Particle-Driven Subaqueous Gravity Processes

M Felix and W McCaffrey, University of Leeds, Leeds, UK

Introduction

Particulate subaqueous gravity flows are sediment-water mixtures that move as a result of gravity acting on the sediment-induced density excess compared with the ambient water. The mixtures can range from densely-packed sediment flows, that are essentially submarine landslides, to very dilute flows carrying only a few kg m^3 of sediment. Gravity flow can take place in lakes and oceans, but some dense flows also occur in rivers. Sediment volumes transported by individual events can range up to thousands of cubic kilometres, although most events are of much smaller magnitude. Due to their infrequent occurrence and destructive nature, much information about subaqueous gravity processes comes from the study of their deposits and from laboratory experiments. Flow initiation mechanisms, sediment transport mechanisms, and flow types are described here separately, to emphasise the sense of process continuum needed to appreciate the development of most natural subaqueous gravity flows. This is followed by a description of internal and external influences on flow behaviour. Finally, the influence of flow regime on individual deposits is outlined.

Flow Initiation Mechanisms

A variety of processes can generate subaqueous gravity currents, with varying initial concentrations.

Direct Formation From Rivers

Currents can be formed when turbid river water flows into bodies of standing water such as lakes or oceans. If the bulk density of the turbid river water (sediment plus interstitial fluid) is higher than that of the receiving body of water, the river outflow will plunge, travelling along the bed as a hyperpycnal flow (or plume) beneath the ambient water. Such sediment-laden underflows may mix with the ambient water and transport sediment oceanward as particulate gravity currents. Although sometimes these river-derived flows are of high concentration (e.g., the Yellow River hyperpycnal plume), mostly they are dilute. Direct formation of subaqueous gravity currents in this way is, however, the exception rather than the rule. More commonly, the bulk density of the turbid river outflow is less than that of the ocean, and turbid surface plumes are generated. Nevertheless, particulate gravity flows can also form from surface plumes if material settling out collects near the bed at high enough concentrations to begin moving. A similar effect results from flow generated by glacial plumes where the sediment is slowly released into the water body.

Where the interstitial fluid in a hyperpycnal plume is of lower density than that of the ambient fluid, as is the case when freshwater rivers flow into brackish or fully saline bodies of water, ongoing sedimentation may induce buoyancy reversal. Thus, the gravity current will loft, in a manner similar to some subaerial pyroclastic density flows, and the flow will essentially cease to travel forwards, resulting in the development of abrupt deposit margins.

Sediment Resuspension

Loose sediment on the seafloor can be resuspended if bed shear stress is high enough. This can occur during storms or during passage of flows caused by density differences as a result of temperature or salinity. The resulting suspended sediment concentrations can be high enough to allow the mixtures to flow under the influence of gravity. As in the case of river-derived flows, resuspension usually generates initially dilute currents.

Slope Failure

Flows of much higher concentration may form as a result of slope failure. Sediment on submarine slopes can become unstable as a result of slope oversteepening during ongoing sedimentation, and during sea-level falls, as a result of high inherited pore fluid pressures and gas hydrate exsolution. Slope failure can alternatively be triggered by externally applied stresses, due to earthquakes, or as a result of loading induced by internal waves in the water column above (which chiefly occur in oceans). Initially, the failing mass becomes unstable along a plane of instability and a whole segment of the slope starts moving. Retrogressive failure and/or breaching can continue, adding material following the initial loss of stability. The concentration of this mass is at packing density but can become more dilute as flow continues.

Terrestrial Input

Not all subaqueous gravity flows need originate under water. Landslides, pyroclastic flows, and aeolian sediment transport originating on land can enter lakes or oceans and continue flowing underwater if the rates of mass flux are sufficiently high.

Grain Transport Mechanisms

Matrix Strength and Particle-Particle Interactions

Within dense flows, grains can be prevented from settling as a result of matrix strength (Figure 1). This strength may arise if some or all of the particles are cohesive. The resulting cohesive matrix prevents both cohesive and non-cohesive particles from settling out. In addition, particles can be supported by matrix strength within flows of non-cohesive grains if the particles are in semi-permanent contact, as is the case for flows whose densities are close to that of static, loose-packed sediment. For slightly lower concentrations, inter-particle collisions will help keep particles in suspension.

Hindered Settling and Buoyancy

Settling of particles can be slowed down by water displaced upwards by other settling particles (Figure 1). Such hindered settling is especially effective in dense mixtures with a range of grain sizes so that the smaller particles are slowed down by settling of the larger particles. The presence of smaller particles also increases the effective density of the fluid that the particles are settling in and thus enhances the buoyancy of the suspended particles and reduces settling rates.

Turbulence

The motion of sediment-laden flows can generate turbulence through shear at the bed, internally in the flow or at the top of a dense layer. The turbulent bursts generated at the bed tend to have an asymmetrical vertical velocity structure, with slower downward sweeps and more rapid upward bursts. This turbulence pattern counteracts the downwards settling of particles, moving them higher up in the flow (Figure 1). Turbulence generation is hindered and dissipation increased, however, if the particle concentration is high, or if the flow is very cohesive or highly stratified.

Flow Types

Broadly speaking, flows can be divided into three main types, depending on density:

Dense, Relatively Undeformed Flows, Creeps, Slides and Slumps

Flows of this type essentially have the same density as the pre-failure material. In each case the sediment moves as one large coherent mass, but with varying amounts of internal deformation. Grains remain in contact during flow and thus matrix strength is the main sediment transport mechanism. Such flows will stop moving or shear stress becomes too low to overcome friction, at which point the entire mass comes to rest. Flow thickness and deposit thickness are essentially the same, although flows may thicken via internal thrusting or ductile deformation as they decelerate prior to arrest. Slope creep caused by gravity moves beds slowly downslope with gentle internal deformation of the original depositional structure. Slides undergo little or no pervasive internal deformation, while slumps undergo partial deformation but the original internal structure is still recognisable in separate blocks. Thicknesses of slides and slumps range from several tens of metres to 1–2 km and travel distances can be up to about 100 km, with displaced volumes of up to 10^{12} m^3, although most flows are considerably smaller.

Dense, Deformed Flows: Rockfalls, Grain flows, Debris Flows and Mudflows

In flows of this type, sediment still moves as one coherent mass, but concentrations can be lower and the mass is generally well mixed, with little or no preservation of remnant structure from the original failed material. Sediment support mechanisms are matrix strength, buoyancy, hindered settling, and grain-grain collisions. Rheologically such flows are plastic (i.e., they have a yield strength). Clast types generally range from purely cohesive in mudflows, to cohesive and/or non-cohesive in debris flows (Figure 2) and purely non-cohesive for grain flows and rockfalls (where movement is by freefall on very steep slopes). These types of flow are formed as a

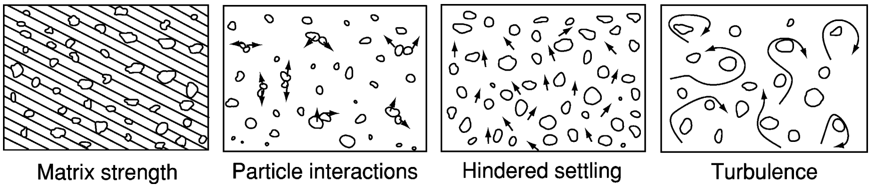


Figure 1 Schematic illustration of the principal grain transport mechanisms, shown in decreasing order of concentration from left to right.

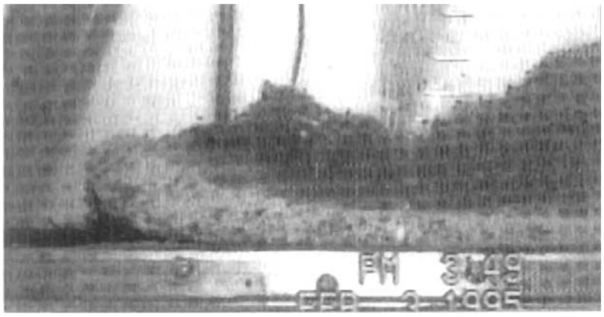


Figure 2 A laboratory debris flow from right to left. Note: a dilute turbidity current has been generated on the upper surface of the debris flow due to erosion of material by fluid shear. (After Mohrig *et al.* (1998) *GSA Bulletin* 110: 387–394.)

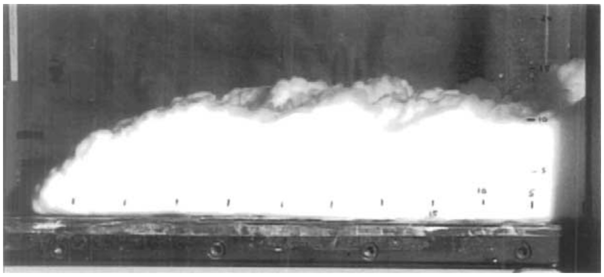


Figure 3 A laboratory turbidity current flow from right to left. Field of view is 55 cm wide. (After McCaffrey *et al.* (2003) *Marine and Petroleum Geology* 20: 851–860, with permission from Elsevier.)

result of rapid internal deformation following slope failure, from high concentration river input or from reconcentration of dilute flows (described below). Flow and deposit thicknesses can be up to several tens of metres with travel distances of several hundreds of kilometres. Erosion can add material to the flow and thus extend both travel distances and size of deposit – neither of which, therefore, necessarily relate to the initial flow mass. Motion will stop once friction is too high and flows will generally deposit *en masse*. Debris flows may develop a rigid plug of material at the top of the flow, where the applied stress falls below the yield strength. Such flows move along a basal zone of deformation, and may progressively ‘freeze’ from the top downwards, ultimately coming to rest when the freezing interface reaches the substrate.

(Partly) Dilute Flows: Turbidity Currents

In flows of this type, the sediment does not move as one coherent mass (Figure 3). These flows are generally dilute although parts of these flows can be of high concentration, especially near the bed. In the dilute parts of these flows, sediment is transported in either laminar or turbulent suspension. In higher concentration areas additional sediment transport mechanisms, such as grain-grain interactions, hindered settling, and buoyancy effects may also play a role. Rheologically, the dense parts of such flows can behave plastically, but the dilute parts are Newtonian. Concentrations in turbidity currents range from only a few kg m^3 to concentrations approaching those of static, loose-packed sediment. The dilute parts of these flows are commonly strongly vertically density-stratified. Turbidity currents can be formed via dilution of debris flows (see below), directly from river input or from resuspension of sediment.

Turbidity current thicknesses can be up to several hundreds of metres and can increase during flow due to turbulent entrainment of ambient water. Velocities can be up to tens of m s, but more commonly are around 1 m s or less. Larger flows, such as the well-documented Grand Banks event of 1929, may travel distances of a few thousand kilometres, even on nearly flat slopes, although distances of tens to hundreds of kilometres are more common. Sediment eroded during flow can add to the driving force and will increase flow duration and travel distance. Flows will gradually slow down as sediment settles out, with coarse material being deposited proximally and fine material distally. Deposit thicknesses generally are significantly smaller than flow thickness and are on the order of cm to dm, but can be up to multi-metre scale for large flows. However, ongoing sedimentation from flows of long duration can result in deposits whose thickness relates principally to flow longevity rather than flow thickness. Consequently, it is generally more difficult to interpret flow properties from analysis of turbidity current deposits (turbidites) than it is for the denser flow types.

Flow Transformations

Transformations of one flow type into another are common. Initially-dense slide masses may be disrupted due to internal shear, liquefaction, and disaggregation on various scales. If this deformation is sufficiently vigorous all the original structure of the failed material will be lost and the slides transformed into debris flows. In turn, these can transform into turbidity currents by erosion of sediment from the front and top of the dense mass due to ambient fluid shear (Figure 2), by disaggregation and dilution, and by deposition of sediment, diluting the flow. Turbidity currents can be transformed into debris flows if they reconcentrate, for example when mud-rich flows slow down. Further transformation into slides is not possible once the original internal structure is broken up.

The extent of transformation depends on flow size, velocity, and sediment content. Variable degrees of transformation can lead to the development of different flow types within one current, both vertically and from front to back. This co-occurrence of different flow types is especially common in flows with a dense basal layer and more dilute upper part. Thus, classification schemes which subdivide flows on the basis of discrete flow types do not recognise the diversity of natural flows, in which different types of flow may occur simultaneously and vary in relative importance in time and space as the flows evolve.

Internal and External Influences on Flow Behaviour

Flow behaviour is influenced both by internal factors such as concentration and grain size distribution and external factors such as input conditions and topography.

Flow Velocity

The driving force, and hence velocity of subaqueous gravity currents increases with both concentration and flow size. However, resistance to internal shear will increase with increasing viscosity due to increasing particle concentrations, and with increasing yield strength caused by cohesive particles. This will inhibit the increase of flow velocities. However, because concentration-induced resistance to shear does not scale with flow size, it can more readily be overcome by the higher gravitational driving forces of larger flows, which are, therefore, faster than smaller flows.

Flow Duration and Run-Out Length

Slope failure-induced slumps and slides that do not transform into debris flows and/or turbidity currents will generally be of short duration and have run-out lengths on the order of the initial failure size. If the failed sediment mass does transform into a debris flow, the duration and run-out length depend on the mobility as described above, with larger flows travelling further. However, because debris flows stretch out as they are flowing and because they may incorporate material by erosion, their run-out length may not be directly related to the initial failure size.

The duration and run-out length of turbidity currents depend on their size and sediment content, and hence also on their formation mechanism. Sustained input from rivers or glacial plumes can result in long duration flows, even if the input concentration is low. Turbidity currents that are generated from slope failures can have a short duration input, but tend to stretch considerably due to turbulent mixing and will thus increase in flow duration provided the transported sediment is kept in suspension. The ability of a flow to keep sediment in suspension, known as the flow ‘efficiency’, directly affects flow run-out lengths. Flow efficiency depends on flow magnitude, with larger flows being more efficient, and on grain size, as finer grains settle out more slowly than coarser grains. The presence of fine sediment in the flow also increases the ability to carry coarse sediment so both types of sediment will be carried further and both flow duration and run-out length will be increased.

Spatial and Temporal Changes to Flow

Flows are influenced both by the input conditions and by the terrain over which flow takes place. Flow behaviour therefore varies both temporally and spatially, causing local areas of erosion and deposition that lead to a deviation from a simple decelerating depositing flow and complicate the depositional pattern. Both spatial and temporal changes in flow behaviour can be caused by changes in sediment content of the flow: erosion adds driving force to the flow and increases velocity, while deposition slows flows down. Temporal changes to flow can also be caused by changing input conditions. River input from floods leads to flows that initially have a progressive increase in velocity followed by a long period of decreasing velocity. In retrogressive failure ongoing detachment of discrete sediment masses will result in pulsed sediment input; the rate of input generally tends to peak rapidly, and then diminish as successive slope failures reduce in size.

Local spatial changes in flow are caused by changes in the topography (Figure 4). The angle of the slope on which flow takes place is obviously important for gravity driven flows; when slope angle increases, the flow will go faster although the velocity increase will be diminished by the increase of friction with the ambient water. Nevertheless, small changes in slope angle can change flow behaviour. If the slope angle decreases, very dense flows can be stopped as the basal friction becomes too high. More dilute flows may undergo hydraulic jumps, in which they abruptly thicken and decelerate. This deceleration can cause coarser sediment to be deposited. Local changes to flow can also be caused by changes in the constriction of the flow path. When a flow goes into a constriction, velocity will increase. Where a flow can expand, as at the end of submarine canyons, velocity will decrease.

Momentum Loss

The evolution of flow behaviour can be different along flow-parallel and flow-transverse directions. Momentum will be greater in the direction of flow

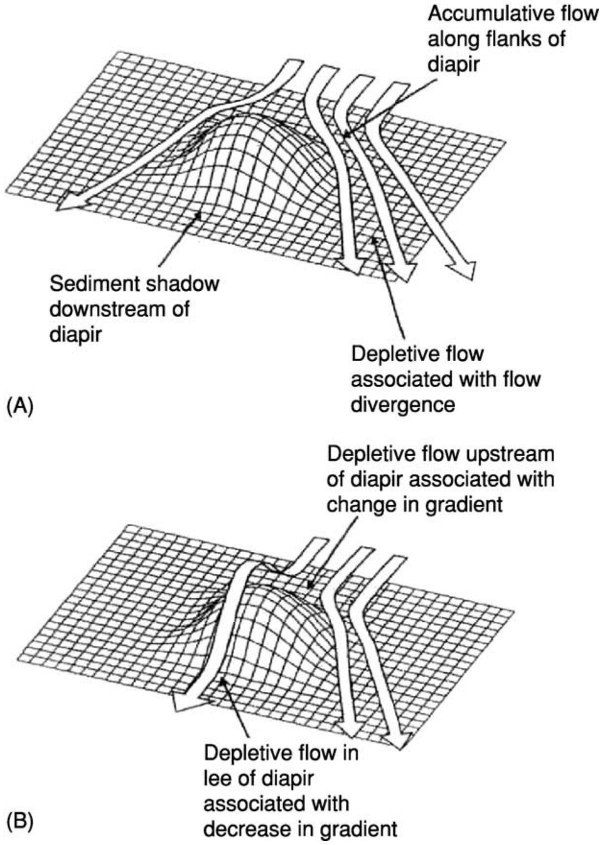

Accumulative flow along flanks of diapir
Sediment shadow downstream of diapir
Depletive flow associated with flow divergence
(A)
Depletive flow upstream of diapir associated with change in gradient
Depletive flow in lee of diapir associated with decrease in gradient
(B)

Figure 4 Schematic illustration of the interaction of turbidity currents with (A) high amplitude and (B) low amplitude bathymetry. Flows are uniform if the velocity does not change with distance and are non-uniform if the velocity does change. Accumulative flows have spatially increasing velocity while depletive flows have decreasing velocity. (After Kneller and McCaffrey (1995) SEPM, Gulf Coast Section, 137–145.) Published with the permission of the GCSSEPM Foundation; Further copying requires permission of the GCSSEPM Foundation.

than in the transverse direction. For coarse sediment in dilute flows, this means transport is principally in the main flow direction as rapid transverse momentum loss results in rapid deposition. This is less the case for fine-grained sediment, which will stay in suspension more easily and will thus generate momentum for flow in the transverse direction. These differences are not so important in restricted parts of the flow path, such as in canyons, but are important in less confined settings.

Channelised flow

If flows are erosive they can create conduits (incisional channels) both for themselves and for later flows. In aggradational systems, dense flows such as debris flows will start to form levees at their edges where flow becomes too thin to overcome the matrix strength. Sideway expansion of coarser-grained turbidity currents may lead to loss of momentum in the transverse direction, and thus greater rates of off-axis than on-axis deposition. This incipient levee formation may lead to the development of aggradational channels (Figure 5). These channels, which are generally sinuous, and often meandering, partly confine flow and can carry sediment downstream for long distances. Dilute parts of the flow can overtop the levee crests resulting in overspill and deposition of thin sheets of relatively fine-grained sediment that decrease in thickness away from the channel. This winnowing process causes the flows progressively to become relatively depleted in fine grained material, resulting in the development of sandy lobe deposits at the end of relatively muddy channel-levee systems. Levee height decreases downstream and flows become less confined. Like subaerial channels,

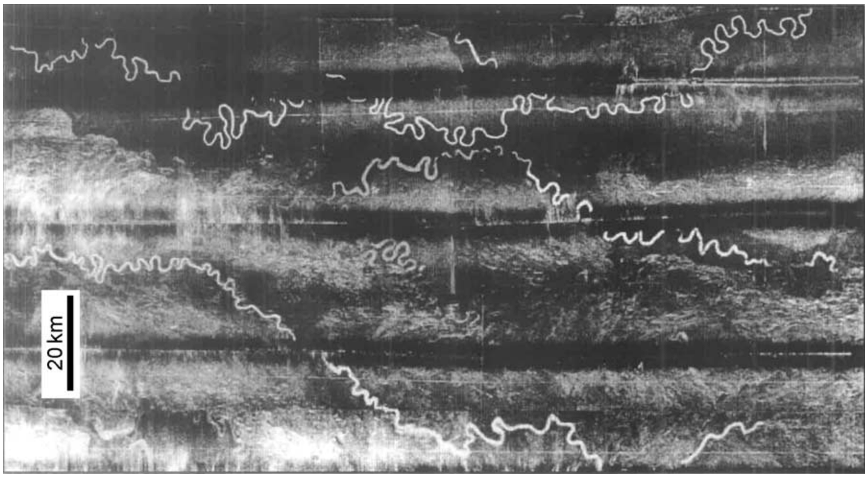


Figure 5 GLORIA image of sinuous submarine channels on the Indus fan. (From Kenyon *et al.* (1995). In: Pickering *et al.* Atlas of Deepwater Environments: architectural style in turbidite systems: London: Chapman and Hall, 89–93.)

aggradational submarine channels may undergo avulsion, resulting in the formation of internally-complex sedimentary fan deposits. Although channels are largely formed by the flows themselves, they can be influenced by pre-existing topography.

Flow in Unconfined Basins

When the basin size is very large compared to the flow, the flows are effectively unconfined. Flows that are not strong enough to erode and that are not captured by antecedent channels can start to spread out. Fine-grained, efficient turbidity currents spread out more evenly in all directions than their coarser-grained counterparts as a result of differing rates of momentum loss. Such unconfined flows can be influenced by Coriolis forces, being deflected to the right in the northern hemisphere, and to the left in the southern hemisphere. Unconfined flows deposit sediment in lobes, with deposit thicknesses decreasing in all directions away from the depocentre. Development of depositional topography may cause subsequent flows to be steered away from depocentres of previous flows and to deposit relatively more of their sediment load in offset positions in a process of autocyclic compensation. Deep-sea fan systems can form in unconfined basins settings through this process.

Flow in Confined Basins

When the basin size is smaller than or of the same size as the flow, the basin margins will prevent flow from expanding and the basin is said to be confined. Processes of topographic interaction induce spatial changes to flow, as detailed above. Flows can overcome small topographic obstacles, but as obstacle height increases relative to the flow height, part or all of the flow will be diverted. Flows in confined basins can be reflected back and forth between different basin margins if enough energy is available, which can result in reworking of the part of the deposit laid down during a previous pass of the flow. If the basin walls are sufficiently high to prevent any of the flow escaping, the basin is said to be ponded. In this case all the sediment is retained in the basin, and any mud present in the flow will be distributed in suspension evenly across the basin and will slowly settle out. The spatial restriction created by confined or ponded basins will hinder flow expansion. Thus, although autocyclic processes can play a role in dictating sedimentary architecture, in general basin fill patterns will be dominated by the confinement. Successive deposits can gradually fill up a basin completely. This can result in flows being able to partially bypass the basin, and enter the next basin downstream, in a process known as fill-and-spill.

Flow Regime Recorded in Depositional Sequences

Erosion and Bypass

If flow power is large enough, erosion can take place, which can remove significant volumes of sediment. This material adds to the driving force of the flow and can lead to acceleration (a process called ignition), and increased flow duration and travel distance. Smaller-scale erosion can form structures that indicate palaeoflow direction, including grooves, where an object is dragged along the bed, and flutes, where turbulent motions erode a characteristic shape that is deeper upstream, and both flares and shallows downstream. Erosion can take place beneath both debris flows and turbidity currents, although flutes require turbulence for their formation, a condition more likely to be met in turbidity currents. Not all flows are capable of erosion, but this does not necessarily mean they deposit their transported load. Bypass of sediment is common in upstream areas and may leave no record in the deposit. This behaviour is closely related to the process of autosuspension, in which sediment is transported by turbulence generated by flow caused by the density difference due to the sediment itself. Strictly speaking, such flows neither erode nor deposit.

Deposition

Eventually all flows, whether they start out as dense or as dilute flows, will lose their momentum and deposit their sediment. Dense flows such as slumps and debris flows will leave deposits whose structure more or less corresponds to that of the flows themselves. This is not the case for turbidity currents, which generally deposit their sediment progressively. Whether deposition takes place at all in turbidity currents depends on local flow competence and capacity. Flow competence indicates which grain sizes can be transported by a flow of a given velocity and flow capacity indicates how much sediment can be carried by a flow of a given velocity. The depositional structures of turbidity current deposits (turbidites) are influenced by the grain sizes carried in the flow, the velocity of the flow, and the sediment fall-out rate. High sediment fall-out rates cause suppression of primary sedimentary structures and lead to the formation of massive (structureless) deposits. The grains in these deposits tend not to be packed at maximum density and commonly re-organise themselves post-depositionally, expelling pore water in the process. This process commonly produces structures that overprint any primary depositional fabric. If fall-out rate is low enough, structures such as ripples and laminations can be formed, depending on grain size and flow regime. Deposit thickness is influenced both by flow size, with larger flows resulting in thicker deposits, and also by flow duration, with sustained flows being able to deposit thick beds, even if the flows themselves are not particularly large.

Various models have been proposed to describe the vertical succession of features in an idealised turbidite. The most widely applied is the model of Bouma, which describes a sequence deposited by a gradually decelerating turbidity current. Because all flows must eventually wane, full or (more commonly) partial Bouma sequences are developed quite frequently, particularly in relatively distal locations. However, the assumption that flows gradually decelerate over the entire flowpath is unlikely to be met, and many deposits will not look like this or other standard sequences. The influence of temporal changes and spatial changes on deposits will be reflected in terms of bed thickness and grain size distribution (grading). These are schematically presented in the diagram of Kneller for turbidity current deposits (Figure 6). This scheme is strictly valid only for flow where concentration does not change, which limits the applicability of the approach, but it illustrates the idea well. Finally, it should be borne in mind that depositional sequences may be reworked by surface currents, dewatering, and/or bioturbation. These processes may obscure any evidence of flow character that was originally recorded in the deposit.

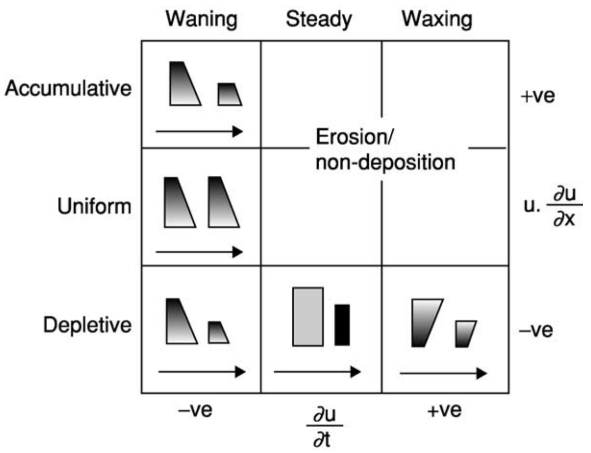


Figure 6 Schematic representation of vertical and lateral grain size variation within single beds as a function of the combined effects of flow steadiness and uniformity. The two logs in each field represent relatively proximal and distal configurations, respectively (arrow indicates flow direction). Flows are steady if velocity does not change with time and are unsteady if the velocity does change with time. Waxing flows have temporally increasing velocity while waning flows have decreasing velocity. Non-uniformity definitions are given in Figure 4. (After Kneller and McCaffrey (1995) SEPM, Gulf Coast Section, 137–145.) Published with the permission of the GCSSEPM Foundation; Further copying requires permission of the GCSSEPM Foundation.

Summary

Subaqueous particulate gravity currents may exhibit a wide range of concentrations, magnitudes, grain size, and type and flow velocities, all of which may change as flow develops. Flow behaviour is dictated both by input conditions (affecting flow magnitude, grain size distribution, and duration), and by the flow pathway (including its bathymetry, and the erodibility of the substrate). Thus, subaqueous particulate gravity currents form a complex and variable range of flow types, which together constitute the principal means by which coarser-grained clastic material is transported into the deep ocean.

Further Reading

Allen PA (1997) *Earth surface processes*. Blackwell Science, Oxford.

Hampton MA, Lee HJ, and Locat J (1996) Submarine landslides. *Reviews of Geophysics* 34(1): 33–59.

Kneller BC (1995) Characters of Deep Marine Clastic Systems. *Geological Society Special Publication 94*.

Kuenen Ph H (1950) Turbidity currents of high density. 8th International Geological Congress, London 8: 44–52.

Kuenen Ph H (1952) Estimated size of the Grand Banks turbidity current. *American Journal of Science* 250(12): 874–884.

McCaffrey WD, Kneller BC, and Peakall J (eds.) (2001) Particulate Gravity Currents *IAS Special Publication* 31: 302.

Schwarz HU (1982) Subaqueous slope failures – experiments and modern occurrences. *Contributions to Sedimentology* 11: 116.

Simpson JE (1997) Gravity currents in the environment and the laboratory, 2nd edn. Cambridge: Cambridge University Press.

Stow DAV, Reading HG, and Collinson JD (1996) Deep seas. In: Reading HE (ed.) *Sedimentary environments: processes, facies and stratigraphy,* 3rd ed., chapter 10, pp. 395–453. Blackwell Science, Oxford.

Walker RG (1992) Turbidites and Deep Sea Fans. In: Walker RG and James NP (eds.) Facies Models, 3rd edition, ch.13, pp. 239–263, *Geol. Soc.* Canada, St John's Canada.

Deposition from Suspension

I N McCave, University of Cambridge, Cambridge, UK

Introduction

Geological treatments of sediment dynamics generally lose sight of the fact that the last event of dynamic importance that happened to the sediment was that it was deposited. Instead, most accounts concentrate on the process of transport. Of course, a fair amount of work deals with the creation of bedforms, many of which are depositional, but occur in the transport regime of 'steady' flow, as well as in the 'unsteady' regime of flow deceleration which leads to deposition. It is not possible to deal sensibly with the topic of deposition from suspension without some mention of how material is transported, and so this article deals briefly with this aspect after giving an outline of the controlling factors and before describing the processes of deposition.

Almost any material, even boulders, can be transported in an aqueous turbulent suspension if the flow is large and sufficiently rapid. Even gravels were in suspension in the flood following the bursting of the glacial Lake Missoula in western Washington State (USA). However, most material deposited from suspension is mud and fine sand. Indeed, most ($>50\%$) of the sedimentary geological record is of silt and finer sizes ($<63\ \mu$m). Fine silt and clay, material of $\leq 10\ \mu$m, has the peculiar property that it can stick together, thereby transforming its settling velocity distribution. This, in turn, affects its response to changes in factors controlling its transport and deposition, such as the boundary shear stress and turbulence intensity.

The term 'suspension' is normally applied to material supported by turbulence in a boundary layer. However, in the oceans, much past work has referred to 'suspended particulate matter' (SPM) or 'total suspended matter' (TSM) obtained by filtration. This material, unless in the ocean bottom mixed layer, is not suspended but sinking, and thus, in a sense, is being deposited, although it may have several kilometres to go to reach the bottom. This material, comprising 'pelagic flux', is also affected by settling velocity transformations and is included in this article.

This article deals with controlling factors, entry into and maintenance in suspension, and aspects of deposition: pelagic flux and deposition from boundary layers on to flat beds and bedforms. It mainly concerns deposition from water, but some of the diagrams in non-dimensional form are applicable to air, and some comparisons are made with dust deposition from wind.

Controlling Factors

Particle Settling Velocity w_s

The still-water settling velocity of spheres collapses nicely on to a single curve when plotted as a dimensionless Reynolds number Re_p ($= w_s d/v$) vs. another dimensionless number used by (amongst others) M.S. Yalin in 1972, and here called Yalin's number: $\Xi = (\Delta\rho_s g d^3/\rho v^2)$ (Figure 1) (w_s is the settling velocity, d is the diameter, $\Delta\rho_s = \rho_s - \rho$ is the solid minus fluid density, and v is the kinematic viscosity). It should be noted that the abscissa Ξ contains material variables only, i.e., particle size and density and fluid viscosity and density. It should also be noted that the curve has two straight line segments and a curved transition joining them. The upper segment is $Re_p = 1.75\Xi^{0.5}$, with a lower limit around $d = 2$ mm, and is thus applicable to gravel in water. The lower segment is Stokes' law, $w_s = \Delta\rho_s g d^2/18\mu$, where μ is the molecular viscosity ($\mu = \rho v$), applicable below $Re_p \approx 0.5$ or $d < 100\ \mu$m in water and air. A further order of complexity is introduced by the fact that particles are not often spheres. Dietrich has developed empirical relations that deal with the varying shapes of solid particles which always sink more slowly than spheres. We can note in passing that Figure 1 divides into a coarse end (gravel), in which $w_s \propto d^{1/2}$, and a fine end (silt and clay), in which $w_s \propto d^2$, whilst the transition is occupied by sand for which, roughly, $w_s \propto d$. Density exerts a strong control, but only for a minority of particles – most solids are quartz-carbonate density (2500–2900 kg m^{-3}). The molecular viscosity of water ranges mainly from 0.9×10^{-3} to 1.5×10^{-3} Pa s (25 to 2°C), and thus gives nearly a factor of two variation. For air, the viscosity is $\sim 1.8 \times 10^{-5}$ Pa s, but as its density is only 1.2 kg m^{-3}, the kinematic viscosity is $\sim 1.33 \times 10^{-5}$ m^2 s^{-1}.

Much greater variation is due to the bulk density variations of particles that are not solid: aggregates, hollow particles, and grains containing gas bubbles. There has been little systematic study of the latter, although they are clearly important in hot volcanic dust suspensions (e.g., ignimbrites). The most important classes of hollow particles are foraminifera, diatoms, and radiolaria. Forams are often partially sediment filled, resulting in saturated bulk densities

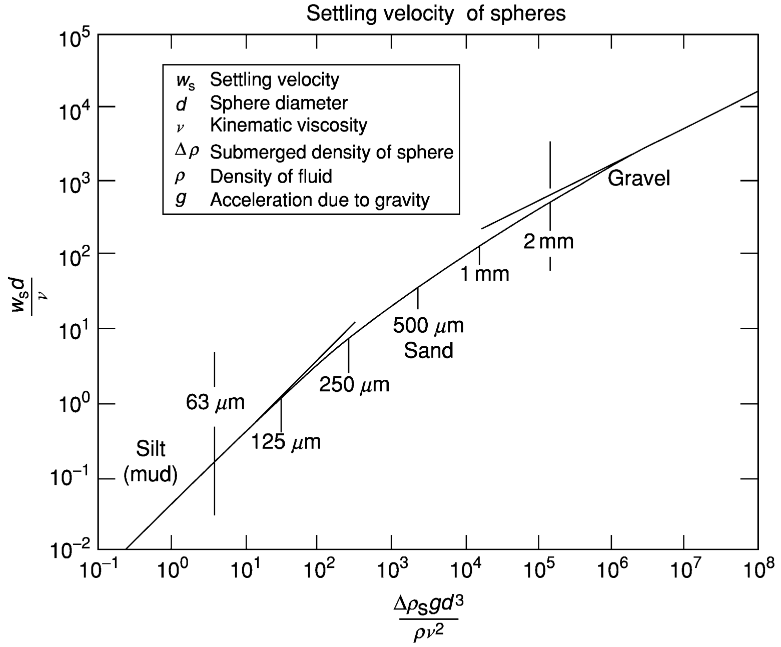


Figure 1 Universal settling velocity curve for spheres. Axes are the particle settling Reynolds number, $Re_p = w_s d/\nu$, and Yalin's number, $\Xi = \Delta\rho_s g d^3/\rho\nu^2$. This allows variable density particle and size fluid viscosity to be accommodated.

(that is the mean density of the foram shell plus sediment and water in the cavities) of 1150–1550 kg m^{-3}. (For simplicity, a deep seawater density of 1050 kg m^{-3} is used.) This gives a large range of $\Delta\rho_s$ from 100 to 500, which translates straight into a settling velocity straddling the Stokes' boundary, but giving a settling velocity of 125–500 m per day for 200 μm forams. With variable size, density, and viscosity, sinking rates can be from 50 to 1000 m per day. In air, the settling velocities are much faster but, because the viscosity is less, the viscous-dominated Stokes' settling region also persists up to 100 μm. However, the fast falling speed means that sand is almost never suspended by wind.

Aggregation

Most fine silt and clay is deposited as aggregates. These aggregates may be formed by physical (often referred to as flocculation, sometimes coagulation) or biological processes, generally involving the feeding and production of real or pseudo-faeces. The finest particles (d below 1 μm) are brought into contact by molecular buffeting known as Brownian motion. Here, for similar sized (diameters d_i, d_j) spherical particles, the probability of an encounter (i.e., number of collisions per cubic metre of suspension per second) is proportional to $TN_0 d_{ij}^2$ and inversely proportional to $\mu d_i d_j$ (where d_{ij} is the sum of the diameters of the two colliding particles). This is clearly favoured by high concentration (N_0) and temperature (T) and opposed by viscosity (μ). Larger particles are brought together either by turbulent shear, where the collision probability is proportional to Nd_{ij}^3 du/dz, or by larger, fast-sinking particles sweeping up finer ones, like rain falling through mist. In both cases, larger particles grow more rapidly because, when d_i is large and d_j is small, the sum cubed is large, whereas, when the size difference is not great, d_{ij}^3 is not large either. (This also takes into account the fact that the number concentration distribution of particles is such that there are far fewer large than small particles. The simplest standard is a flat log volume vs. log diameter distribution, which is equivalent to the cumulative number (log) vs. size (log) with a slope of −3. This means a ratio of 1000 1 μm particles to just one of 10 μm diameter.) The rainfall analogy above is particularly apposite for airborne dust, because one method of deposition is washout in which falling rain droplets form aggregates with dust particles by collision and carry them to earth. The other mechanisms of aggregation, Brownian motion, and shear also act on fine particles in air.

A key feature of much aggregation is the presence of organic mucus which acts as 'glue', allowing particles which get close to stick. Many organisms produce mucopolysaccharides, notably bacteria which sit on particles. Although there has been progress in implementing schemes to calculate particle aggregation, they are not yet simple or robust and represent the

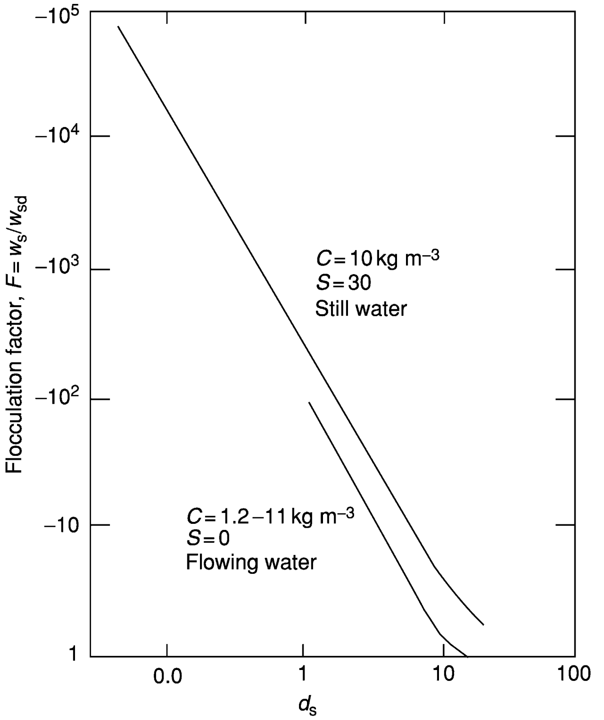


Figure 2 Flocculation factor F (ratio of floc settling velocity to settling velocity of the primary particles from which it is made) vs. diameter of the primary particles. It should be noted that F is negligible in flowing water for $d_{50} > 10\,\mu\text{m}$. Data from Migniot C (1968) Etude des proprietes physiques de differents sediments tres fins et de leur comportement sous des actions hydrodynamiques. *La Houille Blanche* 7: 591–620 and Dixit JG (1982) *Resuspension Potential of Deposited Kaolinite Beds*, MS thesis, U. of Florida, Gainesville.

frontier of research in cohesive sediment dynamics. The importance of aggregation is shown in Figure 2, where the 'flocculation factor', the ratio of the settling velocity of an aggregate to the settling velocity of the primary particles from which it is made, can be up to 10^5.

Aggregates are not stable entities. The organic membrane covering faecal pellets decays and the mucus that holds aggregates together also degrades, and so particles fall apart whilst sinking. Aggregates assembled by moderate levels of turbulence in the outer part of the boundary layer may be broken up by more energetic turbulent eddies close to the boundary. The relationship between aggregate size and boundary shear stress is very poorly known and, if the floc diameter $d_f \propto \tau^{-n}$, then n ranges from 0.25 to 1 (τ is the shear stress in the fluid). The larger aggregates, which can be up to 5 mm in diameter, are thus found in the moderately turbulent, high-concentration environment of the estuarine turbidity maximum above the region within a metre of the bed. Closer to the bed, high shear breaks these large sloppy aggregates into smaller pieces. The density of flocculated sediment decreases as the flocs increase in size. A simple expression based on field data is $\Delta\rho_f = 4.9 d_f^{-0.61}$, where the floc excess density $\Delta\rho_f$ is in kg m^{-3} and the floc diameter d_f is in millimetres. This yields floc excess densities of less than $10\,\text{kg m}^{-3}$ for 300 μm aggregates (compared with solid particles, where it is ~1600 kg m^3). Nevertheless, because the settling velocity increases as the square of the diameter, large flocs settle at the same speed as fine quartz sand grains, or at >200 m per day, which means that they reach the bottom quickly, resulting in significant clearing of the water in a 10 m deep estuary in a slack-tide period of 2 h.

Boundary Layer Turbulence

A fluid flowing over a surface exerts a drag force on it. The drag at the boundary slows the fluid down, but some distance out, known as the boundary layer thickness, the average flow speed no longer changes much with distance. Most rivers are completely boundary layer as are shallow marine tidal flows. In the atmosphere and deep sea, the boundary layer extends several tens of metres above the surface. Boundary layers are intensely turbulent, and the drag force τ_0 exerted on the bed is related to that intensity because the stress is transmitted by eddies. In the vertical plane, $\tau_0 = -\rho\overline{uw}$, where u is the turbulent component in the flow direction and w is the up and down component (actually perpendicular to stream lines which may not be quite vertical). This expression is very important because u and w are related, so that $\tau_0 \propto \overline{w}^2$. This vertical turbulent velocity is responsible for keeping particles in suspension, and the turbulent stress $\overline{uw}$ either causes aggregation or, at higher values, disaggregates fine particles. The term (τ_0/ρ) has the dimensions of a velocity squared and that velocity is called the shear or friction velocity: $U_* = (\tau_0/\rho)^{1/2}$. From the above, it can be seen that $U_* \propto \overline{w}$.

Regions of the Boundary Layer

Flows may be distinguished as laminar or turbulent on the basis of their Reynolds number, Ul/v, where l is some relevant length scale (e.g., depth of a river, diameter of a particle) (*see* **Unidirectional Aqueous Flow**). Low Reynolds number flows are laminar, high Reynolds number flows are turbulent. In a turbulent flow, the speed decreases towards the bed because of the drag, so that very close to the bed the flow becomes laminar, or at least dominated by viscosity, in a layer known as the 'viscous sublayer' of the turbulent boundary layer. This is very thin. In water, for a flow that just moves very fine sand ($U_* = 0.01\,\text{m s}^{-1}$, $v = 10^{-6}\,\text{m}^2\,\text{s}^{-1}$), $\delta_v = 10v/U_*$ is just 1 mm thick. (In

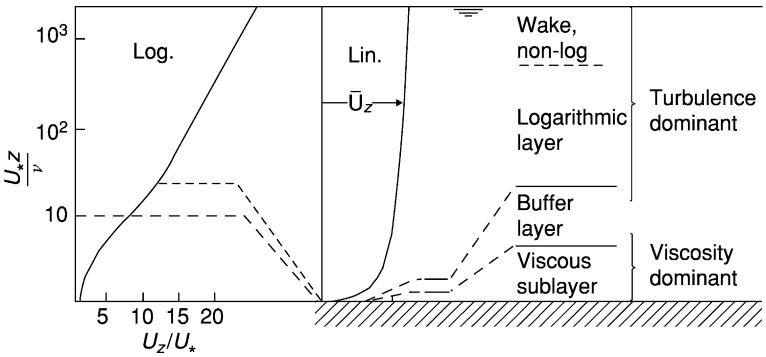


Figure 3 Regions of the turbulent boundary layer for a flow 1–10 m deep. In the centre, the linear representation of flow speed vs. height cannot resolve the viscous sublayer, but the speed vs. log height (*z*, expressed as a Reynolds number) shows it very well.

air, for the same stress, its thickness is similar, about 0.5 mm.) However, this is ten times the diameter of very fine sand. The shear across this layer is very large; for $U_* = 0.01\,\mathrm{m\,s^{-1}}$, the speed goes from 0 to $0.1\,\mathrm{m\,s^{-1}}$ in just 1 mm. Weak aggregates cannot survive this shear and break up. Above this sublayer, there is a transition ('buffer layer') to a region in which the flow speed varies as the logarithm of height above the bed (Figure 3). As the flow speed decreases, U_* decreases and δ_v increases, so that, in deposition, most particles that are going to become part of the geological record have to get through the viscous-dominated layer. Although viscous dominated, this layer is actually not laminar. Spatially, it has a structure of high- and low-speed streaks, and temporally very high-speed 'bursts' of fluid out of the layer and 'sweeps' of fluid into it from outside. These are associated with stresses typically up to 10 times the average (and extremes of 30 times), and so the mean shear example given above is a minimum, and even strongly bound particles may find themselves ripped apart just as they were getting within sight of the bed and posterity. Above the viscous sublayer, the 'buffer layer' is overlain by a zone in which the flow speed varies as the logarithm of distance from the bed (the 'log layer'). This zone is fully turbulent with eddies becoming longer with height above the bed and turbulence intensity becoming smaller.

The roughness of the bed positively influences the drag and turbulence, but also provides quiet regions in between large grains where fine particles can settle. Fine sediment can thus be deposited in the interstices of gravel, affecting several processes, e.g., the spawning of salmon.

