4,170,000	0.31
ground water, deep-lying			1,000,000	4,170,000	0.31
total liquid water in land areas			2,070,000	8,637,000	0.635
GLACIERS	6,900,000	18,000,000	7,000,000	29,200,000	2.15
ATMOSPHERE	197,000,000	510,000,000	3,100	13,000	0.001
WORLD OCEAN	139,500,000	360,000,000	317,000,000	1,322,000,000	97.2
Totals (rounded)			326,000,000	1,360,000,000	100

Figure by MIT OCW

Table 12-1. An accounting of where the Earth’s water resides. (From Strahler, 1975.)

Here’s one interesting fact that underlines the “activeness” of the hydrologic cycle: *the mass of water vapor plus condensed water in clouds plus precipitation that’s in the process of falling to the ground, at any given time, amounts to only about one week’s worth of global precipitation*.Thismeansthat there must be very rapid recycling of water between the liquid reservoirs at the earth’s surface, on the one hand, and the stock of water in the atmosphere, on the other hand.

Two incidental notes:

- The major player in evaporation is the ocean surface: the ocean surface accounts for about 85% of total average evaporation, and the continents account for only about 15%.
- Evaporation of the ocean exceeds precipitation over the ocean, and precipitation over the continents exceeds evaporation. The balance is maintained mainly by *runoff of water from the continents*, in the form of rivers and streams, although in certain places large glaciers deliver water to the oceans, as they calve directly into the ocean.

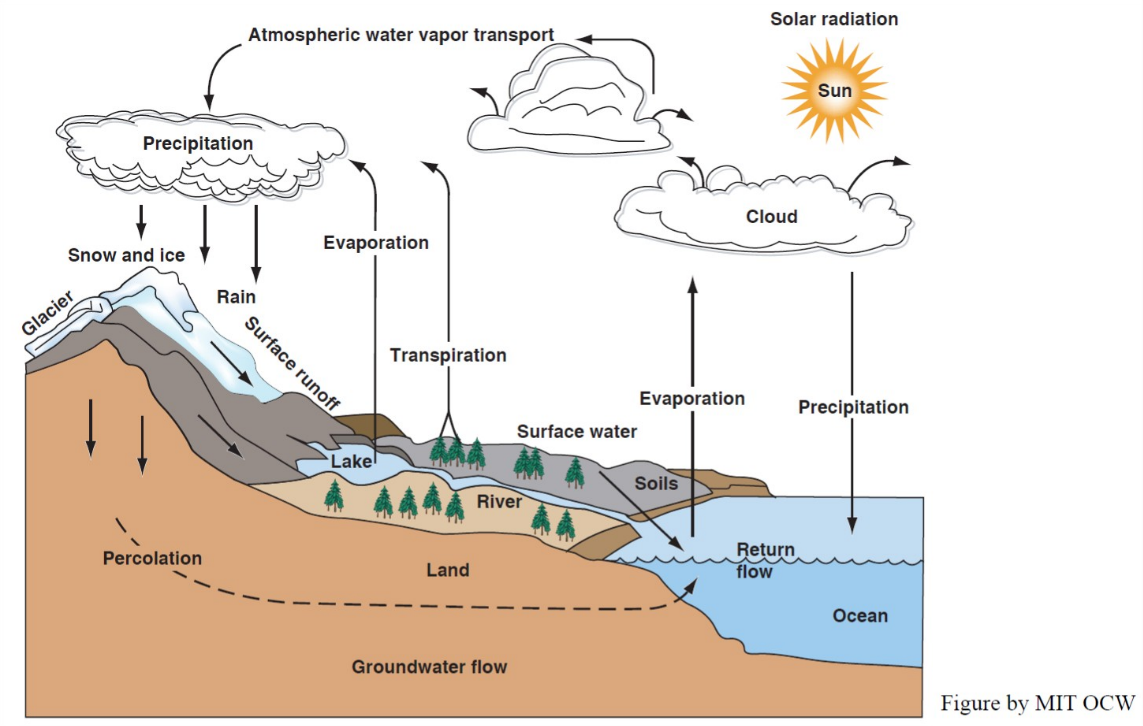


Figure 12-1. The hydrologic cycle. (From Water, The Yearbook of Agriculture 1955.)