Critical Conditions for Suspension

Two views of the critical suspension condition are as follows: (1) at critical movement conditions, the turbulent intensity can hold particles up, and so suspension depends on whether the particles are ejected from the viscous sublayer; and (2) sublayer ejections are fast, and so suspension depends on whether the vertical turbulent velocity can hold the particles up after injection into the flow. The second view was held by many, but recent work suggests that the first view may be correct. This view is based on high-speed video observations of particles close to the bed, which show that there is a threshold level of shear stress for the particles to respond to turbulent ejections of fluid from the viscous sublayer. The second view would mean that fine to very fine sand would immediately go into suspension as soon as it moved. For example, for 100 μm sand, the critical erosion shear velocity U_* is 0.012 $\mathrm{m\,s^{-1}}$, and the settling velocity of this very fine sand is 0.008 m $\mathrm{s^{-1}}$, and so it is capable of being held up by the flow, but video data show that it is not suspended. This means that there is a region of bedload transport for all particles of settling velocity, at least down to ~30 μm silt. This is shown on a conventional non-dimensional erosion diagram in Figure 4.

The significance of this is that, in a decelerating flow, below the suspension threshold, material may continue to move, but not in suspension. Experimentally it has usually been found easier to determine the critical suspension condition with increasing flow, rather than failure of suspension on decreasing flow. It is generally assumed that the two views are equivalent.

Transport in Suspension

Once material is moved out of the near-bed region, it is held in suspension by the action of fluid turbulence. For this, because the vertical turbulent component of velocity is about the same as the shear velocity U_*, the normal suspension criterion is that $w_s/U_* \leq 1$.

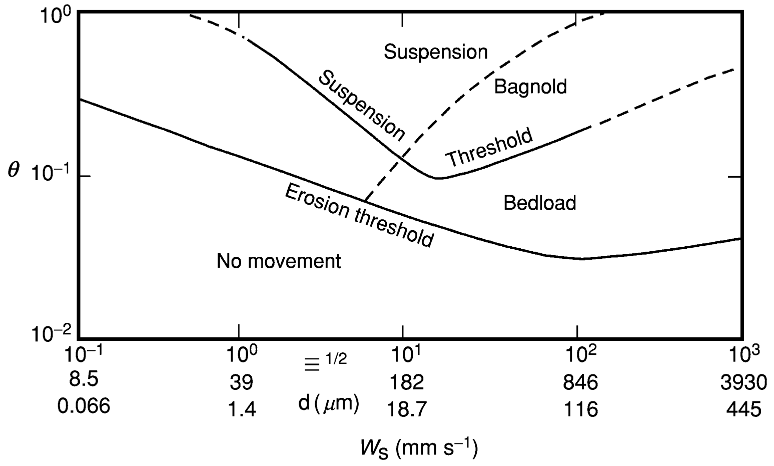


Figure 4 A critical erosion diagram on non-dimensional axes with a critical suspension line added. This divides the diagram into regions of suspension, bedload, and no movement. Below the suspension threshold, material falls out and, as the capacity of a flow to carry bedload is limited, deposition will ensue. Two suspension lines are shown, ''suspension threshold'' results from view (1), while ''Bagnold'' expresses view (2); see text $\theta = \tau_0/\Delta\rho_s gd$, $\Xi = \Delta\rho_s gd^3/\rho\nu^2$.

Particles in 'steady' transport diffuse up from the source at the bed and sink back down under gravity with a balance in steady state. This is expressed as

$$Cw_s + \varepsilon_s \mathrm{d}C/\mathrm{d}z = 0$$

where the first term is gravity settling and the second is upward diffusion (ε_s is the sediment diffusivity). The result of this is that, for a given value of U_*, the faster settling grains are found closer to the bed and the finer slower settling particles are more uniformly distributed over the flow depth (Figure 5). In the bottom of a deep flow, the concentration at height z in the flow is $C_z = C_a(a/z)^\zeta$, where C_a is the concentration at height a (the point near the bed at which a measurement is made) and $\zeta = w_s/\kappa U_*$, where κ is von Karman's constant (0.4). This means that, with our suspension criterion, $w_s \leq U_*$, we would not expect much material in suspension for $\zeta > 2.5$. Figure 5 shows this. Here, it can be seen that relatively fine material (with $\zeta < 0.125$) is distributed throughout the whole flow. This is the fine silt and clay of river 'washload'. Closer to the bed, the relatively coarser sediment is concentrated. Clearly, if the flow slows down, the coarser material will be rapidly deposited because it is only just above the suspension threshold and has very little distance to reach the bed.

In air, there is very little suspension of sand. Above camel height in a 'sandstorm', the suspended material is virtually all silt and clay-sized dust. Saltation (which is bedload) is confined to the lower ~1.5 m, and very little material of >70 μm is carried in suspension.

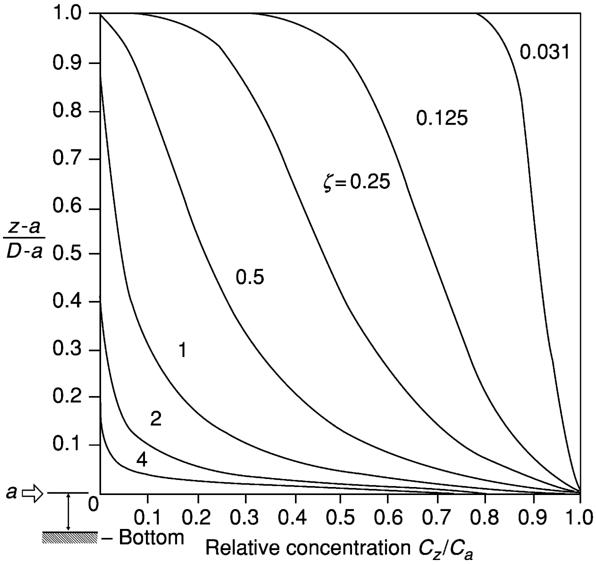


Figure 5 Variation of concentration with height of particles with increasing ratio $\zeta = 2.5w_s/U_*$. This shows material with high values of ζ close to the bed. This could be quite fine-grained material if the flow has become very slow (low U_*).

Sinking Deposition: Pelagic Flux

The oceans are full of particles that are sinking, some slowly, some fast. The origin of most of this material is from biological production in the upper ocean. Thus, it comprises organic matter, calcium carbonate, and opal. It has been shown theoretically and by the use of sediment traps that most of this material would not reach the seabed were it not for the process of aggregation which transforms its settling velocity spectrum. The aggregates are faecal pellets and 'marine snow', loose aggregates based on mucus and gelatinous structures made by zooplankton. This material sinks at ~100 m per day, a huge increase over the 2 m per day of a 5 μm coccolith. Aeolian dust rained out on to the sea surface also becomes incorporated into these aggregates, providing a rapid route to the bed.

Only if there is very slow flow at the bed in the bottom boundary layer will these aggregates plummet down directly on to the bed. This is true for most of the ocean most of the time, but some areas have fast currents which can break up the aggregates and control deposition.

Close to continental margins, the action of waves on the outer shelf and internal waves on the shelf-break and slope leads to the resuspension of material. This resuspended sediment spreads out on surfaces of density contrast as intermediate nepheloid ('cloudy') layers and flows down-slope in bottom nepheloid layers. These turbid layers are found all over the sea bottom, some more concentrated than others. Most fine sediment deposition involves some transport and removal from nepheloid layers, except for coarser (sand-sized) components which simply sink to the bed. Continental margins thus contain much material that is rained out of suspension and moved in bottom nepheloid layers during deposition.

Deposition from Turbulent Boundary Layers

The 'deposition' of bedload is rather straightforward: it stops moving. This occurs in water at a shear stress only slightly lower than the critical erosion stress. In fact, we cannot measure the erosion stress precisely enough to distinguish between erosion and deposition stresses, and so they are effectively the same. In air, high-speed grain bombardment of the bed keeps the bedload moving until the stress has been reduced to ~80% of the critical value. For suspended sediment, once the stress has decreased below the suspension threshold (see Figure 4), material will sink into the near-bed region, thereby increasing the concentration and causing some material to be deposited.

One well-documented consequence of the concentration increase is that the flow becomes density-stratified. This reduces turbulence intensity by absorbing turbulent energy in order to keep grains up. A reduction in turbulence means a reduction in shear stress, and deposition ensues. An extreme case is the suppression of the ability of a highly concentrated flow to sort sediment, resulting in the deposition of the massive graded A division at the base of Bouma-type turbidites. As the concentration decreases, turbulence is sufficient to sort the succeeding B division into laminae.

Deposition from different modes of transport is reflected in sediment size distributions. In a typical S-shaped cumulative size frequency curve for sands, the coarse tail reflects material that was always carried as bedload, the central part of the distribution reflects material carried intermittently in suspension, and the fine tail is made up of the 'washload' – long-distance suspensions. However, for air, the coarse tail is the 'creep' part of the bedload, the central part is the saltation component, and the fine tail is sand and dust from suspension.

Processes of Deposition

Fine sediment may reach the bottom in one of three ways. It may settle to the bed under gravity, it may impinge on the bottom as a result of molecular agitation in Brownian diffusion, or it may be transported downwards by eddy diffusion. The critical region of particle transport for all three of these processes lies within the viscous sublayer of a turbulent boundary layer, because, in most cases in which deposition occurs, a large part of the bed is covered by the sublayer. Simple calculations show that settling, even of 1 μm particles, is several orders of magnitude greater than the diffusive deposition rate. As diffusion is only likely to be important for the smallest particles, we can safely neglect it, and consider deposition to be controlled by particle settling through the sublayer to the bed. In the case of rain washout of dust and falling of 'marine snow' in areas of slow deep-sea currents, particles reach the bed at relatively high speed with no intervention of turbulence or the viscous sublayer.

Rate of Deposition

Experiments in flowing water show that, below a certain shear stress τ_d, for $C_0 < 0.30\,\text{kg}\,\text{m}^{-3}$, the concentration in suspension C_t decreases exponentially with time t

$$C_t = C_0 \exp(-w_s pt/D)$$

where D is the depth of flow (or thickness of the boundary layer), C_0 is the initial concentration, and p is the probability of deposition; the probability is given by $p = (1 - \tau_0/\tau_d)$ (this includes nearly all normal marine conditions; only intense storms, estuaries, and mass flows have higher values). In this expression, τ_d is the limiting shear stress for deposition, the stress below which all the sediment will eventually deposit. This yields $R_d = C_b w_s(1 - \tau_0/\tau_d)$ for the rate of deposition R_d ($\text{kg}\,\text{m}^{-2}\,\text{s}^{-1}$). Here, C_b

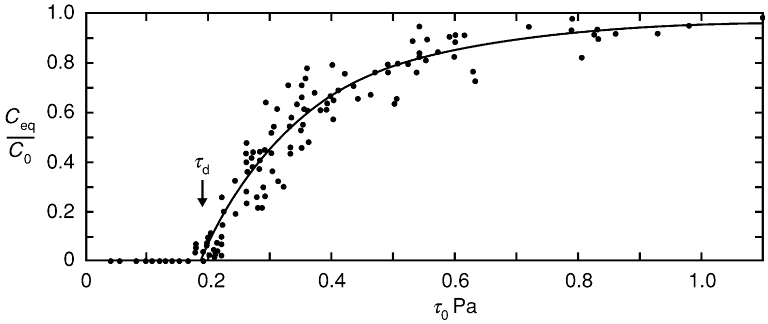


Figure 6 Diagram showing the equilibrium concentration (C_{eq}) of material in suspension as a ratio with the amount initially in suspension (C_0, 1.25 kg m^{-3} in this case) as a function of the bottom shear stress (τ_0). This shows that some material is deposited at shear stresses as high as 0.6 Pa, but that, below $\tau_0 = 0.2$ Pa, all is deposited, thus defining the limiting stress for deposition (τ_d) of this material. Reproduced from McCave IN (1984) Erosion, transport and deposition of fine grained marine sediments. In: Stow DAV and Piper DJW (eds.) *Fine-Grained Sediments: Deep Sea Processes and Facies, Special Publication, Geological Society of London 15*, pp. 35–69.

is the value near the bed. If there is no flow, this reduces simply to the settling flux $C_b w_s$.

At higher concentrations, $C = 0.3$–$10\,\mathrm{kg\,m^{-3}}$, C declines logarithmically with the logarithm of time, $\log C = -K \log t +$ constant, in which $K = 10^3(1-\tau_0/\tau_d)/D$. It has also been noted that, at high concentration, some material is deposited at $\tau_0 > \tau_d$ (Figure 6); an equilibrium concentration is attained, which reduces as τ_0 is reduced to τ_d. This means that some mud can be deposited from relatively fast flows, as long as the concentration is fairly high (>1 kg m^{-3}). The significance of this is that, after storms or under turbidity currents, fine sediment may be deposited under flow speeds of several tens of centimetres per second. The deposition of mud at $\tau_0 = 0.4$–0.5 Pa in Figure 6 is occurring under conditions capable of moving coarse sand of 0.5–1 mm. It is emphatically not the case that mud is only deposited under 'quiet water' conditions when sand cannot be moved.

Limiting Shear Stress for Deposition τ_d

The value of this is not well known, but is probably related to the diameter or, more properly, the settling velocity of the particles, whether aggregates or single grains. The safest assumption is that it is given by the critical erosion stress for non-cohesive grains because, below this value, movement ceases and any grain reaching the bed would be removed from the transport system. This is shown in the critical erosion diagram (Figure 4). An alternative, based on measurements in a laminar flow cell, is $\tau_d = 0.048\Delta\rho_s g d$.

Deposits Formed from Currents

Often, fine sediment deposition may occur from flows of, for example, 0.1–0.2 m s^{-1}. What influence might this have on the character of deposits? A current of 0.1 m s^{-1} in a deep boundary layer having a shear velocity of 4×10^{-3} m s^{-1} allows the deposition of particles larger than 20 μm, but the deposition of finer particles is suppressed. Work on suspended material in nepheloid layers has shown that it comprises aggregates made of silt and clay-sized particles and organic matter. The suppression of the deposition of the finer particles must result in a more silty deposit, but does not eliminate clay completely, because some is caught up in larger fast-settling aggregates, the strongest of which survive stresses in the buffer layer and are deposited. Thus, there is fractionation of a suspension during deposition to yield a more (or less) silty deposit. Some people tend to think that this results from 'winnowing', but this is a process of selective erosion not deposition. Two sorting processes occur: (1) fractionation during deposition, yielding more silty accumulations by deposition under higher shear stress; and (2) fractionation during intermittent erosion, yielding a more compact deposit with (micro) erosion surfaces marked by thin lag layers of terrigenous coarse silt and sand grains and foraminifera.

Above about 10 μm, the flocculation factor in flowing water becomes quite small, as many aggregates are broken up by flow in the buffer layer, particularly by strong flow (but may re-aggregate when away from the bed) (Figure 2). This means that, under stronger flows, this material can be size sorted according to its primary grain size. This part of the size spectrum (10–63 μm) comprises what has been called 'sortable silt' (as opposed to the cohesive material finer than 10 μm), and its mean size has been used as an index of the flow speed of the depositing current. The method gives results showing a striking correspondence to climate change-driven deep circulation changes.

Bedforms from Suspension

Bedforms from sandy suspensions are covered in (*see* **Sedimentary Processes:** Depositional Sedimentary Structures). Some structures seen in the deep sea, but rarely preserved in mudstones, are worth noting.

Mud Waves

Mud waves are regular undulations of the sediment surface with wavelengths of about 0.5–3 km and heights of 10–100 m. Most mud waves are very nearly symmetrical, but they contain subsurface layering that indicates migration. Observed migration usually is up-slope and up-current, but instances of down-current migration have been observed. Under a simple flow, maximum shear stress is expected on the up-stream face of a wavy bedform, and lower shear stress, with a greater deposition rate, on the downstream side. This would give downstream migration of the wave. Commonly observed upstream migration has suggested to some that mud waves are analogous to fluvial antidunes developed under a supercritical flow.

An alternative is that the mud waves form under internal lee waves initially triggered by an upstream topographical disturbance, without the necessity for a supercritical flow speed. Temperature data over mud waves show that such an upstream phase shift does occur, that the implicit internal-wave phase velocity is 0.05 m s^{-1}, and that this is very similar to the measured flow velocity required for the internal wave to be stationary. The flow pattern over the waves has widely spaced stream lines, giving a small velocity gradient and shear stress (= high deposition rate) on the upstream slope and the opposite on the downstream slope. This would give the observed upstream migration.

Longitudinal Ripples

Longitudinal ripples are elongated features parallel to the depositing flow, probably with helical secondary circulation involved in their formation. In the deep sea, they are 5–15 cm high, 0.25–1 m wide, spaced at 1–5 m apart, and up to 10 m long, and have a generally symmetrical cross-section with sides slightly concave upwards (Figure 7A). In many cases, the ripples have a mound of biological origin at the upstream end. Surface markings on some ripples demonstrate the action of oblique flows, with flow separation and a zone of helical reversed flow on the lee side.

Dating by ^{234}Th (half-life, 22 days) suggests that longitudinal ripples form by deposition from suspension, occurring in a few episodes of very rapid deposition following deep-sea storms. Subsequently, the ripple is scoured by flows that may be oblique to its trend, giving the surface markings seen in deep-sea photographs. These structures have not been recorded from shallow marine or estuarine muds, although closely spaced features up to a few tens of centimetres apart have been seen on tidal flats.

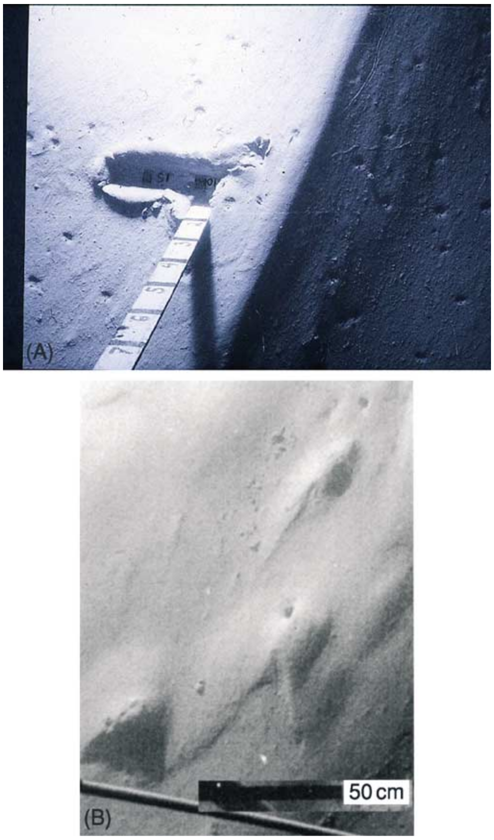


Figure 7 Photographs of bedforms in mud from the deep sea. (A) Longitudinal ripple. Scale bar has penetrated the mud; width of view, ~40 cm; relief of ripple, ~12 cm. (B) Barchan ripples. Both from the Nova Scotian Rise at 4800 m depth.

Smaller, Current-Controlled Bedforms

Smaller, current-controlled bedforms are also revealed by deep-sea photography. The photographed features of the seabed can be arranged in a sequence indicative of increasing flow speed. The progression is from tranquil seafloor (biological mounds, tracks, trails, and faecal pellets), through increasing overprinting by current effects, to features showing clear evidence of erosion. Biological activity is almost ubiquitous, so that a smoothed surface is indicative of an appreciable current, sufficient to remove the surface effects of biota. The most common features are actually biologically produced faecal mounds, tracks and trails, and pelleted surfaces. Mounds are often modified by current activity, the most frequent structure being mound-and-tail formed by lee-side deposition. As suggested above, it may be that longitudinal ripples are very large tails on mounds. Both structures are excellent current direction indicators. Transverse ripples are sometimes seen in muds, but barchan-shaped (crescentic) ripples are more common (Figure 7B). Some are formed rapidly on deposition from high-concentration suspensions in deep-sea storms, and others are winnowed crescentic silt ripples. Both tend to nucleate around biogenic mounds. Unfortunately, the preservation potential of these structures as distinctive stratification patterns is negligible. Mud is so nutritious that it is populated by a rich burrowing infauna (producing the mounds) and the main structure is pervasive bioturbation.

Nomenclature

Parameter

$\Delta\rho_s$	density difference(= $\rho_s-\rho$); unit, kg m^{-3}; dimension, M L^{-3}; value, 1650 (water), 2650 (air)
δ	boundary layer thickness; unit, m (mm); dimension, L
ε	eddy diffusivity; unit, m^2 s^{-1}; dimension, L^2 T^{-1}
μ	dynamic viscosity; unit, Pa s; dimension, M L^{-1} T^{-1}; value, 1×10^{-3} (water), 1.8×10^{-5}(air)
ν	kinematic viscosity (=μ/ρ); unit, m^2 s^{-1}; dimension, L^2 T^{-1}; value, 1×10^{-6} (water),1.5×10^{-5} (air)
ρ	fluid density; unit, kg m^{-3}; dimension, M L^{-3}; value, 1000 (water), 1.2 (air)
ρ_s	sediment density; unit, kg m^{-3}; dimension, M L^{-3}; value, 2650 (water and air)
τ	shear stress; unit, Pa (= N m^{-2}); dimension, M L^{-1} T^{-2}
C	concentration (by mass/volume); unit, kg m^{-3}; dimension, M L^{-3}
D	flow depth; unit, m; dimension, L
d	grain size; unit, m, mm, μm; dimension, L
g	acceleration due to gravity; unit, m s^{-2}; dimension, LT^{-2}; value, 9.8 (air)
J	collision probability; unit, m^{-3} s^{-1}; dimension, L^{-3} T^{-1}
N	concentration (by number/volume); unit, m^{-3}; dimension, L^{-3}
U_*	shear velocity (=$\surd(\tau/\rho)$); unit, m s^{-1}; dimension, LT^{-1}
u,v,w	velocity components (w, vertical); unit, m s^{-1}; dimension, LT^{-1}
x,y,z	space coordinates; unit, m; dimension, L

Subscripts

0	at the bed, or at $t = 0$, e.g., τ_0, U_0, N_0
50	50 percentile value, median (e.g., d_{50})
B	Brownian
b	near bed (e.g., C_b)
c	critical value (e.g., τ_c for critical erosion stress)
d	deposition (e.g., τ_d)
f	floc
i,j	i and j particles
p	particle
S	shear
s	sediment (e.g., ρ_s)
v	viscous sublayer (e.g., δ_v)
z	at height z above the bed (e.g., U_z)

Superscripts

— above	time-averaged value parameter

Constants and Dimensionless Numbers

ζ	Rouse number; $w_s/\kappa U_*$
κ	von Karman's constant; 0.4
Ξ	Yalin's number; $[(\rho_s-\rho)gd^3]/(\rho\nu^2)$
k	Boltzman's constant; 1.38×10^{-23} J°K^{-1}
Re	Reynolds number; U_*d/ν, w_sd/ν, etc.

Equations

$\delta_v = 10\nu/U_*$	Viscous sublayer thickness
$w_s = \Delta\rho g d^2/18\mu$	Stokes' settling speed

See Also

Sedimentary Environments: Contourites; Storms and Storm Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures; Aeolian Processes; Deep Water Processes and Deposits; Particle-Driven Subaqueous Gravity Processes. **Unidirectional Aqueous Flow**.

Further Reading

Allen JRL (1985) *Principles of Physical Sedimentology*, ch. 6 and 7. London: Allen & Unwin.

Dade WB, Hogg AJ, and Boudreau BP (2001) Physics of flow above the sediment–water interface. In: Boudreau BP and Jorgensen BB (eds.) *The Benthic Boundary Layer: Transport Processes and Biogeochemistry*, pp. 4–43. New York: Oxford University Press.

Dietrich WE (1982) Settling velocity of natural particles. *Water Resources Research* 18: 1615–1626.

Friedlander SK (1977) *Smoke, Dust and Haze*. New York: Wiley-Interscience.

Heezen BC and Hollister CD (1971) *The Face of the Deep*, ch. 9. New York: Oxford University Press.

Hill PS and McCave IN (2001) Suspended particle transport in benthic boundary layers. In: Boudreau BP and

Jorgensen BB (eds.) *The Benthic Boundary Layer: Transport Processes and Biogeochemistry*, pp. 78–103. New York: Oxford University Press.

Leeder MR (1999) *Sedimentology and Sedimentary Basins*, ch. 4–6. Oxford: Blackwell Science.

McCave IN (1984) Erosion, transport and deposition of fine grained marine sediments. In: Stow DAV and Piper DJW (eds.) *Fine-Grained Sediments: Deep Sea Processes and Facies, Special Publication of the Geological Society of London 15*, pp. 35–69. London: Geological Society.

McCave IN (2001) Nepheloid layers. In: Steele JH, Thorpe SA, and Turekian KK (eds.) *Encyclopaedia of Ocean Sciences*, vol. 4, pp. 1861–1870. London: Academic Press.

Mehta AJ (ed.) (1993) *Nearshore and Estuarine Cohesive Sediment Transport. Coastal and Estuarine Studies*, vol. 42. Washington DC: American Geophysical Union.

Miller MC, McCave IN, and Komar PD (1977) Threshold of sediment motion under unidirectional currents. *Sedimentology* 24: 507–527.

Pye K (1987) *Aeolian Dust and Dust Deposits*. London: Academic Press.

Fluxes and Budgets

L Frostick, University of Hull, Hull, UK

Introduction

The word ‘flux’ when applied to sediments has come to mean the movement of particles of rock from upland areas down to a receiving basin, the fundamental processes of landscape evolution, and the geological cycle (Figure 1). The basin which receives the sediment can be terrestrial, either a lake, inland sea, or valley, but the most important basins are the seas and oceans which eventually claim more than 99% of the sediment produced on land. Understanding how, where, and why sediment moves into these basins is fundamental to interpreting and predicting the way in which a basin was formed or might evolve in the future and is, therefore, central to the economically important discipline of basin analysis.

Controls on Sediment Fluxes

The rate of sediment flux to basins is governed by a complex series of interactions amongst the physical and chemical processes which bring about rock uplift, weathering, erosion, and transportation (*see* **Sedimentary Rocks:** Mineralogy and Classification). The quantities of sediment that arrive at a receiving basin are the product of these processes integrated across the total area of supply. As the contributing processes can and do alter in both space and time, sediment fluxes also vary at a range of scales in response to geological and climatic changes. Understanding the controlling processes is, therefore, central to predicting fluxes which are, themselves, key to developing accurate models of basin development.

Weathering

The reason why rock fragments are worn away is inextricably linked with the tectonic processes which produce uplift. As rocks rise up to form hills or mountains and successive surface layers are stripped off, the minerals contained in them move from a dry, hot, high pressure environment to one with an abundant water supply from rainfall and where the temperature and pressure are relatively low (*see* **Sedimentary Environments:** Depositional Systems and Facies). Under these new conditions many minerals become unstable and begin to break down. This process is called weathering and it causes what were originally solid rocks to fragment into smaller fragments which are more easily moved. It is these particles, along with the dissolved products of weathering which, once transported and deposited, make up all sediments and sedimentary rocks (*see* **Weathering**). The way in which a rock breaks down is directly related to its composition. Rock forming minerals are stable at different temperatures and pressures and those that form under conditions most unlike those at the Earth’s surface are the most unstable. Quartz is the rock forming mineral that is most stable during weathering and this is the reason why quartz is the predominant mineral in present day beach and river sands and is also common in most ancient sandstones. The rate at which weathering occurs depends on climate, with rapid breakdown favoured by the high rainfall and temperatures of tropical areas and slow weathering occurring where water is absent or solid, i.e., in deserts and arctic zones (*see* **Sedimentary Environments:** Deserts).

Transport

The breakdown products of weathering are transported away from their site of formation and downslope in several ways. On steeper slopes, and for the

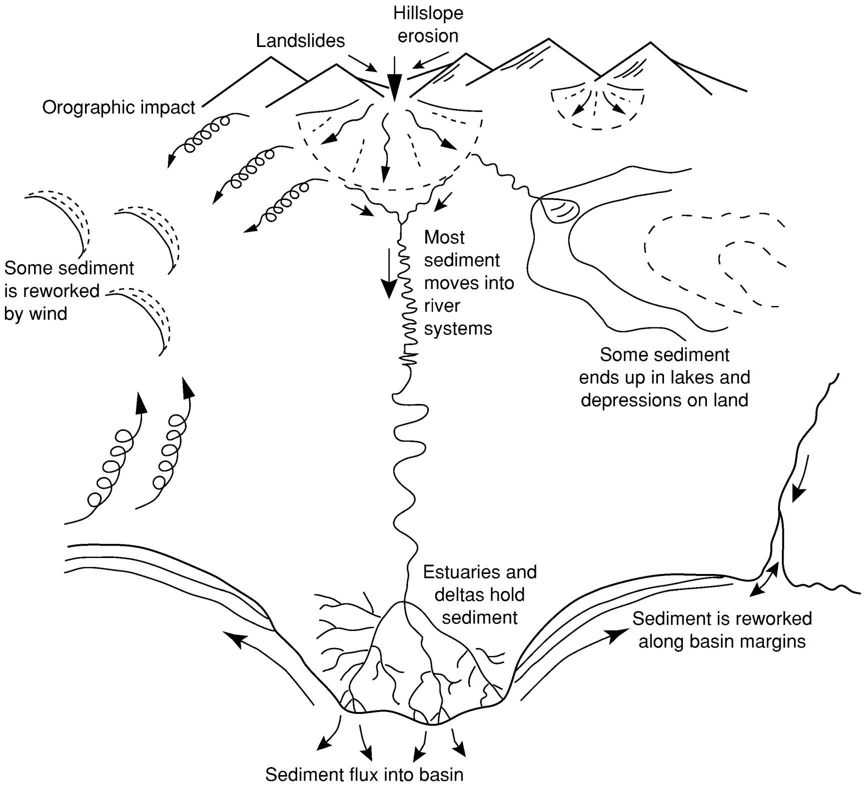

Hillslope erosion
Landslides
Orographic impact
Some sediment is reworked by wind
Most sediment moves into river systems
Some sediment ends up in lakes and depressions on land
Estuaries and deltas hold sediment
Sediment is reworked along basin margins
Sediment flux into basin

Figure 1 Diagrammatic representation of the way in which sediment moves through the landscape from the mountains to the ocean basins. (Adapted from Frostick LE and Jones SJ (2002) Impact of periodicity on sediment flux in alluvial systems: grain to basin scale. *Geological Society of London Special Publication* 191: 81–95.)

larger particles this may occur as a result of gravity alone acting on the particles, which cause rolling, sliding, and avalanching. Such processes are episodic and often linked with the development of instability as rainfall wets the slope (Figure 2). Some material may be removed by wind but this is important only in desert areas and steep rocky slopes which are devoid of the protection offered by vegetation. In all areas with flowing water it is the movement of that medium that induces sediment to move down the slopes and into adjacent valleys. The water sometimes flows as a shallow sheet over wide areas (called overland flow) but more often collects into small channels, generally known as rills. From here the sediment moves into streams and rivers to complete its journey to the receiving basin. Over the majority of the continental landmasses, where temperatures are high enough for water to remain liquid for most of the time, rivers are the main transporting medium for sediment and flux rates, therefore, vary with the character of the river network. Large, long-lived and integrated networks, such as those of the Mississippi, Congo, and Amazon rivers, deliver large volumes of water and quantities of sediment over long periods. The Mississippi, for example, brings to the Gulf of Mexico one tonne of sediment for every 400 tonnes of water during periods of flooding. When the river is flowing less fast, this may diminish to 0.35 tonne. At the present time, nearly 70% of the total sediment supplied by rivers to marine basins comes from five large rivers, the Ganges/Brahmaputra, Amazon, Huang Ho, Irrawaddy, and Mississippi (Table 1).

In arctic and subarctic areas and where mountains are sufficiently high to generate significant quantities of snowfall, glaciers become a major agent for sediment transport. In the past, shifts in the climatic balance have led to major glacial events when ice-sheets spread out from the polar caps and engulfed large areas of previously temperate landmasses. During these periods, soil and other surface deposits are scoured from the landscape and accumulate in basins.

Climate and Tectonism

Both the character and quantity of sediment carried by a river will reflect the climate and geology of the drainage basin. Even in adjacent river basins, differences in these factors can lead to huge contrasts in sediment fluxes. One example of this can be seen

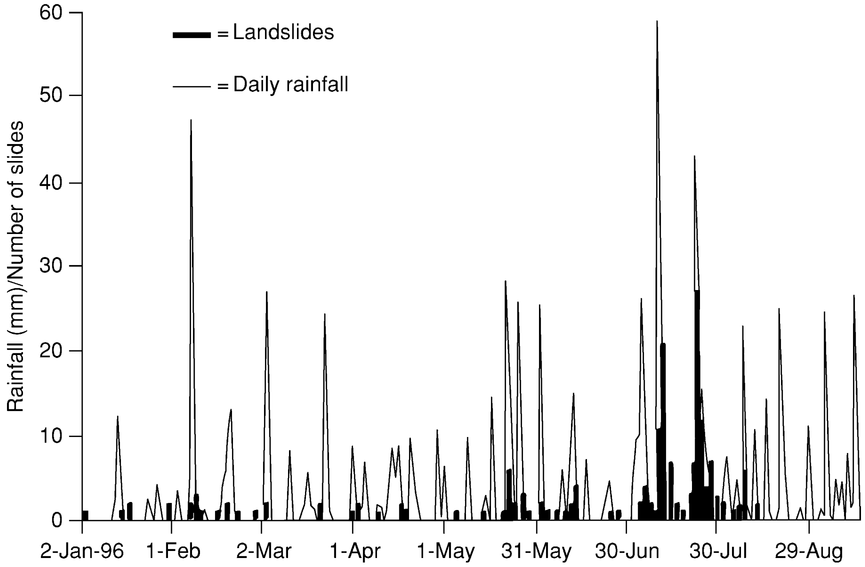


Figure 2 Diagram of the relationship between avalanche frequency and the occurrence of rainfall in the mountains near Wellington, New Zealand. The process is very episodic and each event will produce a 'slug' of sediment which will move down adjacent streams. (Adapted from Crozier MJ (1999) Prediction of rainfall-triggered landslides: a test of the antecedent water status model. *Earth Surface Processes and Landforms* 24: 825–833.)

Table 1 The 24 rivers of the world with the largest annual sediment fluxes. Major Rivers Ranked by Sediment Load

Ranking	*River*	*Mean water discharge,* $10^3\,m^3\,s^{-1}$	*Mean sediment load,* $10^6\,ta^{-1}$
1	Ganges/Brahmaputra	31	1821
2	Amazon	200	1190
3	Huang Ho	1	922
4	Irrawaddy	13.5	356
5	Mississippi	18	352
6	Magdalena	7	240
7	Mekong	21	219
8	Orinoco	36	181
9	Indus	7.5	179
10	Mackenzie	8	144
11	Danube	6.5	136
12	Paraná	15	118
13	Rhone	2	96
14	Yukon	7	94
15	Congo	41	85
16	Volga	8	81
17	Yenisei	17.5	75
18	St Lawrence	13	75
19	Lena	160	68
20	Ob	14	59
21	Zambezi	2.5	45
22	Niger	5	39
23	Murray-Darling	1	38
24	Columbia	6	36

in Brazil, where the sediment laden Rio Solimoes meets the sediment poor Rio Negro to form the Amazon River in a spectacular river confluence where the contrast is so marked that it can be seen from space (Figure 3). Sediment fluxes are at a maximum where tropical weathering results in rapid release of particles, large rivers are generated by high rainfall, and tectonic activity promotes uplift and steep slopes. One such area is the Himalayas where the monsoon rains sweep sediment into the vast Indus, Ganges, and Bramaputra rivers. The courses of these rivers and the locations of their outlets into the Indian Ocean are also controlled by tectonic activity, areas of uplift shedding water into adjacent lowland areas. At the mouths of these rivers much of the sediment load is deposited and deltas may build out into the adjacent basin. Large accumulations of sediment on the continental shelf can also lead to the development of turbidity currents which sweep material offshore into large submarine fans (e.g., the Bengal fan) (*see* **Sedimentary Environments:** Shoreline and Shoreface Deposits). Such localised deposits are characteristic of all areas where rivers debouch into larger water bodies and their morphology and sedimentology are largely controlled by the nature of basin processes actively redistributing the sediment.

Basin Processes

The character of the basin receiving the material removed from the land surface will control its dispersion. The balance between subsidence rates and sediment supply is particularly important, since if a deposit is buried rapidly it is less likely to be removed and reworked by waves and currents. In some areas sediments form large river deltas, e.g., the Mississippi (Figure 4), in others the deposit is submerged and

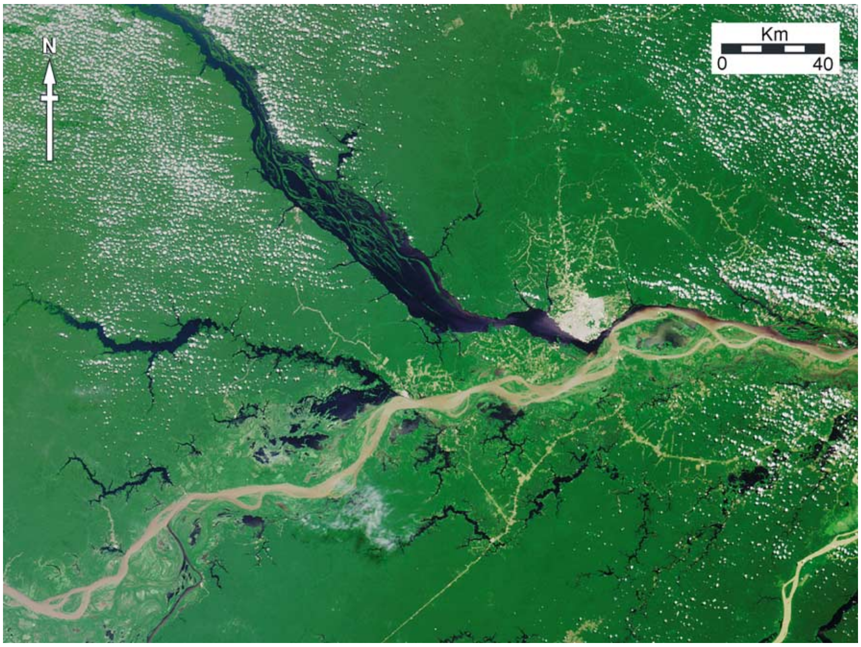


Figure 3 Satellite remote sensing image of the confluence between the sediment laden Rio Solimoes (showing brown) and the clear waters of the Rio Negro (showing black) to form the Amazon River. The Solimoes crosses soft glacial silt and sand, whereas the Negro passes through old hard rocks resistant to weathering. The two water bodies remain essentially separate for tens of kilometres downstream of the confluence. (Picture taken from the TERRA satellite using the multi angle imaging spectroradiometer on 23/7/2000.)

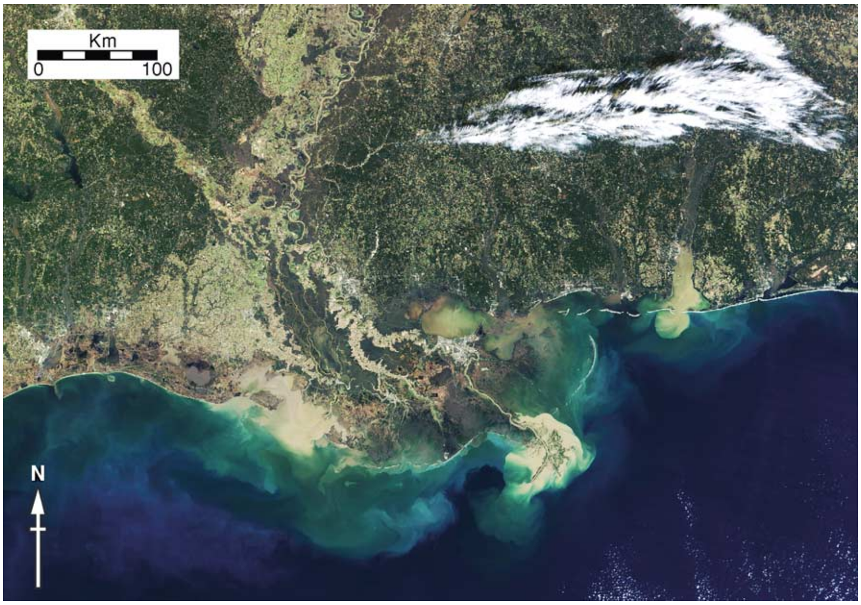


Figure 4 Satellite remote sensing image of the Mississippi Delta on the south-east coast of the USA, showing a sediment plume issuing from the mouth of the main channel. This sediment will settle out in the adjacent ocean basin. (Picture taken from the TERRA satellite using the moderate resolution imaging spectroradiometer.)

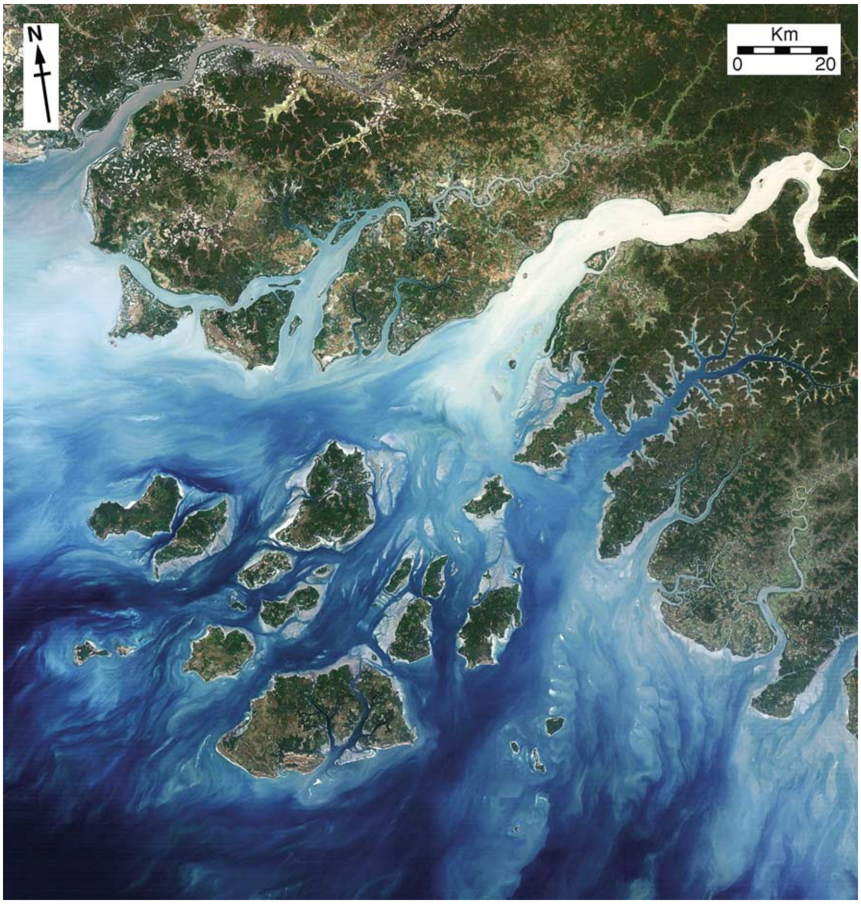


Figure 5 Satellite remote sensing image of tidal sand-banks off the coast of Guinea in West Africa. Note the plumes of sediment moving along the channels between the banks. The high tidal velocities are reworking sediment supplied by local rivers. (Picture taken from LANDSAT-4TM using bands 3, 2, and 1 for red green and blue colour channels, respectively.)

forms either a single fan or breaks up into a series of sand-banks (see, e.g., the coast of Guinea, West Africa, shown in Figure 5). As with the supply of sediment, the intensity of basin processes, such as tides, waves, and currents will also vary over distances, giving a complex spectrum of interactions that can lead to a plethora of deposit types.

Wind Blown Sediment

The flux of sediment in wind may be a minor factor in most areas today, except from in our major deserts, e.g., the Sahara, but in the past it has been significant and has led to the accumulation of extensive loess deposits, e.g., in China. Dust loadings in the atmosphere have changed in response to climate change, both on the scale of millennia, e.g., during the last Ice Age (Late Glacial Maximum) and at a decadal scale, e.g., in response to natural oscillations in climate associated with changes in ocean currents (e.g., the North Atlantic Oscillation). At present, most dust is derived from the worlds desert areas where reduced vegetation allows high-speed winds to pick up predominantly silt-sized particles and transport them for thousands of kilometres before finally depositing them downwind. The Sahara is the world's major source of wind-blown dust, producing between 400 and 700×10^6 tons.a^{-1}, approximately half the total amount of wind blown or aeolian dust estimated as being supplied to all the worlds oceans (*see* **Sedimentary Processes:** Aeolian Processes). At times of major dust storms, the dust being blown out of the Sahara can travel long distances and penetrate far out into the Atlantic Ocean. The composition of this dust is important as it adds nutrients to the deep oceans and influences their productivity (*see* **Sedimentary Processes:** Deep Water Processes and Deposits). Although quartz is the dominant mineralogy, dusts may contain compounds with appreciable proportions of aluminium, iron, magnesium, and calcium.

Flux Variations Over Time

Sediment fluxes at a point will vary over both short and long time scales as a result of temporal changes in any of the controlling variables. At the longer time-scales of geology, both climate change and tectonic uplift can bring about large-scale changes in flux rates. One example of the impact of climate change is the large-scale fluctuations in sediment movement during glacial and interglacial periods of the Quaternary, approximately 1.8 million years before present. Evidence of the fluctuations is found in the preserved deposits of this time, particularly the oceanic deposits, which received rock fragments ranging in size from flour to the size of a house carried by ice as it ground its way in glaciers across the barren landscape.

Periodicity in tectonic uplift has also been linked to major shifts in sediment fluxes. In the Himalayas, for example, there have been four major periods of uplift over the past 12 million years, each linked with a higher rate of sediment accumulation in adjacent basin areas (Figure 6). Local changes in surface elevation, as a result of tectonic activity, will impact on both where a river flows and how fast it flows. Higher flow speeds associated with steeper slopes will allow a river to pick up more sediment from its bed, thereby causing it to cut down and increasing fluxes. When slopes get less steep, less sediment can be carried by lower energy flows and material accumulates in the river valleys.

At shorter time-scales, sediment fluxes change in response to fluctuations in water discharge (Figure 7), the majority of the material being carried during flood periods when the rivers are more energetic. Any changes which impact on the delivery of water to the river system, for example, changes in rainfall patterns surface vegetation and land use, can result in shifts in flood frequency and impact on sediment fluxes. One example is in the American midwest during the late nineteenth and early twentieth centuries where the removal of vegetation, as a result of intensifying agriculture, led to floods becoming more intense (*see* **Sedimentary Processes:** Catastrophic Floods). This, combined with the removal of the protection from erosion offered by plants, led to rapid erosion and the development of 'badlands'.

The Importance of Geology

The rock types within the river system will control the rate at which sediment can be produced and supplied to the river system. This can lead to the development of very different environments in otherwise identical basins. For example, in the African rift, lakes in areas where the rocks are mainly old, hard, and resistant to weathering are deep and sediment starved (e.g., Lake Tanganyika), whereas those in areas containing volcanic rocks and old sediments are sand- and mud-rich and generally more shallow (e.g., Lake Baringo). The same factors are important in controlling the way in which sediments filled up ancient basins. For

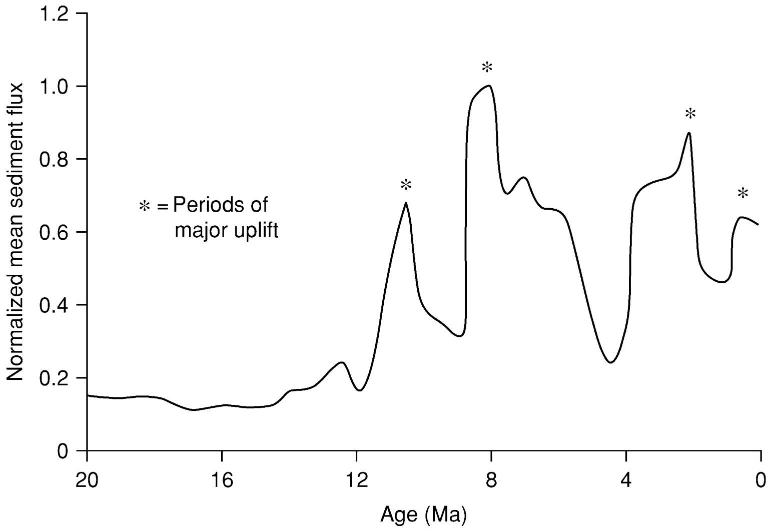


Figure 6 Variations in sediment fluxes from the Himalayas. High fluxes relate to periods of active uplift. (Adapted from Hovan SA and Rea DK (1992) The Cenozoic record of continental mineral deposition on Broken and Ninetyeast ridges, Indian Ocean: southern African aridity and sediment delivery from the Himalayas. *Paleoceanography* 7: 833–860.)

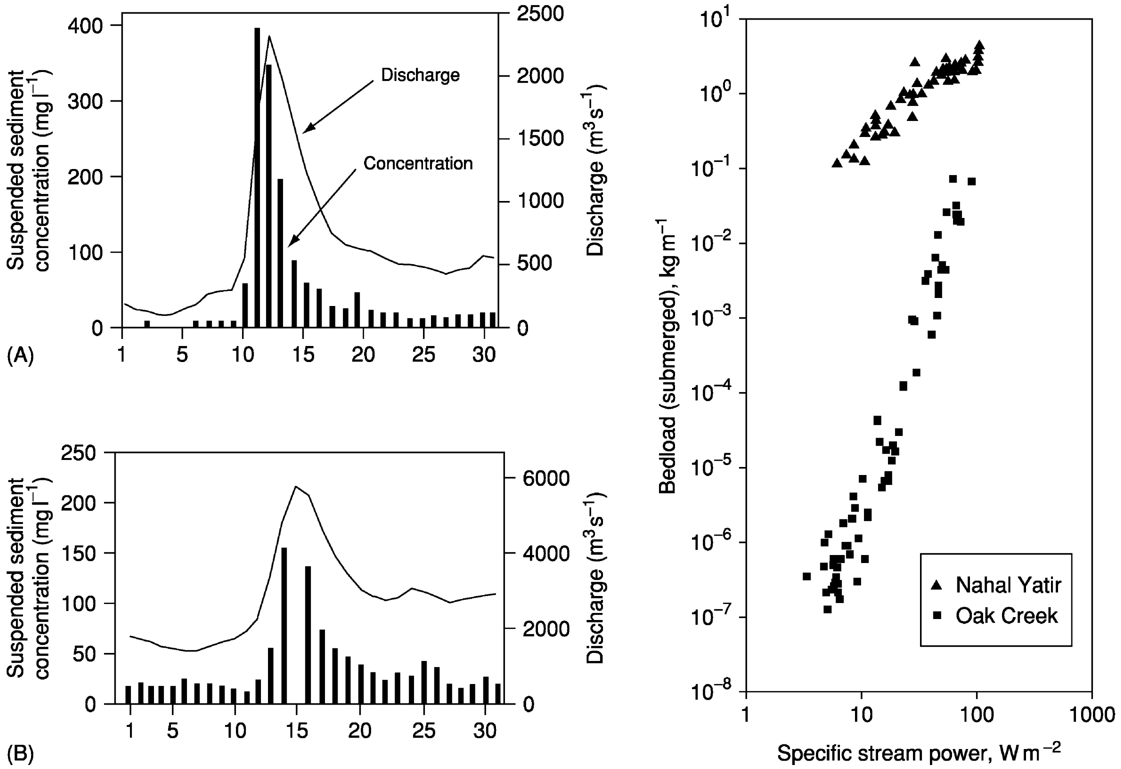


Figure 7 Variations in suspended sediment transport for the Meuse and the Rhine Rivers (left) showing how sediment flux varies over flood events. (Adapted from Asselman NEM and Middlekoop H (1998) Temporal variability of contemporary floodplain sedimentation in the Rhine-Meuse delta, The Netherlands. *Earth Surface Processes and Landforms* 23: 595–609.) and bed material transport for Nahal Yatir, Israel and Oak Creek, USA (right) showing the difference between sediment fluxes in ephemeral (Nahal Yatir) and perennial (Oak Creek) rivers. (Adapted from Laronne JB and Reid I (1993) Very high rates of bedload sediment transport by ephemeral desert rivers. *Nature* 366: 148–150.) plotted against water discharge over flood periods. Fluxes generally increase with increasing water discharge, reaching maxima at flood peaks.

example, the Cretaceous to Tertiary Tucano and Reconcavo basins of Brazil are similar in size and structure but whereas the Reconcavo has a low sediment flux and is oil retaining the Tucano Basin contains interconnected sands that allow any oil generated to escape (*see* **Petroleum Geology:** The Petroleum System). When a geological map of the area is consulted, it is evident that the Reconcavo Basin was formed on ancient and hard gneisses which could only be broken down very slowly, whereas the Tucano Basin was cut into soft sedimentary rocks. This example highlights the economic importance of understanding the controls on sediment flux.

Sediment Budgets: Modelling the Past and Predicting the Future

The complexity of controls on sediment fluxes conspires with changes in rates that occur at all timescales to make the interpretation of the past and predictions of the future equally problematic. Added to this, the role of river drainage basin in storing and releasing sediment, thereby modulating the impact of environmental change on sediment flux, is still poorly understood. Even the direct measurement of present-day fluxes is fraught with difficulties since the costly nature of making such measurements usually means that data are collected only from a small number of points and for very short periods of time. As rivers are the main means of sediment delivery to basins, many of the flux calculations that exist are based on short periods of measuring water and sediment discharges which are then used to extrapolate to larger areas and longer periods, ignoring the limitations of the data. This makes the calculation of sediment budgets for any area very difficult; even under ideal circumstances with good data they generally only give a crude approximation of what is really happening.

Estimations of sediment fluxes and the understanding of sediment budgets are central to the science of basin analysis. Basin analysts endeavour to determine the controls on where and how sediments are deposited in a basin and they provide essential information to those wishing to exploit natural mineral resources, for example, oil, gas, and diamonds. In the case of oil and gas, it is important to locate sands and other permeable and porous rocks that might act as oil and gas reservoirs (*see* **Petroleum Geology:** The Petroleum System). In addition, identifying parts of the basin and periods of time when sediment fluxes are low and the organic debris from which oil is formed can accumulate undiluted by other sediments, is a vital part of exploration. For placer deposits such as diamonds, rivers are the main transport system and they deliver these important heavy minerals to the basin for reworking into economic beach deposits, e.g., in South Africa diamonds are delivered by the Orange River to the coast of Namibia.

See Also

Geomorphology. **Petroleum Geology:** The Petroleum System. **Sedimentary Environments:** Depositional Systems and Facies; Deserts; Shoreline and Shoreface Deposits. **Sedimentary Processes:** Aeolian Processes; Catastrophic Floods; Deep Water Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification. **Weathering**.

Further Reading

Allen PA and Allen JR (1990) *Basin Analysis Principles and Applications*. Oxford: Blackwells.

Burbank D, Leland J, Fielding E, *et al.* (1996) Bedrock incision, rock uplift and threshold hill slopes in the northwestern Himalayas. *Nature* 379: 505–510.

Descroix L and Mathys N (2003) Processes, spatio-temporal factors and measurements of current erosion in the French southern Alps: a review. *Earth Surface Processes and Landforms* 28: 993–1011.

Hovius N (1996) Regular spacing of drainage outlets from linear mountain belts. *Basin Research* 8: 29–44.

Hovius N and Leeder MR (1998) Clastic sediment supply to basins. *Basin Research* 10: 1–5.

Jansson MB (1988) A global survey of sediment yields. *Geografiska Annaler* 70: 81–98.

Jones SJ and Frostick LE (eds.) (2002) Sediment Flux to Basins: Causes, controls and consequences. *Geological Society Special Publication* 191: 284.

Macklin MG (1999) Holocene river environments in prehistoric Britain: human interaction and impact. *Journal of Quaternary Science* 14: 521–530.

Middleton NJ and Goudie AS (2001) Saharan dust: sources and trajectories. *Transactions of the Institute of British Geographers* 26: 165–181.

Milliman JD and Meade RH (1983) Worldwide delivery of river sediments to the ocean. *Journal of Geology* 91: 1–21.

Milliman JD and Syvitski JPM (1992) Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *Journal of Geology* 100: 525–544.

Shanley KW and McCabe PJ (eds.) (1998) Relative role of Eustacy, climate and tectonics in continental rocks. *Society of Economic Palaeontologists and Mineralogists Special Publication, 59*.

Studies in Geophysics (1994) *Material Flux on the Surface of the Earth*. Washington, USA: National Academy Press.

Syvitski JP, Morehead MD, and Nicholson M (1998) HYDROTREND: a climatically-driven hydrologic transport model for predicting discharge and transport load to lakes and oceans. *Computers and Geosciences* 24: 51–68.

Tipper JC (2000) Patterns of stratigraphic cyclicity. *Journal of Sedimentary Research* 70: 1262–1279.

Tucker GE and Slingerland R (1996) Predicting sediment flux from fold and thrust belts. *Basin Research* 8: 329–350.

SEDIMENTARY ROCKS

Contents

Mineralogy and Classification

R C Selley, Imperial College London, London, UK

Introduction

Rocks and Their Classification defined the three main classes of rock: igneous, metamorphic, and sedimentary. It described the main features by which the three types of rock may be distinguished, and presented anomalous examples of each. The classification of igneous and metamorphic rocks is described in **Igneous Processes** and **Metamorphic Rocks:** Classification, Nomenclature and Formation, respectively. This article describes the mineralogy and classification of sedimentary rocks.

Sedimentary rocks are formed from the detritus of pre-existing rocks: igneous, metamorphic, or sedimentary. The way in which rock is weathered, eroded, transported, and deposited is discussed in detail elsewhere (*see* **Weathering, Sedimentary Processes:** Fluxes and Budgets, **Unidirectional Aqueous Flow**, and **Sedimentary Environments:** Depositional Systems and Facies). Sediments possess a wide range of particle size, ranging from boulders to clay, and of chemical composition, including silica, lime, or ferromagnesian volcanic detritus. These parameters of particle size and composition are used to classify sedimentary rocks. Sedimentary rocks commonly exhibit two properties that may be used to differentiate them from igneous and metamorphic rocks.

1. Where they crop out at the surface of the Earth, sedimentary rocks generally show stratification (layering). The strata indicate successive episodes of deposition. Layering is usually absent from igneous rocks, but is found in some metamorphic rocks.
2. When examined under the microscope, sedimentary rocks are generally seen to consist of particles. Void space (porosity) is commonly present between the constituent grains. Interconnected pores give the rock permeability. Permeability allows fluids to migrate through rock, and enables rock and soil to drain. Additionally, fossils are only found in sedimentary rocks, some of which are, indeed, made up of nothing else.

Mineralogical Basis for Sedimentary Rock Classification

In the earliest classifications of rock, such as that proposed by Charles Lyell (*see* **Famous Geologists:** Lyell), four classes were recognized: volcanic, plutonic (these two are now grouped as igneous), metamorphic, and aqueous. The aqueous rocks were subdivided into three groups: arenaceous, argillaceous, and calcareous. The term 'aqueous' has long been abandoned, as it is now known that sediments are deposited by aeolian, gravitational, and glacial processes, as well as by purely aqueous ones.

The processes of weathering, erosion, transportation, and deposition are nature's way of chemically fractionating the Earth's surface, and lead to a logical classification of sedimentary rocks based largely on their chemistry and mineralogy. Serendipitously, this fractionation correlates broadly with their mode of formation. Although geologists broadly agree about the definition of the main classes of sedimentary rocks, there is no unanimity about all of them.

In 1937, Goldschmidt proposed a classification based on five chemical groups, namely: (1) resistates, (2) hydrolysates, (3) oxidates, (4) carbonate precipitates, and (5) evaporites. In 1950, Rankama and Sahama added a sixth: reduzates. The resulting classification highlights the chemical fractionation that results from sedimentary processes, but produces some very strange and uncouth names.

Meanwhile, a practical field-based classification had been widely adopted, although with some variation in the fine detail. It had long been noted that fractionation on the surface of the earth naturally divided rocks into those that had never gone into solution, and might therefore be termed 'allochthonous' or 'detrital', and those formed from minerals that had been dissolved in surface water, and had precipitated out. These are termed 'autochthonous' or 'chemical' rocks. The allochthonous or detrital sediments are subdivided by grain size. The autochthonous or chemical sediments are subdivided by mineralogical (chemical) composition (Table 1).

Geopedants will already notice the inconsistency of this classification. The detrital sediments are composed of a wide range of minerals, and thus exhibit a diversity of chemistry, which is ignored for the purposes of their classification. Similarly, the chemical rocks may occur in a wide range of particle size, from boulders of limestone to sapropelic muds. Nonetheless, the classification displayed in Table 1 is not for the benefit of geopedants, but for practical use by geologists. There is no consensus on the classification of sedimentary rocks proposed in Table 1. Note that the caption reads 'A classification of sedimentary rocks' not 'The classification of sedimentary rocks'.

The main groups of sedimentary rocks are now described briefly, pointing the way to articles in this encyclopedia that describe them in more detail.

Table 1 A classification of sedimentary rocks

Allochthonous or detrital sediments
Classified by grain size
Gravel/conglomerate
Sand/sandstone
Silt/siltstone
Clay/claystone (sometimes also termed 'mudrocks' or 'shales')
Autochthonous or chemical sediments
Classified by mineralogy
Carbonates (limestone and dolomite)
Evaporites (gypsum/anhydrite, halite, etc.)
Residual (bauxite, laterite, kaolinite)
Kerogenous (peat, lignite, coal)
Ironstones (haematitic, chamositic, and sideritic)
Phosphates (guano)
Siliceous (chert, opal)

Allochthonous or Detrital Sediments

As defined earlier, the allochthonous or detrital sediments are the insoluble residue of weathering of pre-existing rocks: igneous, metamorphic, or sedimentary. The mineralogy is very varied, depending on the source material and the type and duration of the weathering process (*see* **Weathering**). The mineralogy also correlates crudely with the grain size. Conglomerates tend to be polymineralic, sandstones are dominated by quartz, and mudrocks are dominated by clay minerals. Table 1 shows the subdivision of the allochthonous or detrital rocks by grain size, gravel, sand, silt, and clay being the basis for conglomerate, sandstone, siltstone, and mudstone, respectively. The terms 'rudaceous', 'arenaceous', and 'argillaceous' have also been applied to conglomerates, sandstones, and shales, but are little used now. The main groups of detrital sediments are now described briefly.

Conglomerate

Conglomerate is composed of particles of gravel, that is to say of particles of greater than 2 mm in diameter, consisting, with increasing size, of granules, pebbles, cobbles, and boulders. Collectively, conglomerates have also been known as rudaceous rocks. Conglomerates are distinguished from breccias by the fact that the clasts are rounded, whilst those of breccias are angular. Because of their large size, conglomerate clasts are composed of many grains or crystals (depending on whether they were derived from earlier sediments or from crystalline igneous or metamorphic rocks). They may thus be composed of a wide range of minerals. When derived from igneous or metamorphic rocks, conglomerates may be composed of the wide range of minerals found in the parent rock. By contrast, conglomerates derived from sediments will reflect their source mineralogy, but will tend to be composed of a higher percentage of minerals that are stable at the Earth's surface, rather than in the parent rock. The concept of sediments as the insoluble residue of pre-existing rocks is again a useful one to recall. This is illustrated by the vast volumes of flint (chert) gravels to be found on the beaches of north-west Europe. These flints originate in rare horizons in the Cretaceous Chalk. As Charles Lyell wrote: “The entire mass of stratified deposits in the Earth’s crust is at once the monument and measure of the denudation which has taken place.” Gravels and conglomerates are described in greater detail in **Sedimentary Rocks:** Rudaceous Rocks.

Sandstones

Sandstones are composed of particles with an average size of between 2.00 and 0.0625 mm in diameter. They have four constituents: grains, matrix, cement, and, sometimes, porosity (Figure 1). Sand-sized particles form the framework of the rock. Matrix, the finer grained material that may infill space between the framework grains, was deposited at the same time as the framework grains. Cement is the term that describes minerals precipitated in pores after the deposition of the sediment. Thus, matrix is syndepositional and cement is postdepositional. Cement and porosity are described in **Sedimentary Rocks:** Sandstones, Diagenesis and Porosity Evolution, which deals with the diagenesis of petroleum reservoirs. Framework grains and matrix are described below.

The framework grains of sandstones are normally composed of varying amounts of the mineral quartz (silica, SiO_2). In order of decreasing abundance, sandstones also contain feldspar (a suite of calcium, potassium, and sodium silicates), micas (sheet silicates, with varying amounts of iron, magnesium, and aluminium), a complex of ferromagnesian minerals, informally termed ‘mafics’, and heavy minerals (those with densities significantly greater than that of quartz (2.65 g/cc), examples of which include iron ores, mica, siderite, zircon, and apatite). Sandstone may also contain sand-sized grains composed of more than one mineral or crystal. These are termed ‘rock fragments’ or ‘lithic grains’. Sand-sized rounded green grains of the complex mineral glauconite are a common constituent of shallow marine sands (glauconite is described in **Minerals:** Glauconites). Sandstones often contain fossil fragments. Teeth, fish scales, and bone are largely phosphatic. The most common fossils, however, are shells, mainly composed of lime, calcium carbonate ($CaCO_3$). With increasing lime content, sandstones grade into calcareous sandstones, then to sandy limestones, and finally to pure limestone, composed entirely of calcium carbonate, and with negligible quartz. Thus, although typically composed of quartz, sandstones also contain a range of other minerals. These are used as a basis for naming and classifying sands and are important because of their impact on geophysical well-log interpretation.

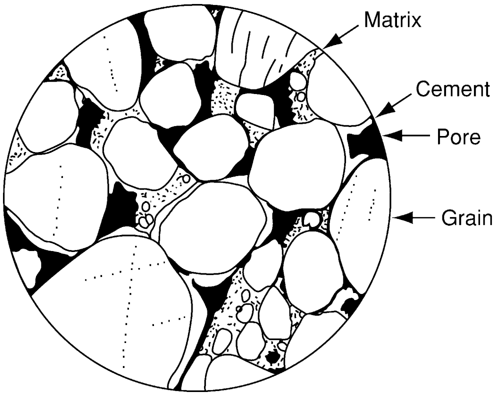


Figure 1 Diagram of a thin section of sandstone showing the four components of framework grains, matrix, cement, and pores. Only the first of these is always present. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

The syndepositional matrix that may occupy some of the space between the framework sand grains consists of silt, clay, and organic matter. Heavy mineral grains are commonly silt sized, and so technically they may form part of the matrix.

The composition of typical sandstone may be as follows. Framework grains: quartz, 45%; rock fragments, 5%; feldspar, 10%; mafics, 5%; mica, 5%; heavy minerals, 2%. Matrix: clay, 7%. Cement: calcite, 5%. Porosity, 16%. Total: 100%.

Two parameters are used to name and classify sandstones: chemical mineralogy and physical texture. When sediment is first eroded from its parent outcrop, it is generally immature in both its composition and texture. That is to say, it will still contain a range of chemically unstable mineral grains that surface processes have yet to break down and dissolve. Similarly, the debris first transported down a hillside will be very poorly sorted, consisting of a range of particles, varying in size from boulders to clay. When looking at an ancient lithified sandstone, its maturity may be described in terms of its chemical and physical properties (mineralogy and texture). Four main types of sandstone may thus be recognized as shown in Table 2. This table also employs four commonly used names to describe sandstones.

Table 2 A classification of sandstones based on textural and mineralogical maturity

	Mineralogical maturity	
	Immature	*Mature*
Texturally immature	Greywacke Figure 2	Quartz wacke Figure 3
Texturally mature	Arkose Figure 4	Quartzite Figure 5

'Greywackes' are poorly sorted sandstones with a large component of chemically unstable grains, not only feldspars, but also rock fragments and ferromagnesian minerals (Figure 2). 'Quartz wackes' are also texturally immature, but the framework grains are composed largely of quartz and lithic (rock) fragments (Figure 3). 'Arkoses' are texturally mature, but contain a large percentage of chemically unstable grains, principally feldspar (Figure 4). Quartzites, also termed quartz arenites (from the Greek '*arenos*' for sand), are texturally and mineralogically mature, being well sorted, and composed of little but quartz (Figure 5).

Mudrocks

There is little unanimity over the terminology for the argillaceous detrital sedimentary rocks. The easiest to name objectively are siltstone and claystone, being

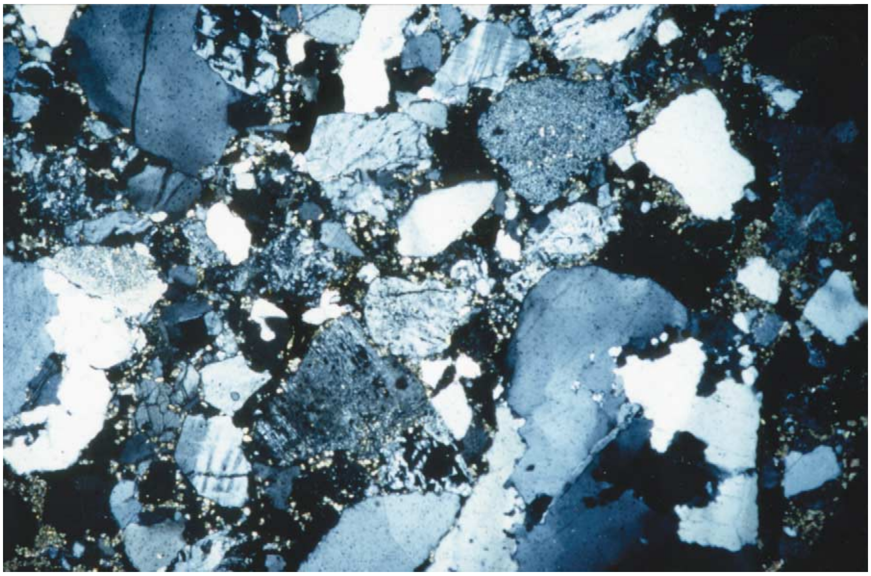


Figure 2 Photomicrograph of a greywacke under polarized light. Jurassic, UK North Sea. Note the poorly sorted texture and abundance of matrix and twinned feldspar. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

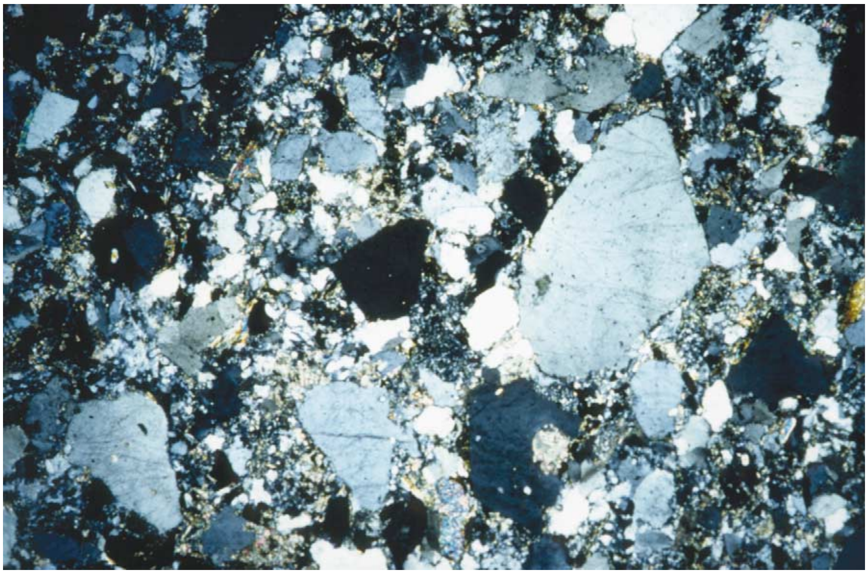


Figure 3 Photomicrograph of quartz wacke under polarized light. Carboniferous, Chios, Greece. Note the poorly sorted texture and abundance of matrix. The framework grains are almost entirely composed of quartz and chert. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

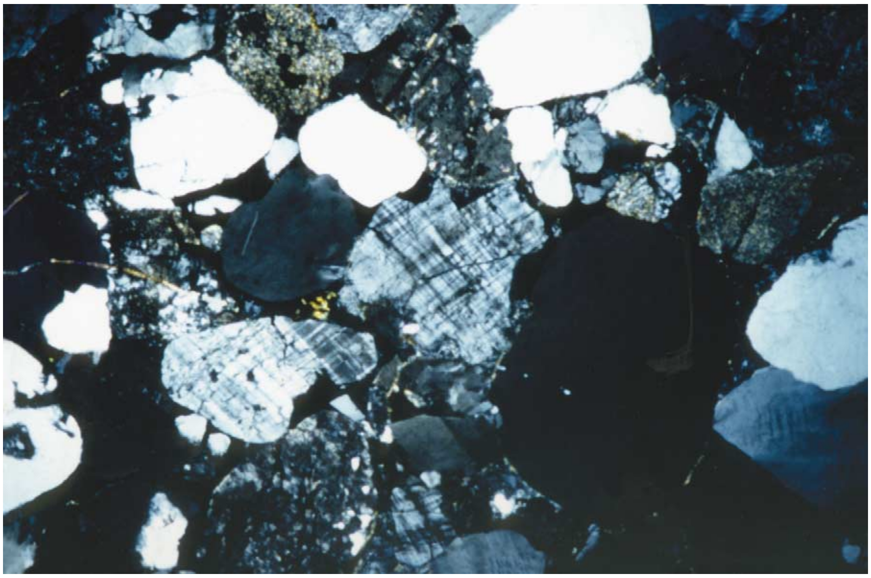


Figure 4 Photomicrograph of arkose under polarized light. Torridonian, Precambrian, Scotland. Note the abundance of twinned feldspar and the better sorted texture than in **Figures 2** and **3**. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

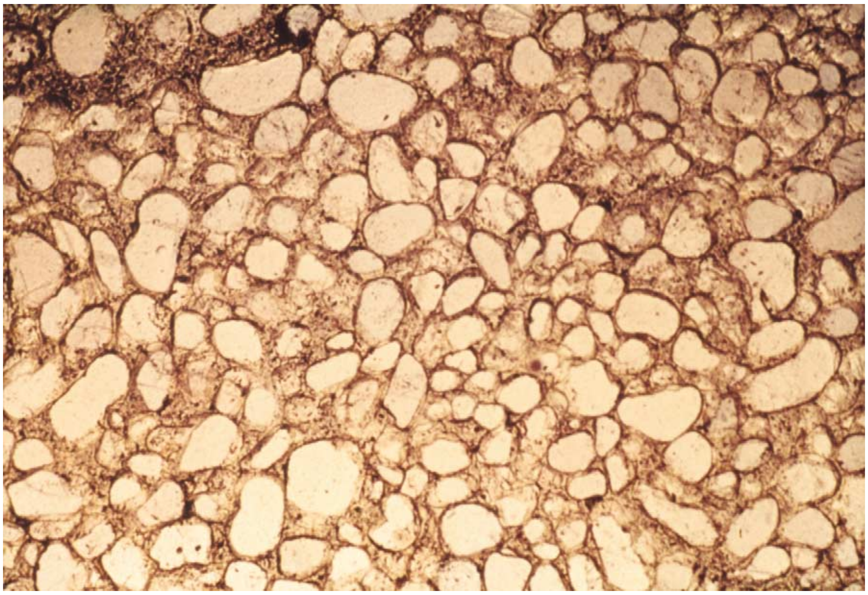


Figure 5 Photomicrograph of a quartz arenite under ordinary light. Simpson Group, Ordovician, Oklahoma, USA. Note the well-sorted texture. The framework grains are almost entirely composed of well-rounded and well-sorted quartz. There is neither matrix nor cement. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

composed, by definition, of predominantly silt or clay, respectively. They are easy to identify in the field. Siltstones have an unpleasant gritty feel between the teeth, clays a pleasing plastic texture. When sufficiently indurated, both siltstones and claystones may be termed ‘shales’. The term ‘shale’ was defined by Pettijohn as “a laminated or fissile claystone or siltstone”. This is a widely used term, and with good reason: siltstones are commonly argillaceous and claystones silty, and so the term shale covers all variations. The term ‘mudrock’ or ‘mudstone’ has been applied to siltstones and claystones (and their admixtures) that do not possess the fissility of shale. Siltstones are composed of detrital quartz, shell fragments, assorted heavy minerals, and, often, mica, which imparts the fissility to shale. Siltstones also commonly contain varying amounts of clay matrix and organic matter.

Claystones are composed largely of the clay minerals, kaolin, illite, montmorillonite, and chlorite. Clay mineralogy is described in detail in **Clay Minerals**. Lime and organic matter may also be present in claystones. With increasing lime content, shales grade into marls, argillaceous limestones, and limestones. With increasing organic content, shales grade into sapropelite. Detrital grains of silt, mica, plant debris, and shell fragments may occur in mudstones as impurities. Clays and their diagenesis are described in **Sedimentary Rocks:** Clays and Their Diagenesis.

Autochthonous or Chemical Sediments

It has already been noted that the classification of sediments into detrital and chemical categories is somewhat artificial. Both are composed of chemical or biochemical components, and of particles of varying sizes. Nonetheless, it is a convenient grouping. The chemical sediments are those that principally precipitate out of solution, although thereafter they may become detrital in some instances. Table 1 shows that seven types may be recognized: carbonates, evaporites, residual deposits, kerogen, ironstone, phosphate, and silica. These are now described briefly.

Carbonates

The carbonate chemical sediments include a wide range of rocks, of which limestone and then dolomite are volumetrically the most important, whilst siderite and magnesite, although rare, are economically important. The mineralogy of carbonates is described in detail in **Minerals:** Carbonates. Limestone is composed largely of the mineral calcium carbonate ($CaCO_3$). Limestones may be made up of many different types of grain that originate in different ways in a range of depositional environments (Figure 6). Thus grain type is one of the keys to interpreting their depositional environment. Limestones form, almost without exception, from the aqueous precipitation of calcium carbonate, aided by some organic process or other, most obviously as shells secreted by invertebrates, but also as nodules, laminae, and clouds whose origins owe much to biochemical reactions. Limestones are described in greater detail in **Sedimentary Rocks:** Limestones.

Dolomite is composed of the mineral dolomite, $CaMg(CO_3)_2$, named by and from the eponymous French Count Dolomieu (1750–1801). Geopedants restrict the term dolomite to the mineral, and dolostone to the rock. This was not the Count's original intent. Like limestone, dolomite forms in several different ways, from penecontemporaneous cryptocrystalline mudstones in sabkhas, to coarsely crystalline varieties during late diagenesis. In the latter case, dolomite is virtually a metamorphic rock (Figure 7). Dolomite is an important petroleum reservoir, and occurs as a gangue mineral with lead–zinc sulphide ores (*see* **Mining Geology:** Hydrothermal Ores). It is described more fully in **Sedimentary Rocks:** Dolomites.

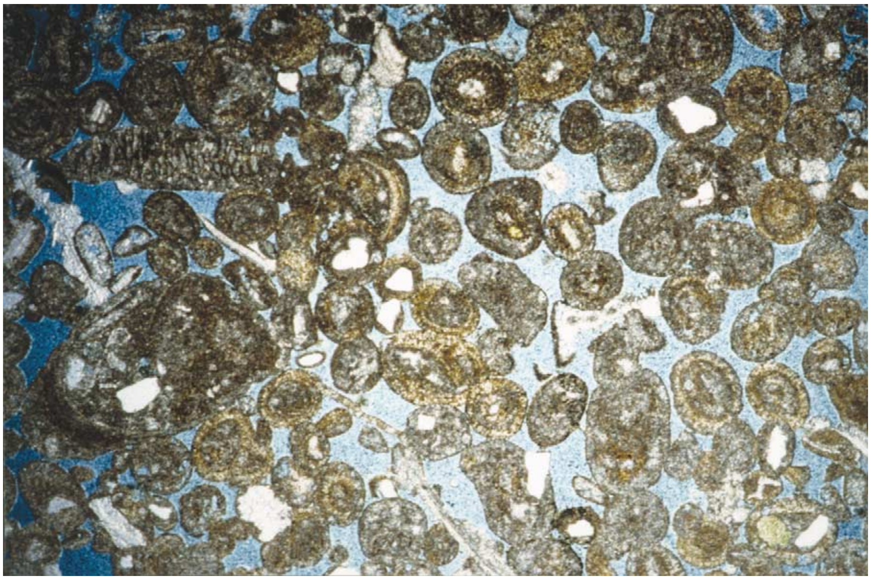


Figure 6 Photomicrograph of limestone under ordinary light. This is a well-sorted oolite grainstone from the Upper Jurassic Portland Limestone, Dorset, UK. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

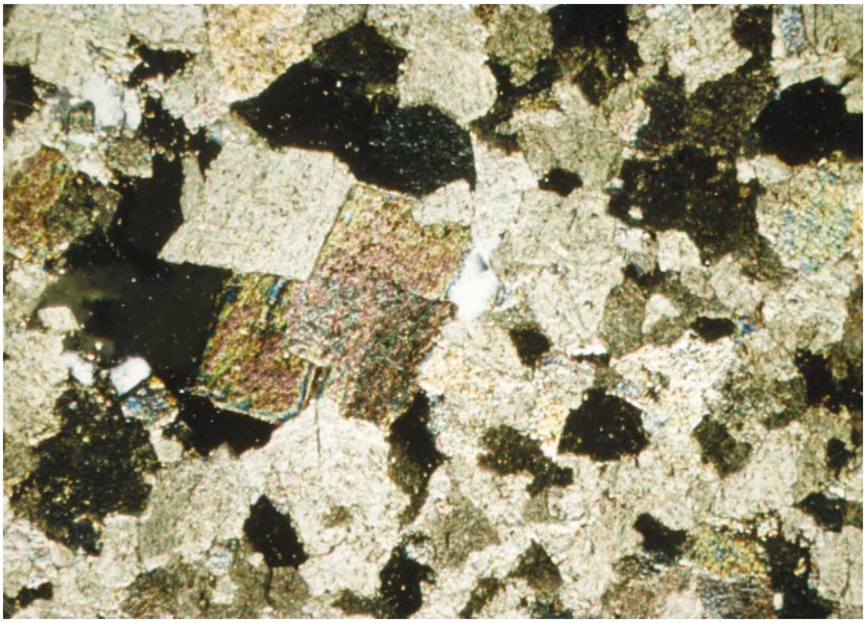


Figure 7 Photomicrograph of dolomite under ordinary light. This is a coarsely crystalline variety from the Zechstein (Upper Permian) of the UK North Sea. Some porosity (pale blue) is visible. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Siderite is composed of iron carbonate ($FeCO_3$). It occurs commonly in shales as early cement and as concretions. It occurs as crystalline cement in sandstones, and occasionally in spherulites (sphaerosiderite) in lacustrine deposits. Here, it may be sufficiently abundant to become an iron ore (described below).

Magnesite ($MgCO_3$) is the name for the mineral magnesium carbonate, as well as the rock. It forms both as an alteration product of dolomite, and from the action of magnesium-rich fluids on limestone. Magnesite is rare, but occurs in commercially important deposits at Radenhein (Austria), Liaotung (China), and Clark County, Nevada (USA).

Evaporites

The evaporite chemical sedimentary rocks are rare, but extremely important commercially as the raw materials for the chemical industry. As the name suggests, the evaporites consist of a suite of minerals formed from the evaporation of sea water. They tend to occur in restricted sedimentary basins in cyclic sequences that begin with carbonates (limestone and/or dolomite), overlain by sulphates (gypsum and/or anhydrite), halite (sodium chloride), and then a range of potassium salts, including carnallite and polyhalite (Figure 8).

As the name suggests, it was once thought that evaporites formed exclusively from the drying out of enclosed marine basins. This required improbably large volumes of sea water to provide the resultant evaporites. It is now realized that many evaporites actually form in sabkhas (Arabic for salt marsh) from the replacement of pre-existing rocks, principally carbonates, by circulating brines. Evaporites should thus more correctly be termed 'replacementites'. Evaporites are described in more detail in **Sedimentary Rocks:** Evaporites.

Residual Deposits

Residual deposits are a variety of rocks produced by *in situ* chemical alteration or weathering (*see* **Weathering**). They include three economically important rocks: laterite, china clay (kaolin), and bauxite. These are now described in turn.

The word 'laterite' is derived from the Latin '*later*', a brick, as this rock has been widely employed for this purpose, being soft when quarried, but hardening on exposure. The term was first employed by a British geologist of the Raj, working in India, where laterites are exceptionally well developed. Laterites result from the intense weathering in many parts of the world of rocks of diverse ages and types, but particularly iron-rich rocks such as basalts. Laterites thus occur as laterally extensive residues up to 10 m in thickness above the bedrock. They require thousands of years to form, a humid climate, and a well-drained terrain. The resultant laterite is rich in hydrated iron and aluminium oxides, and low in humus, silica, lime, silicate clays, and most other minerals. Laterites are red and argillaceous in appearance, but often possess

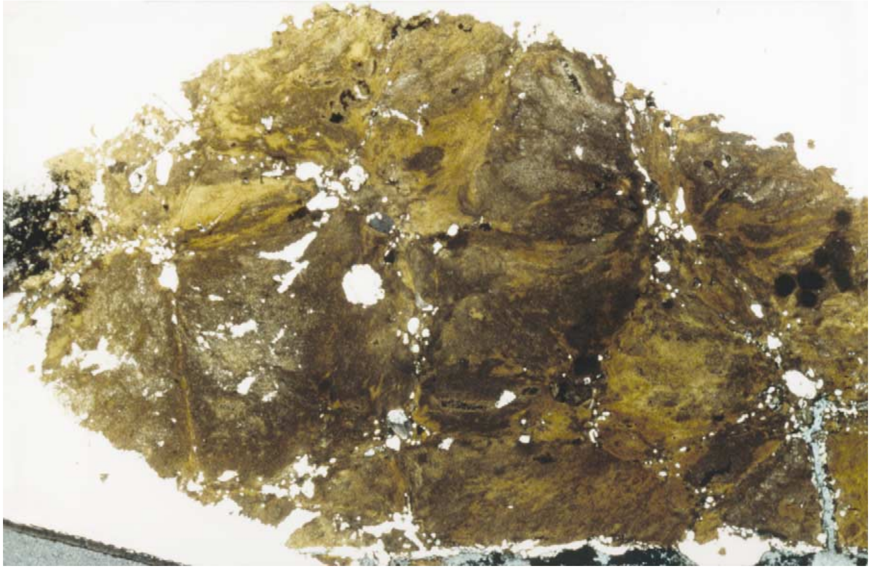


Figure 8 Photomicrograph of nodular anhydrite ($CaSO_4$) from the sabkha of Abu Dhabi, United Arab Emirates. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

a nodular pisolitic texture. They are often termed ferricrete (etymologically, although not genetically, comparable with silcretes and calcretes). Laterites are important sources of iron in West Africa and Western Australia, and of nickel in Cuba. Further details are given in **Soils:** Modern and **Soils:** Palaeosols.

China clay, or kaolin, is the name given to rock composed almost entirely of the clay mineral kaolinite, $Al_2Si_2O_5.(OH)_4$ (*see* **Sedimentary Rocks:** Clays and Their Diagenesis). Kaolin and kaolinite are occidental corruptions of Kauling, a hill in China, from whence the first samples to enter Europe were shipped by a Jesuit missionary in 1700. Some kaolinite is produced by the *in situ* hydrothermal alteration of feldspar in granites, as for example that of south-west England. Kaolinite may then be reworked from such a source, and re-deposited in lacustrine environments, as for example the Oligocene 'Ball Clay' deposits of Bovey Tracey, Devon.

Kaolin also forms, however, as a residual deposit due to the intense weathering of aluminosilicate-rich rocks. These include feldspar-bearing igneous rocks, such as granites and gneisses. Kaolin can also be produced from sedimentary rocks, including arkosic sandstones and shales. The general chemical reaction leading to the production of kaolinite is:

$$KalSi_3O_8 + H_2O \rightarrow Al_2Si_2O_5(OH)_4$$
$$\text{Feldspar} + \text{Water} \rightarrow \text{Kaolinite}$$

Kaolin forms as a residual deposit in the soil horizons of warm humid climates, where erosion rates are low, and there is plenty of time for leaching to take place. Kaolin has many important industrial uses (see **Clays, Economic Uses**). Notable commercial deposits occur in China, naturally, south-west England, Saxony (Germany), Bohemia (Czech Republic), and Georgia, USA.

The third type of residual deposit is bauxite, hydrated aluminium hydroxide ($Al_2O_4.nH_2O$). Bauxite takes its name from Le Baux, near Arles in France. Bauxite is the end result of the intensive and prolonged weathering of soils that commences with laterite, and proceeds, via kaolin, to bauxite. These changes reflect the progressive leaching of silica, iron, and kaolinite (Figure 9).

The chemical reaction that finally leads to the formation of bauxite is:

$$H_2O + Al_2Si_2O_5(OH)_4$$
$$= Al_2O_4.nH_2O + 2SiO_2.2H_2O$$

$$\text{Water} + \text{Kaolinite}$$
$$= \text{Aluminium Hydroxide} + \text{Silicic Acid}$$

Bauxites tend to be reddish or pink in colour due to some residual iron oxide. They may also possess a pisolitic texture inherited from an earlier lateritic phase (Figure 10).

Bauxite is very important as the ore for aluminium. Bauxite occurs as residual deposits on limestone, as for example in France and Jamaica. It also occurs as a residual soil on Precambrian igneous and metamorphic rocks, as in Surinam.

Kerogenous Chemical Sediments

Kerogen is defined as hydrocarbons that are insoluble in normal solvents, such as carbon tetrachloride, but which yield liquid or gaseous petroleum when heated. Chemically, kerogen includes a range of complex hydrocarbons, with traces of many other elements, including sulphur, nitrogen, and various radioactive and heavy metals. Kerogen is generally deposited in anoxic reducing stagnant conditions, most commonly found in marshes, swamps, meres, salt marshes, and lagoons, and is particularly characteristic of deltas (*see* **Sedimentary Environments:** Deltas). In these environments, vegetation may accumulate as laterally extensive horizons of peat many metres thick. During subsequent burial, peat undergoes extensive compaction and diagenesis, changing first into brown coal (lignite), then bituminous coal, then anthracite, and finally graphite, as it enters the realm of metamorphism. The variety of kerogen termed coal is, of course, a very important source of energy. Less conspicuously, although equally important, kerogen occurs in varying amounts in mudstones. Originating as plant and animal detritus, this disseminated kerogen is the mother of petroleum (Figure 11). When kerogen constitutes >1.5%, or thereabouts, of a shale, the shale becomes a potential petroleum source rock, subject to sufficient thermal maturation. Depending both on its chemistry and the level of maturation, kerogen generates petroleum gas and oil (*see* **Petroleum Geology:** The Petroleum System).