12.3: The Sediment Cycle

You learned earlier that sediment is material, in either particulate or dissolved form, that is produced by weathering of rocks on the continents and then transported, by some agent (water is by far the most important, in general, but wind and moving glacier ice are important as well, at certain times and places) to come to rest as a sediment deposit. A whole field of Earth science, *sedimentary geology*, is devoted to the study of sediments, as well as the sedimentary rocks that are formed when sediment is buried and becomes lithified.

In a sense, the sediment cycle is simpler than the water cycle, because after sediment is formed it inevitably moves downhill toward places of rest. From the perspective of the Earth's surface, sedimentary processes are basically a matter of source, transport, and sink. (Scientists like to use the term ***sink*** for a kind of place to which matter moves and accumulates, or a kind of place to which energy flows and is there stored or dissipated.)

To have a full appreciation of the sediment cycle, you need to think beyond the Earth's surface and the processes and kinds of environments of reservoirs associated with it. Although sediment can be stored in lakes and in river valleys for geologically long periods of time, and become buried deeply enough to be converted to sedimentary rocks, most sediments end up in the world's oceans. The oceans don't fill up, though: plate tectonics operates, in ways that are beyond the scope of this course, to recycle the sediments and sedimentary rocks back to the continents, there to be exposed once again to weathering and transport, to complete the cycle.

12.4: The Carbon Cycle

Introduction

The element carbon holds a distinctive place among the chemical elements: it can form covalent bonds with itself (these are called carbon–carbon double bonds and carbon–carbon triple bonds, depending upon how many of the bonding positions are involved) as well as with other elements, especially hydrogen, to form an enormous number of chemical compounds, called ***organic compounds***.

The term "organic" here is a bit misleading: great numbers of such compounds are synthesized by organisms, but even greater numbers can be synthesized in the laboratory, from inorganic raw materials, without the involvement of organisms, living or dead. The field of chemistry devoted to the study of organic compounds is called ***organic chemistry***.

Carbon also forms various ***inorganic compounds***. In such compounds, carbon bonds covalently with oxygen. The two most important kinds of such carbon-containing inorganic compounds are carbon dioxide (CO_2), invariably a gas under Earth-surface conditions, and carbonate minerals, mainly calcite, $CaCO_3$, and dolomite, $CaMg(CO_3)_2$.

Carbon, as both organic compounds and inorganic compounds, is present at and near the Earth's surface in a great variety of environments. The study of where carbon resides in and on the Earth, together with the modes and rates at which carbon moves from one to another of such environments, is known as the ***carbon cycle***.

The concept of a "cycle" here carries the implication that there is an unceasing movement of carbon, in all of its various forms, in and on the Earth. In this respect the Earth can be viewed as a closed system with respect to carbon, in the sense that although the carbon undergoes a great variety of transformations the total amount does not change with time.

A full consideration of the Earth's carbon cycle would necessitate casting a wide net: we would need to consider not only the Earth's surface environment, including the atmosphere, the land surface, and the oceans, but also the Earth's deep interior—because, by various processes, carbon becomes buried within the Earth, and, by various other processes, it is released from deep storage back into the surface environment. The carbon cycle is truly integrative, in the sense that its study draws upon a wide range of scientific disciplines: geology, biology, and climatology. Here we'll concentrate on the Earth-surface part of the carbon cycle.

Elements of the Carbon Cycle

There are many reservoirs of carbon at and near the Earth's surface. Table 12-2 shows the important carbon reservoirs, along with rough estimates of the mass of carbon in those reservoirs. (In ecology, the term ***biomass*** is used for *the total mass of living material, in plants and animals*. It can be expressed as mass per unit area in a particular region of the Earth's surface, or as the total over the entire Earth.) You can see that the mass of carbon in the atmosphere, the oceans, and biomass is tiny in comparison to the carbon that resides in sedimentary rocks in the earth's crust. Note also from Table 12-2 that the mass of carbon in the atmosphere is approximately equal to total biomass carbon.

In Figure 12-2 you can see the distinction between organic carbon and inorganic carbon. It's useful to consider these two forms of carbon separately. It's easier to deal with the latter than with the former.

As noted earlier, the two important forms of inorganic carbon are carbon dioxide and carbonate minerals and rocks. Carbon dioxide is present in the Earth's atmosphere in an average concentration of close to 0.4 percent by volume.