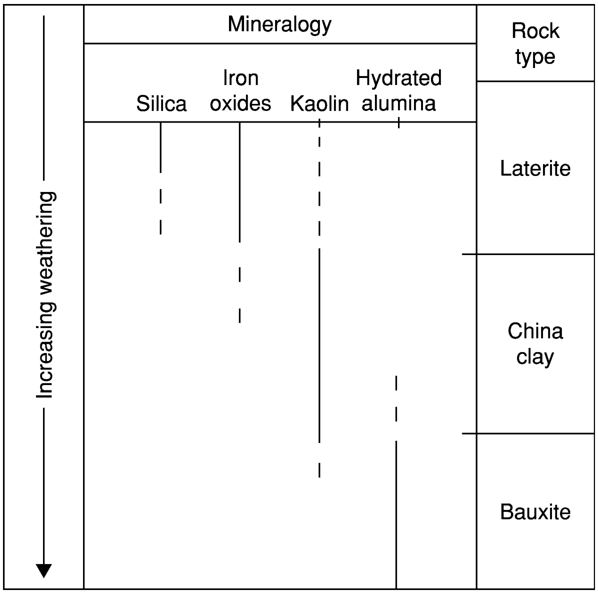


Figure 9 Diagram to show the progressive formation of the residual deposits laterite, kaolin, and bauxite that result from prolonged chemical weathering. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Ironstones

Iron is widespread in sedimentary rocks, but is concentrated in economic amounts in very few. A distinction is made between the Banded Ironstone Formations (colloquially referred to as BIFs), and ironstones *sensu stricto*. Banded ironstones are widespread around the Earth, but they are all of Precambrian age, and are curiously interbedded with

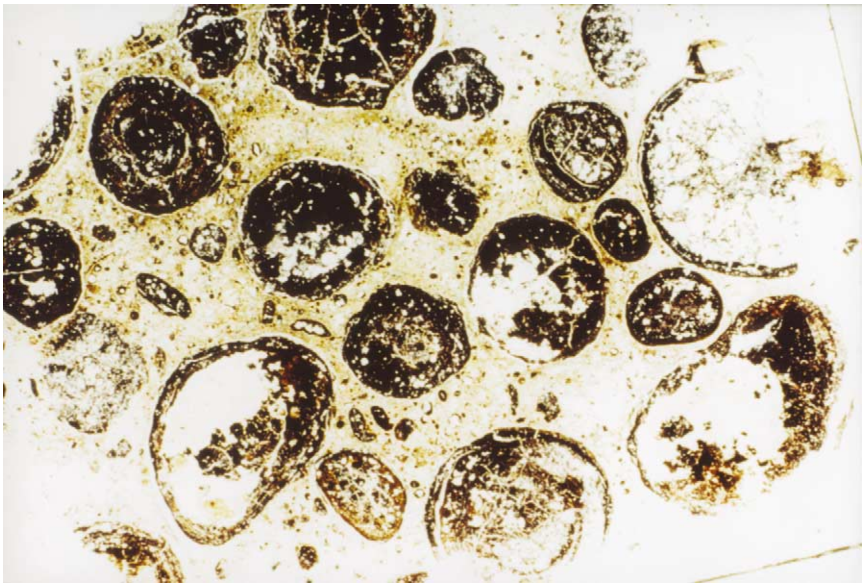


Figure 10 Photomicrograph of pisolitic bauxite under ordinary light. Individual pisoids are approximately 0.5 cm in diameter. Le Baux, France. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

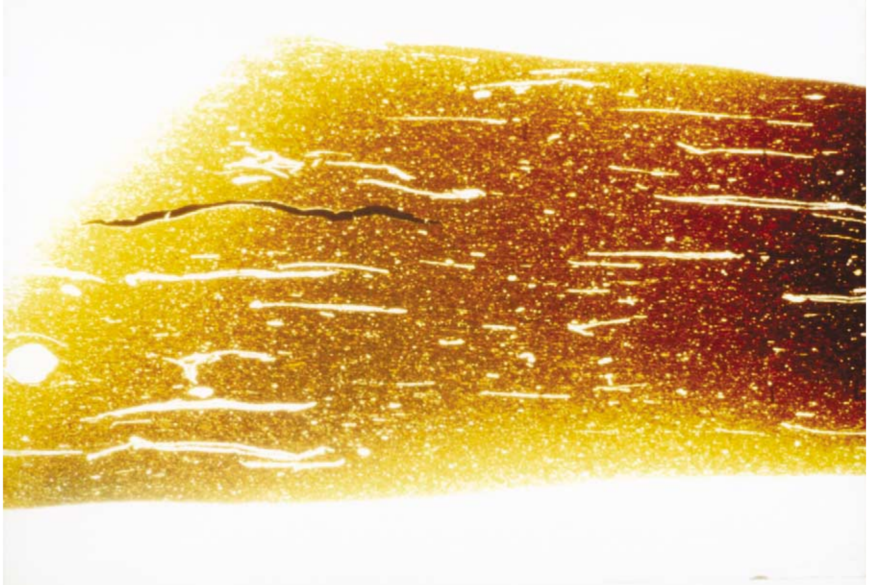


Figure 11 Photomicrograph of kerogen. This is the sapropelic Kimmeridge Coal (Upper Jurassic) from Dorset, UK. Cross-sections of bivalves are ubiquitous, and carbonized plant detritus is also visible. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

chert. They formed when the Earth's atmosphere was significantly different from that of today (*see* **Sedimentary Rocks:** Banded Iron Formations). The term ironstone is now restricted to Phanerozoic sedimentary rocks consisting of at least 15% by weight of iron, either 19% FeO, or 21% Fe_2O_3, or an equivalent admixture.

Ironstones consist of a range of iron minerals, including oxides (magnetite, haematite, and goethite/limonite), carbonates (siderite), and silicates (chamosite and berthierine). Three main types of ironstone are recognized: blackbands, claystones, and ooidal. Blackband and claystone ironstones are organic-rich sideritic mudstones. They commonly occur in deltaic deposits associated with peat and coal. Intraformational conglomerates composed of ironstone clasts and horizons of brittle fractured ironstone in slumped delta slope shales indicate that the ironstones formed during early shallow burial.

The origin of the third type of ironstone is more controversial. The ooidal ironstones are composed of several types of iron mineral (Figure 12). Ooid formation is normally associated with high-energy depositional environments (*see* **Sedimentary Rocks:** Limestones). The ooidal ironstones, however, are often poorly sorted wackes. Thus, argument has raged as to whether ferruginous ooids formed in high-energy environments, and were then dumped as poorly sorted wackes. Alternatively, did iron minerals replace quotidian lime ooids during subsequent diagenesis?

The data and arguments are examined more fully in **Sedimentary Rocks:** Ironstones, but it is probably as true today as it was in 1949, when Taylor wrote in his seminal memoir on the Northampton Sand Ironstone, that: "Conditions of deposition of the sedimentary iron ores are still to some extent a mystery".

Phosphates

The penultimate group of chemical sediments to consider are the phosphates. Phosphates are an extremely important mineralogically complex group of rocks that are essential as plant fertilizers. About three-quarters of the world's supply of phosphates comes from sedimentary deposits. Sedimentary phosphate deposits are of three types: bedded, placer, and guano. Bedded phosphates are formed by the replacement of limestone, and of teeth, bones, and coprolites, to form the mineral phosphorite. Factors that favour phophoritization include a broad shelf adjacent to an ocean, slow shallow marine sedimentation, low terrigenous input, and high organic productivity. Once phosphorite has formed, it is stable in sea water, and sparingly soluble in fresh water. Thus, phosphate pebbles occur, not only as intraformational conglomerates intimately associated with bedded phosphate, but also as placer deposits that are sometimes far removed from their parent body. Bedded phosphates formed during several geological periods on ancient shelves around the world. The Cretaceous phosphate belt of the southern shores of Tethys is noteworthy. This stretches

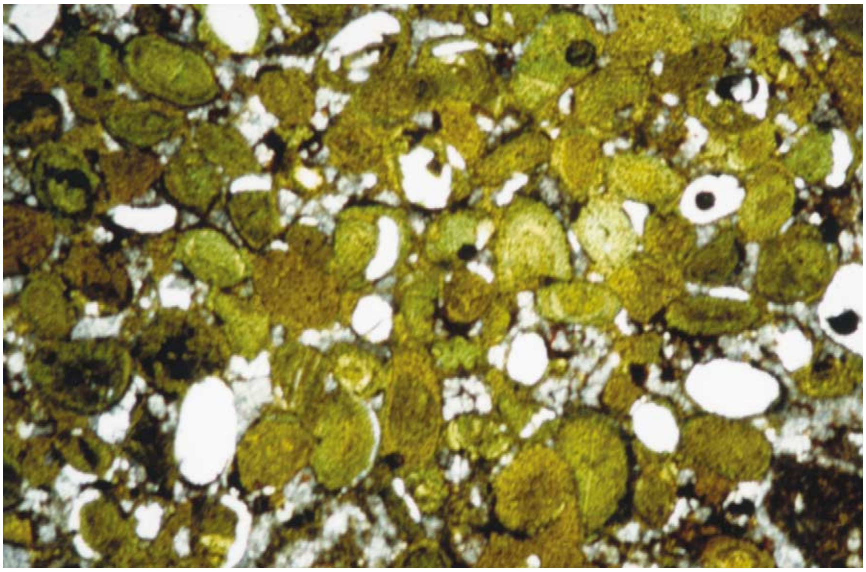


Figure 12 Photomicrograph, under ordinary light, of Northampton Sand Ironstone, Middle Jurassic, UK. This ironstone is composed of chamosite and siderite ooids with concentric growth rings. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

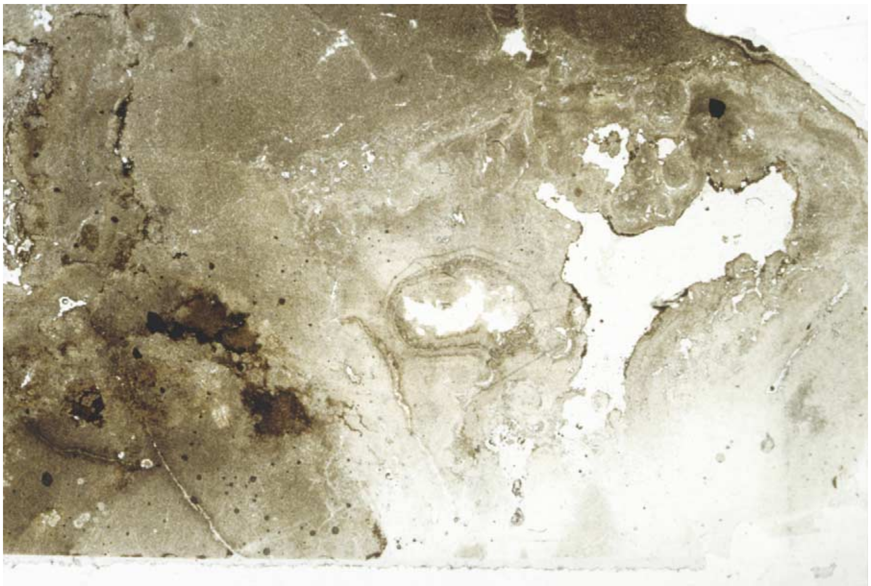


Figure 13 Photomicrograph, under ordinary light, of guano phosphate deposit. This is formed by the phosphatization of bird droppings. St Helena, South Atlantic. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

from Bu Craa in the western Sahara, through Morocco and Algeria to the flanks of the Red Sea, Jordan, and Jabal ash-Sharki in Syria. Younger placer phosphate deposits of note include the alluvial phosphate gravels of South Carolina and Florida in the USA.

Guano is the youngest phosphate rock (Figure 13). This is a fertilizer rich in phosphates and nitrates that forms from the accreted excreta of birds and bats. Whole Pacific Islands, such as Nauru, are, or rather, were composed of guano that has subsequently been quarried away.

Phosphates are described in greater detail in **Sedimentary Rocks:** Phosphates.

Siliceous Deposits

The last group of chemical sediments to describe are those composed of silica (SiO_2), not the detrital sands, but those that formed by organic secretion, replacement, or, possibly, by direct precipitation from water.

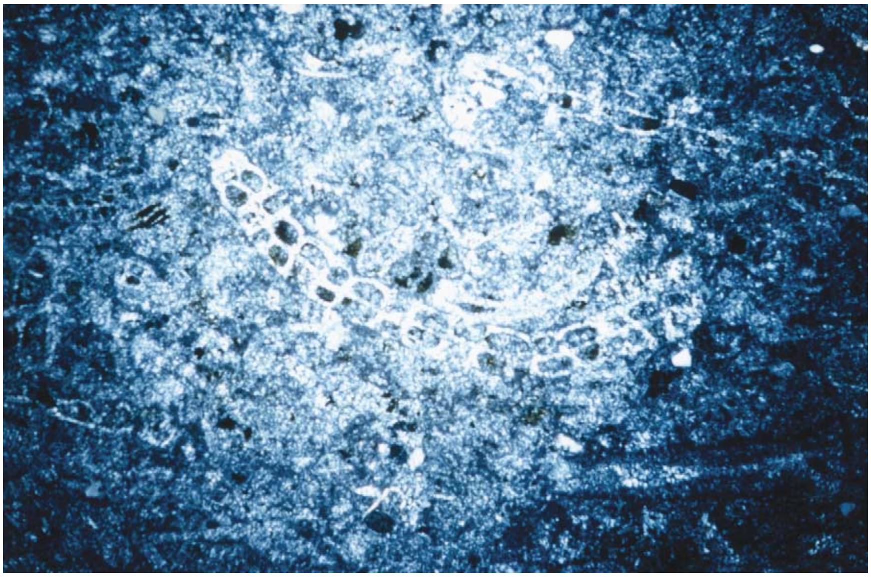


Figure 14 Photomicrograph, under polarized light, of chert (colloquially, flint) from the Chalk (Upper Cretaceous), Dover, UK. This formed by replacement of pre-existing limestone. Some silicified bioclasts can be detected. Reproduced with permission from Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Various organisms secrete silica. Radiolaria and diatoms secrete silica shells, whilst some sponges secrete internal spicules of silica. Deposits of radiolarian and diatom shells may accumulate under water in sufficient amounts to constitute sedimentary rocks in their own right, termed 'radiolarite' and 'diatomite', respectively. Organically precipitated silica disseminated in sediments may undergo dissolution and reprecipitate as cryptocrystalline silica. This variety of silica is termed 'chert', often referred to colloquially as 'flint' by the aboriginal inhabitants of the Cretaceous Chalk downlands of south-east England. Chert is a non-clastic cryptocrystalline variety of silica that occurs as nodules and horizons in limestones and sandstones (Figure 14). Chalcedony is a radial or fibrous variety of chert that infills fossils and other pores. Opal, sometimes termed opaline silica, is a hydrated variety of silica, amorphous and isotropic, with the chemical formula $SiO_2.nH_2O$, sometimes also written $SiO_2.nSi(OH)_4$. It occurs only in Tertiary to Recent sediments, having dehydrated to chert in older rocks. Opal is a semiprecious stone.

Cherts pose three main problems. What is the source of the silica? What is the environment of deposition of bedded cherts? How do nodular cherts replace their host sediments? These questions are discussed in **Sedimentary Rocks:** Chert.

Conclusion

The primary aim of this overview of the sedimentary rocks has been to simultaneously point the reader in two different directions: to the tranche of articles dealing with mineralogy, and to the tranche of articles that describe the various sedimentary rocks in more detail. The article offers a simple classification of the sedimentary rocks, but at the same time points out the problems and inconsistencies of any scheme of rock classification. Sediments can be broadly grouped either by their physical attributes, principally grain size, but also by their chemistry or mineralogy. A detailed analysis of any scheme shows its inconsistencies. To call a group of rocks 'chemical' is nonsensical, because it implies that others are 'non-chemical', whatever that might mean. Evaporites should really be called 'replacementites,' but who would dare to argue for such a change. Siliceous chemical rocks include detrital sediments, primary precipitates, and also some of diagenetic origin, yet they are placed in one single group. Classification is purely a pedagogical framework, like the chrysalis from which exegesis may emerge like a butterfly.

See Also

Clay Minerals. **Clays, Economic Uses**. **Famous Geologists:** Lyell. **Metamorphic Rocks:** Classification, Nomenclature and Formation. **Minerals:** Carbonates; Glauconites. **Mining Geology:** Hydrothermal Ores. **Petroleum Geology:** The Petroleum System. **Rocks and Their Classification**. **Sedimentary Environments:** Depositional Systems and Facies; Deltas; Storms and Storm Deposits. **Sedimentary Processes:** Fluxes and Budgets. **Sedimentary Rocks:** Banded Iron Formations; Chert;

Clays and Their Diagenesis; Dolomites; Evaporites; Ironstones; Limestones; Phosphates; Rudaceous Rocks; Sandstones, Diagenesis and Porosity Evolution. **Soils:** Modern. **Unidirectional Aqueous Flow**. **Weathering**.

Further Reading

Leeder MR (1999) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Science.

Lyell C (1842) *(and many subsequent editions) Elements of Geology*. London: John Murray.

Pettijohn FJ (1957) *Sedimentary Rocks*, 2nd edn. New York: Harper Geoscience.

Selley RC (2000) *Applied Sedimentology*, 2nd edn. London: Academic Press.

Taylor JH (1949) *Petrology of the Northampton Sand Ironstone Formation. Memoirs of the Geological Survey*. London: HMSO.

Banded Iron Formations

A Trendall, Curtin University of Technology, Perth, Australia

Introduction

Banded Iron Formation, normally abbreviated to BIF, is a type of sedimentary rock commonly present in Precambrian sedimentary successions of specific ages. Apart from this time restriction it has a number of characteristics which make it unique among sedimentary materials. Firstly, there is nothing chemically similar to it being laid down in the waters of the modern Earth, so that its origin cannot be deduced by directly observing the formation of similar materials in the present-day environment. Secondly, it has virtually none of the textural features common in other sedimentary rocks that give a clue to the conditions under which it was laid down; theories for its origin, therefore, have to be argued from various indirect lines of evidence, and there is still no complete agreement on its significance in the stratigraphic record. A third feature is that it is exceptionally hard, tough, and dense, making it instantly recognisable in field exposures, where it commonly forms resistant ridges standing out from more easily eroded rocks. Finally, it is a sedimentary rock in which fine details of stratification have been shown to persist over exceptionally large distances, arguing for depositional conditions free from disturbance by currents. BIF also has great economic importance because it is the source, either directly or indirectly, of most of the iron ore presently mined; and as the main source of raw material for the world's iron and steel industries, underpins the physical fabric of the developed world.

Nomenclature, Classification, Definition

Before describing the characteristics of BIF in more detail, it is worth reviewing the different names that have been applied to it on different continents and at different times, so that there is a clear understanding of how the name is applied in this article. The first formal geological descriptions of BIF were made in the latter part of the nineteenth century, focusing on occurrences in parts of the USA south of Lake Superior, where iron ore mining was established in areas of outcrop known as 'ranges', such as the Mesabi, Marquette, Cuyuna, Gogebic, and Menominee Ranges. The rock was then called 'jasper', 'jaspilite', or 'iron-bearing formation', which was later shortened to 'iron-formation'. The name 'taconite' was also used, and because many of the rocks had conspicuously multicoloured banding the term 'banded iron-formation' also became common. During the early twentieth century, other names came to be used for similar rocks of other continents. For example, 'itabirite' was used in Brazil, 'ironstone' in South Africa, and 'BHQ' ('banded haematite quartzite') in India. Studies in all these places tended to assume that the iron-rich rocks to which these names were applied were identical with those that had been well documented from the Lake Superior ranges.

As more detailed studies were made later in the twentieth century, it was realized that many extensive occurrences of BIF, and especially those of South Africa and Australia, were distinctively different from those of the Lake Superior area. In particularly, they lacked the granular structure that was generally present in the latter, and had a different pattern of banding. The term 'granular iron-formation', or GIF, is now preferred for the type of BIF to which the name was first applied in the USA. In a perfectly rational classification it would probably be best to restrict the names BIF and GIF to two types of iron-rich sedimentary rock with separate textural features, and to use the name 'iron-formation' (IF) as a generic name for both. But because the use of BIF is so strongly entrenched, it is used here as the general term, and GIF is regarded as a particular type of BIF.

BIF is therefore defined as a chemically precipitated sedimentary rock containing at least 15% of iron, typically thin-bedded or laminated, and commonly containing layers of chert.

Chemical and Mineralogical Composition

Although the definition above requires a minimum iron content of 15%, the great majority of rocks that would be called BIF contain between 25% and 35% Fe. About half of the rock by weight consists of iron oxides, while the remainder is mainly silica. Carbon dioxide (as carbonate) is present as a minor constituent in many BIFs, and is a major component in some, but all other oxides (e.g., Al_2O_3, MgO, Na_2O, K_2O, P_2O_5) are relatively minor, and trace elements are insignificant. Haematite (Fe_2O_3) and magnetite (Fe_3O_4) are the most abundant iron minerals. Others that may be present are carbonates (ankerite, siderite) and silicates (stilpnomelane, greenalite, riebeckite). The silica normally occurs as microcrystalline quartz, usually called chert. Neither the chemical nor mineralogical composition of GIF differs significantly from that of BIF.

The Banding

The banding of BIF, with the exception of the type known as GIF, which is dealt with later, may best be described by reference to that of the well preserved BIFs which were laid down in the Hamersley Basin of north-western Australia between about 2650 Ma and 2450 Ma. Apart from very low-grade metamorphism, open folding, and gradual uplift and exposure, the rocks of this basin remain essentially unaffected by post-depositional events; the basin had a depositional area of at least 10^5 km^2. The 1.2 km (compacted thickness) of BIF present within the 2.5 km-thick median sequence of this basin forms five main units interstratified with shales and carbonate sediments. One of these units, the *ca.* 140 m-thick Dales Gorge Member, is particularly useful for stratigraphic study in that it has 33 internal alternations, termed macrobands, of BIF and shale. The banding of the BIF of each of the 16 BIF macroband is termed 'mesobanding', and is defined by sharply defined alternations, on a centimetric scale, of dark iron-rich (silica-poor) mesobands and light silica-rich (iron-poor) mesobands; the latter are usually called chert. Chert mesobands of the Dales Gorge Member BIF are normally between 5 mm and 15 mm thick, with an average thickness about 8 mm; they constitute about 60% of the total BIF volume; the intervening iron-rich mesobands have a mean thickness of about 10 mm, and make up about 20% of the total BIF volume. A minor proportion of mesobands of magnetite and/or carbonate make up the remainder of the volume. All the mineral components of all mesoband types are made up of fine-grained closely crystalline material that shows no evidence of a clastic contribution to the formation of the rock.

The monotonously repetitive alternation of iron-rich and iron-poor mesobands, on the same centimetric scale as those of the Dales Gorge Member, is an easily identifiable characteristic of BIF wherever it occurs, and is the scale of banding from which the name BIF derived. In field exposures the alternating bands may be coloured black and white, red and white, or red and black, largely according to the weathering of the rock, and often give it a spectacularly striped appearance (Figure 1).

Within many chert mesobands of the Dales Gorge Member there is a regular small-scale lamination defined by a concentration of some Fe mineral within the pervasive matrix of microcrystalline quartz. The Fe mineral may be either hematite, magnetite, carbonate (siderite or ankerite), stilpnomelane, or some combination of these. This lamination within individual chert mesobands has been called microbanding. The iron-rich laminae that define microbands are normally separated by thicknesses of between 0.2 mm and 1.6 mm of virtually iron-free chert. Within a microbanded chert mesoband the microband thickness tends to vary only slightly, but from one such mesoband to another there may be significant variation. Since its identification in the Dales Gorge

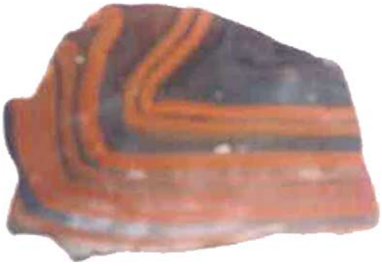


Figure 1 Hand specimen (15cm wide) of folded Archaean BIF from the Yilgarn Craton, Western Australia. The conspicuous red and black banding is mesobanding; the red mesobands are chert and the black mesobands consist mainly of iron oxides and microcrystalline quartz. (Photograph by GJH McCall).

Member, it has been identified in a number of other well-preserved BIF units elsewhere.

The characteristic mesobanding of BIF is not present in GIF. GIF also has alternations of iron-rich and iron-poor material, but these are typically coarser and much less regular. The coarsely crystalline chert bands are commonly wavy or lenticular. Both iron-rich and silica-rich bands may be granular, more particularly the latter. The iron-rich bands of GIF, as the name implies, often consist of a close-packed and lithified mass of granules or ooliths, averaging about a millimetre in diameter. They are made up of iron oxides, with or without quartz, and the intergranular material consists mainly of the same minerals, but usually with a lower iron content.

Continuity of Banding

A remarkable stratigraphic feature of Hamersley Basin BIF is its exceptional lateral continuity. Thus the Dales Gorge Member, used above as a model for the description of banding, and the other main BIF units, are easily identifiable over the entire basin area. BIF macrobands within the Dales Gorge Member are similarly recognisable through their constant relative thicknesses, and within BIF macrobands both individual centimetre-scale mesobands, as well as sub-millimetre microbands within them, have been correlated over 300 km. This degree of fine-scale lateral stratigraphic continuity is reminiscent of evaporites, particularly in the Permian of Europe and the United States.

Metamorphic and Tectonic Modification

Because of their exceptional freedom from post-depositional metamorphism and tectonic deformation, the BIFs of the Hamersley Basin have been used as 'type examples' for the introductory descriptions above; the BIFs of the Transvaal and correlative Griqualand West basins of South Africa, which are of similar age, also have the same excellent preservation. But in this respect they are atypical of the majority of BIF occurrences. Many Early Precambrian examples have suffered significant metamorphic modification, which begins with coarse-grained recystallisation (annealing) of both the initially microcystalline chert and fine-grained iron oxides, and continues with the growth of iron-rich silicates (e.g., grunerite, ferrohypersthene, fayalite). Even in early metamorphic stages the delicate fine textures within the banding tend to become blurred, but the essential iron-rich/iron-poor alternation of the mesobanding shows a robust resistance to complete obliteration. Most metamorphism of BIF appears to be isochemical, but there is some evidence of chemical modification at higher grades.

Older Precambrian BIFs also tend to have undergone significant tectonism, particularly those of Archaean greenstone belts of all continents, where the BIFs often form curvilinear steeply dipping units which are useful in deciphering complex structures. It is characteristic of deformed BIF in these belts that it appears to have reacted sensitively to tectonic stress, forming complex internal flowage folds defined by the banding.

Distribution Over the Earth

BIFs are widely distributed throughout the Precambrian areas (cratons and shields) of all continents except Antarctica, where only one occurrence is so far known. BIFs of the older cratons include the oldest known example, at Isua, in Greenland, aged about 3.8 Ga. They are consistently present in the greenstone belt sequences of all the main old cratons. Examples include the Abitibi Belt of the Superior Province of Canada, the greenstone belts of the Yilgarn and Pilbara Cratons of Australia, and the those of the Baltic Shield, the North China Craton, the Amazon Craton of Brazil, the Kaapvaal and West African Cratons, and the BIFs of the Ukraine Craton, notably at Krivoi Rog; most of these are relatively thin, tectonised, and metamorphosed. Most of these greenstone-associated BIFs have ages between 2.8 Ga and 2.5 Ga. A different style of BIF occurrence is present in four of the Gondwana continents (South America, southern Africa, India, and Australia). In this type the BIFs occur as well preserved, gently dipping supracrustal sequences, which may form conspicuous topographic plateaus. The Carajás Formation of the Amazon Craton, the Cauê Itabirite of the São Francisco Craton (both in Brazil), the Kuruman Iron Formation and Penge Iron Formation of South Africa, and the Mulaingiri Formation of the Indian Karnataka Craton all belong to this group, which includes the BIFs of the Australian Hamersley Basin, already mentioned above; these occurrences have been called the Great Gondwana BIFs, and are mostly younger than those present in greenstone belt occurrences. Further mention of a final distinct category of BIF occurrences, which includes Rapitan in the Yukon, and those of the Damara Belt, in Namibia, is made below the next heading.

Distribution in Time

Mention has already been made not only of the general restriction of BIFs to the Precambrian, but also to specific intervals within that Eon. This topic needs closer attention, since it is critical for understanding the genesis of BIF, and has been a subject of strong controversy. A relevant point to be made is, of course, that the topic could not be discussed at all without the availability of the precise isotopic dating techniques that have established, only in the last couple of decades, a reliable time-scale for the entire span of Precambrian time.

An age of about 3.8 Ga has already been noted for the oldest BIF, from Isua. This is still the oldest known sedimentary (or more accurately metasedimentary) succession, so that BIF is included among the oldest sedimentary rocks. The abundance of BIF in Archaean greenstone belts indicates that it continued to be deposited intermittently until the end of that Era, which by convention is accepted as 2.5 Ga. By far the two largest basins of BIF deposition on Earth, in terms of contained Fe, are the Hamersley Basin and the Transvaal/Griqualand West Basin, and in these, massive deposition of BIF continued until ~2450 Ma. There is then a paucity of BIF occurrences until the Lake Superior GIFs, at ~1850 Ma, which represents a global peak for this type. A long interval in which no BIF deposition is known then followed, and it was not until the end of the Precambrian, in the Late Neoproterozoic, that there is a final burst of BIF deposition, which includes the Rapitan and Damara examples.

In summary, the overall picture is of a long early period of intermittent BIF deposition, culminating in a peak at ~2500 Ma, after which there was a smaller burst at ~1800 Ma, and finally, after a long hiatus, another significant depositional event at the end of the Precambrian. It has been pointed out that some iron-rich Phanerozoic rocks closely resemble Precambrian BIFs, but these exceptions only prove the rule, and indicate that there were special environmental conditions during the Precambrian, and particularly during the Early Precambrian, which led to the deposition of BIF.

Association with Other Rocks, and with Volcanism

BIFs do not, of course, exist in isolation, but occur in association with other lithologies as components of sedimentary, or volcanosedimentary, sequences. Two general stratigraphic points are worth making: in the first place, BIF tends to form discrete, clearly demarcated, units of significant thickness and wide lateral extent – it does not form thin or discontinuous beds interlaminated, or interdigitated, with other lithologies; and in the second place, BIF has never been demonstrated to grade imperceptibly into another lithology, whether laterally or vertically. It is as though the conditions for BIF deposition, in any given basin, were intermittently switched on and off abruptly, rather than by any gradual change in the depositional environment.

Within those two overarching generalisations, BIF has not been demonstrated to occur in immediate sedimentary contact with any preferred lithology. Examples can be found of association with both fine-grained (shale) and medium-grained (quartzite) epiclastic sediments, with turbidites and with carbonates. For BIFs in Archaean greenstone belts, there is a common association with mafic volcanic rocks, but an immediate and local genetic link between the two has never been unequivocally demonstrated. While it is true that volcanic rocks, both felsic and mafic, are commonly present as components of depositional basins in which BIF occurs, the fact remains that no volcanic rocks at all are closely associated with one major BIF of Brazil – the Cauê Itabirite of the Quadrilátero Ferrífero. There is abundant trace element and isotopic evidence that BIFs have an igneous connection, but an immediately local association cannot be inferred from this.

Theories of Origin

From direct observation of the present environment it is easy to see that lithostratigraphic units of, say, sandstone are likely to have been formed by the transport of sand in rivers to the sea. There is no such direct way to determine how BIF formed, and a credible hypothesis for its origin must be built up step-by-step, from various lines of evidence. An early suggestion in the Lake Superior area was that BIF may be an acid lava, and more recently it has been proposed that BIF was initially a carbonate rock which had been replaced by iron and silica.

Most workers now accept that BIF is a chemically precipitated sedimentary rock, but that acceptance does not remove the need to explain many associated problems, including:

i. what was the source of the materials (both iron and silica)?
ii. how were they transported to the basin?
iii. did deposition occur in a lake or in the ocean?
iv. what caused precipitation?
v. what does the banding represent?
vi. to what degree does the final rock differ from the precipitate?

The first two of these questions were at one time debated in terms of two radically opposite concepts. One suggested that the iron and silica were derived from deep weathering of continental crust, and were selectively carried to the adjacent basin by rivers; the other proposed that the source of the the iron and silica was fumarolic volcanism very close to or within the basin, obviating the need for significant transport. As the debate progressed both ideas were generally abandoned, in favour of a model in which both iron and silica were in very dilute solution in ocean water, and were precipitated from the water of marginal basins kept supplied with both materials by appropriate circulation; in this model the primary source of iron and silica was ocean-floor volcanism or volcanic rocks, either at mid-ocean ridges or more generally from the ocean floors. As far as the third question is concerned, the abundance of BIF in the Precambrian stratigraphic record, its frequent association with clearly marine lithologies, and the difficulty of forming such immense thicknesses of iron-rich material in lakes, jointly make the lacustrine hypothesis untenable.

The fourth question then arises from a model involving precipitation from the water of an ocean margin basin which has low levels of dissolved iron and silica. Attention here has focused on the iron component, and specifically on the fact that dissolved iron will be in the ferrous state, and that oxidation will lead to precipitation in some ferric form. A variety of mechanisms for this oxidative precipitation has been proposed, including algal photosynthesis, anoxygenic bacterial photosynthesis, photo-oxidation by sunlight, and decomposition of water by ^{40}K radiation. Current research increasingly involves attention by microbiologists to the detailed mechanisms by which early biota could have effected precipitation. From a sedimentological viewpoint, the key point is that such mechanisms are quantitatively viable for precipitation of iron from small basin water concentrations.

The fifth question is one which received little early attention. Although it was generally agreed to represent the stratification of the BIF there was virtually no analysis of how the mesobanding was generated. A structured model was first proposed in respect to the BIFs of the Hamersley Basin. There it was argued that the microbands, whose presence has been noted within the chert mesobands, probably represent annual layers, or chemical varves: the thin iron-rich laminae of microbands may be presumed to represent summer seasons of high photosynthetic activity. Acceptance of this hypothesis permits calculation of the quantity of iron precipitated in the Hamersley Basin per unit area each year ($225\,\mathrm{t \cdot km^{-2} \cdot yr^{-1}}$), and hence, from knowledge of the bulk composition of the rock, an estimate of the compacted depositional rate. But these arguments have still not addressed the fifth question, which asks about mesobanding, not microbanding. So an additional step of this model added the proposal that the initial precipitate was a collloidal silica ferrihydrate gel which was compressed to about 10% of its initially deposited thickness during burial and diagenesis; and its final step was the suggestion that the mesobands developed during this stage by differential compaction. These last two steps, of course, address the sixth question also.

Some aspects of this depositional model have been challenged. Most workers on BIF agree that an annual (varve) significance for microbands is likely, although a diurnal or even tidal origin has been suggested. Others have preferred to see a direct link between mesobanding and iron supply from mid-ocean ridge activity, the mesobands representing pulsed variations in supply, with microbanded chert mesobands representing stable periods of perhaps of tens of years of relatively low iron availability. The origin of mesobanding still remains uncertain.

Unsolved Problems

Apart from the origin of mesobanding, the uncertainty of whose origin has just been noted, three other still-unsolved questions of BIF genesis are worth emphasizing in conclusion. The possible involvement of biotic processes in the precipitation of the iron of BIFs has already been noted; further studies related to this question are clearly not only significant for understanding BIF but for understanding important processes of biochemical evolution. The distribution of BIF in time has already been described, but its significance was not discussed. An early hypothesis tied the Late Archaean peak closely into a step in the biochemistry of photosynthesis, but additional geochronological work has lessened support for that proposal. A recent suggestion that the deposition of BIF is related to deep-water phases in basin development, and that this development shows systematic tectonically controlled secular changes, has yet to be fully debated. And finally, the place of BIF in the chemical evolution of the atmosphere and oceans has yet to be fully understood. Research on this uniquely puzzling sedimentary rock still holds many challenges.

See Also

Precambrian: Overview. **Rocks and Their Classification**. **Sedimentary Rocks:** Chert; Ironstones.

Further Reading

Appel PWU and La Berge GL (eds.) (1987) Precambrian iron-formations. Athens: Theophrastus Publications.

Beukes NJ (1980) Lithofacies and stratigraphy of the Kuruman and Griquatown Iron Formations, Northern Cape Province, South Africa. *Transactions of the Geological Society of South Africa* 83: 69–86.

Beukes NJ and Klein C (1992) Models for iron-formation deposition: Section 4.3. In: Schopf JW and Klein C (eds.) *The Proterozoic Biosphere: a multidisciplinary study,* pp. 146–151. Cambridge: Cambridge University Press.

Isley AE (1995) Hydrothermal plumes and delivery of iron to banded iron formations: *Journal of Geology* 103: 169–185.

James HL and Sims PK (eds.) (1973) Precambrian iron-formations of the world. *Economic Geology* 68(7): 913–1179.

Klein C and Beukes NJ (1992) Time distribution, stratigraphy and sedimentologic setting and geochemistry of Precambrian banded iron-formations: Section 4.2. In: Schopf JW and Klein C (eds.) *The Proterozoic Biosphere: a multidisciplinary study,* pp. 139–146. Cambridge: Cambridge University Press.

Morris RC (1993) Genetic modelling for banded iron-formation of the Hamersley Group, Pilbara Craton, Western Australia. In: Blake TS and Meakins A (eds.) *Archaean and Early Proterozoic Geology of the Pilbara Region, Western Australia, Precambrian Research:* 60 243–286.

Simonson BM (1985) Sedimentological constraints on the origins of Precambrian iron-formation. *Geological Society of America Bulletin* 96: 244–252.

Trendall AF (2002) The significance of iron-formation in the Precambrian stratigraphic record. In: Altermann W and Corcorane PL (eds.) *Precambrian Sedimentary environments: a modern approach to depositional systems.* International Association of Sedimentologists Special Publication 44: 33–66.

Trendall AF and Blockley JG (1970) The iron formations of the Precambrian Hamersley Group, Western Australia, with special reference to the associated crocidolite. *Western Australia Geological Survey Bulletin* 119: 365.

Trendall AF and Morris RC (eds.) (1983) *Iron-formation: Facts and Problems.* Elsevier: Amsterdam.

Chalk

J R Ineson and L Stemmerik, Geological Survey of Denmark and Greenland, Geocenter Copenhagen, Copenhagen, Denmark
F Surlyk, University of Copenhagen, Geocenter Copenhagen, Copenhagen, Denmark

Introduction

Chalk is a familiar rock type, particularly amongst Europeans, forming spectacular white cliffs along coastlines flanking the North Sea, the English Channel and the Baltic Sea. The essential characteristic of a true chalk is its microscopic composition – being composed predominantly of the skeletal remains of tiny calcareous marine algae known as coccolithophorids (Figure 1). Following their appearance in the Jurassic, these haptophycean algae became a common constituent of marine sediments and remain important components of the marine ecosystem today. Only during specific periods of Earth history and in certain palaeogeographic areas, however, were conditions such that pure carbonate oozes accumulated and were preserved on continental shelves and in vast epeiric seas. Chalk is thus most characteristic of the Upper Cretaceous (and in places the Danian) of north-west Europe and North America. This review focuses on these typical chalks, particularly from north-west Europe (Figure 2).

Chalk as a Sediment

Composition

A typical chalk is a fine-grained carbonate rock (a lime mudstone or micrite), the lithified equivalent of pelagic carbonate oozes recorded from ODP boreholes in present-day oceans. The sediment is dominated by debris derived from coccolithophorid algae that comprise a spherical calcareous test (coccosphere) up to several tens of microns in diameter. The test is made up of overlapping circular or elliptical discs or rings (coccoliths), which are 1–20 μm in diameter and are, in turn, constructed of tiny calcite platelets, laths or rays ranging from 0.1 to 2.5 μm across (typically 0.5–1 μm; Figure 1). The complete coccosphere is only rarely preserved and the chalk consists largely of coccoliths and their disaggregated platelets and spines (rhabdoliths). Despite being dominated by coccolithophorid debris, planktonic foraminifers and calcispheres are also common in chalks, together with coarser skeletal elements from both pelagic and benthic organisms such as belemnites, bryozoans, echinoids, bivalves, brachiopods, serpulids and sponges.

Mineralogically, pure chalks are composed of low magnesium calcite; this is the stable form of calcite

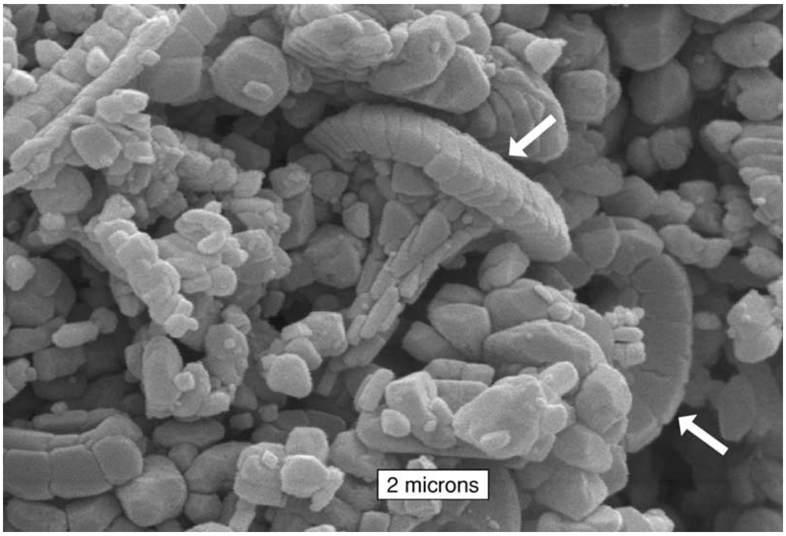


Figure 1 Scanning electron micrograph of chalk showing disc-shaped coccoliths (arrowed) in a mass of disaggregated platelets and cement (e.g., top right). Uppermost Maastrichtian, eastern Denmark. Photo: P Frykman.

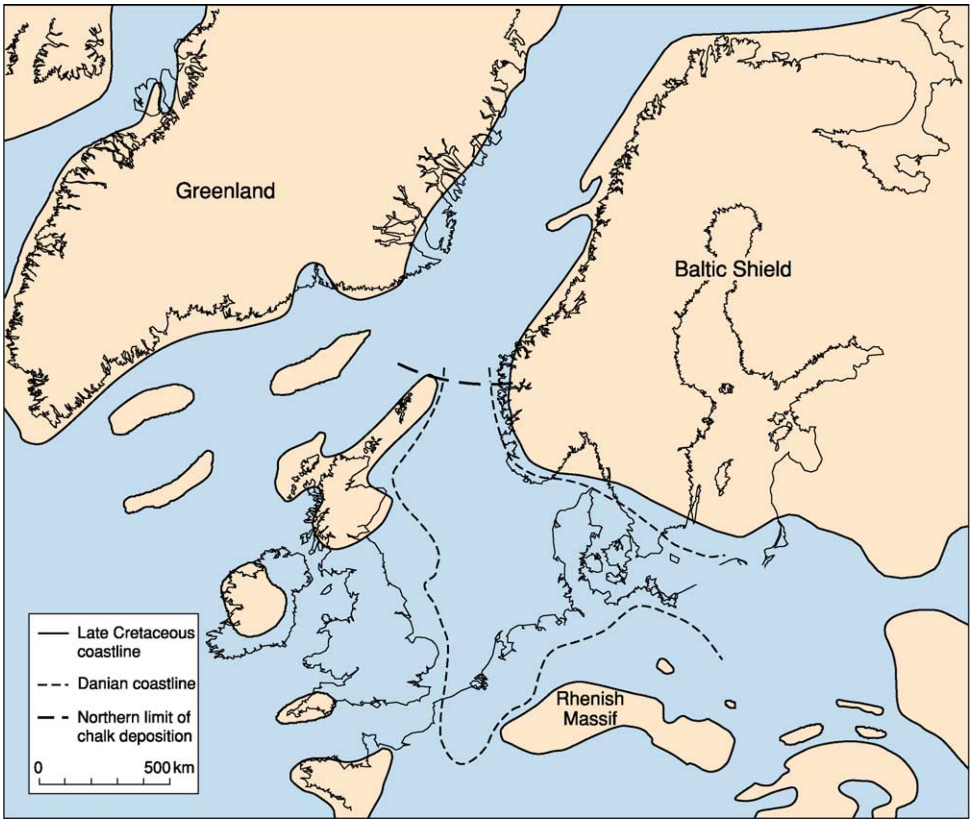


Figure 2 Late Cretaceous–Danian palaeogeography of the North European – North Atlantic region showing the land:sea distribution and the northward limit of the chalk facies in the North Sea region (at a palaeolatitude of *ca.* 50° N).

seen in most ancient limestones and is commonly the result of diagenetic modification of unstable carbonate precursors (e.g., high magnesium calcite, aragonite). In the case of chalks, however, this composition is essentially primary since coccolithophorids secrete low magnesium calcite. The chalk is thus not prone to significant diagenesis under normal marine conditions, imparting a chemical stability to the chalk that is unique amongst limestones. This is a major factor in its utility as a reservoir both for hydrocarbons and water (see below). The content of siliciclastic detritus in chalk is typically low, but clay-rich chalks are observed at certain stratigraphic and palaeogeographic positions and clay content is often instrumental in highlighting a distinctive cyclicity in the chalk succession. Primary biogenic silica may be present in the form of opaline radiolarians, diatoms and sponge spicules, but more characteristically silica in the chalk is represented by bands of nodular flint or its precursor, cristobalite lepispheres. Glauconite and phosphorite are also important authigenic minerals in chalks, particularly over structural highs or associated with hiatal surfaces.

Facies and Processes

Typical fine-grained coccolith-dominated chalk forms one, deeper-water variant within a spectrum of facies referred broadly to the chalk family (Figure 3). The faunal content increases towards the basin margins and over structural highs, and in the shallowest parts of the north-west European chalk sea coarse-grained, shallow marine carbonates were deposited. These are dominated by skeletal carbonate sands with abundant bivalves, brachiopods and echinoderms. Basinwards, this facies belt may pass into a zone dominated by bryozoan mounds before entering the area of true chalk deposition (Figure 3). The marginal facies of the chalk sea are known mainly from onshore outcrops in Denmark and southern Sweden; the Danian-age bryozoan mound complexes exposed in eastern Denmark are particularly impressive (Figure 4).

The pelagic chalk was initially deposited as an ooze consisting of coccoliths with a variable content of foraminifers and calcispheres, and a landward increasing content of invertebrate fossils, including bryozoans, echinoderms, brachiopods, and bivalves. The chalk seafloor was a unique, long-lived macrohabitat and a remarkably well-adapted, highly specialized fauna gradually developed, dominated by millimetre-sized suspension-feeding invertebrates. This reached a climax in the Late Campanian–Maastrichtian with a diversity of several thousand benthic species. Most epifaunal species were very small allowing attachment to very restricted hard substrates such as individual skeletal fragments. Other organisms developed ‘snowshoe’ strategies (a flat profile, often with long marginal spines, or hemispherical with the convex valve downwards) permitting the organism to ‘float’ on the soft substrate.

The coccolithophorid algae lived largely within the photic zone near the sea surface, and their skeletal debris settled slowly to the seafloor from suspension, most likely in the form of faecal pellets. At the sea floor, the ooze was watery with a primary porosity of 70–80%. The grain size was extremely fine, probably about 1 μm, since the coccolithophorid tests readily disaggregate into their component coccoliths and platelets. The pelagic ooze typically accumulated under well-ventilated conditions on the sea floor where sufficient oxygen was available to support a diverse fauna of burrowing benthic invertebrates – the pelagic chalks are thus characteristically intensely bioturbated. Studies of the onshore chalk exposures have revealed composite ichnofabrics that reflect the succession of diverse benthic communities that occupied the uppermost layers of the ooze as it experienced gradual dewatering and changed from a soupground to a softground (Figure 5). The trace fossils in pelagic chalks reveal much information about substrate conditions, sedimentation rates and oxygenation as well as evidence of non-deposition and the development of firmgrounds and hardgrounds. This is based on the recognition of characteristic groups of trace fossils, known as tiers, that characterize different levels in the ooze from the sea bed down to a metre or more below the sediment surface (Figure 6). The shallowest tier completely obliterated the primary fabric and only rarely are discrete trace fossils recognizable (e.g., diffuse *Planolites*). Downwards, trace fossils are better preserved and define a succession of tiers characterized by forms such as *Thalassinoides*, *Zoophycos* and *Chondrites*. Nodular chalks and hardgrounds are distinctive features of shelf-sea chalks and record decreasing rates of sedimentation and consequent increasing intensity of cementation at or near the sea floor. True hardground surfaces (i.e., cemented layers exposed on the seafloor) may show evidence of encrustation of the hardened surface by bivalves, serpulids and bryozoans and boring by sponges, algae and bivalves. The hardground surface may also be impregnated by phosphorite or glauconite.

Resedimentation of pelagic chalks has been documented on all scales from both onshore and offshore areas, and it is widely recognized that intrabasinal slides, slumps, turbidity current and debris flow deposits form an important part of the chalk depositional system (Figure 3). The largest slides occur close to tectonic inversion or salt structures and involved downslope movement of slabs of semi-lithified chalk, tens of metres thick. They may be identifiable on seismic data, but can be difficult to recognize in core since there is little or no internal deformation within the slide masses. Such allochthonous sediment slices have, however, been recognized on the basis of anomalous biostratigraphic data, a Maastrichtian interval sandwiched between Danian chalks for example. In contrast, slumps are more readily recognized by the presence of pervasive deformation

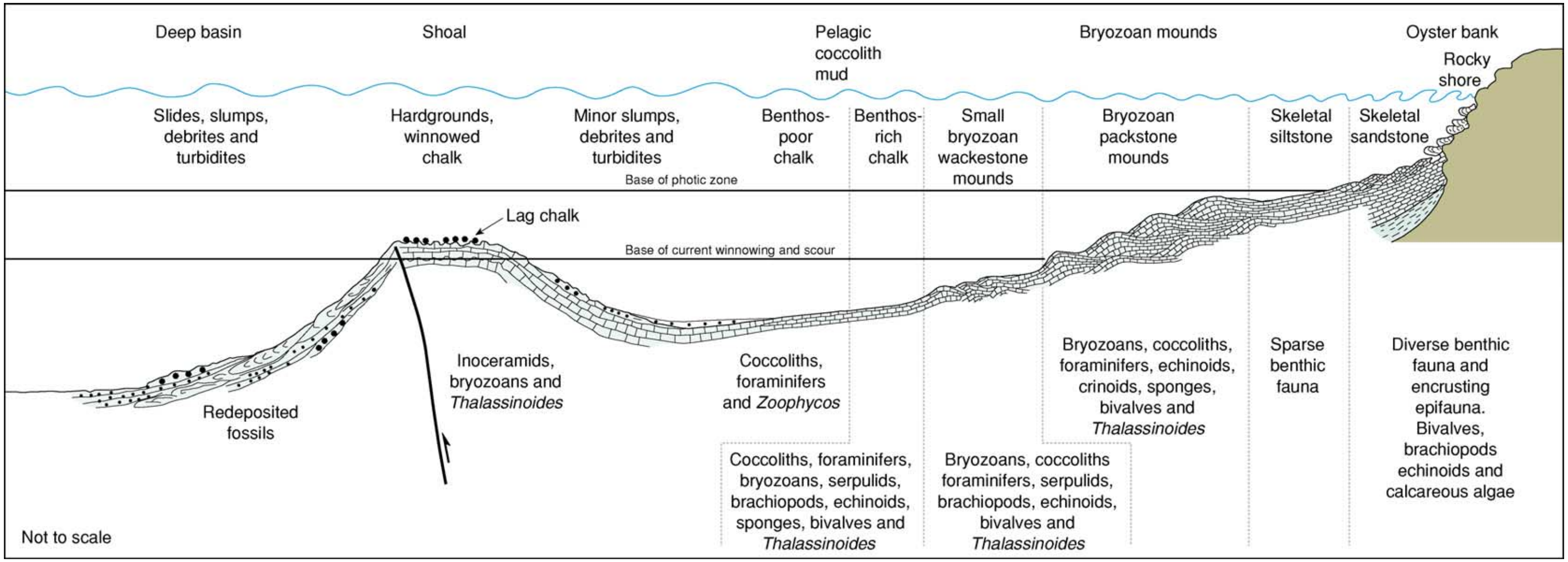

Deep basin
Shoal
Pelagic coccolith mud
Bryozoan mounds
Oyster bank
Rocky shore
Slides, slumps, debrites and turbidites
Hardgrounds, winnowed chalk
Minor slumps, debrites and turbidites
Benthos-poor chalk
Benthos-rich chalk
Small bryozoan wackestone mounds
Bryozoan packstone mounds
Skeletal siltstone
Skeletal sandstone
Base of photic zone
Lag chalk
Base of current winnowing and scour
Redeposited fossils
Inoceramids, bryozoans and *Thalassinoides*
Coccoliths, foraminifers and *Zoophycos*
Bryozoans, coccoliths, foraminifers, echinoids, crinoids, sponges, bivalves and *Thalassinoides*
Sparse benthic fauna
Diverse benthic fauna and encrusting epifauna. Bivalves, brachiopods echinoids and calcareous algae
Coccoliths, foraminifers, bryozoans, serpulids, brachiopods, echinoids, sponges, bivalves and *Thalassinoides*
Bryozoans, coccoliths foraminifers, serpulids, brachiopods, echinoids, bivalves and *Thalassinoides*
Not to scale

Figure 3 Schematic Late Cretaceous–Danian facies model for the NW European chalk sea showing the main facies belts passing from shoreline skeletal sands through bryozoan mounds to the area of 'true' chalk deposition. Based mainly on a NE–SW transect from the western margin of the Baltic Shield to the central North Sea. Reproduced from Surlyk F, Dons T, Clausen CK, and Higham J (2003) Upper Cretaceous. In: Evans D, Graham C, Armour A, and Bathurst P (eds.) *The Millenium Atlas: Petroleum Geology of the Central and Northern North Sea*, pp. 213–233. London: Geological Society.

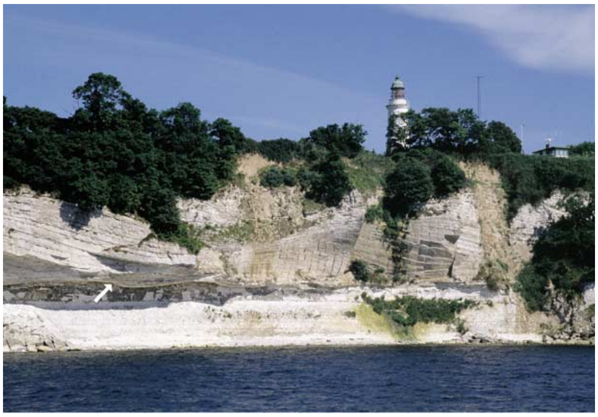


Figure 4 Upper Maastrichtian–Danian chalks exposed at Stevns Klint, eastern Denmark. The lighter coloured, lower third of the cliff (up to the prominent overhang) is the uppermost Maastrichtian. The K/T boundary (arrow) is gently undulating and the boundary clay layer is only preserved in the depressions. The Danian bryozoan-rich succession above shows well-developed mounds. Height of cliff *ca.* 40 m. Photo: F Surlyk.

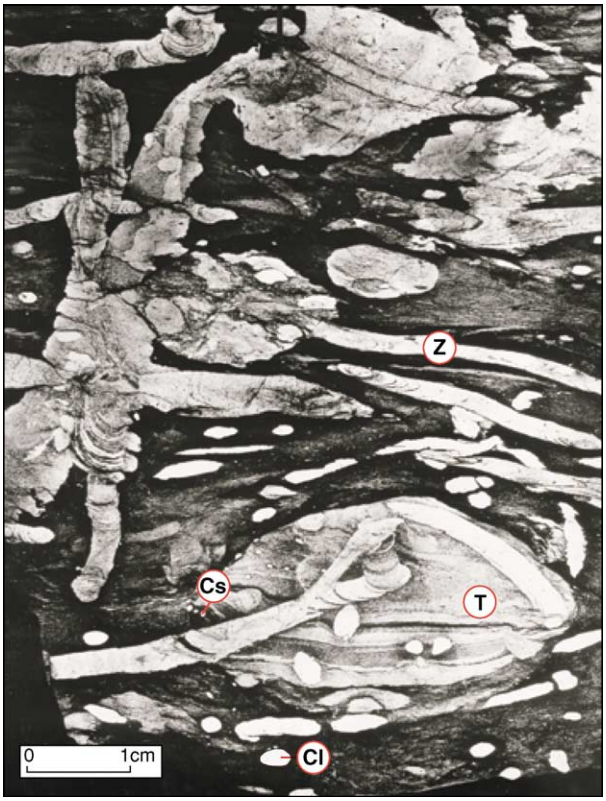


Figure 5 Photograph of slabbed core (Maastrichtian, Denmark) in which the ichnofabric is enhanced by oil staining. Note the complex cross-cutting relationships recording the overprinting of successive tiers (see Figure 6). Cl, large *Chondrites*; Cs, small *Chondrites*; T, *Thalassinoides*; Z, *Zoophycos*. Photo courtesy of RG Bromley.

structures such as isoclinal folds and stratigraphically inverted successions. Chalk debrites, comprising chalk pebbles or slabs supported in a fine-grained chalk matrix, form a significant part of the Maastrichtian–Danian succession in the North Sea Central Graben. Most resedimented chalk clasts are plastically deformed, implying that they were poorly lithified at the time of deposition. However, the presence of angular clasts in some debrites indicates that some of the material originated from lithified chalk, either from penecontemporaneous firmgrounds/hardgrounds or from exhumed more deeply buried chalks, for example at fault scarps. Sand-grade ‘classical’ turbidites are uncommon in the chalk, most likely due to the scarcity of sand- and silt-sized material, although dilute low-density turbidity currents were important in the redistribution of mud-grade sediment.

The Chalk Sea

In the Late Cretaceous, pelagic carbonate oozes extended far onto the European craton and formed the dominant facies for tens of millions of years. This was the result of a unique coincidence of global and regional factors. The chalk sedimentary record attests both to such long-term controlling factors as eustatic sea-level and regional tectonics and to the influence of short-term climatic variation controlled by orbital forcing mechanisms.

Palaeogeography

The chalk sea of north-west Europe existed for more than 35 Ma, from the Cenomanian to the Danian, at a time when global sea-level was at its highest during the Phanerozoic and relative tectonic stability prevailed in the region. Much of the north-west European craton was flooded to depths in excess of 50 m. Hinterland relief was low and potential source areas were restricted in extent so siliciclastic supply was limited and a pelagic carbonate drape accumulated, extending from a palaeolatitude of 35° N northwards to 50° N where the carbonates passed into siliciclastic muds (Figure 2). The biogenic components largely belonged to the heterozoan association that today characterizes cool-water, temperate carbonate systems; typical Cretaceous tropical organisms, such as reef corals, large foraminifers and rudist bivalves are absent or rare in the chalk of north-west Europe. However, direct latitudinal comparison with present-day seas are invalid since the Cretaceous was one of the ‘greenhouse’ phases of Earth history when equable temperatures extended further poleward than in our present ‘icehouse’ situation. The chalk sea is thus probably best characterized as ranging from warm temperate to sub-tropical, despite its mid-latitude

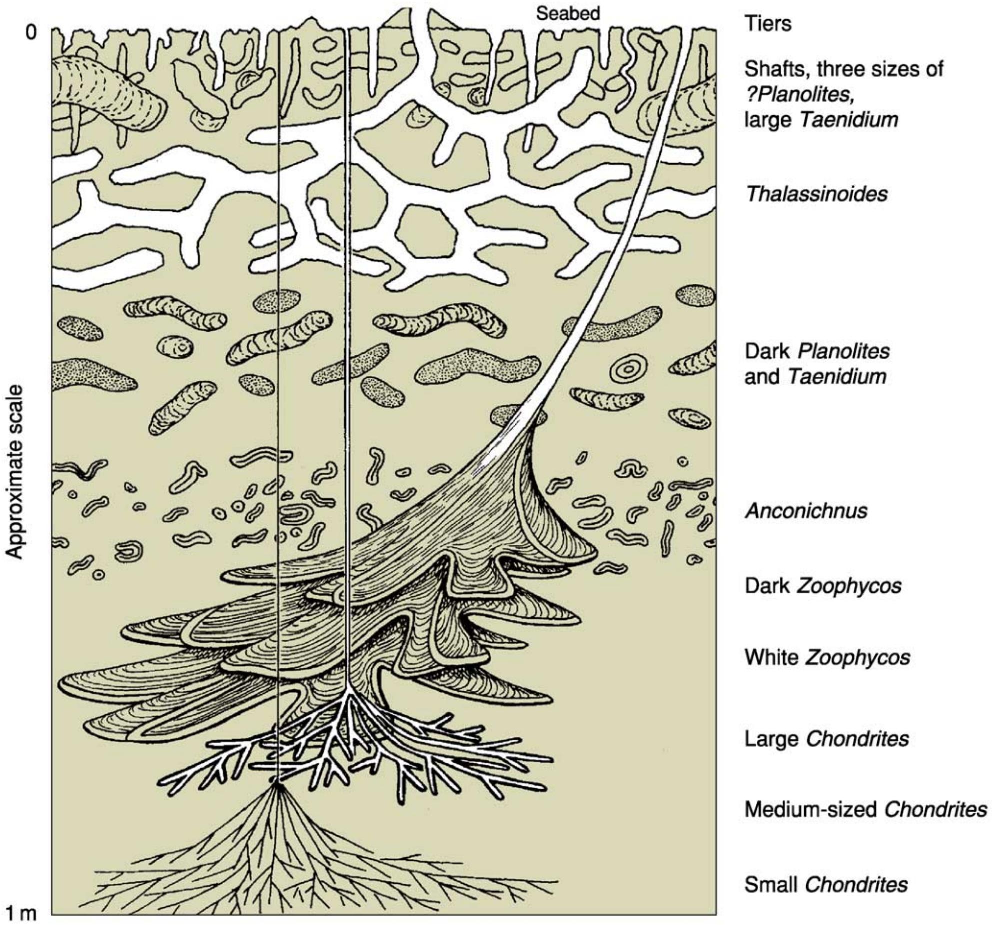


Figure 6 Schematic 'snapshot' of a Cretaceous ooze profile showing the various trace fossil tiers at different levels beneath the sea floor. Modified from Ekdale AA and Bromley RG (1991) Analysis of composite ichnofabrics: an example in uppermost Cretaceous chalk of Denmark. *Palaios* 6: 232–249.

setting. The overwhelming dominance of coccolithophorid skeletal material suggests that overall the chalk sea was a low nutrient (oligotrophic) setting. Today, shelf seas are separated from the open ocean by shelf break fronts that isolate inshore waters from the open ocean. During maximum sea-level highstand in the Late Cretaceous, the high water depths over the shelf break precluded the development of an effective shelf front and oceanic conditions extended far onto continental shelves and into epeiric seas.

Late Cretaceous pelagic sedimentation rates are estimated at 2–2.5 cm per thousand years. The Upper Cretaceous–Danian chalk succession is typically a few hundred metres thick where exposed in countries bordering the North Sea (Figure 2), but can be over a kilometre thick within major graben structures in the central North Sea and thicknesses in excess of 2 km are found in the Danish Basin. Chalk deposition was interrupted at the Maastrichtian–Danian boundary due to the mass extinction, which included coccolithophorids and foraminifers with only a few surviving species. The mass extinction severely affected all carbonate-shelled organisms and some, such as the ammonites, became totally extinct. The boundary is marked by a thin clay layer in all complete sections (Figure 4); this layer shows enrichment in iridium, forming the basis for the hypothesis that the mass extinction owed its origins to the impact of a giant meteorite. Carbonate deposition, however, rapidly resumed; the surviving microplankton and benthic invertebrates soon regained high diversities and the Danian ecosystem closely resembled that of pre-extinction, Cretaceous times.

Sea-floor relief in the NW European chalk sea was subdued and the carbonate system is best considered overall as a gently shelving ramp (Figure 3). However, significant depositional relief was developed along structures inherited from Jurassic rift events or related to localized Cretaceous inversion or salt movements. The North Sea Central Graben, for example, was a north–south-trending trough with a complex morphology formed both by the marginal slopes and by

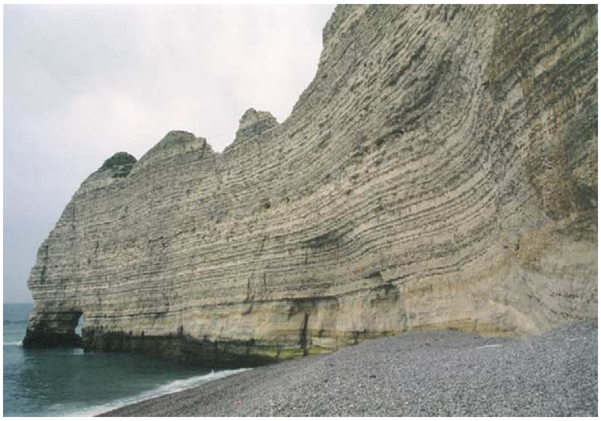


Figure 7 Cliff section, *ca.* 50 m high, at Port d'Amont, NE of Etretat, France, showing prominent chalk–flint cycles. Note slump sheet (right) and coalescing hardgrounds at beach level. Photo: F Surlyk.

intra-basinal ridges and domes along inversion axes and atop salt structures, respectively. Such relief led to sediment instability and instigated sediment slumps and gravity flows, resulting in redeposition of the coccolith ooze in deeper depocentres. The depositional relief may also have inhibited bottom water circulation and promoted the periodic development of anoxia/dysoxia in the deeper parts of the Central Graben.

In marginal settings, over intrabasinal highs and in areas of focussed, amplified bottom currents, the chalk sea floor locally developed marked depositional relief, both in the form of aggradational mounds and ridges, and erosional features. Thus, seismic data from the Danish Basin, the North Sea and onshore UK reveal ridges and valleys with a relief of up to 150 m and width of several kilometres. These features are combined constructional/erosional structures and were probably controlled by long-lived contour current systems. On a smaller scale, well-developed bryozoan-rich mounds in the Maastrichtian–Danian of eastern Denmark show amplitudes of 50–100 m, heights of 5–9 m and flanks dipping up to 20°; they show a marked asymmetry recording lateral (southwards) migration of the mounds (Figure 4). At Etretat, in northern France, Coniacian–Santonian chalks exposed in dramatic cliff sections (Figure 7) display a complex array of erosional and constructional architectures that record enhanced current activity in a tectonically constrained setting close to the margin of the chalk sea.

Cyclic Sedimentation and Orbital Forcing

Chalk successions are often overtly cyclic in nature, the typical decimetre- to metre-scale cyclicity being picked out either lithologically, as in chalk/flint and chalk/marl cycles (Figure 7), or due to changes in the fabric of the chalk, as in laminated/bioturbated cycles and those revealed by variation in the intensity or type of bioturbation. Detailed correlation of such small-scale cycles in the Cenomanian, constrained by biostratigraphy, has demonstrated their lateral persistence across the basin – at certain levels, individual cycles have been correlated from southern England, over northern Germany to southern Crimea, a distance of nearly 4000 km! This small-scale cyclicity records recurrent change in a number of interrelated factors including carbonate productivity, the balance between productivity and siliciclastic input, and bottom-water conditions, factors that are thought to have been ultimately controlled by subtle fluctuations in climate dictated by orbital fluctuations in the Milankovitch frequency band. The precession signal (mode at 21 ka) dominates, at least in the Cenomanian where the most detailed studies of small-scale cyclicity have been undertaken. Sequence-scale sea-level changes were driven by the long eccentricity cycle of 400 ka, allowing sequence stratigraphic correlation from north-west Europe to Kazakhstan and south-east India.

Chalk as a Hydrocarbon Reservoir and Aquifer

The Upper Cretaceous–Danian chalk forms significant reservoirs both for hydrocarbons, as in the North Sea Central Graben, and for groundwater, for example in Denmark, England, France and Belgium. Hydrocarbons are also produced from older, Barremian–Aptian marly chalk facies in the Danish Valdemar Field and from the Upper Cretaceous Austin Chalk in Texas. Indeed, the first hydrocarbon discovery in the North Sea, in 1966, was in chalk – the A-1 well of the Danish Kraka Field. Since that first discovery, the number of chalk fields in the North Sea has increased to nearly 30, containing almost 5 billion barrels of recoverable oil and more than 16 000 billion cubic feet of gas. Production of hydrocarbons from these fields is still a major challenge, since the chalk forms a unique family of very fine-grained reservoir rocks, characterized by high matrix porosity and low permeability, differing from most other carbonate reservoirs. The minute coccolith platelets making up the chalk are typically 0.5–1 μm across, and pores and pore throats are on the order of a micron in size, reducing the matrix permeability to a few millidarcy even at porosities of more than 35%.

The North Sea chalk is composed almost entirely of the stable low magnesium variant of calcite and has not been subjected to freshwater diagenesis.

Diagenetic modifications of the chalk were controlled by early processes at or near the sea floor and by the later burial history. The chalk ooze had an initial porosity of 70–80%, but dewatering due to bioturbation rapidly reduced porosity to about 50%. The porosity declined further to 35–40% at depths of around 1000 m due to compaction. At greater burial depths, the effects of pressure solution became more important and under normal conditions the matrix porosity is around 10% at burial depths of 2000 m. The permeability of the chalk is directly related to the porosity, so reduced porosity also means lowering of permeability and thereby hydrocarbon productivity (Figure 8). The relationship between porosity and permeability is not constant, but varies with the stratigraphic age of the chalk (Figure 8). The best chalk reservoir properties in the North Sea are found in the Maastrichtian Tor Formation where the matrix permeability for a given porosity is almost 10 times that of the Danian Ekofisk Formation. As seen in the figure, Lower Cretaceous chalks form even poorer reservoirs due to their high content of clay. The differences in reservoir properties between the Maastrichtian and Danian chalks are not fully understood, but it has been suggested that a change in the coccosphere flora across the Cretaceous–Palaeogene boundary resulted in changes in the detailed texture of the coccolith-rich sediment thereby affecting the size/geometry of pores and pore throats.

Other factors that are known to influence the quality of the chalk reservoirs are the content of clay and silica, and the mode of deposition. It was long a common belief that reworking of the chalk was the key to preservation of anomalously high porosities, a belief possibly driven by the fact that allochthonous chalks form the main reservoirs in the Norwegian North Sea sector. This view has changed, however, and although resedimentation locally has a positive effect on porosity preservation this is not always the case. On a metre-scale, however, it has been shown that facies have a major control on porosity in cyclically interbedded successions of bioturbated and laminated chalk, the highest porosities being in the laminated units. The development of firmgrounds and hardgrounds is also important since early cementation reduces primary porosity, so these layers form characteristic low-porosity zones in the chalk and may create barriers to fluid flow in the reservoir. On structural highs, several hardgrounds may coalesce, leading to a major negative effect on reservoir quality. This is particularly true for the complex hardground that developed at the Maastrichtian–Danian boundary over most of the North Sea highs. The best reservoir properties are found in the purest chalks, and the Maastrichtian Tor Formation is particularly pure with less than 5%, commonly only 1–3%, non-carbonate fraction. The Danian Ekofisk Formation has a more variable content of clay and silica throughout the North Sea; the lower Danian forms a more clay- and silica-rich (up to 20%), non-reservoir interval known as the 'Danian tight zone'. The reason for the strong negative influence of clay on reservoir quality seems to be that its presence inhibits the growth of early cement between carbonate grains, thereby preventing early lithification. As a consequence, the more weakly lithified clay-rich intervals are more easily compacted during deeper burial.

North Sea chalks are extremely fine-grained and have low permeabilities so that porosities of more than 25% are required to allow commercial

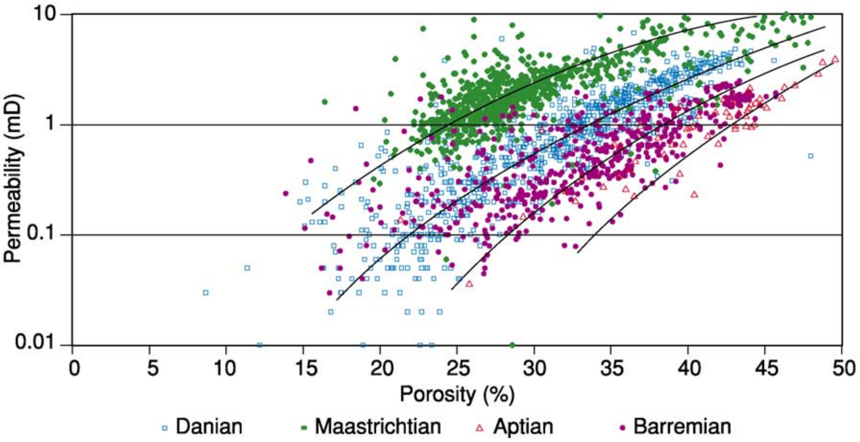


Figure 8 Plot showing the relationship between porosity and permeability for four different chalk units in the North Sea (Danish sector): the Barremian Tuxen Formation, the Aptian Sola Formation, the Maastrichtian Tor Formation and the Danian Ekofisk Formation. Reproduced with permission from Jakobsen F, Ineson JR, Kristensen L, and Stemmerik L (2004) Characterization and zonation of a marly chalk reservoir: the Lower Cretaceous Valdemar Field of the Danish Central Graben. *Petroleum Geoscience* 10: 21–33.

production. Since matrix porosity is reduced to 10% at burial depths of around 2000 m, preservation of such substantial porosity requires unusual conditions, and the North Sea chalk fields are all located in areas with significant overpressure. Another process that may result in the retention of high porosities is early oil migration into the reservoir, since the oil prevents further cementation during burial. The overpressured chalk reservoirs maintain porosities of up to 45–50% and matrix permeabilities in the 3–10 mD range at burial depths of 1700–3300 m. The effective permeability is commonly much higher due to fracturing since many North Sea chalk fields are localized over salt structures.

During the last two decades, the recovery factors in the chalk fields have increased significantly as the result of horizontal drilling, water injection and stimulation of artificial fractures. In the Danish Dan Field, the initial recovery factor of 10% has increased to 35%, and in the Norwegian Ekofisk Field estimates have increased from 18% to 38%.

In north-west Europe, chalks and other Upper Cretaceous–Danian limestones form important aquifers, both where they outcrop at the surface and in areas where they are covered by thin Cenozoic or Quaternary deposits. The chalk is the most important unconfined aquifer in the Paris Basin, both in terms of areal extent and size of resources, and production from the aquifer is *ca.* $10^9\,m^3$ per year. In Denmark, about 35% of the annual water consumption (*ca.* $0.4 \times 10^9\,m^3$) is derived from the chalk, largely in the north and east of the country, whereas south-east England is particularly dependent on chalk aquifers, accounting for 55% of the groundwater utilized in the UK. The uppermost 50–60 m of the saturated chalk forms the principal aquifer since water flow at reasonable rates relies on the presence of open, commonly solution-modified fractures and fissures. The permeability of fissured chalk is 10^{-5}–$10^{-3}\,ms^{-1}$ whereas that of the chalk matrix is in the order of 10^{-9}–$10^{-8}\,ms^{-1}$ and thus has a negligible contribution to the transmissivity of the aquifer. As in hydrocarbon reservoirs, therefore, the matrix porosity provides the volume for storage of groundwater and the fractures provide the distribution system that drains the matrix and allows the water to flow.

See Also

Diagenesis, Overview. **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills. **Mesozoic:** Cretaceous. **Petroleum Geology:** Overview. **Sedimentary Environments:** Reefs ('Build-Ups'). **Sedimentary Rocks:** Chert. **Trace Fossils**.

Further Reading

Bromley RG and Ekdale AA (1986) Composite ichnofabrics and tiering of burrows. *Geological Magazine* 123: 59–65.

Downing RA, Price M, and Jones GP (eds.) (1993) *The Hydrogeology of the Chalk of North-West Europe,* p. 300. Oxford: Clarendon Press.

Ekdale AA and Bromley RG (1984) Comparative ichnology of shelf-sea and deep-sea chalk. *Journal of Paleontology* 58: 322–332.

Ekdale AA and Bromley RG (1991) Analysis of composite ichnofabrics: an example in uppermost Cretaceous chalk of Denmark. *Palaios* 6: 232–249.

Gale AS (1995) Cyclostratigraphy and correlation of the Cenomanian stage in Western Europe. In: House MR and Gale AS (eds.) *Orbital Forcing Timescales and Cyclostratigraphy,* Geological Society Special Publication 85, pp. 177–197. London: Geological Society.

Hancock JM (1976) The petrology of the chalk. *Proceedings of the Geological Association* 86: 499–535.

Hancock JM (1993) The formation and diagenesis of chalk. In: Downing RA, Price M, and Jones GP (eds.) *The Hydrogeology of the Chalk of North-West Europe,* pp. 14–34. Oxford: Clarendon Press.

Hay WW (1995) Cretaceous paleoceanography. *Geologica Carpathica* 46: 257–266.

Kennedy WJ (1987) Late Cretaceous and early Palaeocene Chalk Group sedimentation in the Greater Ekofisk Area, North Sea Graben. *Bulletin du Centre Recherche Exploration Production Elf-Aquitaine* 11: 91–126.

Kennedy WJ and Garrison RE (1975) Morphology and genesis of nodular chalks and hardgrounds in the Upper Cretaceous of southern England. *Sedimentology* 22: 311–386.

Quine M and Bosence D (1991) Stratal geometries, facies and sea-floor erosion in Upper Cretaceous chalk, Normandy, France. *Sedimentology* 38: 1113–1152.

Scholle PA (1977) Chalk diagenesis and its relation to petroleum exploration: oil from chalks, a modern miracle? *American Association of Petroleum Geologists Bulletin* 61: 982–1009.

Surlyk F (1997) A cool-water carbonate ramp with bryozoan mounds: Late Cretaceous–Danian of the Danish Basin. In: James NP and Clarke JAD (eds.) *Cool-Water Carbonates. SEPM (Society for Sedimentary Geology) Special Publication,* 56, pp. 293–307. Tulsa, Oklahoma: SEPM.

Surlyk F, Dons T, Clausen CK, and Higham J (2003) Upper Cretaceous. In: Evans D, Graham C, Armour A, and Bathurst P (eds.) *The Millenium Atlas: Petroleum Geology of the Central and Northern North Sea,* pp. 213–233. London: Geological Society.

Chert

N H Trewin and S R Fayers, University of Aberdeen, Aberdeen, UK

Introduction

The term chert is currently used for any microcrystalline siliceous rock containing only minor impurities. In the early nineteenth century, hornstone and chert were names used for rather non-descript splintery siliceous rocks, but chemically similar material with specific features of colour and texture received a plethora of names, largely based on ornamental value. It is usually the minor impurities that impart colour, such as the red of haematite in jasper, and the green of chrysoprase assigned to nickel. Differing textural features produce the banding seen in agates. Porcellanite is a white variety containing clay inclusions and resembling porcelain, and flint refers to the nodules from the Cretaceous Chalk that produce superb conchoidal fractures, and were certainly one of the earliest geological industrial materials, being used by Palaeolithic man.

Cherts occur in low abundance in a variety of geological settings, and have a variety of origins. In many cases, the chert product is the result of maturation by time and the diagenesis of precursor silica phases. The diagenetic pathway of silica is controlled by phase solubility, a function of crystal structure and size, and usually proceeds from amorphous opal through intermediate stages to quartz by dissolution–reprecipitation reactions. Amorphous silica may have its origin in biogenic skeletal material (radiolaria, diatoms, siliceous sponge spicules), or in volcanic glass, or in sinter deposited from hydrothermal solutions.

Chert Composition

Chert is seldom uniform in texture and, in young material, transitions from opal are frequent. The grain size and texture are related to the origin of the silica and the diagenetic history of the rock.

The Main Constituents of Chert

Microquartz Microquartz is the main constituent of chert. It occurs as equant crystals in a solid mosaic (Figure 1), and often has a porous, spongy texture. Spongy forms are frequently dark in thin section due to included pores. Crystals are generally less than 5–20 μm in diameter. Microcrystalline quartz comprising individual crystals below the resolution of a standard petrographical microscope is termed cryptocrystalline.

Megaquartz Megaquartz occurs as a mosaic quartz fill to cavities and veins in chert, often displaying a drusy fabric (Figure 1A and B). It may also occur as a replacement fabric in carbonate fossils or fossil wood. Crystals are tens to hundreds of micrometres in diameter.

Chalcedony Chalcedony has a fibrous texture, with fibres tens to hundreds of micrometres in length. It occurs in radiating spherulitic textures and has overlays, frequently brown and zoned in thin section. Chalcedonic quartz is length-fast. Quartzine is length-slow. Leutecite is intermediate between chalcedonic quartz and quartzine, with the fibre axis oriented approximately 30° to the crystallographic *c*-axis. Chalcedonic quartz and quartzine both form cement and replacement fabrics (Figure 1C and D), whereas leutecite typically occurs as a replacement fabric in carbonate fossils. Together with quartzine, other fibrous types, namely zebraic chalcedony and microflamboyant quartz, are most common in chert-replaced evaporites.

Opal Opal occurs as amorphous to cryptocrystalline forms. The crystallinity increases from opal-A to opal-CT to opal-C; these phases can be distinguished by X-ray diffraction. Opal-A is amorphous, opal-CT comprises disordered interlayers of cristobalite and tridymite, and opal-C comprises cristobalite.

Silica Solubility and Precipitation

Biogenic opal-A has a solubility of 120–140 ppm in normal marine sediment pore water, cristobalite 25–30 ppm, and quartz 6–10 ppm. The dissolution of opal-A generally results in supersaturation with respect to opal-CT, which is precipitated in preference to quartz. Opal-CT typically consists of small bladed crystals that form spherical lepispheres 5–10 μm in diameter. Quartz precipitation generally takes place from dilute solutions over longer periods of time. The solubility in water of both amorphous silica and quartz increases rapidly above pH 9, and thus a sharp reduction in alkalinity from $\mathrm{pH} > 9$ results in silica precipitation.

The precipitation of silica, particularly around hot springs, occurs due to the evaporation of silica-saturated water. Silica solubility is higher in hot water, and thus cooling results in silica deposition. Boiling is also a trigger for silica precipitation in hydrothermal systems.

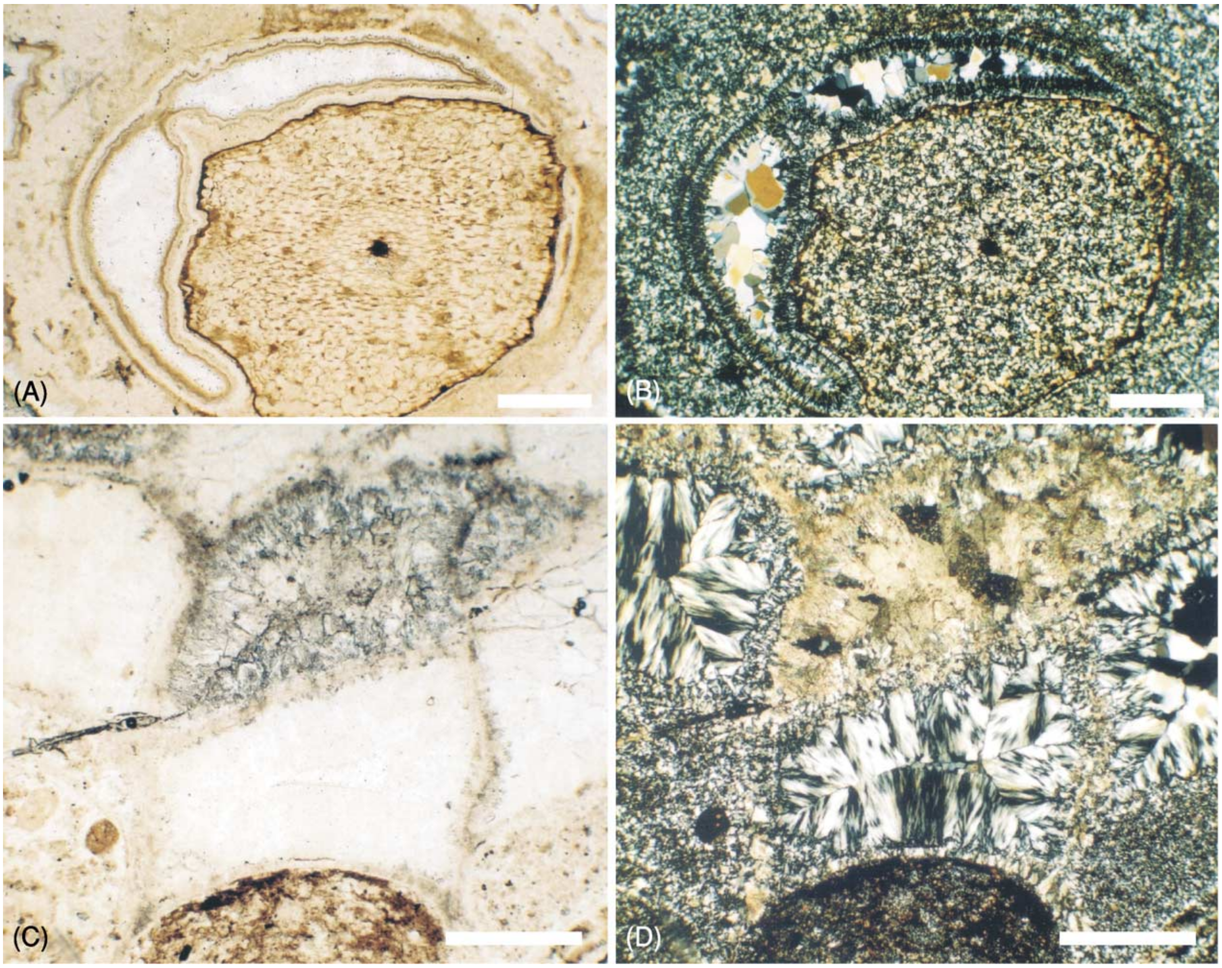


Figure 1 Examples of chert constituents and textures in thin section. (A) Transverse section through a shrunken plant stem (*Rhynia gwynne-vaughanii*) in Early Devonian Rhynie Chert (Rhynie, Scotland), observed in PPL (plane polarized light) (plane polarized light). (B) Same view as (A) in XPL (cross polarized light). The plant and surrounding organic-rich matrix is preserved in microcrystalline quartz. The void created by the shrunken stem is lined by zoned chalcedonic quartz and occluded by later megaquartz cement. (C) Cross-section through dissepiments of the coral *Dibunophyllum* viewed in PPL. Hob's House Coral Bed, Viséan, Carboniferous, Derbyshire, UK. (D) Same view as (C) in XPL. Sediment-filled voids are replaced by microcrystalline quartz (below); those above are occluded by chalcedonic quartz cement. One dissepiment is occluded by calcite spar (top, centre). Scale bars = 500 μm.

Textures

In many cherts, there is a clear textural sequence in which microquartz, a transformation of original silica gel or opal, is overlain by chalcedonic layers, and the remaining space is filled by megaquartz. The full maturation sequence of opal to quartz is not the only route to chert; amorphous opal can transform directly to quartz and, in diagenetic cherts, microquartz or chalcedony may be the primary crystallization phase.

Sources of Silica

Biogenic

Two groups of common micro-organisms and one group of sponges build skeletons of opal, thus fixing silica from solution. The main groups with their ranges and environments are as follows:

- Radiolaria – Cambrian–Holocene – plankton, marine.
- Diatoms – Jurassic–Holocene – plankton, marine.
- Diatoms – Tertiary–Holocene – planktic–benthic, non-marine.
- Siliceous sponges – Cambrian–Holocene – benthic, dominantly marine.

These organisms can extract silica from water that is undersaturated with respect to silica by one or two orders of magnitude. For example, diatoms have been recorded to reduce the Si content of aquarium water from 0.95 ppm to 0.075 ppm. Thus, these organisms are important in fixing silica in a particulate form that can be sedimented, and provide a concentration of metastable biogenic opal. This opal can then be converted to chert *in situ* to produce bedded cherts, or be dissolved and transported to a site of diagenetic deposition to produce chert cement and replacement nodules.

Volcanic

There is a strong association between chert and submarine volcanics in the geological record. Thus, it has been postulated that silica is derived from the devitrification of volcanic glass, leading to the production of smectite and silica. However, it is likely that the higher Si contents of water in volcanically active areas result in population explosions ('blooms') of diatoms and radiolaria, and that the silica is fixed by organisms.

Hydrothermal Silica

Hydrothermal systems, developed in the waning phase of volcanicity, produce hot springs and geysers that frequently deposit silica in the form of amorphous siliceous sinter. The silica is dissolved from hot rocks at depth and, as water is circulated through the convecting hydrothermal system, hot silica-saturated water is brought to the surface where silica is deposited due to cooling.

Silica Precipitation in Lakes

The sources of silica resulting in cherts in non-marine lakes are various. In sediment-starved Tertiary to Holocene lakes, diatoms can accumulate to form a siliceous sediment (diatomite) that can be converted to chert through time and diagenesis. Lakes (e.g., Lake Magadi) in volcanically active areas of the African Rift Valley contain sodium carbonate brines with $\mathrm{pH} > 10$. Silica is leached from volcanic rocks and Si concentrations can rise to 2500 ppm. Seasonal evaporation and dilution of the brine by river waters causes the deposition of hydrated sodium silicates that are converted to chert during diagenesis. In the Coorong region of South Australia, the pH in some Mg-rich carbonate lakes can rise above pH 10 due to algal photosynthesis. Silica is derived by the corrosion of detrital minerals, resulting in Si supersaturation of the lake waters. Subsequently, the silica is deposited in lake carbonates as a gel, giving the potential for conversion to chert during diagenesis.

Occurrence of Chert

There are a number of modes of occurrence of chert, the most common, and volumetrically the most important, being bedded cherts and nodular cherts in limestone sequences.

- Bedded cherts in ocean basins.
- Nodular cherts in limestone sequences.
- Cherts of hydrothermal origin, both surface and subsurface.
- Cherts in lake basins.
- Silcrete, chert in palaeosols.
- Silicified wood.

Bedded Cherts

Bedded cherts have been formed through the burial and diagenesis of siliceous oozes throughout Phanerozoic time (Figure 2). However, the Palaeozoic is dominated by radiolaria, and diatoms do not make a significant contribution until the Late Mesozoic. There are also extensive bedded cherts in the Precambrian, at a time from which no silica-secreting organisms are known. Thus, it is pertinent to consider the mechanisms and environments of accumulation of siliceous oozes through time.

At the present time, siliceous oozes are accumulating in deep ocean basins, in areas starved of detrital supply. A broad band of siliceous diatom-dominated deposits surrounds Antarctica, and similar deposits are accumulating between North America and Asia in the northern Pacific to the south of the Aleutian Island chain. An equatorial belt of radiolarian-dominated ooze is present in the Pacific and Indian Oceans. Drilling by the Deep Sea Drilling Project has shown that, in some oceanic areas, the siliceous oozes are converted at depth to bedded cherts, the chert generally being of Tertiary age.

The conditions considered to be favourable for the accumulation of siliceous ooze are summarized below (Figure 3).

- High organic productivity in surface waters due to upwelling of nutrient-rich oceanic currents.
- Lack of significant input of land-derived detritus that would dilute the deposit. Such material is carried by the wind from deserts, and by ocean currents from sources of clastic input.
- Limited presence of calcareous plankton. Calcareous oozes are accumulating at present at rates of 10–50 mm Ka^{-1}, compared with <10 mm Ka^{-1} for siliceous oozes. Modern siliceous oozes tend to occur in oceans below the carbonate compensation depth (CCD), the depth at which carbonate dissolution balances the oceanic carbonate rain. The CCD varies in the oceans, but is generally at about 4 km. Siliceous oozes are associated with, and diluted by, red clays derived from aeolian desert dust and volcanic fallout.