Carbon is *put into* the atmosphere, in the form of carbon dioxide, mainly from respiration by organisms and by decomposition of organic matter, and also by burning of plant matter. Before the discovery of fire by human beings, burning was ignited by lightning strikes. Nowadays, of course, enormous quantities of carbon are put into the atmosphere by burning of fossil fuels. Carbon is also put into the atmosphere in the form of carbon dioxide released during volcanic eruptions.

Carbon is *removed from* the atmosphere mainly by photosynthesis by plants and by becoming dissolved in the oceans. Some of the added carbon dioxide from burning of fossil fuels results in increase in the concentration of carbon dioxide in the atmosphere, but that increase is in part offset by dissolution into the oceans and in part by increase in total plant biomass.

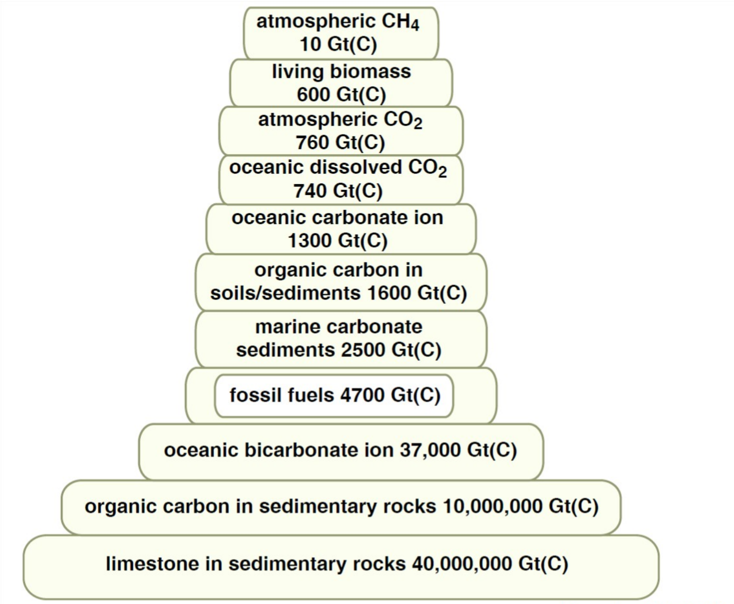


Figure by MIT OCW

Figure 12-2. The Earth's carbon reservoirs. (From Kump, Kasting, and Crane, 2004.)

Incidentally, not all of the carbon in the atmosphere is in the form of carbon dioxide. Methane, CH4, is present in the atmosphere in concentrations of about 1.7 x 10-4 percent by volume. Methane is added to the atmosphere in two ways, mainly: upward seepage from shallow and deep subsurface, where methane is generated both from by the activity of microorganisms at shallow depths, and by complex reactions involving deeply buried organic matter at greater depths. Methane is unstable chemically in the atmosphere: it is converted to carbon dioxide and water vapor.

Inorganic carbon in the Earth' crust is largely in the form of carbonate minerals calcite, aragonite, and dolomite. Enormous volumes of carbonate rocks, mainly limestone and dolostone, reside in the crust. As you learned in an earlier chapter, carbon in the form of carbon dioxide is released as carbonate rocks are weathered at the Earth's surface. At the same time, carbonate minerals are precipitated, mainly in the shallow ocean, in places where the ocean water is oversaturated with respect to calcium carbonate. Most of that precipitation is in the tissues of a great variety of carbonate-secreting organisms, although some is precipitated inorganically.

The *residence time* of carbon in the atmosphere is about thirteen years. To figure that out, you need to know, in addition to the 760 Gton of carbon that's present in the atmosphere, that the inflow of carbon into the atmosphere, and the outflow of carbon from the atmosphere, which from year to year are not far from being the same, is about 60 Gton carbon.