Siliceous oozes can also accumulate in shallower oceanic settings if conditions are suitable. Small,

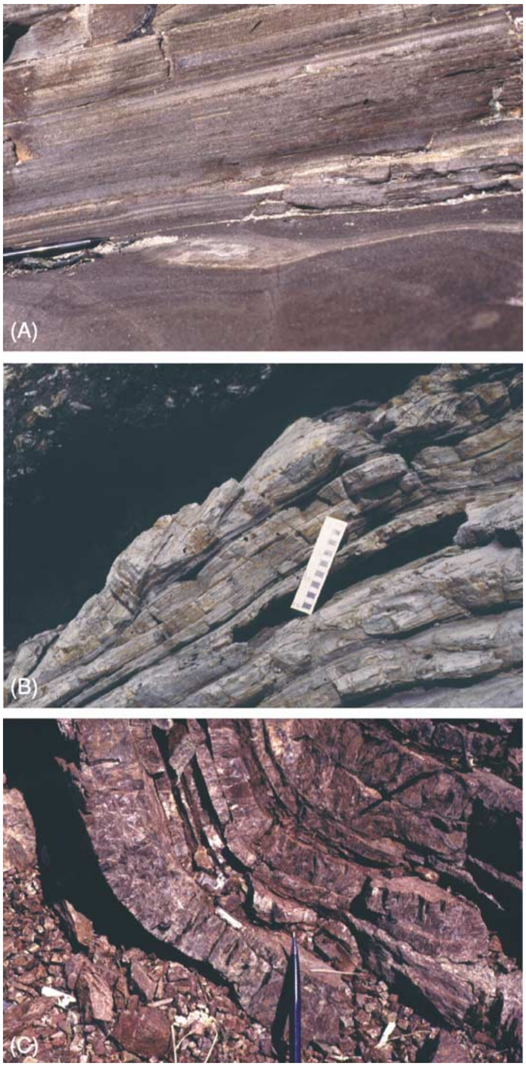


Figure 2 Bedded cherts. Laminated diatomite (A) and laminated diatomite with slump fold (B). Monterey Formation, Miocene, Gaviotta Beach, near Ventura, California, USA. (C) Deformed bedded radiolarian cherts within the Khabarovsky Complex, exposed on the banks of the Amur River in the city of Khabarovsk, Russian Far East. These Upper Triassic–Lower Jurassic cherts are associated with black mudstone and metabasalt units of the same age, and are tectonically interleaved with Cretaceous turbidite units. This complex association formed in a Late Mesozoic accretionary complex during collision of the North China Block with Siberia. Photographs courtesy of David Macdonald.

sediment-starved basins and submerged carbonate platforms on passive continental margins are potential sites for deposition when upwelling currents bring nutrients and calcareous plankton is scarce. The Gulf of California (<1.5 km deep) is an area in which siliceous oozes are accumulating in association with distal turbidites and organic-rich shales.

Tertiary bedded chert Tertiary bedded chert derived from diatomaceous ooze occurs in the Miocene–Pliocene of the Pacific margin, and formed in small rifted and back-arc basins where there was strong nutrient supply, high phytoplankton activity, and a lack of detrital sediment input. The Monterey Formation (Miocene, California) contains chert derived from diatomaceous sediments (Figure 2A and B), and is associated with hydrocarbon source rocks. The maturation process from diatomaceous sediment to chert involves porosity loss to the extent that the cherts can form a diagenetic reservoir seal at depth.

In the Mediterranean, cherts occur in the Late Miocene at the time when the area was characterized by small, restricted basins in the build-up to the Messinian salinity crisis at the end of the Miocene. In general, cherts of Tertiary age are widespread in oceanic areas, and occur in a wide variety of settings satisfying the conditions described above.

Mesozoic and Palaeozoic bedded cherts Mesozoic and Palaeozoic bedded cherts can be divided into those that are associated with volcanics, usually of ocean floor origin, and those that have no relationship to volcanicity. Prior to the expansion of planktonic diatoms in the Late Mesozoic, the radiolaria were the main contributors to siliceous ooze. The major calcareous planktonic organisms, coccoliths and planktonic foraminifera, did not appear in abundance until the Late Mesozoic. Thus, it can be postulated that siliceous radiolarian ooze would have been more abundant in the Palaeozoic and Early Mesozoic, and that radiolaria provided the main source for biogenic silica. It is likely that radiolarian ooze accumulated over a greater depth range, extending into shallower water above the current level of the CCD, in the absence of, or at least the reduction of, the diluting effect of calcareous plankton.

Volcanic association There is a common association of bedded cherts with black shales, pillow lavas, and volcaniclastic rocks. In some cases, sheeted dykes and ultramafic rocks are present, and the whole assemblage is typical of the ophiolite suite: a preserved ocean floor succession. The cherts are usually dark in colour, and may contain recognizable ghosts of radiolaria (Figure 4). The presence of radiolaria points to a biogenic origin for the silica, although elevated Si in seawater from the volcanics may have been responsible for the proliferation of radiolaria. Mesozoic examples occur in the Troodos Massif in Cyprus, the Franciscan of California, and the Khabarovsk area of the Russian Far East (Figure 2C). Palaeozoic examples associated with the margin and closure of the Iapetus Ocean occur in the Ordovician of Scotland, from the Girvan–Ballantrae area in the

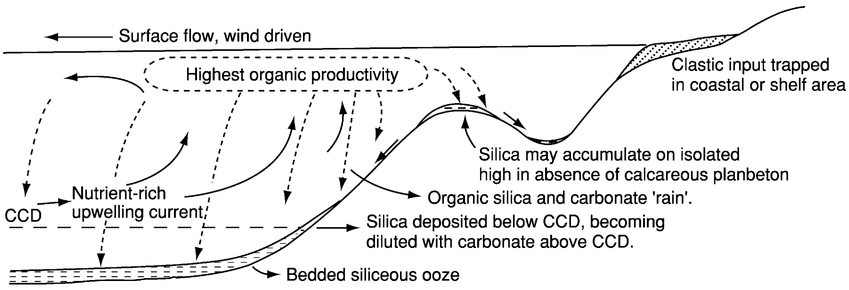

Surface flow, wind driven
Highest organic productivity
Clastic input trapped
in coastal or shelf area
Silica may accumulate on isolated
high in absence of calcareous planbeton
Organic silica and carbonate 'rain'.
Nutrient-rich
upwelling current
CCD
Silica deposited below CCD, becoming
diluted with carbonate above CCD.
Bedded siliceous ooze

Figure 3 Diagram illustrating the production and deposition of organic silica under the influence of an upwelling nutrient-rich current at an ocean margin. CCD, carbonate compensation depth. Not to scale.

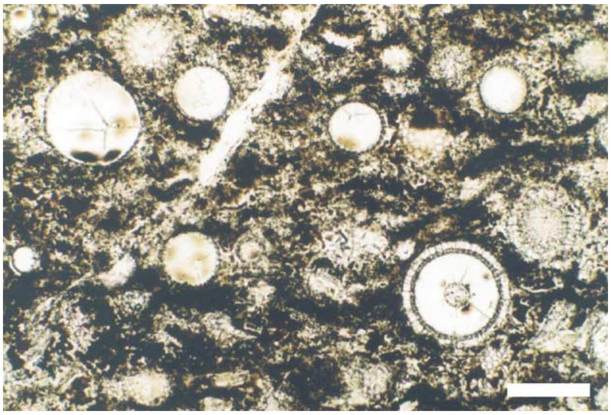


Figure 4 Photomicrograph of radiolarian chert in thin section. Viewed in PPL (plane polarized light). Scale bar = 250 μm.

west, through to Stonehaven in the east, where the chert is present as jasper. Similar developments occur in Newfoundland (Canada) and Maine (USA). The jasper in these instances lacks recognizable biogenic material and is frequently attributed to the hydrothermal alteration of sediments and hydrothermal vent deposits. It is not always possible to distinguish biogenic and volcanic/hydrothermal origins.

No volcanic association Many Palaeozoic and Mesozoic cherts have no volcanic association, and occur with black shales, pelagic limestones, and turbidites. Passive continental margins with small rifted basins are a common setting. Water depths were typically less than oceanic. Radiolarian cherts in the Lower Carboniferous of south-west England and Germany provide an example. Submarine rises, starved of land-derived detritus, formed a typical location. Resedimentation of siliceous material from shallow-water areas by turbidity currents provided a concentration mechanism in deeper basinal areas. The Caballos Formation of the Marathon Basin, Texas, USA, is a widespread bedded chert (part termed novaculite). It is of Devonian age and marine in origin, containing sponge spicules and radiolaria. The chert is interbedded with shale, and some sandstone beds and chert conglomerates are present. Arguments have been presented for both shallow marine to restricted lagoonal and deep marine environments.

Precambrian cherts Precambrian cherts are abundant, particularly in association with Precambrian 'iron formations'. However, no silica-secreting organisms are known from the Precambrian and it cannot be assumed that Precambrian cherts have a biogenic origin. There is a concentration of these deposits dated at around 2 Ga, and these cherts contain fossils of coccoid and filamentous bacteria. In the absence of silica-fixing organisms, it is likely that Si concentrations in water were high, and Si precipitation may have occurred initially as a gel in the large iron formation basins. It is possible that photosynthetic cyanobacteria played a significant role in silica deposition as in Coorong-type lakes. Texturally, the presence of algal stromatolites, ooids, and intraclasts implies shallow water, and Si replacement of carbonates. Shallow extensive shelves and large lake basins are postulated environments.

Classic examples at around 2 Ga are the Gunflint Chert of Ontario (Canada) (Figure 5), and the Biwabik Chert of South Africa. These deposits may mark the period in the Earth's history when oceans were changed from reducing to oxygenated conditions by the action of photosynthetic bacteria, and consequently vast amounts of iron were deposited that had previously been held in solution in ocean waters. Even older, at about 3.5 Ga, are the cherty rocks of the north pole region in Western Australia containing silicified stromatolites that are possibly the oldest morphological fossils on Earth (there is older geochemical evidence).

Nodular Cherts

Chert is common as a nodular and, occasionally, 'bedded' replacement feature in limestone sequences,

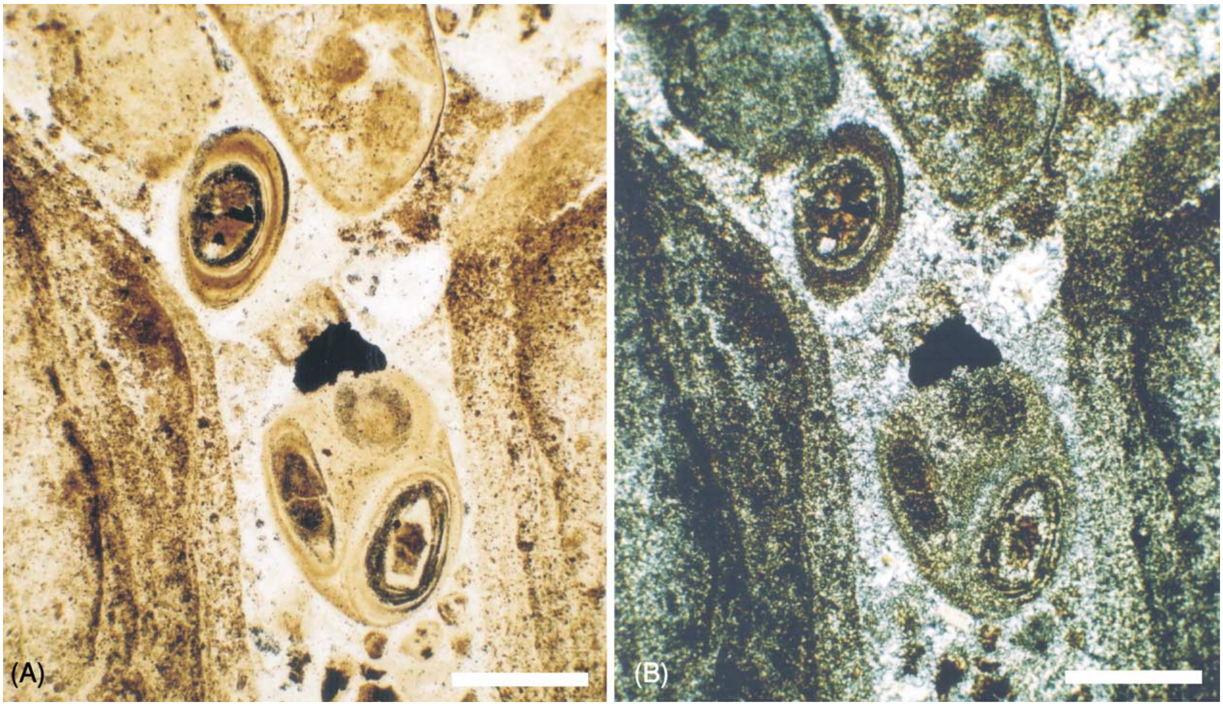


Figure 5 Photomicrographs of Gunflint Chert in thin section showing chert-replaced intraclasts and ooliths within a crack between stromatolites. (A) Viewed in PPL (plane polarized light). (B) Viewed in XPL (cross polarized light) showing replacement mainly in the form of microcrystalline quartz. Precambrian, Ontario, Canada. Scale bars = 500 μm.

and occurs to a lesser extent in mudstones and evaporites. The apparently bedded cherts in this situation are the result of the replacement of original sedimentary beds by chert (Figure 6). The source of the silica is considered to be biogenic, with the dissolution products of biogenic opal being redistributed in solution and precipitated as cement and replacement during diagenesis. The replacement nodules show a great variety in form, ranging from irregular forms with smooth curved margins (Figure 7), to more tabular and diffuse cherts seen in Carboniferous limestones, and the generally spherical nodules (geodes) representing the replacement of original anhydrite nodules.

Both chalcedony and microquartz are present (Figure 6), and there is frequently evidence for the direct precipitation of quartz in the rock in the form of isolated bipyramidal crystals. The general process of formation of the nodules involves the dissolution of biogenic opal, present in low abundance in the deposited sediment (*ca.* 1%). The mobilized silica is then deposited at suitable nucleation sites, probably as opal-CT. Such sites are controlled by rock texture and biogenic content; hence, silica deposition may favour specific beds. The opal-CT fills the pore space and replaces carbonate, and, with burial, is converted to chert. The silicification appears to be a relatively early diagenetic event, taking place during shallow burial.

In marine phreatic conditions, silica precipitation and replacement of carbonate tend to occur along redox boundaries between aerobic surface sediments and underlying sediments dominated by sulphate-reducing bacteria. The degradation of organic material by sulphate-reducing bacteria releases carboxyl and sulphide ions. Many carbonate sediments contain very little iron; therefore, very little sulphide is precipitated as pyrite. The rest is hydrolysed to hydrogen sulphide which then diffuses to more oxic conditions. Oxidation produces sulphate and hydrogen ions; the former diffuse back into the sulphate reduction zone, whilst the increased acidity causes carbonate dissolution at the redox boundary. The high concentration of carbonate ions, organic matter, and the reduced pH promote silicification.

Early silicification may also take place in emergent areas where marine pore waters in carbonate sediments mix with meteoric pore waters. In these ‘mixing zones’, the mixing of waters with suitable differences in $P\mathrm{CO}_2$ provides ideal conditions for carbonate dissolution with contemporaneous silica replacement and precipitation.

The concentration of chert replacement nodules at specific horizons, often on a basin-wide scale, can be

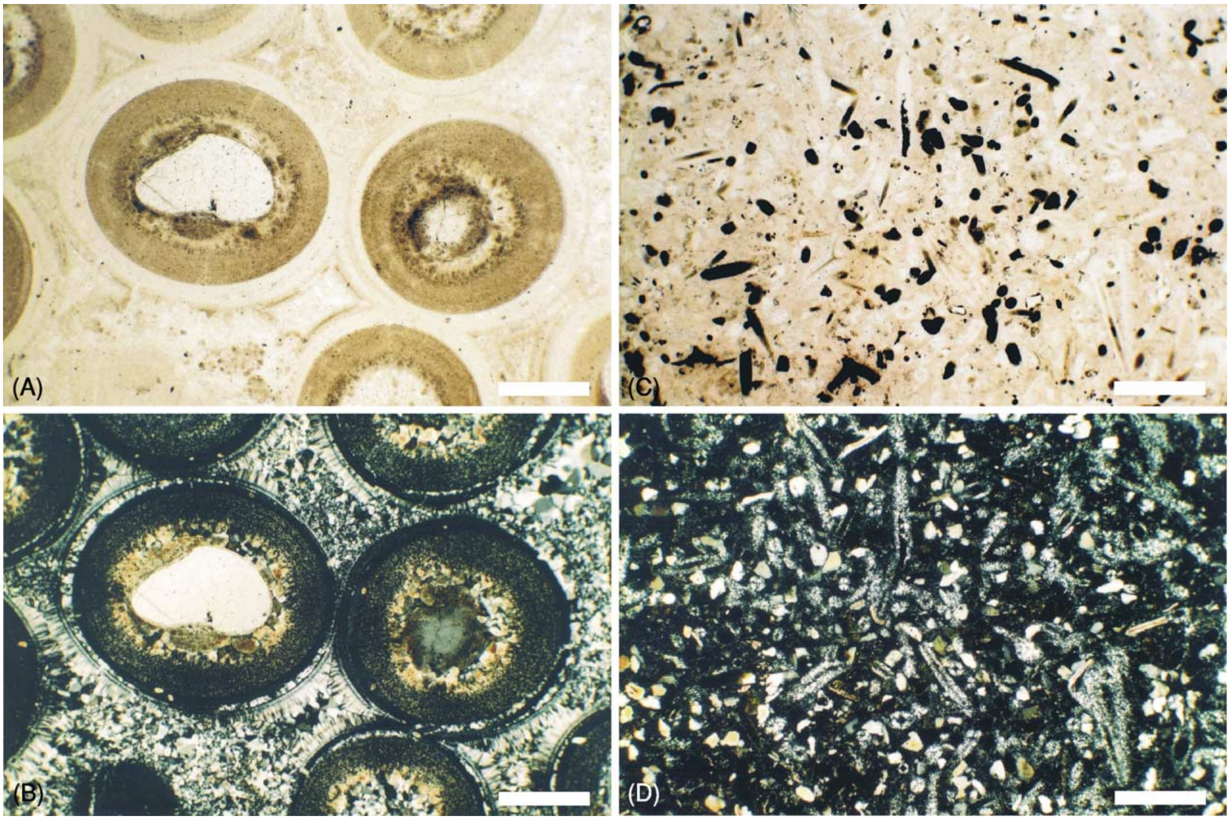


Figure 6 Photomicrographs of chert-replaced and cemented sedimentary rocks in thin section. (A) Silicified oolite in PPL (plane polarized light). (B) Same view as (A) in XPL (cross polarized light) showing a variety of silica cement and replacement fabrics after the original carbonate sediment. Trenton Series, Ordovician, Centre County, Pennsylvania, USA. (C) Silicified bioclastic sandstone with abundant sponge spicule tetraxons, viewed in PPL. (D) Same view as (C) in XPL. Both the matrix and sponge spicules are replaced primarily by microcrystalline quartz. Upper Greensand, Cretaceous, Ventnor, Isle of Wight, UK. Scale bars = 500 μm.

related to redox boundaries in the original sediment (Figure 8). The spacing of chert bands reflects abrupt, stepwise rises of the redox boundary related to pulses in sedimentation and hiatuses. The geometry of the redox boundary (governed by permeability and porosity contrasts in the sediment) generally determines the chert morphology, accounting for the spectrum of burrow-form and tabular cherts commonly observed in the field.

The replacement by chert can be remarkably selective, with preferential replacement of limestone matrix, or of biogenic debris of a particular original composition, such as high-Mg calcite. Very often the earliest stages of silicification in carbonate sequences occur in shell material in which localized silica precipitation/carbonate dissolution is promoted by the bacterial breakdown of organic matter, particularly conchiolin within the shell matrix. Thus, as well as the nodular form of chert, selective silicification may result in scattered silicified fossils within limestone.

In the Cretaceous Chalk of Europe and the USA, flint nodules of irregular form occur at specific widespread stratigraphical horizons, but also in sheets and pipes that cross-cut bedding. Flint is generally dark grey, and contains carbonate inclusions, particularly of bivalves and echinoderms. Flint nodules have a thin white crust, or patina. Many echinoids from the Upper Chalk are filled with flint, the silica having nucleated within the urchin, but without replacing the shell. Sponges and burrows are also selectively silicified, with the shape of many flints in the chalk reflecting the morphology of *Thalassinoides* burrows in which they nucleated (Figure 9). Siliceous sponge spicules probably provided much of the biogenic silica for the formation of flint.

There are also nodular chert-bearing beds within the Portlandian (Late Jurassic) limestones of southern England (Figures 7 and 8), where bioturbation textures and diagenetic redox boundaries controlled silica precipitation and replacement. Siliceous sponges were the main biogenic silica source.

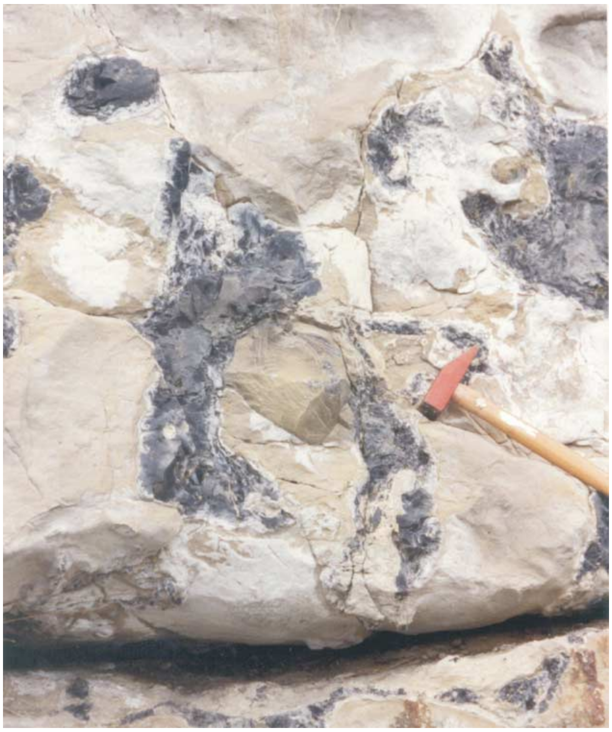


Figure 7 Nodular cherts replacing bioclastic and spicule wackestones. Cherty Beds, Portland Stone Formation, Upper Jurassic, Isle of Portland, Dorset, UK.

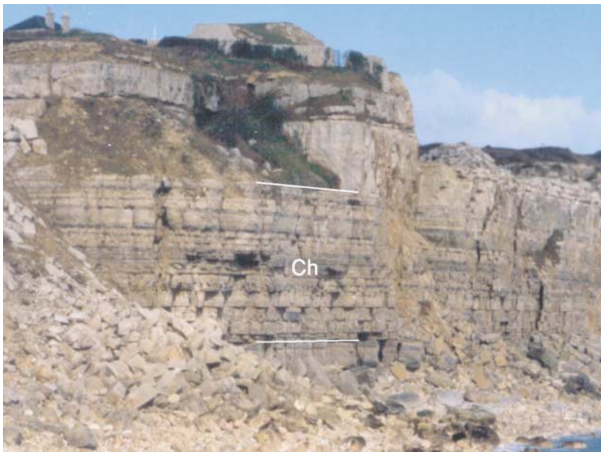


Figure 8 Laterally persistent beds of nodular and tabular chert (dark bands) concentrated within the Cherty Beds (Ch) of the Upper Jurassic Portland Stone Formation. Isle of Portland, Dorset, UK.

The Carboniferous limestones of Europe and the USA contain abundant chert as nodules and as silicified fossils (Figures 1C,D, and 10). The chert is generally black and has a splintery fracture, rather than the conchoidal fracture of flint. Silicification can affect specific beds, such that chert nodules may link up to form a diagenetically bedded chert.

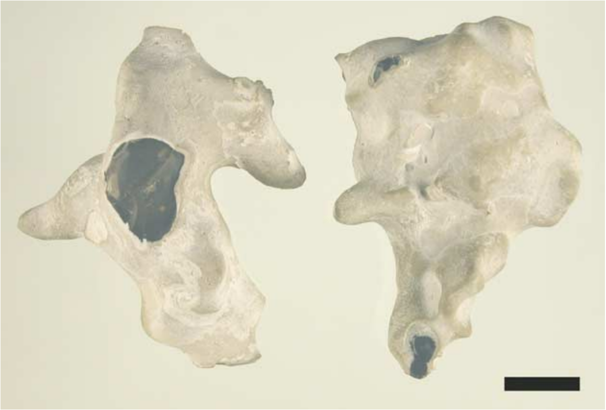


Figure 9 Flint nodules after *Thalassinoides* burrows. Chalk, Upper Cretaceous, UK. Scale bar = 25 mm.

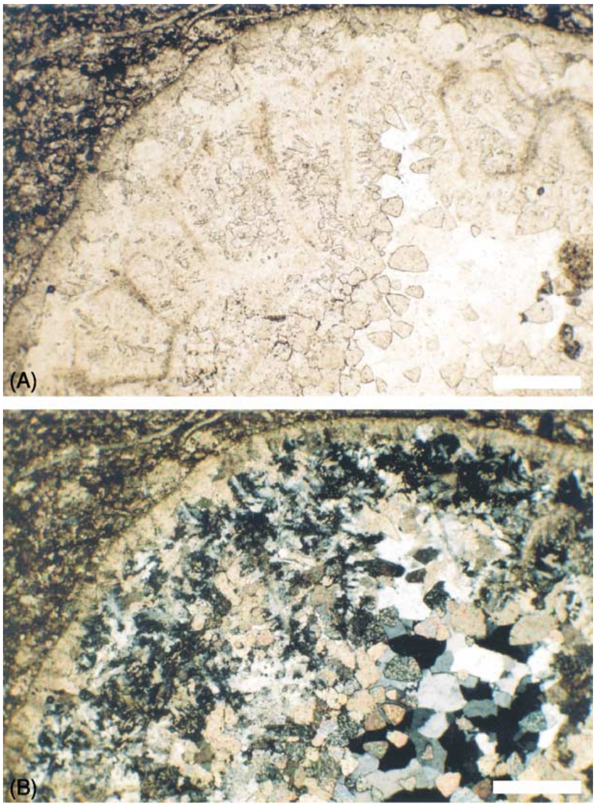


Figure 10 Transverse cross-section through a partially silicified corallite of *Siphonodendron junceum*. (A) Viewed in PPL (plane polarized light). (B) Viewed in XPL (cross polarized light). Hob's House Coral Bed, Viséan, Lower Carboniferous, Derbyshire, UK. Scale bars = 500 μm.

Chert in Lakes

Modern examples of lacustrine chert deposition comprise the Lake Magadi type, where the lake is Na-rich and alkaline and has pH > 9 in the dry season, leading to silica dissolution, and a pH that fluctuates below pH 9 in the wetter months, resulting in silica precipitation. Silica is initially deposited as magadiite (hydrated sodium silicate), which is subsequently replaced by silica. Thus, the controlling factors are evaporation and freshwater input to the lake.

In the Coorong type from South Australia, Mg-rich carbonate lakes acquire a high pH due to the seasonal activity of photosynthetic algae, resulting in the dissolution of silicates; with a seasonal reduction in pH, direct precipitation of mixed opal and cristobalite takes place.

The Cretaceous Uhangri Formation of southwest Korea was deposited in an alkaline lake surrounded by alkaline volcanics. The sequence includes couplets of sandstone overlain by chert, and of laminated chert with black shale. The sandstone/chert couplets were deposited following episodic influxes of fresh, less alkaline water. The influxes carrying sand produced density-current underflows in the stratified lake, depositing sand followed by opaline silica, caused by the fall in pH due to the influx of freshwater. The laminated cherts are interpreted to be the result of interflows causing silica precipitation. The chert beds show soft-sediment deformation and injection features, indicating a gelatinous consistency for the deposited silica. Thus, with regard to the feature of direct silica deposition, this example has similarities with the Coorong type.

Ancient deposits interpreted as belonging to the Magadi type are more common, and range in age from the Precambrian Reitgat Formation, Hartbeesfontain, South Africa, to the present day. In typical examples, there is an association with contemporaneous volcanics, and evidence for evaporite minerals.

Chert of Hydrothermal Origin

Silica-rich fluid expulsion from basins Basin marginal faults are commonly the site of chert deposition as veins and porosity-filling cement. Chert is deposited as a result of the cooling of silica-rich water expelled from the basin and rapidly rising up marginal fault zones. Silica is more soluble at high temperatures, and hence cooling results in silica precipitation. Chert may seal a fault, and subsequent fault movement may result in new fractures, which themselves become sealed; the result is a chert-cemented and veined fault zone.

Cherts resulting from hydrothermal systems Hydrothermal systems associated with volcanic activity are seen today at Geysir in Iceland, Yellowstone National Park in the USA, and North Island, New Zealand. At these, and many other localities, hot springs and geysers deposit large quantities of silica both in the subsurface and at the point of eruption, which may be on land or under water. Silica is deposited from cooling waters that have dissolved silica from hot rocks at depth (Figure 11).

In the subsurface, the result is the silicification of country rocks, particularly along fluid pathways such as faults. Cherty rock may develop on a large scale in the subsurface above a hydrothermal system, resulting in chert cement and cherty veins. The silica is initially deposited as amorphous silica, and this matures to chert with time, heat, and burial.

Hot springs and geysers bring hot water to the surface that cools rapidly on eruption, resulting in the instant deposition of amorphous silica in the form of sinter. Sinter may form mounds around geyser vents, or the outflow from a hot spring may result in sinter terraces or a low-angle sinter outwash apron (Figure 12). Under water, sinter chimneys may form above vents as occurs in Lake Yellowstone. The silica is deposited as highly porous amorphous opal-A, which is transformed to opal-CT, and later to chert, with a loss of porosity.

In New Zealand, the stages of mineral transformation are well documented. The Umikiri sinter is up to 15 m thick, can be dated to between 27 000 and 200 000 years BP, and shows a preserved silica maturation stratigraphy of opal-CT to opal-C to quartz with depth, all original opal-A having already been converted to opal-CT. Thus, the textural features associated with phase changes and solution–precipitation phenomena occur in a geologically short period of time in near-surface environments.

Probably the best-known fossil hot spring deposit is the Early Devonian Rhynie Chert of north-east Scotland (Figures 1A,B, and 13). The chert beds were deposited as sinters on a low-angle run-off apron from hot springs fed along a marginal fault to the Rhynie Basin of Old Red Sandstone. The beds are up to 0.5 m thick, laterally non-persistent, and with interbedded shale and sandstone of an alluvial plain environment. The chert is generally bluish to brown in colour, and is remarkable for the early terrestrial and freshwater biota it contains. The plants in some beds are preserved in three dimensions, with perfect cellular preservation, with plant axes still in the position of growth (Figure 13) to a height of 15 cm. This chert has yielded the most diverse terrestrial and freshwater arthropod fauna of any locality of similar age in the world. The detail of preservation is remarkable, including germinating plant spores and even sperm in the process of release from the male fertile organ of a gametophyte plant. Such features require virtually instant preservation, and point to a silica gel as the primary silica deposit. The presence of framboidal pyrite and the preservation of organic matter suggest reducing conditions during silicification. The textures within the Rhynie Chert are closely comparable with

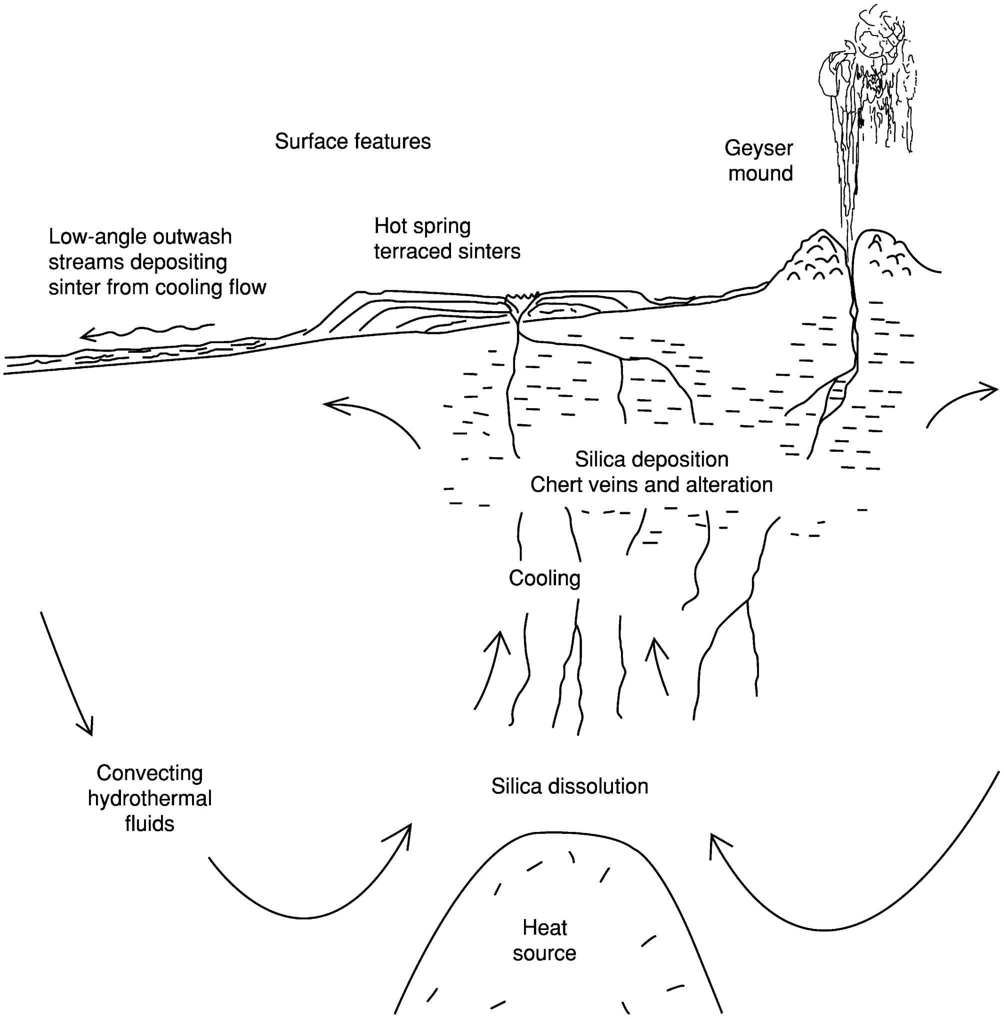

Surface features
Geyser mound
Low-angle outwash streams depositing sinter from cooling flow
Hot spring terraced sinters
Silica deposition
Chert veins and alteration
Cooling
Convecting hydrothermal fluids
Silica dissolution
Heat source

Figure 11 Diagram of a convecting hydrothermal system above an igneous heat source to illustrate surface and subsurface deposition of silica. Not to scale.

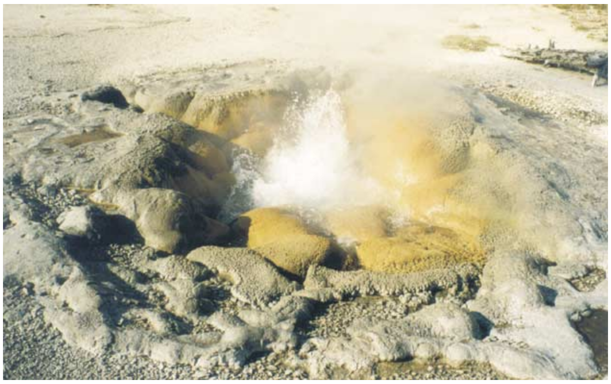


Figure 12 Geyserite mound and outlying sinter apron surrounding a small active geyser vent. Shell Spring, Lower Geyser Basin, Yellowstone National Park, Wyoming, USA.

those of modern siliceous sinters, but the maturation process to quartz is complete, and the chert comprises microcrystalline quartz, chalcedony, and macroquartz (Figure 1A and B).

Also found in the same area is a remnant of a geyser vent, with the typical geyserite texture preserved in chert. The country rocks in the area of this ancient hot spring system are also silicified, and a cherty breccia occupies the hot spring feeder zone along the marginal fault. More uncommon are silica deposits resulting from submarine exhalations; examples are the Cretaceous ochres of Cyprus, and the cherty ironstones of Tynagh, Ireland.

Agates with concentric and layered textures of microcrystalline quartz, chalcedony, and megaquartz

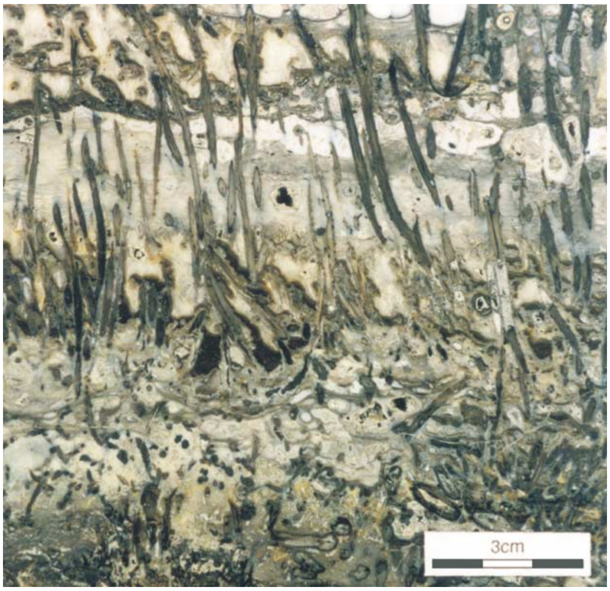


Figure 13 Vertical section through a bed of Rhynie Chert showing abundant stems of the plant *Rhynia gwynne-vaughanii* preserved in a three-dimensional, upright, growth position. Vague horizontal laminae, draped between the plant axes, represent silicified microbial mats. Chert-filled fenestrae within the laminae and between the plants represent gas bubbles trapped within the microbial layers. Pragian, Early Devonian, Rhynie, Scotland.

are mainly found in gas cavities (amygdales) in lavas. The silica is sourced from the volcanic rock, and deposited from hydrothermal and later diagenetic solutions migrating through the rock.

Chert in Soil Profiles – Silcrete

Silica is precipitated in soil profiles forming a hard chert-cemented rock known as silcrete. Most silcretes form in arid to semiarid climatic regions in which silica-saturated, alkaline groundwater with $\text{pH} \geq 9$ is evaporated from the surface, or mixes with surface water of lower pH. The presence of iron, aluminium, and magnesium oxides, and also NaCl, appears to encourage silica precipitation. The microquartz occurs as a mosaic cementing any detrital material present. Some replacement by chert is usual, particularly affecting carbonate and micaceous minerals. Silcretes are present in parts of Australia and in both northern and southern Africa. Silica is present in the form of microcrystalline quartz, chalcedony, and, to a lesser extent, opal. Silcrete horizons up to 10 m thick occur; the degree of silicification decreases with depth, with isolated silcrete nodules in the lower part of the profile, and massive silcrete at the top.

The 'Hertfordshire Puddingstone' of southern England is a Tertiary silcrete containing rounded pebbles of flint derived from the chalk, and cemented with chert. It closely resembles examples from the Lake Eyre region of Australia. Silcrete has also been described from the Proterozoic of north-west Canada, where it formed from the weathering of acid volcanics.

Silicified Wood

At many localities throughout the world, fossil wood is preserved in microcrystalline quartz with excellent preservation of the cellular structure of plant tissue. Woody material is a favoured site for silica deposition; in some cases, organic cell walls are preserved; in others, all organic material is lost. The Eocene fossil forests of Yellowstone National Park comprise a succession of 27 forests that were buried by volcanic ash, and occur in a 400 m thick sequence. The silicified trees are preserved as upright stumps several metres high. The silica source was the volcanic ash. In contrast, the Petrified Forest in the Painted Desert region of Arizona represents logs transported to the depositional site, where they occur in alluvial mudstones of Late Triassic age. Silica was probably derived from migrating groundwater, and nucleated in the acidic environment of the decaying wood structure.

Acknowledgments

SRF acknowledges funding from the Lyon Bequest to the University of Aberdeen. We would like to thank Professor David Macdonald (University of Aberdeen) for providing the images and caption information for Figure 2.

See Also

Geysers and Hot Springs. **Minerals:** Quartz. **Sedimentary Environments:** Lake Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification; Deep Ocean Pelagic Oozes; Evaporites. **Tectonics:** Hydrothermal Activity.

Further Reading

Carson GA (1991) Silicification of fossils. In: Allison PA and Briggs DEG (eds.) *Taphonomy: Releasing the Data Locked in the Fossil Record,* pp. 25–70. New York: Plenum Press.

Hesse R (1989) Silica diagenesis: origin of inorganic and replacement cherts. *Earth Science Reviews* 26: 253–284.

Knauth LP (1979) A model for the origin of chert in limestone. *Geology* 7: 274–277.

McBride EF (compiler) (1979) *Silica in Sediments: Nodular and Bedded Chert. Society of Economic Palaeontologists and Mineralogists, Reprint Series No. 8*. Tulsa, OK: Society of Economic Palaeontologists and Mineralogists.

Sieveking G de G and Hart MB (eds.) (1986) *The Scientific Study of Flint and Chert*. Cambridge: Cambridge University Press.

Trewin NH, Fayers SR, and Anderson LI (2002) *The Rhynie Chert: A Web-Based Teaching and Learning Resource*. http://www.abdn.ac.uk/rhynie.

Tucker ME (1991) *Sedimentary Petrology: An Introduction to the Origin of Sedimentary Rocks*. London: Blackwell Scientific Publications.

Williams LA and Crerar DA (1985) Silica diagenesis, II. General mechanisms. *Journal of Sedimentary Petrology* 55: 312–321.

Williams LA, Parks GA, and Crerar DA (1985) Silica diagenesis, I. Solubility controls. *Journal of Sedimentary Petrology* 55: 301–311.

Clays and Their Diagenesis

J M Huggett, Petroclays, Ashtead, UK and The Natural History Museum, London, UK

Introduction

Clay diagenesis is the process of clay transformation (layer by layer replacement) and authigenesis (or neoformation) in buried sediments. Diagenesis commences with the onset of burial and ends with the onset of metamorphism. These boundaries are defined in a variety of ways, including clay mineral crystallinity. Although, in the nineteenth century, microscopists were able to observe euhedral stacks of kaolinite platelets in sandstones, it was not until the early 1970s that it became apparent just how widespread and significant are clays that form after burial. Until this time, geologists had argued over 'the greywacke problem', i.e., how was it possible that clay, and sometimes quite high proportions of clay, could be present in sandstones deposited in high-energy environments? The realization that many sandstones contain authigenic clays came about largely as a result of the arrival of the scanning electron microscope in geological research. This allowed the examination of rocks in three dimensions at magnifications (typically <1000×) ideal for imaging clay particles. Clay diagenesis is not restricted to sandstones; similar processes occur in mudrocks, although transformation reactions and reaction pathways may differ. At around the same time, X-ray diffraction (XRD) was being used to measure changes with depth in clay and non-clay mineralogy with depth in mudrocks, and it was through the pioneering work of John Hower that this was shown to be due to diagenesis (Figure 1). This difference in approach resulted in some rather different perceptions regarding the nature of diagenesis in sandstones and mudrocks.

Factors controlling clay diagenesis include the detrital sediment composition, environment of deposition, temperature, permeability, and burial history (rate of burial, overpressuring, faulting, uplift). Diagenesis typically involves a simplification of the mineralogical suite of a sedimentary rock unit. As the temperature and pressure increase, so too does the tendency towards an equilibrium assemblage, provided that the pore fluid remains aqueous. Most diagenesis occurs at less than 160°C (although the cut-off is not determined by temperature; it is more likely to be determined by the exhaustion of reactive minerals, loss of permeability, or the cessation of movement of aqueous fluids for other reasons).

The movement of ions in solution from argillaceous sediment to coarser sediment has been widely invoked to account for the lack of an obvious internal source for the authigenic minerals present. This has partly come about through a lack of petrographical studies of mudrocks. However, it is now apparent that interbedded mudrock/sandstone can contain the same authigenic minerals, but in very different proportions. This is because, although the same or similar detrital minerals are present in both lithologies, their proportions can be quite different. In argillaceous rocks, there may be more organic matter diagenesis driving particular reactions, and the higher proportions of clay will result in much lower fluid flow rates and water/rock ratios than exist in sandstones. In Tertiary sediments from the North Sea, qualitative mass balance calculations have demonstrated that cross-lithology (sandstone/mudrock) flow can be insignificant for either clay or quartz diagenesis. Backscattered scanning electron microscopy (SEM) provides unequivocal evidence for coarsely crystalline authigenic clays in mudrocks, as well as in sandstones, and with the high resolution and magnifications possible by field emission SEM, it is possible to image overgrowths on clays and small packets of authigenic clay enclosed in detrital clay.

Clay Diagenesis in Mudrocks

Detailed investigations of diagenesis in argillaceous sediments in a wide variety of sedimentary basins have shown some consistent patterns of clay diagenesis. Variations in diagenetic clay assemblages result from differences in detrital assemblages and burial history. The bulk composition of most mudrock results in illite as the predominant end product of diagenesis. For

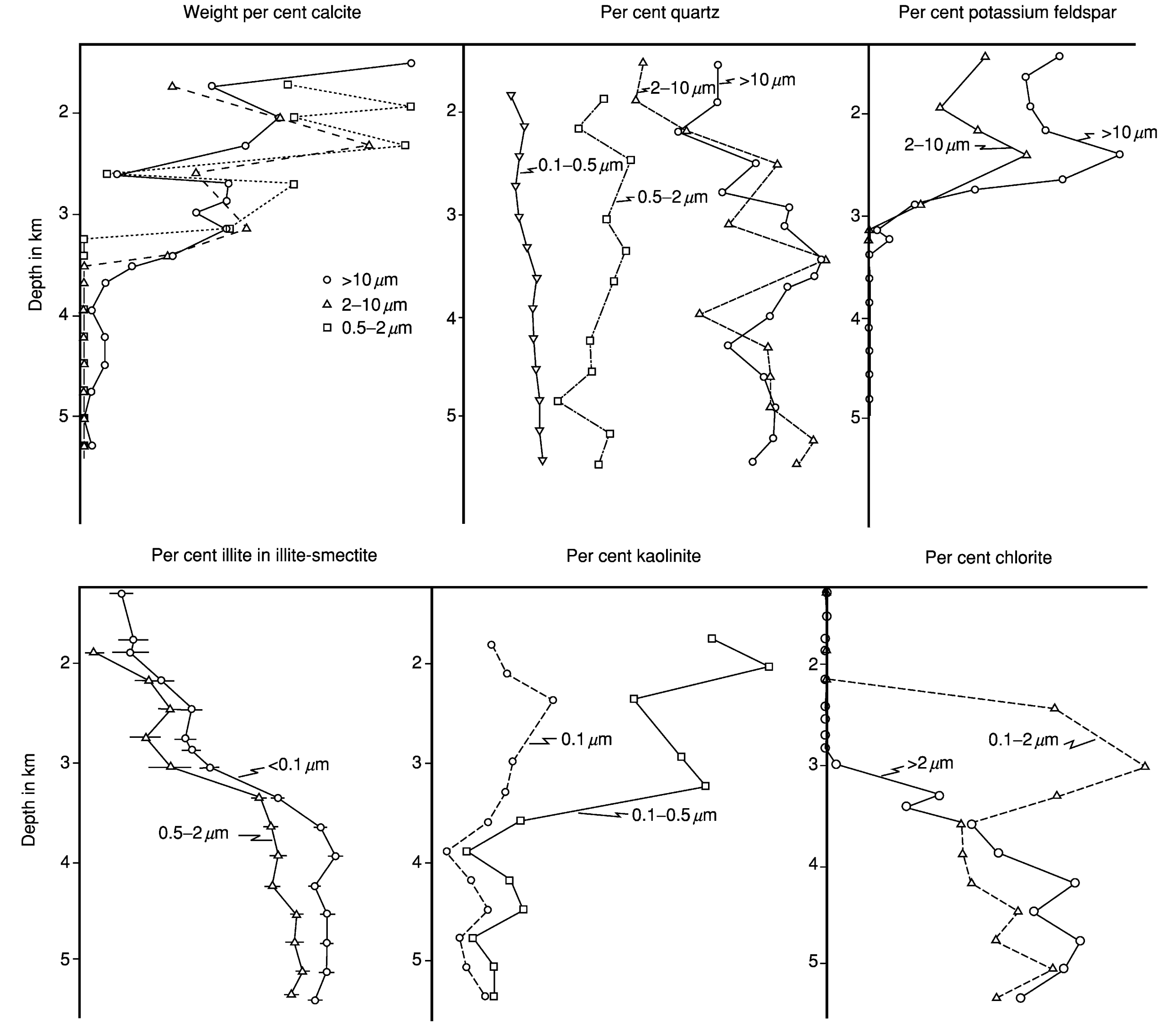


Figure 1 Change in percentage of calcite, quartz, K feldspar, illite, kaolinite, and chlorite with depth. Adapted from Hower J, Eslinger EV, Hower ME, and Perry EA (1976) Mechanism of burial metamorphism of argillaceous sediments: 1: Mineralogical and chemical evidence. *Geological Society of America Bulletin* 87: 725–737.

kaolinite or chlorite to be the predominant clay, special conditions are required, whilst the preservation of smectite at depth requires the inhibition of diagenesis, usually through overpressuring.

Illitization of Smectite

The most studied and first recognized aspect of clay diagenesis is the illitization of smectite with increasing depth. This has been most intensively studied in the Gulf Coast region of the USA, where Tertiary smectite-rich argillaceous sediments have undergone progressive burial. With increasing depth (i.e., increasing temperature), K feldspar, kaolinite, and smectite decrease, whilst illite and chlorite increase. In its simplest form, this reaction can be written as the dissolution of K feldspar to yield Al^{3+} and K^+, which react with smectite to form illite–smectite and ultimately illite:

$$\text{smectite} + Al^{3+} + K^+ \rightarrow \text{illite} + Si^{4+} + Fe^{2+} + Na^+ + Mg^{2+}$$

or

$$\text{smectite} + K^+ \rightarrow \text{illite} + Si^{4+} + Fe^{2+} + Na^+ + Mg^{2+}$$

The second reaction conserves aluminium and requires the dissolution of some smectite. The rock evidence suggests that both reactions are possible. K feldspar may also react with kaolinite to form illite. It is widely claimed that the K^+ for these reactions is a result of acid dissolution, especially by organic acids. However, feldspar dissolution can occur at low, neutral, or high pH. Indeed, the rate of feldspar dissolution is kinetically controlled. Hence feldspar (and mica) dissolution increases with increasing temperature, and therefore depth. Moreover, organic acids have low buffering capacities and therefore do not influence pH greatly. It should be noted that the reaction results in the release of silica and, through smectite dehydration, the release of water. The source of aluminium is probably mostly feldspar, but the means of transporting sufficient dissolved aluminium to the reaction site has not been entirely resolved, as the aluminium solubility varies enormously with pH, but in most geological situations is rather low. Carboxylic acids are claimed to have the capacity to complex with aluminium, thereby increasing the amount that can be held in solution. However, such acids are unlikely to have much effect on aluminium solubility in complex (i.e., natural) systems. It should also be noted that the reaction yields a potential source of quartz cement. It is largely agreed that the reaction is kinetically controlled, although, if the proposed kinetic equations are applied to the older sedimentary basins, the amount of illitization is vastly overestimated; this may be because the total heating to which they have been exposed has been overestimated. How close K feldspar needs to be to the site of illitization probably depends on the fluid flow, the degree of sandstone/mudrock interbedding, and the overpressure. In the Mahakam Delta Basin in Indonesia, it has been shown that the K feldspar alteration in both sandstone and mudrock is restricted to the upper 2 km of sediment, whereas illitization occurs at greater depths, thus necessitating an open system for K^+ transfer at depth. In contrast, in the Gulf Coast and the Tertiary mudrock/sandstones of the North Sea, diagenesis may be a nearly closed system.

The illitization of smectite commences at approximately 70°C, and peaks at approximately 120–130°C. However, in sedimentary basins with high geothermal gradients, this will occur at shallower depths than in those with lower geothermal gradients. Time, overpressure, pore fluid composition, and hydrothermal activity are also important factors in clay burial diagenesis. In general, there is sufficient K feldspar and mica for this not to be an inhibiting factor; clay diagenesis in clay-rich basins is most likely to be inhibited by overpressure which restricts fluid movement. If a source of K^+ is lacking, illite will not be formed, except where intense organic diagenesis releases NH_4, which is able to form ammonium illite. The illitization of smectite in mudrocks proceeds via random, mixed-layer, smectite-rich, illite–smectite to ordered, illite-rich, illite–smectite. Ordering commences at about 35% expandable layers. The maximum illite content is typically 80%. This sequence has been recognized in a wide variety of settings. The analysis of the expandability and thermal histories of basins ranging in age from Precambrian to Quaternary has indicated that the composition of illite–smectite in mudrocks is primarily controlled by the maximum palaeotemperature. Hence, illite–smectite may be used as a geothermometer for mudrocks, although, as pore fluid overpressure and a lack of K^+ may occasionally be more than minor controls on the illitization process, this should always be performed with caution. The interpretation of mixed-layer illite–smectite in terms of fundamental particles and interparticle diffraction, in the early 1980s, triggered much research into the true nature of illite–smectite. The concept of interparticle diffraction implied that, during illitization, mixed-layer crystals were not two chemically distinct clay minerals, but single illite layers 10 Å thick (Figure 2). When these fundamental particles were analysed by XRD, diffraction between particles created the illusion of smectite interlayers. With increasing diagenetic maturity, these particles grow in three dimensions and the apparent smectite layers decrease. This interpretation is not intended to imply that smectite does not exist! The mechanism implies the dissolution of smectite and the precipitation of illite, which initially exists as fundamental particles. However, it is still argued from transmission electron microscopy (TEM) data that layer-by-layer replacement of smectite by illite occurs. High-resolution investigations of the illitization of smectite have shown coherently scattering domains of interstratified illite-expandable layers (Figure 3), expandable layers within coherently scattering domains of illite, and domains of illite within

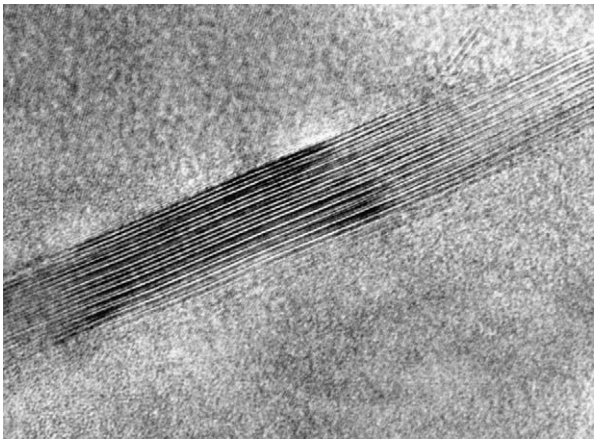


Figure 2 Lattice fringe transmission electron microscopy (TEM) image of interstratified illite-expandable clay (the expandable layers have been fixed to prevent collapse in the electron beam).

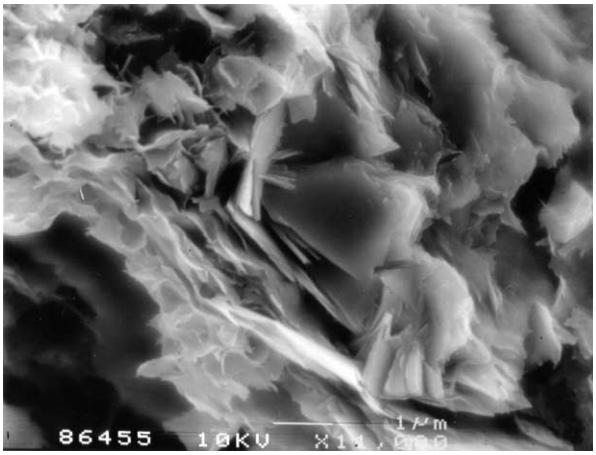


Figure 3 Field emission scanning electron microscopy (SEM) image of authigenic chlorite enclosed by detrital clay in mudrock.

smectite. It is also probable that both dissolution/precipitation and layer transformation occur in different settings. In both bentonites and mudrocks, most smectite is dioctahedral, which appears to be more prone than trioctahedral smectite to illitization. The latter is more likely to react to form chlorite via interstratified chlorite/smectite minerals.

Illitization in Bentonites

Bentonites are ash falls that have undergone extensive devitrification to dioctahedral smectite (usually montmorillonite). Because they have a very simple mineralogical assemblage (most mudrocks contain not only more than one clay type, but a mixture of smectites and illite–smectites), and are often almost monomineralic, ancient bentonites have been extensively used to study the process of illitization of smectite. Comparison of different bentonites, or single bentonites which have undergone variable heating during burial, shows that Si^{4+}, Ca^{2+}, and Na^{+} are lost from the bed and K^{+} is gained as the smectite is illitized. It should be noted that the supply of K^{+} is the rate-limiting step in the illitization of most bentonites because they are K^{+} deficient. Thus, the most potassic (illitized) portions of many bentonite beds are frequently the margins. Where the enclosing sediment is limestone, illitization will be restricted to any K^{+} present within the bentonite bed.

Illite Crystallinity and Illite Sharpness Ratio

With increasing temperature, illite in mudrocks undergoes an increase in crystallinity, as measured by the 001 reflection sharpness on XRD traces. This is due to the loss of smectite layers, increased particle size, and a reduction in abundance of crystal lattice defects. This property has been widely used as an indicator of diagenetic grade, and the results may be correlated with vitrinite reflectance. The Kubler crystallinity index is a measure of the width at half height of the glycol-solvated illite 001 reflection. The Weaver sharpness ratio is the ratio of the illite 001 (10 Å) reflection to the height of the low-angle side of the reflection at 10.5 Å, also on the glycol-solvated trace. The validity of using illite crystallinity or the sharpness ratio has been much debated, but the recent finding that the thickness of fundamental illite particles follows a unique evolution has permitted the refinement of illite crystallinity into a precise measurement of mean crystal thickness. It should be noted that the inclusion of detrital metamorphic mica in the analysis will result in an overestimation of crystallinity.

Chlorite

Si^{4+}, Al^{3+}, Fe^{2+}, and Mg^{2+}, released from the dissolution of smectite and kaolinite, may react to form chlorite. This is usually detected as a down-hole increase in chlorite, although it can be argued that mineralogical changes may also result from a shift in provenance or climate. To obtain unequivocal evidence of authigenesis, it is usually necessary to use an imaging technique, such as SEM or TEM analysis, to demonstrate the face-to-edge arrangement of euhedral platelets. The chlorite shown in Figure 4 was investigated because XRD patterns for a Tertiary mudrock sequence showed unusually high chlorite concentrations at around 2.6 km burial depth. A sample from the same depth, analysed by TEM, confirmed the presence of 14 Å and 7 Å lattice fringes, and X-ray analyses confirmed that it was an iron-rich chlorite. Authigenic chlorite in mudrocks can also form by the replacement of biotite; commonly, replacement is partial, resulting in chlorite–biotite 'stacks'.

Kaolinite

Authigenic kaolinite in mudrocks is much more abundant than was previously thought before backscattered electron imaging made mudrock petrography a real possibility. Previously, a high kaolinite content in an argillaceous rock was assumed to indicate that the clay was formed through tropical weathering, and that it was consequently a climatic indicator. Kaolinite in mudrocks typically replaces muscovite and phengite mica, and cements microfossil cavities. Replacement of detrital mica by kaolinite is a hydrolysis reaction that releases K^{+}. It characteristically occurs during early diagenesis, whilst the

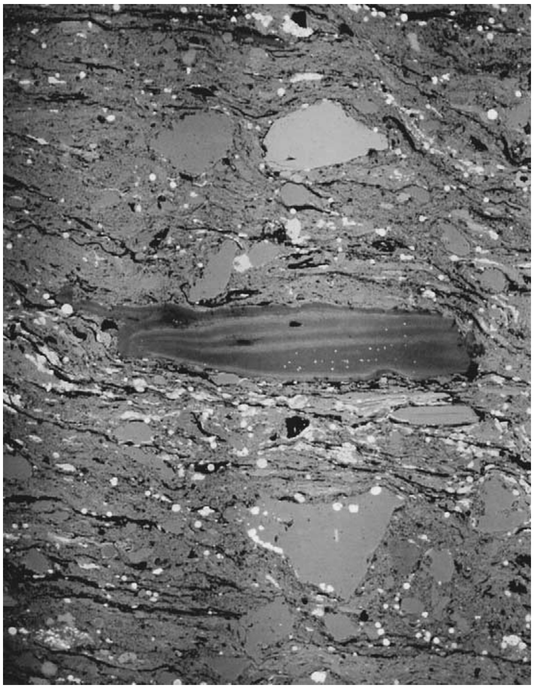


Figure 4 Backscattered scanning electron microscopy (SEM) image of kaolinite pseudomorph after mica in the London Clay Formation (Eocene).

fluid/rock ratio is still high, and the K^+ can be removed from the site of dissolution, allowing the reaction to continue. A tentative reaction is deduced from the common association of kaolinite pseudomorphs after mica with mica/quartz pressure solution contacts:

$$\text{muscovite} + H_4SiO_4 + K^+ + H^+ \rightarrow \text{kaolinite} + \text{K feldspar} + Na^+ + Mg^{2+} + Fe^{2+} + H_2O$$

However, in the Eocene of the London Basin, where burial has never been greater than 1 km, probably much less, kaolinite pseudomorphs after mica are widespread (Figure 5), and evidence for pressure solution is minimal. This particular reaction mechanism may be more applicable to sandstones than mudrocks. The fate of the K^+ that is released at shallow depths is not clear; certainly, it is not normally needed for the illitization of smectite at such low temperatures.

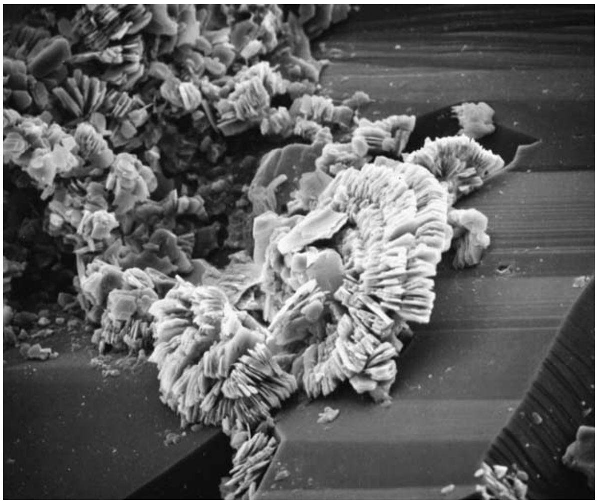


Figure 5 Fracture surface scanning electron microscopy (SEM) image of pore-filling kaolinite.

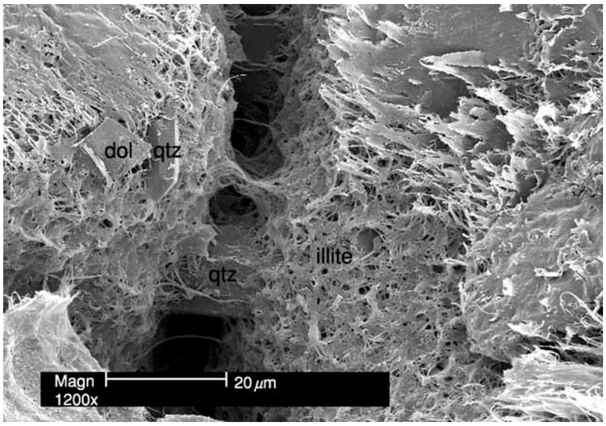


Figure 6 Fracture surface scanning electron microscopy (SEM) image of wispy illite in Rotliegendes Sandstone (Permian). The clay has constricted the pore throat.

Clay Diagenesis in Sandstones

Kaolin Clays

Authigenic kaolinite in sandstones forms stacks of euhedral pseudohexagonal platelets, with the *c* axis parallel to the stacking direction (Figure 6). Very long stacks are called vermicules (Figure 6).Typically, this clay has a pore-filling habit. Kaolinite also forms pseudomorphs after detrital mica, usually muscovite and chlorite. It should be noted that this kaolinite forms particles far larger than the 2 μm maximum defined for clay particles. Pore-filling authigenic kaolinite is often linked to the dissolution of feldspar at temperatures of less than 100°C. Although it is commonly assumed that a high aH^+ is necessary for significant feldspar and mica leaching, the stability field of kaolinite extends to the greatest range of $[K^+]/[H^+]$ values at a pH close to neutral. It should be noted that, for significant kaolinite to precipitate, K^+ and Na^+ from feldspar and mica need to be removed, or illite rather than kaolinite will become the stable clay mineral. The implication is that kaolinite will form at higher fluid flow rates (or a higher water/rock ratio) than illite, i.e., in the most porous parts of a sandstone, in the most coarse-grained beds, and at a time when a sandstone is less cemented. This is likely to be an important reason why kaolinite almost invariably precipitates before illite in a paragenetic sequence, and why it is often interpreted as a product of meteoric flushing. Meteoric flushing can be an effective mechanism for the replacement of feldspar by kaolinite if the water is sufficiently acidic, which effectively restricts the mechanism to the tropics. Early authigenic kaolinite in sandstones is most abundant in nearshore sediments, because these are most prone to meteoric flushing. Late (or relatively late) authigenic kaolinite is often associated with unconformities, and uplifted fault blocks. A meteoric origin for authigenic kaolinite can be demonstrated through $\delta^{18}O$ stable isotope measurements.

Dickite in sandstones typically forms larger crystals than kaolinite, which are thicker in the direction of the *c* axis. Dickite has been observed to replace kaolinite in reservoir sandstones at approximately 120°C (typically at 2500–4000 m burial depth), most notably where a high water/rock ratio has been preserved. The scarcity of dickite relative to kaolinite is probably due to the rarity of a high water/rock ratio in deeply buried sediment.

Smectite

Authigenic smectite in sandstones is fairly uncommon. Most smectite forms in surface sediments or through the alteration of ash layers. Where it does occur, it characteristically forms early diagenetic rims of crenulate, 'honeycomb', interlocking crystallites. The fusing of adjacent crystallites and the undulose morphology serve to distinguish smectite rims from chlorite rims in SEM images. It should be noted that SEM qualitative X-ray analyses are not always sufficiently accurate to distinguish between smectite and Mg chlorite. Smectite formation is associated with the dissolution of acid volcanic rock fragments and biogenic silica, because smectite formation is favoured by a high Si^{4+} activity. During burial, dissolved Si^{4+} is removed by quartz precipitation (greater than approximately 65°C), making smectite unstable.

Illite

Illite in sandstones has a range of morphologies, from undulose platelets ('cornflake' texture) at one end of the spectrum to laths, fibres, wisps, or ribbons ('hairy' illite) at the other. In fact, the elongate form is not fibrous, but sometimes the particles are so long that they appear so. Wispy illite particles are typically only a few 100 Å thick and 0.1–0.4 μm wide. Their length varies enormously, but can be tens of micrometres. There is no general agreement on any relationship between authigenic illite morphology, the timing of precipitation, chemistry, or structure (although this does not preclude the existence of such relationships). Certainly, with increasing burial depth, the morphology becomes increasingly that of rigid laths, i.e., the width and thickness increase (reflected in an increase in illite crystallinity). Wispy illite is bad news for hydrocarbon reservoir quality. Its high surface area and pore-throat constricting habit (Figure 7) can

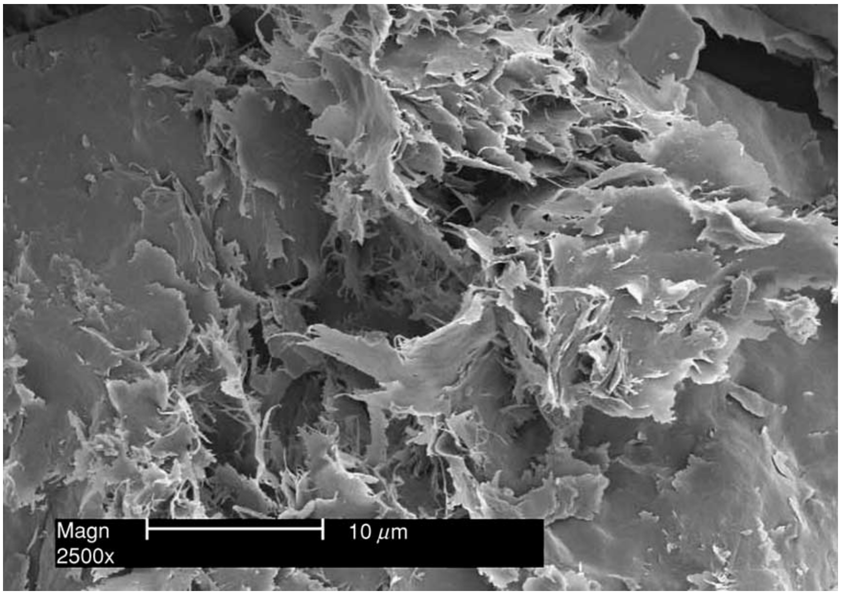


Figure 7 Fracture surface scanning electron microscopy (SEM) image of illite lath overgrowths on infiltrated illitic clay in Rotliegendes Sandstone (Permian).

drastically reduce the permeability (without causing much porosity reduction, because the actual volume of illite is small in proportion to the pore volume). Furthermore, injection wells may suffer a loss of permeability if the injected water breaks up the illite, which then migrates to pore-throats. Not uncommonly, authigenic illite forms short wispy overgrowths on platey detrital illite or illite–smectite, whilst, in some non-marine settings, authigenic illite preferentially nucleates on infiltrated illitic clay (Figure 8). With increasing temperature, pH, and $[K^+]/[H^+]$, kaolin and smectite minerals become unstable and are replaced by illite (Figure 9). In simple terms, the reactions may be shown as:

$$\text{smectite} + K^+ \rightarrow \text{illite} + \text{silica (approximately } 80 - 100^\circ\text{C)}$$

$$\text{kaolinite} + \text{K feldspar} \rightarrow \text{illite} + \text{quartz (greater than } 120^\circ\text{C)}$$

This reaction involves an increase in layer charge to accommodate the K^+ in place of the more weakly bonded exchangeable cations in the smectite. This is achieved through substitution of Al^{3+} for Si^{4+} in tetrahedral sites and a reduction in octahedral iron. In fact, many sandstones have pore fluid in thermodynamic equilibrium with illite, but little precipitation occurs due to an extremely low kinetic precipitation rate at temperatures of less than 120°C. Kaolinite is illitized at burial depths in the region of 3–3.5 km, whilst, at greater depths, the thermodynamically more stable dickite is also replaced by illite. Sometimes the mass balance (based on petrographical data) shows that sufficient K^+ can be obtained from locally dissolved K feldspar; in other instances, insufficient K feldspar dissolution has occurred (K feldspar may even be absent) to account for all the illite present, and external sources need to be invoked. It should be noted that the illitization of kaolin minerals is not inevitable above 120°C; there are many instances in which kaolinite coexists with authigenic illite and unleached K feldspar at burial depths ranging from 3 to 4 km. This is because significant illite precipitation requires a higher $[K^+]/[H^+]$ value than that which can occur in a closed system, in which $[K^+]/[H^+]$ is controlled by K feldspar solubility. Hence, the fact that sufficient K^+ can be locally derived does not mean that it is. In the North Sea, authigenic illite abundance is often highest close to major faults, suggesting that the K^+ is derived from the dissolution of K feldspar in deeper parts of the basin. In parts of the North Sea Basin, extensive illite cementation of the Rotliegend Sandstone has been linked to the dissolution of Zechstein salt deposits. Hence, in sandstones, the degree of illitization may reflect the temperature of migrating fluid rather than, as is the case with mudrocks, the maximum burial temperature. Consequently, illite geothermometry is not so reliable a tool for sandstones as it is for mudrocks. The dissolution of K feldspar releases more silica than is required for either kaolinite or illite precipitation. This is thought to be one of the main sources of quartz cement in sandstones. Indeed, textural relationships between kaolinite, illite, and quartz are frequently suggestive of the coprecipitation of one or other clay with quartz cement.

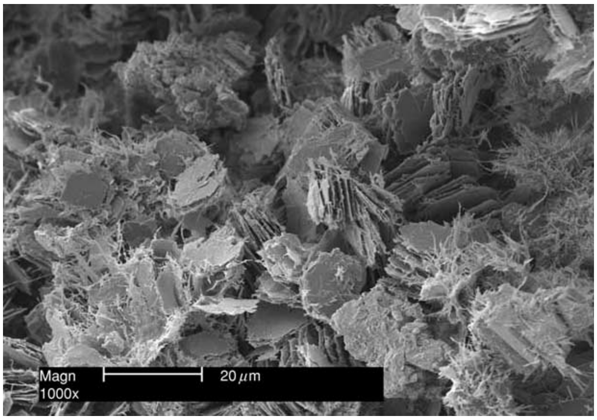


Figure 8 Fracture surface scanning electron microscopy (SEM) image of illite pseudomorph of kaolinite.

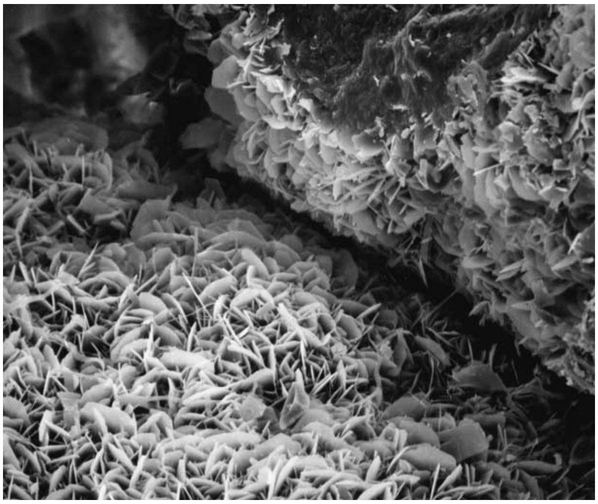


Figure 9 Fracture surface scanning electron microscopy (SEM) image of grain-rimming Fe chlorite.

Table 1 Differences between Fe chlorite and Mg chlorite. Adapted from Hillier (1984)

	Fe chlorite	*Mg chlorite*
Morphology	Pseudohexagonal, planar	Cornflake-like
Arrangement	Individual plates and rosettes, face-to-edge contacts	Boxwork, face-to-face and edge-to-edge contacts
XRD	Interstratified with 7 Å layers at <200°C	Often interstratified with corrensite or smectite
Polytype	Ib transforming to IIb	IIb only
Occurrence	Marginal marine sandstones offshore from major river systems in tropical climates	Coastal aeolian dunes and sandy sabkhas, any facies associated with evaporite brines
Facies associations	Oolitic ironstone	Evaporites
Associated early authigenic minerals	Siderite, calcite, phosphates	FeO rims, anhydrite, K feldspar, calcite, dolomite
Precursor		

XRD, X-ray diffraction.

K/Ar dating allows the determination of the mean age of an authigenic illite occurrence. Although contamination of data by detrital illite and K feldspar can be a problem, this technique is much used in unravelling burial histories. The method is based on the known rate of decay of radioactive ^{40}K to stable ^{40}Ar. Important advantages of this method are that it is generally reasonable to assume that no argon was present in the mineral at the time of its formation, whilst illite has excellent Ar retention at burial temperatures of less than 175°C. Illite ages generally decrease with increasing burial depth. In sandstones, smaller sized fractions may give younger dates than coarser ones, as contamination by detrital illite decreases with decreasing particle size. However, in bentonites, smaller sized fractions can give older dates than coarser ones, which is consistent with the growth of fundamental particles as described above. In the Jurassic of the North Sea, most illite dates are in the range 50–30 Ma, coincident with the onset of rapid burial in the Late Cretaceous–Early Tertiary.

Glauconite

Although glauconite forms only at the sediment–water interface, with increasing burial temperature, aluminium is partly substituted for iron.

Chlorite

In sandstones, authigenic chlorite typically forms grain-coating rims of radial, interlocking platelets. Such rims have been the focus of petroleum company research, due to the inhibiting effect they have on quartz overgrowth cementation. The result of this inhibition can be sandstones with excellent reservoir quality at 4–5 km burial depth. Textural relationships indicate unequivocally that such chlorite forms very early in diagenesis. As with chloritic green clay pellets, early Fe chlorite rims are associated with sandstones offshore from major river systems within the tropics. Examples of this occur in the Jurassic Åre Formation in the Norwegian sector of the North Sea and the Miocene off the Niger Delta. It is now thought that Fe chlorite (chamosite) originates as odinite or a similar 7 Å iron-rich clay, although recent examples of odinite are rare. One line of evidence for a 7 Å clay precursor is that the proportion of 7 Å interlayers in 14 Å chlorite decreases with depth and burial temperature. This is not the only change exhibited by chlorite with depth: a gradual Ib to IIb polytype transition has been demonstrated for Fe chlorite, and there is a strong linear relationship between tetrahedral Al in authigenic chlorites and present-day temperature in hydrothermal systems. However, there is no simple relationship between chlorite polytype and temperature, and attempts to develop a universal chlorite geothermometer have largely failed as authigenic chlorite composition is influenced by detrital sediment composition. Chlorite in sandstones also occurs as a pore-filling replacement of ferromagnesian igneous rock fragments; the composition of this type of chlorite will reflect the mineral being replaced. Chlorite–smectite and corrensite can form by a similar process. These clays typically have more undulose platelets than true chlorite and are most frequently found in volcaniclastic sediments. Mg-rich chlorite, chlorite, and corrensite also form by diagenetic replacement of Mg smectite in evaporite basins. Nonchemical differences between Mg chlorite and Fe chlorite are summarized in Table 1.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray); Geochronological Techniques; Mineral Analysis. **Clay Minerals**. **Colonial Surveys**. **Sedimentary Rocks:** Ironstones; Sandstones, Diagenesis and Porosity Evolution.

Further Reading

Bjørlykke K (1998) Clay mineral diagenesis in sedimentary basins – a key to the prediction of rock properties. Examples from the North Sea. *Clay Minerals* 33: 15–34.

Burley SD and MacQuaker JHS (1992) Authigenic clays, diagenetic sequences and conceptual diagenetic models in contrasting basin-margin and basin-centre North Sea Jurassic sandstones and mudstones. In: Houseknecht DW and Pittman ED (eds.) *Origin, Diagenesis and Petrophysics of Clay Minerals in Sandstones, Society of Economic Paleontologists and Mineralogists Special Publication 47,* pp. 81–110. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Ehrenberg SN and Nadeau PH (1989) Formation of diagenetic illite in sandstones of the Garn Formation, Haltenbanken area, mid-Norwegian continental shelf. *Clay Minerals* 24: 233–253.

Hower J, Eslinger EV, Hower ME, and Perry EA (1976) Mechanism of burial metamorphism of argillaceous sediments: 1: Mineralogical and chemical evidence. *Geological Society of America Bulletin* 87: 725–737.

Huggett JM (1995) Formation of authigenic illite in Palaeocene mudrocks from the central North Sea: a study by high resolution electron microscopy. *Clays and Clay Minerals* 43: 682–692.

Huggett JM (1996) Aluminosilicate diagenesis in a Tertiary sandstone–mudrock sequence from the central North Sea, U.K. *Clay Minerals* 31: 523–536.

Kisch HJ (1990) Calibration of the anchizone: a critical comparison of illite 'crystallinity' scales used for definition. *Journal of Metamorphic Geology* 8: 31–46.

Lanson B, Beaufort D, Berger G, Bauer A, Cassagnabere, and Meunier A (2002) Authigenic kaolin and illitic minerals during burial diagenesis of sandstones: a review. *Clay Minerals* 37: 1–22.

Longstaffe F (1989) Stable isotopes as tracers in clastic diagenesis. In: Hutcheon IE (ed.) *Short Course in Burial Diagenesis,* pp. 201–277. Toronto: Mineralogical Association of Canada.

Nadeau PH, Wilson MJ, McHardy WJ, and Tait J (1984) Interstratified clays as fundamental particles. *Science* 225: 923–925.

Pollastro R (1993) Considerations and applications of the illite/smectite geothermometer in hydrocarbon-bearing rocks of Miocene to Mississippian age. *Clays and Clay Minerals* 41: 119–133.

Srodon J (1999) Use of clay minerals in reconstructing geological processes: recent advances and some perspectives. *Clay Minerals* 34: 27–38.

Deep Ocean Pelagic Oozes

R G Rothwell, Southampton Oceanography Centre, Southampton, UK

Introduction

Deep-ocean pelagic (from the Greek *pelagios*, meaning 'of the sea') sediments are areally and volumetrically the dominant sediment type found on the ocean floor. They comprise three main types, depending on their primary composition: deep-sea siliceous oozes, calcareous oozes, and deep-water red clays (Figure 1). Pelagic sediments mixed with terrigenous material derived from continental weathering are termed 'hemipelagic'. Siliceous and calcareous oozes are composed largely of test and test debris of planktonic micro-organisms such as foraminifera, coccolithophores, pteropods, diatoms, and radiolaria. The formation of pelagic sediments involves settling of material, that is commonly derived from biological surface productivity, but also includes wind-derived material that travels through the water column ('pelagic rain') to the seafloor (Figure 2). This process occurs throughout the world's oceans and true pelagic deposits tend to blanket seafloor topography. Local remobilization of pelagic sediments on topographic highs due to slope instability may result in pelagic turbidites (redeposited units) that pond in adjacent lows. However, the distribution of pelagic sediments is strongly depth controlled, because calcium carbonate shows increasing solubility with depth. In contrast to terrigenous sediments (the other main type of deep-sea sediment, largely composed of detrital material derived from continental weathering), pelagic sediments are characterized by low sedimentation rates and frequently contain a high proportion of authigenic minerals, extraterrestrial material, and, where physicochemical conditions allow, a substantial biogenic component.

History of Research

Although mention of marine sediments has been found in Greek and Roman texts, it was not until 1773 that the recovery of sediment from the deep sea was first recorded. In that year, Captain John

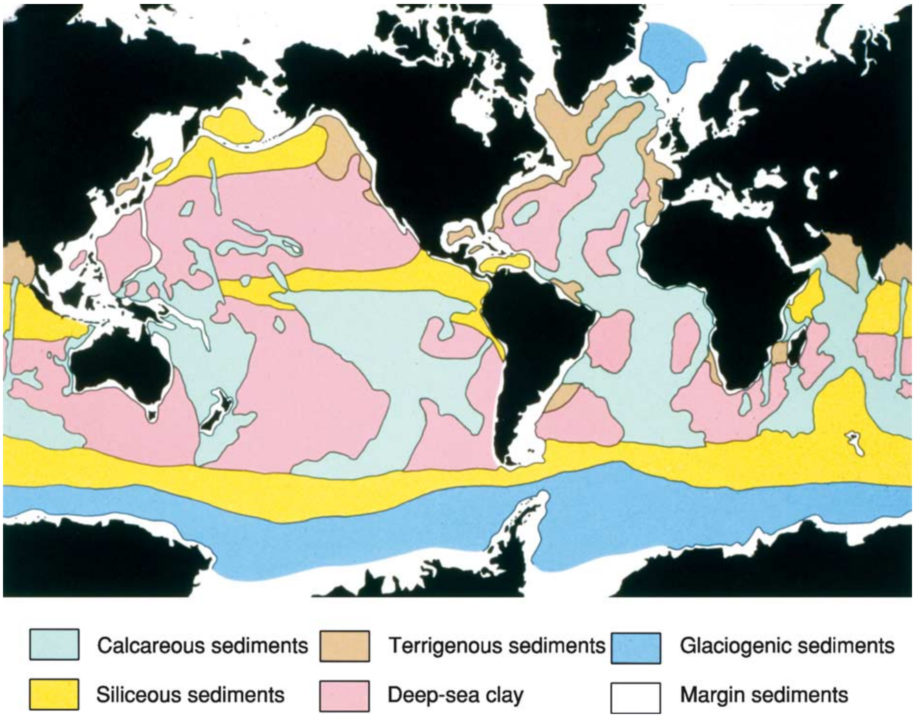


Figure 1 Distribution of the main sediment types on the ocean floor. Reproduced with permission from Davies TA and Gorsline DS (1976) Oceanic sediments and sedimentary processes. In: Riley JP and Chester R *Chemical Oceanography*, vol. 5, 2nd edn., pp. 1–80. London: Academic Press.

Phipps, on *HMS Racehorse*, bought up "fine soft blue clay" from a depth of 1250 m, the first recorded successful deep-sea sounding. The sample was taken in water on the southern edge of the Voring Plateau in the Norwegian Sea. In 1818, Sir John Ross recovered 6 lb (2.7 kg) of greenish mud from a depth of 1920 m in Baffin Bay, offshore Canada, using a deep-sea grab, representing one of the first recorded successful substantial deep-sea sediment recoveries. The laying of the first functioning submarine telegraph cable across the Straits of Dover in 1851 led to rapid expansion in the collection of deep-sea soundings and samples, driven by the prospect of the new means of intercontinental communication. However, it was not until the voyage of *HMS Challenger* (1872–76) that enough deep-sea samples were recovered to produce the first global seafloor sediment map. The voyage of *HMS Challenger*, led by Professor Charles Wyville Thomson, professor of natural history at the University of Edinburgh, was the first large-scale expedition devoted to oceanography. A wealth of seafloor samples were recovered from 362 observing stations, spaced at uniform intervals, along the 128 000-km track traversed during the voyage. John Murray, a naturalist on the *Challenger* expedition, oversaw the initial analysis of the recovered samples; Murray later edited the *Challenger Reports*, following the death of Wyville Thomson in 1882. The milestone *Challenger Report* on 'Deep-Sea Deposits' represented the first comprehensive volume on sediments of the deep-ocean seafloor. Published in 1891 with the assistance of Murray's co-worker AF Renard, this volume introduced many of the descriptive terms used today, such as 'red clay' and '*Globigerina* ooze', and provided the basis for further deep-ocean sediment studies. Murray also correctly related the distribution of shell-bearing plankton in the surface waters to the calcareous and siliceous sediments of the deep-ocean seafloor.

A major step forward in the investigation of seafloor sediments was the invention of the gravity corer by German researchers; the corer allowed continuous samples of extended lengths of sediment to be collected (but generally restricted to 1–2 m in length). The German South Polar Expedition (1901–03) collected several 2-m cores that were described by E Philippi in 1910. These cores showed that some deep-water sediments were stratified. During the period 1925–38, Germany ran a series of oceanographic expeditions using the ship *Meteor*, which recovered several 1-m-long cores from the southern Atlantic and Indian oceans. These cores, studied by Wolfgang Schott, showed changes in foraminifer species with depth and initiated the first

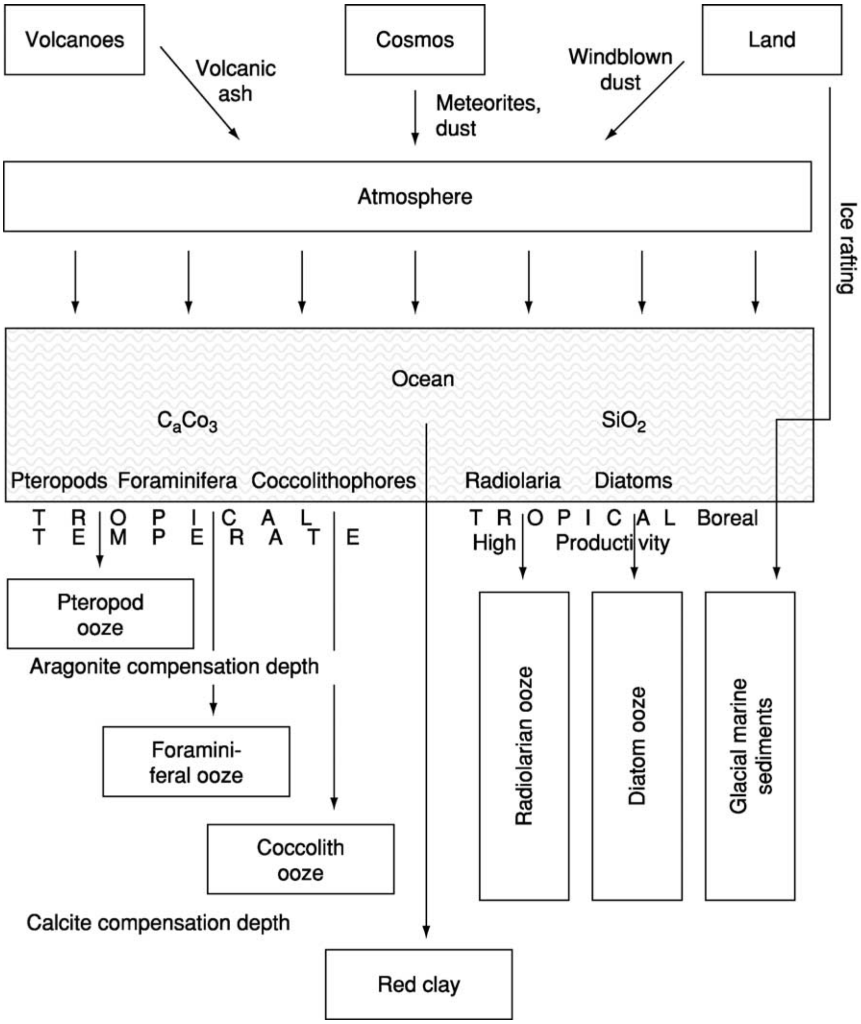


Figure 2 Sources and pathways for pelagic sedimentation in the oceans. Reproduced with permission from Hay WW (1974) Studies in Paleoceanography. Tulsa, Oklahoma: Society of Economic Paleontologists and Mineralogists.

palaeoceanographic studies. The Swedish Deep-Sea Expedition (1947–49) provided another fundamental advance in the study of deep-sea sediments. This expedition on the *Albatross* deployed a new kind of coring device, called a piston corer, which was developed by Börje Kullenberg, a marine geologist working in Gothenburg, Sweden. An innovative modification of the traditional gravity corer allowed the coring tube to fall past a stationary piston at the end of the wire, so that water was expelled from the falling tube above the piston and sediment was admitted from below. This allowed retrieval of much longer (10 m or more) and much less disturbed sediment cores. Acquisition of such long piston cores made possible the study of Pleistocene ocean history. The era of modern deep-sea sediment sampling had begun. Piston coring remains the main contemporary method of sampling the deep-sea sedimentary record, although sizing-up of the coring apparatus has led to development of giant piston corers, now capable of obtaining sediment cores up to 60 m in length. A giant piston core providing a continuous 54-m-long sediment record covering 4 million years of sedimentation was recovered from the Indian Ocean by the French ship *Marion Dufresne* in 1990; this is one of the longest piston cores ever recovered.

The advent of the Deep Sea Drilling Project (DSDP) in 1968 took deep-sea sediment sampling technology further. This programme, using the dynamically positioned drillship *Glomar Challenger*, set to recover continuous or semicontinuous sediment records from the ocean basins and heralded another new era in the exploration of the deep-ocean sedimentary record. Prior to the DSDP, the global inventory of cores recovering pre-Quaternary sediments numbered fewer than 100. The DSDP and its successor, the Ocean Drilling Program (ODP), which began drilling operations in 1985 with a new drillship *JOIDES Resolution*, with improved capabilities, have led to major advances in our understanding of the processes of plate tectonics, of Earth's crustal structure and composition, of past changes in climate, and of conditions in ancient oceans. A vast amount of core material has been collected by the DSDP and ODP from all ocean basins (except the Arctic Ocean) and continuous sediment sequences back to Early Jurassic times have now been recovered.