Index

Detailed Licensing

Overview

Title: The Environment of the Earth's Surface (Southard)

Webpages: 108

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 80.6% (87 pages)
- Undeclared: 14.8% (16 pages)
- CC BY-NC-SA 1.3: 4.6% (5 pages)

By Page

- The Environment of the Earth's Surface (Southard) - *CC BY-NC-SA 4.0*
    - Front Matter - *CC BY-NC-SA 4.0*
        - TitlePage - *CC BY-NC-SA 4.0*
        - InfoPage - *CC BY-NC-SA 4.0*
        - Table of Contents - *Undeclared*
        - Introduction - *CC BY-NC-SA 4.0*
        - Licensing - *Undeclared*
    - 1: Introduction - *Undeclared*
        - 1.1: Some Basic Physics Relevant to The Earth's Surface - *Undeclared*
        - 1.2: Heading and Cooling of the Earth's Surface - *Undeclared*
        - 1.3: Some Elementary Chemistry Relevant to the Earth's Surface - *Undeclared*
        - 1.4: The Nature of Water - *Undeclared*
        - 1.5: Some Essential of the Chemistry of the Surface Zone of the Earth - *Undeclared*
        - 1.6: The Flow of Water - *Undeclared*
    - 2: Introduction and Geology - *CC BY-NC-SA 4.0*
        - 2.1: The Earth - The Ground Beneath Your Feet - *CC BY-NC-SA 4.0*
        - 2.2: Minerals - *CC BY-NC-SA 4.0*
        - 2.3: Rocks - *CC BY-NC-SA 4.0*
        - 2.4: Bedrock - *CC BY-NC-SA 4.0*
        - 2.5: Weathering - *CC BY-NC-SA 4.0*
        - 2.6: The Mineral Particles of the Earth's Surface Materials - *CC BY-NC-SA 4.0*
        - 2.7: Regolith - *CC BY-NC-SA 4.0*
        - 2.8: Sediment - *CC BY-NC-SA 4.0*
        - 2.9: Soils - *CC BY-NC-SA 4.0*
    - 3: Topography - *Undeclared*
        - 3.1: Latitude and Longitude - *CC BY-NC-SA 4.0*
        - 3.2: Maps - *CC BY-NC-SA 4.0*
        - 3.3: Topographic Maps - *CC BY-NC-SA 4.0*
        - 3.4: Stream Networks, Drainage Basins, and Divides - *CC BY-NC-SA 4.0*
        - 3.5: Geologic Maps and Cross Sections - *CC BY-NC-SA 4.0*
    - 4: Groundwater - *CC BY-NC-SA 4.0*
        - 4.1: Introduction- How Water Gets to Be Groundwater - *CC BY-NC-SA 4.0*
        - 4.2: A Home Experiment on Infiltration - *CC BY-NC-SA 4.0*
        - 4.3: The Physics of Groundwater Movement - *CC BY-NC-SA 4.0*
        - 4.4: Qualitative Aspects of Groundwater and Groundwater Flow - *CC BY-NC-SA 4.0*
        - 4.5: Some Practical Things About Water Wells - *CC BY-NC-SA 4.0*
        - 4.6: Groundwater in Coastal Regions - *CC BY-NC-SA 4.0*
    - 5: Rivers - *CC BY-NC-SA 4.0*
        - 5.1: Introduction - *CC BY-NC-SA 4.0*
        - 5.2: Fluvial Hydrology - *CC BY-NC-SA 4.0*
        - 5.3: Open-Channel Hydraulics - *CC BY-NC-SA 4.0*
        - 5.4: The Energy of Rivers - *CC BY-NC-SA 4.0*
        - 5.5: The Morphology of Rivers - *CC BY-NC-SA 4.0*
        - 5.6: Classifying Rivers - *CC BY-NC-SA 4.0*
        - 5.7: Variables Involved In Rivers - *CC BY-NC-SA 4.0*
        - 5.8: Fluvial Sediment Transport - *CC BY-NC-SA 4.0*
        - 5.9: Morphology and Dynamics of Meandering Streams - *CC BY-NC-SA 4.0*
        - 5.10: Drainage Networks - *CC BY-NC-SA 4.0*
        - 5.11: Fluvial Deposits - *CC BY-NC-SA 4.0*
        - 5.12: Floods - *CC BY-NC-SA 4.0*
        - 5.13: Some Practical Aspects of Rivers - *CC BY-NC-SA 4.0*
    - 6: Lakes - *CC BY-NC-SA 4.0*
        - 6.1: The Origin of Lakes - *CC BY-NC-SA 4.0*
        - 6.2: The Geometry of Lakes - *CC BY-NC-SA 4.0*

  - 6.3: Residence Time - *CC BY-NC-SA 4.0*
  - 6.4: The Water Balance of Lakes - *CC BY-NC-SA 4.0*
  - 6.5: The Vertical Temperature Structure of Bath Water - *CC BY-NC-SA 4.0*
  - 6.6: The Thermal Structure of Real Lakes - *CC BY-NC-SA 4.0*
  - 6.7: Classification of Lakes by Thermal Regime - *CC BY-NC-SA 4.0*
  - 6.8: The Lifetime of Lakes - *CC BY-NC-SA 4.0*
- 7: Glaciers - *CC BY-NC-SA 4.0*
  - 7.1: Introduction to Glaciers - *CC BY-NC-SA 4.0*
  - 7.2: Classification of Glaciers - *CC BY-NC-SA 4.0*
  - 7.3: Distribution of Glaciers - *CC BY-NC-SA 4.0*
  - 7.4: Glacier Ice - *CC BY-NC-SA 4.0*
  - 7.5: The Budget of Glaciers - *CC BY-NC-SA 4.0*
  - 7.6: Movement of Glaciers - *CC BY-NC-SA 1.3*
  - 7.7: Deformation of Ice - *CC BY-NC-SA 1.3*
  - 7.8: The Thermal Structure of Glaciers - *CC BY-NC-SA 1.3*
  - 7.9: Glacial Meltwater - *CC BY-NC-SA 1.3*
  - 7.10: Glacial Erosion - *CC BY-NC-SA 1.3*
  - 7.11: Glacial Sediment Transport - *Undeclared*
  - 7.12: Glacial Deposits - *Undeclared*
- 8: Coasts - *CC BY-NC-SA 4.0*
  - 8.1: Introduction to Coasts - *CC BY-NC-SA 4.0*
  - 8.2: Classification - *CC BY-NC-SA 4.0*
  - 8.3: Tides and Tidal Currents - *CC BY-NC-SA 4.0*
  - 8.4: Beaches - *CC BY-NC-SA 4.0*
  - 8.5: Deltas - *Undeclared*
- 9: Deserts - *CC BY-NC-SA 4.0*
  - 9.1: Introduction to Deserts - *CC BY-NC-SA 4.0*
  - 9.2: Classification of Deserts - *CC BY-NC-SA 4.0*
  - 9.3: A Brief Summary of Desert Features - *CC BY-NC-SA 4.0*
  - 9.4: Saltation - *CC BY-NC-SA 4.0*
  - 9.5: Eolian Ripplers and Eolian Dunes - *CC BY-NC-SA 4.0*
- 10: Mass Wasting - *CC BY-NC-SA 4.0*
  - 10.1: Introduction to Mass Wasting - *CC BY-NC-SA 4.0*
  - 10.2: The Controls on Downslope Movement - *CC BY-NC-SA 4.0*
  - 10.3: Kinds of Mass Wasting - *CC BY-NC-SA 4.0*
  - 10.4: Creep - *CC BY-NC-SA 4.0*
  - 10.5: Landslides - *CC BY-NC-SA 4.0*
  - 10.6: Debris Flows - *Undeclared*
- 11: Landscapes - *CC BY-NC-SA 4.0*
  - 11.1: Introduction to Landscapes - *CC BY-NC-SA 4.0*
  - 11.2: Mountains and Valleys - *CC BY-NC-SA 4.0*
  - 11.3: Davisian Geomorphology - *CC BY-NC-SA 4.0*
  - 11.4: How Fast Are the Continents Worn Down - *CC BY-NC-SA 4.0*
  - 11.5: Drainage Development in Newly Emergent Regions - *CC BY-NC-SA 4.0*
  - 11.6: Some Common Geomorphic Features Produced by Fluvial Erosion - *CC BY-NC-SA 4.0*
- 12: The Hydrologic Cycle, the Sediment Cycle, and the Carbon Cycle - *CC BY-NC-SA 4.0*
  - 12.1: Introduction - Systems, Cycles, Reservoirs, and Fluxes - *CC BY-NC-SA 4.0*
  - 12.2: The Earth’s Water, and the Hydrologic Cycle - *CC BY-NC-SA 4.0*
  - 12.3: The Sediment Cycle - *CC BY-NC-SA 4.0*
  - 12.4: The Carbon Cycle - *CC BY-NC-SA 4.0*
- Back Matter - *CC BY-NC-SA 4.0*
  - Index - *CC BY-NC-SA 4.0*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*