Control of the Distribution of Pelagic Deposits

Pelagic sediments are defined as those formed of settled material that has fallen through the water column; their distribution is controlled by three main factors, distance from major landmasses (which affects their dilution by terrigenous, or land-derived, material), water depth (which affects sediment preservation), and ocean fertility (which controls surface water productivity). Pelagic sediments are composed largely of the calcareous or siliceous remains of planktonic micro-organisms or wind-derived material or mixtures of these. Several types of pelagic deposits can be identified on compositional grounds, and because seawater is increasingly corrosive with pressure and depth, the distribution of pelagic sediment types is strongly controlled by the calcite compensation depth (CCD), which is that depth at which the rate of supply of biogenic calcite equals its rate of dissolution (Figure 3). Therefore, below the CCD, only carbonate-free sediments accumulate. Thus the calcite compensation depth marks a major boundary defining the deposition of pelagic clays and calcareous sediments.

Another important depth in the water column is the lysocline, which lies above the CCD and is the depth at which the degree of undersaturation with respect to calcium carbonate is sufficient for dissolution of calcareous particles to become significant (Figure 3). Therefore, the lysocline is a depth that separates well preserved from poorly preserved solution-etched calcareous particles. The depth of the lysocline varies but it generally lies between 3000 and 5000 m. At water depths less than the lysocline, calcareous particles accumulate without loss through dissolution. The depth of the CCD varies as a function of a number of variables that reflect oceanic productivity patterns and the shoaling of the lysocline near continental margins. The CCD varies between 3500 and 5500 m in the Atlantic and Pacific oceans but has a mean depth of around 4500 m. The difference between the lysocline and CCD depths is also not constant, because this depends on the gradient of the concentration of carbonate ions in the water column overlying the sediments. However, the depth at which the calcite content of sediments falls to only a few percent is typically about 700 m deeper than the lysocline.

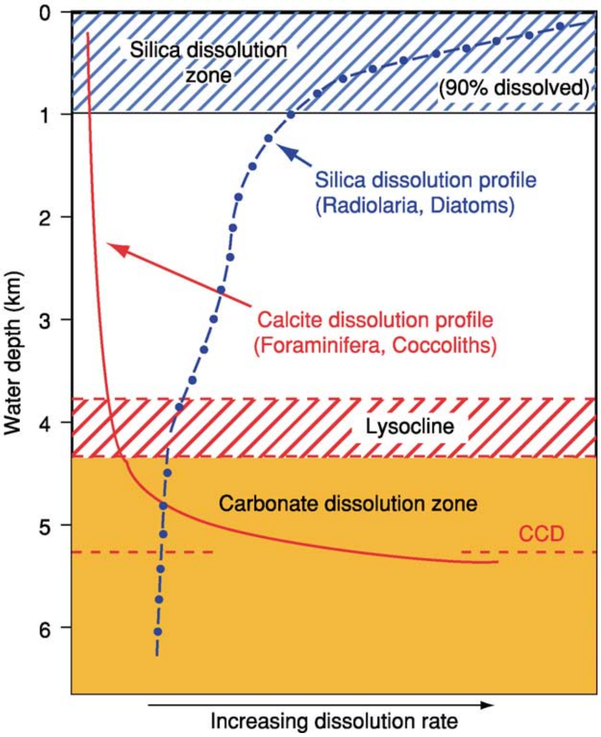

0
1
2
3
4
5
6
Water depth (km)
Silica dissolution zone
(90% dissolved)
Silica dissolution profile (Radiolaria, Diatoms)
Calcite dissolution profile (Foraminifera, Coccoliths)
Lysocline
Carbonate dissolution zone
CCD
Increasing dissolution rate

Figure 3 Generalized dissolution profiles of silica and calcite. Note that the depth of the lysocline, taken as the level below which there is rapid increase in calcite dissolution, and the calcite compensation depth (CCD), which is the depth at which the rate of supply of biogenic calcite equals its rate of dissolution, vary within and between ocean basins. Reproduced with permission from Douglas RG (2003) Oceanic sediments (Figure 5). In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks*, pp. 481–492. Dordrecht: Kluwer Academic Publishers.

Types of Pelagic Deposits

Pelagic sediments fall into two broad groups based on composition, deep-water pelagic clay and biogenic oozes. Deep-water pelagic clay is found in deep-ocean areas far from land, where solution has removed the biogenic component and only insoluble inorganic material, much of it wind-derived, remains. Biogenic oozes are composed largely of biogenic planktonic debris derived from surface water productivity (they contain more than 30% biogenic debris). Biogenic oozes are largely composed of the remains of zoo- and phytoplankton such as foraminifera, coccolithophores, pteropods, diatoms, and radiolaria. In the upper water column, these remains are biologically 'packaged' and 'repackaged' into larger particles, which hastens their descent to the seafloor (e.g., as faecal pellets or phytoplankton aggregates). Indeed, most of the organic and skeletal matter produced in the euphotic zone is consumed and only a fraction is exported, and a fraction of this reaches the deep seafloor, where more is destroyed by dissolution. The distribution of biogenic oozes is strongly depth controlled due to dissolution of calcium carbonate with depth. Two main types are recognized, calcareous oozes, the composition of which is dominated by the remains of calcareous plankton, and siliceous oozes, which are dominated by the remains of siliceous plankton. Siliceous oozes lithify into radiolarites, diatomites, and cherts, whereas calcareous oozes lithify into pelagic chalks and limestones, and examples are well known from the geological record, well-documented examples occurring in the Troodos Massif, Cyprus, and the Ligurian Apennines in Italy.

Calcareous Oozes

Calcareous oozes may be dominated either by the tests and test debris of planktonic foraminifera (termed 'foraminiferal ooze') (Figure 4) or by the remains of planktonic plants (coccolithophores; termed 'nannofossil ooze'). In either type of calcareous ooze, the other component will often be the second most important constituent. In the modern world ocean, ~50% of the seafloor is blanketed by foraminiferal ooze (Table 1). Calcareous oozes commonly also contain a terrigenous fraction (which may amount to 10–15%), composed mainly of quartz and clay minerals, but may contain trace amounts of pyrite, iron and manganese precipitates, mica, chert, rock fragments, glauconite, feldspar, ferromanganese minerals, detrital carbonate, zeolites, volcanic glass, and cosmic spherules. Minor biogenic components may include benthonic (bottom-dwelling) foraminifera, ostracods, echinoid remains, radiolaria, silicoflagellates, diatoms, sponge spicules, pteropod shells and shell debris (in shallow water), phosphatic vertebrate remains and fish teeth.

Pteropods (pelagic gastropods) are relatively common zooplankton, especially in warm-water latitudes, and some forms secrete delicate aragonitic shells. Pteropod shells may range up to 30 mm in length, although most are in the range 0.3 to 10 mm. Aragonite is unstable and dissolves as ocean waters become undersaturated in respect to carbonate with depth. Consequently, pteropod-rich oozes are only found at depths shallower than 2500 m in the Atlantic Ocean and shallower than 1500 m in the Pacific Ocean.

Foraminifera comprise a group of protozoans characterized by a test of one to many chambers composed of secreted calcite or agglutinated grains. Test sizes are generally in the range 0.05–1 mm. Forms with agglutinated tests are typically benthonic (bottom-dwelling) and make only a very minor contribution to pelagic sediments, which are overwhelmingly dominated by the remains of globular planktonic forms. Modern species show clear latitudinal distribution patterns related to water temperature. Oxygen isotope analysis of planktonic foraminifera tests can provide estimates for past

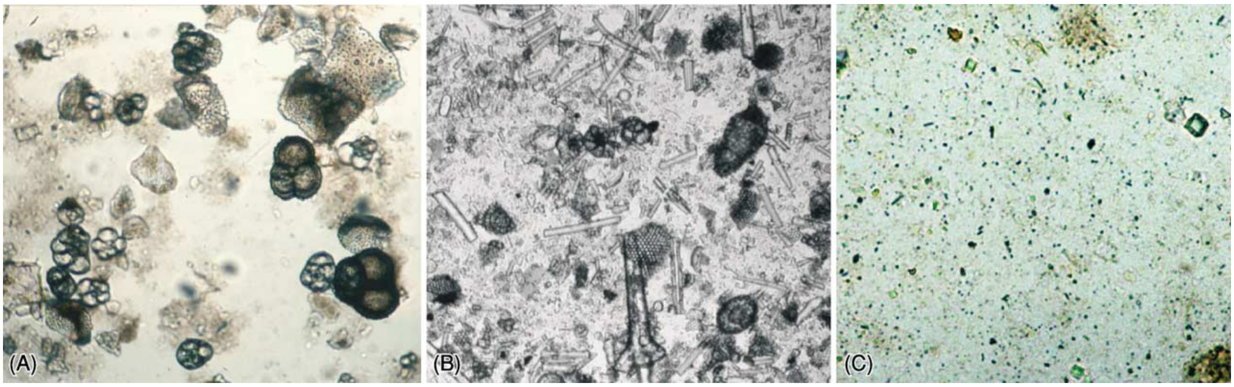


Figure 4 Illustration showing the three main types of pelagic sediments as seen under the microscope in plane-polarized-light. Left: Calcareous ooze from the North Atlantic Ocean, comprising mainly planktonic foraminifer tests and test fragments. The larger complete foraminifer tests are about 0.1 mm across. Centre: Siliceous ooze from the South Atlantic Ocean, comprising mainly silica sponge spicules (tubular forms), radiolaria (high-relief bell-shaped and circular forms, right of centre), and broken centric diatom frustules (lower left and centre). Two planktonic foraminifera can be seen in the upper centre field. The foraminifera are about 0.05 mm across. Right: North-east Atlantic Ocean pelagic red clay containing rhomboid dolomite crystals. The red colour is due to the presence of amorphous or poorly crystalline iron oxide minerals and grain coatings. The largest dolomite rhomb (upper right) is about 0.01 mm across.

Table 1 Coverage of the deep-ocean floor by pelagic sediments[a]

	Seafloor coverage (%)			
Sediment type	*Atlantic Ocean*	*Pacific Ocean*	*Indian Ocean*	*Total World Ocean*
Foraminiferal and nannofossil ooze	65.1	36.2	54.3	47.1
Pteropod ooze	2.4	0.1	—	0.6
Diatom ooze	6.7	10.1	19.9	11.6
Radiolarian ooze	—	4.6	0.5	2.6
Pelagic clay	25.8	49.0	25.3	38.1

[a]Data from Open University (1991) *Ocean Chemistry and Deep-Sea Sediments*, 2nd edn. Oxford: Pergamon Press.

sea-surface temperatures and salinities. Isotope data from benthonic forms allow reconstruction of bottom-water mass histories. Foraminifera hence can provide important information on thermohaline structure and circulation patterns in ancient oceans.

Coccoliths are minute, usually oval, calcite plates produced by unicellular planktonic algae (family Coccolithophoridae); because of their small size, coccoliths are referred to as nannoplankton. In life, coccolith plates, eight or more in number, depending on the species, are attached to a membrane surrounding a living cell. Each organism (i.e., the cell surrounded by coccolith plates) is termed a 'coccosphere'. Coccospheres are generally spherical, usually 5–30 μm in diameter. The individual coccolith plates are usually around 3 μm in diameter, although some forms can be as large as 35 μm. On the death of the organism, the membrane holding the coccolith plates disintegrates, releasing the coccoliths to contribute to calcareous oozes. Coccoliths are single calcite crystals and are more resistant to dissolution than the tests of foraminifera or pteropods are. Globally, their diversity increases from a minimum in subpolar seas to a maximum in tropical and equatorial waters; and species distribution is closely linked to water masses.

Siliceous Oozes

Siliceous oozes are largely composed of the opaline silica tests and test fragments of siliceous plankton (Figure 4). Again, there are two main varieties: radiolarian ooze, composed mainly of the tests of radiolarians, and diatom ooze, dominated by the siliceous remains of unicellular plants (diatoms). Both types may contain minor amounts of silicoflagellates. Some sediments (for example, in some high-latitude abyssal environments and near spreading ridges) may also contain significant numbers of siliceous sponge spicules (Figure 4). Typically, siliceous ooze is present only in regions of high biological surface water productivity (such as the equatorial and polar belts and areas of coastal upwelling), where the depth of the seafloor is deeper than the calcite compensation depth. In the North Pacific and Antarctic belts of siliceous oozes, diatoms make up as much as 95% of the bulk sediment. The mineralogical composition of the detrital fraction of siliceous oozes is commonly similar to that of calcareous oozes, with quartz and clay minerals being the dominant detrital minerals.

Radiolaria are a diverse group of planktonic, pseudopod-bearing protozoans characterized by transparent opal skeletons. These exquisitely structured lattices are often of great complexity. Radiolarian tests show a great variety of shapes, but most are based on conical, spherical, or helmet-shaped forms. Most radiolarians are within the size range 20–400 μm. They are particularly abundant and diverse in equatorial latitudes (especially in areas of upwelling) and in subpolar seas. Radiolarian oozes occur mainly in the equatorial Pacific.

Diatoms are single-celled algae that secrete a test (called the frustule) of opaline silica. They are a major part of the phytoplankton and typically occur as pinnate (spindle-, rod-, or wedge-shaped) or centric (discoidal, spherical, elliptical, or oblong) forms. Most planktonic diatoms are centric types, although in Antarctic Seas, planktonic pinnate forms occur. They generally fall within the size range 10–100 μm. Diatoms represent most of the suspended silica in the water column and are the main contributors to deep-sea siliceous sediments. However, although in productive areas diatom concentrations are many millions of frustules per cubic metre, most tests are redissolved in the water column, because surface waters are greatly undersaturated in respect to silica due to high biological demand (Figure 3). Indeed, in areas of low silica supply, diatom assemblages in sediments are commonly biased to dissolution-resistant robust forms rather than to more fragile species. Diatoms are particularly abundant in regions of high productivity, especially in high latitudes and areas of upwelling.

Silicoflagellates are small unicellular flagellated marine plankton with internal skeletons of opaline silica. These skeletons consist of hollow rods arranged in a lattice, a common arrangement comprising a basal ring from which rods arise on one side to form an arch or dome, resulting in an overall hemispherical shape. Most silicoflagellates are in the size range of 20–50 μm. Although silicoflagellates are widespread in sediments, they are seldom abundant, so do not make a significant contribution to marine sediments.

Pelagic Clays

Deep-water pelagic clays (sometimes called 'red clay') are found only in deep-ocean areas, generally below water depths of 4000 m, far from land. Such clays cover large areas of the seafloor, particularly in the Pacific, southern Atlantic, and southern Indian oceans (Figure 1), in areas remote from terrigenous sources and below the calcite compensation depth. The reddish-brown appearance of these clays, first noted on the *Challenger* expedition, is due to the presence of amorphous or poorly crystalline iron oxide minerals and grain coatings. Pelagic clays usually contain less than 10% biogenic material and are mainly composed of fine-grained quartz and clay minerals, the bulk of which is derived from aeolian fallout and has been slowly deposited from fine suspensions. Typically, 75–95% of pelagic clay deposits consists of clay minerals with a grain size of less than 3 μm (Figure 4). These clay minerals are dominated by illite, smectite, kaolinite, and chlorite, with illite as the main type. Illite is, in fact, largely land derived and is transported to the ocean by rivers and glaciers and as windblown dust. Both kaolinite (a product of humid tropical weathering) and chlorite (typically derived from low-grade metamorphic rocks) are also mainly land derived. Smectite, however, is a low-temperature alteration product of volcanic ash and is particularly widespread on the Pacific Ocean floor. Wind transport is the major mechanism by which land-derived clay, fine-grained silt (commonly quartz), and dust reach the ocean surface, ultimately to be deposited in pelagic clays. The highest rates of aeolian dust deposition (up to 1000 mg cm^{-2} ky^{-1}) are in the north-western Pacific downwind of far-east Asia. Substantial fluxes of windblown dust also enter the deep ocean offshore of the Sahara, South Africa, the Arabian peninsula, and the Horn of Africa and around Australia. The origin of wind-derived material in pelagic clays can be determined by rare earth geochemistry and study of Sr and Nd isotopes. Pelagic clays also commonly contain significant amounts of authigenic minerals, such as chert, zeolites, apatite, phosphorite, volcanic glass, and manganese micronodules, as well as indicators of slow sedimentation, such as fish debris and cosmic spherules. Pelagic clays may also contain varying amounts of feldspar, pyroxenes, and mica. In total, pelagic clays cover about 38% of the modern seafloor (Table 1).

Ferromanganese Deposits

New mineral phases may be formed on the seafloor (a process known as authigenesis) either by direct precipitation from seawater or by the alteration of pre-existing minerals or grains. Ferromanganese deposits are the most common and probably the most widely known authigenic deposits found on the deep-ocean seafloor. They occur as encrustations or crusts on submarine rock outcrops, or as discrete nodules and concretions on the seafloor. Ferromanganese crusts, which grow on exposed rock surfaces, acquire the elements necessary for their growth directly from seawater. Ferromanganese nodules occur throughout the sediment column, but the greatest concentrations are found on the surface; the nodules range in size from the microscopic (called 'micronodules', usually in the silt-sand size range) to the macroscopic, reaching several centimetres across (Figure 5). There are two main controls on nodule abundance: (1) the rate of accumulation of the host sediment, with the highest number of nodules being found on sediments with low accumulation rates (e.g., a few millimetres per thousand years), and (2) the presence of suitable accretion nuclei for the nodules to grow around; the nuclei may be small clumps of sediment, fragments of volcanic rock, shark teeth or teeth fragments, or even foraminifer tests. Ferromanganese nodules can show a great variety of shape and size and are found in all oceans, but are particularly common on the deep Pacific seafloor. Nodules may form by precipitation from the overlying seawater and from elements supplied from interstitial porewaters below the sediment surface, or through a combination of both element sources. The shape of the nodule may reflect the dominant source of elements available for its precipitation and growth. For circular nodules, it is thought that the dominant supply of metals is from the overlying seawater, whereas for discoidal, flattened nodules the dominant supply of metals may be via interstitial porewaters below the seabed. Nodule growth rates are slow, varying from a few to a few hundred millimetres per million years. Ferromanganese nodules are rich in iron and manganese as their name implies, but also contain relatively high concentrations of a number of trace metals, including cobalt, molybdenum, thorium, nickel, silver, iridium, and lead. Deep-sea ferromanganese nodules may someday become an important economic resource.

Metalliferous sediments, including iron and manganese-rich mudstones (termed 'umbers') and iron-rich sediments (termed 'ochres'), are frequently associated with ophiolites (fragments of oceanic crust that have been tectonically emplaced onto continental margins), well-known examples of which occur in Cyprus and Oman. These record past oceanic sediments that have contained hydrothermal minerals or authigenic ferromanganese deposits.

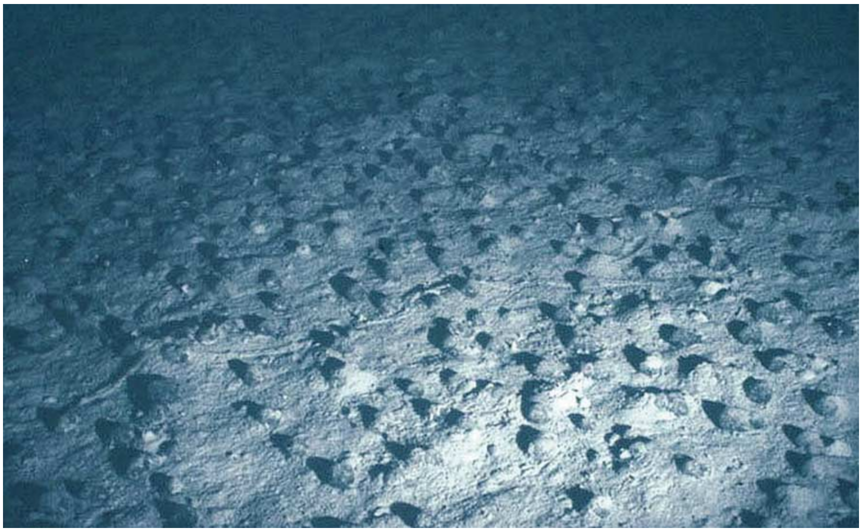


Figure 5 Deep-sea photograph, showing a field of ferromanganese nodules over part of the Madeira Abyssal Plain, north-east Atlantic Ocean. Individual nodules are a few centimetres across; water depth is 5400 m.

Biogenic Sedimentation in the World Ocean

Early researchers such as John Murray believed that pelagic sediments must accumulate slowly, but it was Wolfgang Schott, studying cores collected by the German *Meteor* expeditions of 1925–27, who was first able to demonstrate that Atlantic calcareous oozes had accumulated at rates of several centimetres thickness per thousand years. He found that the distinctive tropical foraminifer *Globorotalia menardii* was absent in glacial-period sediments in the North Atlantic, and that its appearance correlated with the start of the Holocene. Because the age of the base of the Holocene was known from land sections, Schott was able to determine accumulation rates for Atlantic pelagic calcareous oozes for this time period. Today, a wide variety of dating techniques (for example, radiocarbon dating and uranium series dating) can be used to determine accumulation rates. Pelagic sedimentation rates do vary considerably, but pelagic clays accumulate the most slowly (typically 0.1–0.5 cm ky^{-1}), whereas sedimentation rates for calcareous oozes are typically in the range 0.3–5 cm ky^{-1} and siliceous oozes are in the range 0.2–1 cm ky^{-1}.

Biogenic sediments show considerable variation in both space and time. In the present-day Atlantic Ocean, pelagic sediments are predominantly calcareous and siliceous sediments are virtually absent in the North Atlantic. In the Pacific, however, calcareous sediments are limited to oceanic ridges, plateaus, and seamounts (at water depths less than 3500 m) and also occur as a broad belt in the southern central Pacific. Siliceous sediments are widespread in the North Pacific, along the equator and adjacent to Antarctica. Calcareous sediments occur along the mid-ocean ridges in the Indian Ocean and siliceous sediments are widespread in the northern and southern Indian Ocean (Figure 1). Pelagic sediment distribution reflects both seafloor depth and ocean fertility. Where nutrient supply is low and surface waters are nutrient poor (especially in dissolved silica), sinking particles deliver to the seafloor more carbonate than silica (low Si/Ca ratio), which will be preserved, providing the seafloor lies above the CCD. Low nutrient supply favours production of coccolithophorids, which are fed on by small foraminifera, and long food chains develop in the euphotic zone. Foraminifera and coccoliths therefore dominate export to the seafloor. Where nutrient concentrations in surface waters are high, such as at upwelling areas and ocean divergence zones, diatoms will be the primary producers. Diatoms can reproduce rapidly and produce dense blooms (10^7 frustules per cubic metre). Food chains in these regions tend to be short because large diatoms are eaten by large zooplankton and fish (high trophic-level consumers). Export to the seafloor is high in silica and organic carbon, and flux rates are high, leading to siliceous sediment deposition. Further, bacterial decomposition of the organic carbon results in production of carbonic acid, which dissolves carbonate grains. In this way, carbonate is removed and the siliceous content as a proportion of total sediment increased. Deep-ocean circulation also leads to fractionation of silica and carbonate between ocean basins. In the North Atlantic, deep-water outflow is exchanged for surface water inflow and bottom waters are young, nutrient poor, well oxygenated and saturated in respect to calcium carbonate. Sediments deposited here tend to be calcareous. In the North Pacific, in contrast, deep water is old and poorly oxygenated but nutrient rich, because surface water outflow is exchanged for deep-water inflow. Before reaching the North Pacific, the deep water has flowed through the southern Atlantic and Indian oceans, hence its age and low oxygen content. However, during this long passage, microbial breakdown of organic matter (which has depleted oxygen) produces CO_2 and regenerates nutrients. These waters therefore become undersaturated in regard to calcium carbonate but are enriched in nutrients and dissolved silica. Upwelling of this water will cause high surface productivity and diatom production, resulting in deposition of siliceous oozes with little calcareous content. Thus, pelagic sediment distribution is determined by bottom water circulation, which controls both the rate of particle dissolution and the productivity of surface waters through upwelling. In this way, in the modern ocean, the Atlantic is depositing carbonate and exporting silica, whereas in the Pacific, the reverse is happening. However, changes in climate and continental positioning and ocean connectivity, caused by plate motion, will affect ocean chemistry and fertility, and hence pelagic sediment deposition and distribution.

Data from the DSDP and ODP have shown that the distribution and relative abundance of seafloor sediment types have changed with time. Biogenic sediments were even more widely distributed in Cretaceous and Early Tertiary time. The deep-ocean sedimentary record provides a most important source for our knowledge of the past Earth, particularly regarding ocean fertility, geochemistry, evolution of marine biota, and past wind regimes and patterns.

See Also

Fossil Plants: Calcareous Algae. **Microfossils:** Foraminifera. **Sedimentary Processes:** Deposition from Suspension. **Sedimentary Rocks:** Oceanic Manganese Deposits.

Further Reading

Burton JD (1996) The ocean: a global geochemical system. In: Summerhayes CP and Thorpe SA (eds.) *Oceanography – An Illustrated Guide,* pp. 165–181. London: Manson Publishing.

Chester R (2000) *Marine Geochemistry,* 2nd edn. [particularly chs. 13, 15, and 16]. Oxford: Blackwell Science.

DeMaster DJ (2004) The diagenesis of biogenic silica: chemical transformations occurring in the water column, seabed and crust. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry,* vol. 7, pp. 87–98. Oxford: Elsevier-Pergamon.

Douglas RG (2003) Oceanic sediments. In: Middleton GV (ed.) *Encyclopedia of Sediments and Sedimentary Rocks,* pp. 481–492. Dordrecht: Kluwer Academic Publishers.

Li Y-H and Schoonmaker JE (2004) Chemical composition and mineralogy of marine sediments. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry,* vol. 7, pp. 1–35. Oxford: Elsevier-Pergamon.

Martin WR and Sayles FL (2004) The recycling of biogenic material at the seafloor. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry,* vol. 7, pp. 37–65. Oxford: Elsevier-Pergamon.

Morse JW (2004) Formation and diagenesis of carbonate sediments. In: MacKenzie FT (ed.) *Sediments, Diagenesis and Sedimentary Rocks, Treatise on Geochemistry,* vol. 7, pp. 67–85. Oxford: Elsevier-Pergamon.

Open University (1991) *Ocean Chemistry and Deep-Sea Sediments,* 2nd edn. Oxford: Pergamon Press.

Rothwell RG (1989) *Minerals and Mineraloids in Marine Sediments – An Optical Identification Guide.* Barking: Elsevier Applied Science.

Seibold E and Berger WH (1993) *The Sea Floor – An Introduction to Marine Geology,* 3rd edn. [particularly chs. 3 and 6–9]. Berlin: Springer-Verlag.

Stow DAV, Reading HG, and Collinson JD (1996) Deep seas. In: Reading HG (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy,* 3rd edn., pp. 395–453. Oxford: Blackwell Science.

Whitmarsh RB, Bull JM, Rothwell RG, and Thomson J (1996) The evolution and structure of ocean basins. In: Summerhayes CP and Thorpe SA (eds.) *Oceanography – An Illustrated Guide,* pp. 113–135. London: Manson Publishing.

Dolomites

H G Machel, University of Alberta, Edmonton, Alberta, Canada

Introduction

Dolomite was first described in 1791 as a rock by Deodat de Dolomieu, who investigated samples from the Italian Alps. Dolomites are of special interest because they often form hydrocarbon reservoir rocks. Despite intensive research for more than 200 years, the origin of dolomites is subject to considerable controversy. This is because some of the chemical and/or hydrological conditions of dolomite formation are poorly understood, and because the available data often permit more than one viable genetic interpretation. This article covers the thermodynamic and kinetic conditions that favour dolomitization, mass balance considerations for the generation of massive dolostones, dolomite textures and pore spaces in dolostones, geochemical methods that are used in dolomite case studies, an overview of the various dolomitization models, and a brief section on secular variations in dolomite abundance.

Basics

Ideal, ordered 'dolomite' has a formula of $CaMg(CO_3)_2$ and consists of alternating layers of Ca^{2+}–CO_3^{2-}–Mg^{2+}–CO_3^{2-}–Ca^{2+}, etc., perpendicular to the crystallographic c axis. Most natural dolomite contains up to a few per cent Ca surplus (and a corresponding Mg deficit), as well as less than ideal ordering. 'Protodolomite' contains about 55–60% Ca, is poorly ordered, i.e., the alternating cation layer structure is poorly developed, and is common as a metastable precursor of well-ordered, nearly stoichiometric dolomite in both laboratory experiments and in nature. Good arguments have been made to abandon the term protodolomite or to restrict it to laboratory products, yet the term is useful to describe metastable precursors of dolomite in nature. The term 'dolostone' refers to a rock that consists largely (>75%) of the mineral dolomite. This term has been rejected by some, but has gained wide acceptance during the last 20 years. The term 'dolomites' is the best term to use to refer to types of dolomite that vary in texture, composition, genesis, or a combination thereof.

Two types of 'dolomite formation' are common, i.e., 'dolomitization', which is the replacement of $CaCO_3$ by $CaMg(CO_3)_2$, and 'dolomite cementation', which is the precipitation of dolomite from aqueous solution as a cement in primary or secondary pore spaces. Dolomites and dolostones that originate via replacement of $CaCO_3$ are called 'replacement dolomites' or 'secondary dolomites', especially in the older literature. A third type of dolomite formation is direct precipitation from aqueous solution to form sedimentary deposits. Dolomites that form in this way may be called 'primary dolomites'.

Genetically, all natural dolomites can be placed into two major families, i.e., 'penecontemporaneous' dolomites and 'postdepositional' dolomites. Penecontemporaneous dolomites may also be called 'syndepositional' dolomites. They form while a carbonate sediment or limestone still resides in the original environment of deposition as a result of the geochemical conditions that are 'normal' for that environment. Such dolomites are also called 'primary' or 'early diagenetic', although these terms are not strictly synonymous with penecontemporaneous. True penecontemporaneous dolomites appear to be relatively rare. Most known cases are of Holocene age, and are restricted to certain evaporitic lagoonal and/or lacustrine settings. It is quite possible, however, that such dolomites are much more common in the geological record than presently known, but their presence is hard to prove because of later diagenetic overprinting.

'Postdepositional' dolomites may also be called 'postsedimentary'. They form after a carbonate sediment has been deposited and removed from the active zone of sedimentation, which may happen via progradation of the sedimentary surface, burial and subsidence, uplift and emergence, eustatic sea-level fluctuations, or any combination of these. Such dolomites and dolostones are often called 'late diagenetic', although this term is not synonymous with postdepositional. Almost all known examples of massive, regionally extensive dolostones are postdepositional.

One aspect that transcends the above genetic grouping is that of hydrology. Whether syndepositional or postdepositional, the formation of large amounts of dolomite requires advection, i.e., fluid flow, because of chemical mass balance constraints. On the other hand, small amounts of dolomite can be formed without advection. In such cases, the Mg for dolomite formation is locally derived and redistributed, or supplied via diffusion. Examples include dolomite formed from Mg that was contained in (high-)Mg calcite, adsorbed to the surfaces of minerals, organic substances, or biogenic silica, or that was contained in older primary or secondary dolomites.

Thermodynamic and Kinetic Constraints

The thermodynamic conditions of dolomite formation have been known quite well since at least the 1970s. The kinetics, however, i.e., the catalysts and inhibitors of dolomite formation, are relatively poorly understood and continue to be a source of controversy.

According to the present state of knowledge, dolomite formation is favoured chemically, i.e., thermodynamically and/or kinetically, under the following conditions: (1) low Ca^{2+}/Mg^{2+} ratios, (2) low Ca^{2+}/CO_3^{2-} ratios (high carbonate alkalinity), (3) high temperatures, and (4) salinities substantially lower or higher than that of seawater. These constraints translate into four essential and common conditions for the formation of dolostones in natural settings:

- Settings with a sufficient supply of Mg^{2+} and CO_3^{2-}. This condition favours marine settings and burial diagenetic settings with pore fluids of marine parentage, because seawater is the only common Mg-rich natural fluid in sedimentary/diagenetic settings.
- Settings with a long-lasting and efficient delivery system for Mg^{2+} and/or CO_3^{2-} (also exporting Ca^{2+} in the case of calcite replacement). This favours settings with an active and long-lasting hydrological drive.
- Carbonate depositional settings and/or limestones that can be replaced.
- Settings in which fluids suddenly release CO_2, i.e., from hydrothermal solutions that ascend rapidly via fault systems.

Considering that the above chemical constraints allow dolomite formation in almost the entire range of surface and subsurface diagenetic settings, the question arises as to why there are so many undolomitized limestones. The essential conditions for the common lack of dolomitization appear to be:

- Ion pair formation (especially hydration), inactivating much of the Mg^{2+} and CO_3^{2-} in solution.
- Insufficient flow because of the lack of a persistent hydraulic head, too small a hydraulic head, or insufficient diffusion, resulting in insufficient Mg^{2+} and/or CO_3^{2-} supply.
- The limestones are cemented and not permeable enough, inhibiting or prohibiting the throughput of Mg-rich waters.
- The diagenetic fluids are incapable of forming dolomite because of kinetic inhibition, e.g., because the environment is too cold (most kinetic inhibitors of dolomite nucleation and growth are rather potent at temperatures below about 50°C), and the Ca^{2+}/Mg^{2+} ratio of many cold diagenetic fluids is not low enough for dolomitization.

The last point leads to kinetic factors, three aspects of which deserve special mention. Firstly, almost all researchers agree that most kinetic inhibitors that lower the nucleation rate and growth rate of dolomite are especially potent at temperatures below about 50°C. Hence, dolomite formation is easier at higher temperatures. Secondly, it is also generally acknowledged that dolomite forms via metastable precursors, but the significance of this phenomenon for the formation of massive dolostones is not clear. Thirdly, sulphate has been shown to increase as well as decrease the rate of dolomitization, and thus the role of sulphate is not clear, and may vary from place to place, depending mainly on fluid composition and temperature.

Mass Balance Constraints

Within the chemical constraints outlined in the previous section, the amount of dolomite that can be formed in a given diagenetic setting depends on the stoichiometry of the reaction, temperature, and fluid composition. Dolomitization can be represented by two equations, i.e.,

$$2CaCO_3(s) + Mg^{2+}(aq) \rightarrow CaMg(CO_3)_2(s) + Ca^{2+}(aq) \quad [1]$$

where '(s)' is solid and '(aq)' is aqueous, or by

$$CaCO_3(s) + Mg^{2+}(aq) + CO_3^{2-}(aq) \rightarrow CaMg(CO_3)_2(s) \quad [2]$$

Reactions [1] and [2] are end members of a range of possible reaction stoichiometries, i.e.,

$$(2-x)CaCO_3(s) + Mg^{2+}(aq) + xCO_3^{2-}(aq) \rightarrow CaMg(CO_3)_2(s) + (1-x)Ca^{2+}(aq) \quad [3]$$

Reaction [3] can be used to represent dolomitization in general, as it encompasses reactions [1] and [2]. For $x=0$, reaction [3] becomes reaction [1] and, for $x=1$, reaction [3] becomes reaction [2]. Dolomite cementation is most simplistically represented by

$$Ca^{2+}(aq) + Mg^{2+}(aq) + 2CO_3^{2-}(aq) \rightarrow CaMg(CO_3)_2(s) \quad [4]$$

If dolomitization proceeds via reaction [1], and if the dolomitizing solution is normal seawater, about 650 m^3 of solution is needed to dolomitize 1 m^3 of limestone with 40% initial porosity at 25°C. Dolomitization may not take place with 100% efficiency, however, and some Mg in excess of saturation is carried away by the dolomitizing solution. In such cases, larger water/rock ratios are required for complete dolomitization. If seawater is diluted to 10% of its original concentration, as is the case in a typical seawater–freshwater mixing zone, ten times as much water is needed. On the other hand, only about 30 m^3 of brine is needed per cubic metre of limestone at 100% dolomitization efficiency in the case of a halite-saturated brine. The role of increasing temperature in the underlying thermodynamic calculations is to reduce the amount of Mg necessary for dolomitization, because the equilibrium constant (and hence the equilibrium Ca/Mg ratio) is temperature dependent. For example, at 50°C, only about 450 m^3 of seawater is needed for complete dolomitization of 1 m^3 of limestone with 40% initial porosity at 100% efficiency. The amounts of dilute and hypersaline waters change accordingly.

These calculations have two major implications. Firstly, large water/rock ratios are required for complete dolomitization, and the more dilute the solution, the larger the water/rock ratio. This necessitates advection for extensive and pervasive dolomitization, which is why all models for the genesis of massive dolostones are essentially hydrological models. The exceptions are natural environments in which carbonate muds or limestones can be dolomitized via diffusion of magnesium from seawater rather than by advection. Secondly, variable reaction stoichiometries result in variable porosity development during dolomite formation (see below).

Rock and Pore Classifications

Crystal size distributions are classified as ‘unimodal’ and ‘polymodal’, whereas crystal shapes are classified as ‘planar-e’ (euhedral), ‘planar-s’ (subhedral), and ‘nonplanar-a’ (anhedral). Using this semantic scheme, almost all other dolomite texture types can be named, i.e., planar-c (cement), planar-p, and nonplanar-p (both porphyrotopic). Saddle dolomite, with its distinctive warped crystal faces, is simply called nonplanar (Figure 1). A complete textural description includes recognizable allochems or biochems, matrix, and void fillings. Particles and cements may be unreplaced, partially replaced, or completely replaced. Replacement may be mimetic or non-mimetic, which can be added to a rock description, such as ‘unimodal, non-mimetic, planar-s dolomite’.

Pores in dolostones are commonly addressed using the same classification as for limestones, with pore types such as mouldic, vuggy, shelter, etc. This

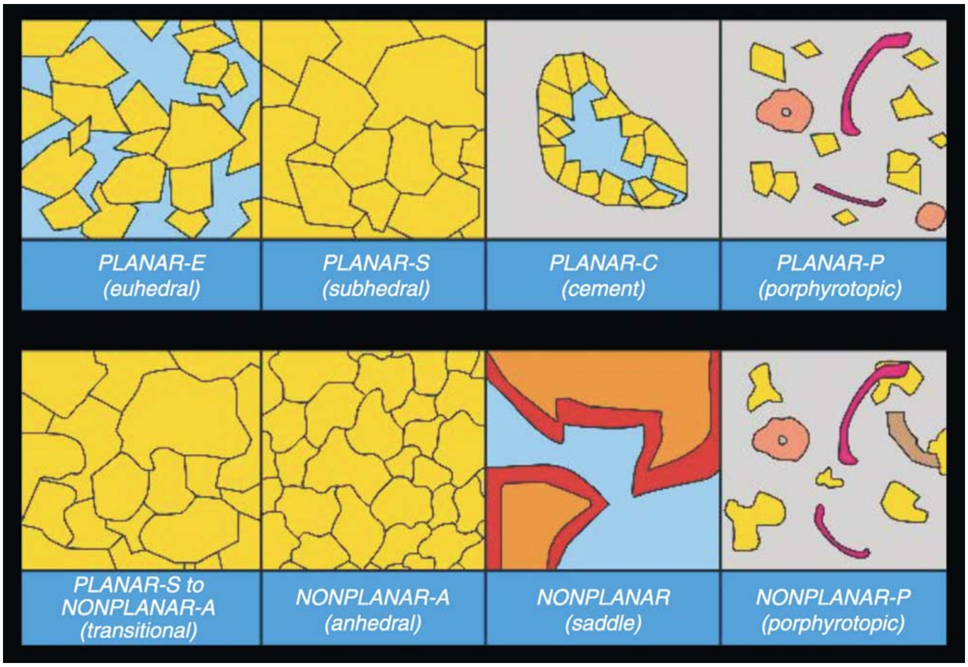


Figure 1 Dolomite textural classification combined from Gregg and Sibley (1984) and Sibley and Gregg (1997), supplemented by a ‘transitional’ form. Reproduced with permission from Wright WR (2001) Dolomitization, fluid-flow and minerlization of the Lower Carboniferous rocks of the Irish Midlands and Dublin Basin. Unpub. Ph.D. thesis, Univerisity College Dublin, Belfield, Ireland, 407 p.

classification is independent of pore size. The latter, however, is of special interest for the petroleum industry. Another classification contains categories in size/magnitude from the very smallest to the very largest, i.e., from mercury injection capillary measurements (MICPM) and scanning electron microscopy (SEM) to karst caverns, respectively (Figure 2).

Textural Evolution

The textures and reservoir characteristics of natural dolostones are highly variable. On the microscopic scale, a unimodal size distribution generally results from a single nucleation event and/or a unimodal primary (pre-dolomite) size distribution of the substrate. Polymodal size distributions result from multiple nucleation events and/or a differential nucleation on an originally polymodal substrate. Planar crystal boundaries tend to develop at growth below about 50°C (the so-called 'critical roughening temperature'), whereas nonplanar boundaries tend to develop at $T > 50°\mathrm{C}$ and/or high degrees of supersaturation.

Within this framework, observations from many dolostone occurrences show that dolomitization often proceeds in a certain sequence of steps that correspond to certain textural types on the macroscopic scale. Within limits, these steps correspond to certain types of dolomitizing fluids (especially seawater and its derivatives) and/or meteoric water incursion. The most common sequence includes:

1. *Matrix-selective dolomitization*. Dolomitization begins as a selective replacement of the matrix (Figure 3).
2. *Vugs and moulds*. Holes resulting from the dissolution of undolomitized fossils and allochems (Figures 4 and 5).
3. *Emplacement of calcium sulphate*. Commonly anhydrite, both as a replacement and as a cement during advanced dolomitization from seawater (Figures 5 and 6).
4. *Development of two dolomite populations*. A smaller sized population with 'cloudy' centres with or without clear rims (overgrowths), and a larger population (Figures 7 and 8) resulting either from recrystallization or inherited from primary textural features.
5. *Dolomite cementation ('overdolomitization')*. Dolomite cement as overgrowth on the earlier formed dolomite crystals.

Furthermore, outcrop evidence shows that there is a distinct difference in the textures resulting from 'low-temperature' versus 'high-temperature' dolomitization of limestones. Empirical evidence suggests that the range of 50–80°C approximately marks the boundary between these two temperature realms. In the low-temperature settings, dolomitization commonly is matrix selective and at least partially fabric retentive, as discussed earlier, whereas dolomitization tends to be fabric destructive in the high-temperature settings (Figures 9 and 10). However, there are counterexamples.

Saddle dolomite (Figures 11 and 12) is a special type of dolomite. Its crystallographic, geochemical, and paragenetic characteristics suggest formation at temperatures above about 80°C. Saddle dolomite forms from stylolitization of older dolomites, as a

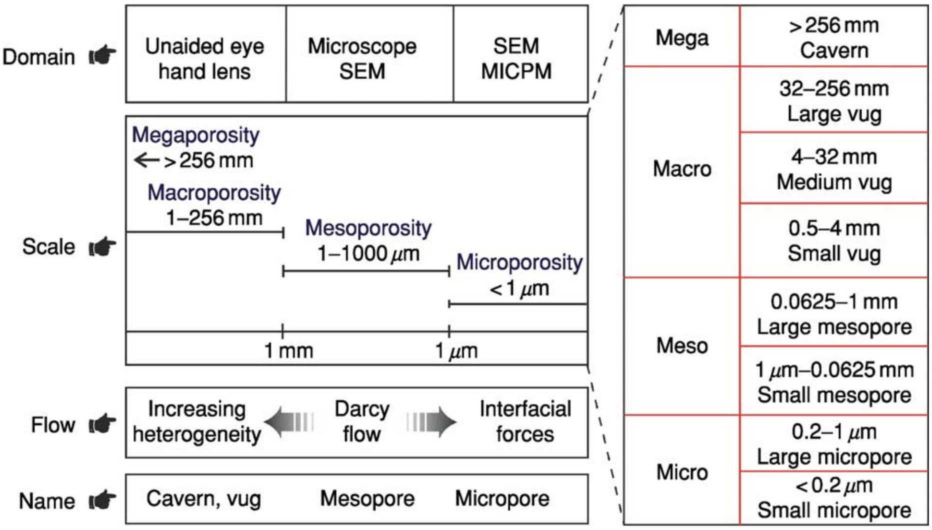


Figure 2 Pore size classification for carbonates. Measurements under 'scale' refer to pore diameters. MICPM, mercury injection capillary measurements; SEM, scanning electron microscopy. Reproduced from Luo and Machel (1995). Reprinted by permission of the AAPG whose permission is required for further use. AAPG ©1995.

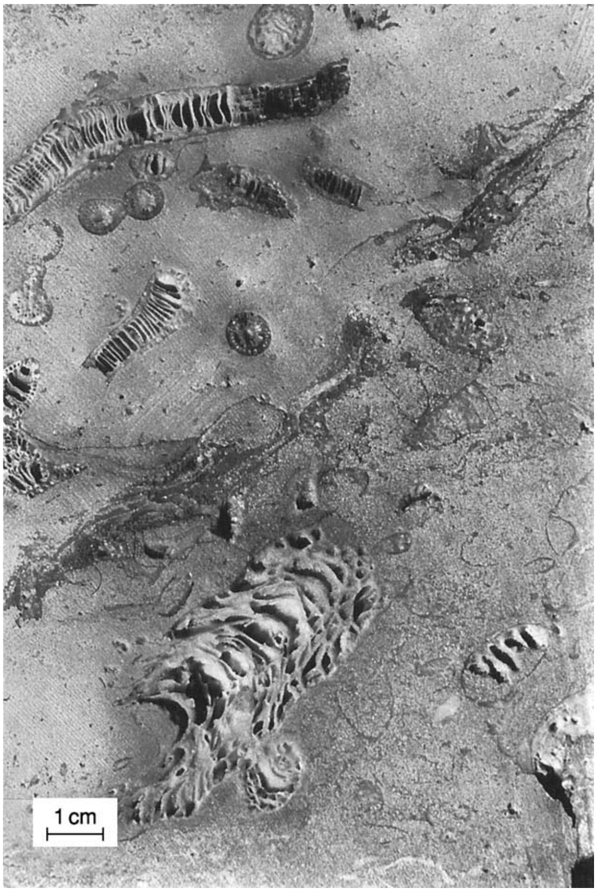


Figure 3 Uncemented *Smithiphyllum* and *Phacelophyllum* with calcite preservation of the delicate chamber walls (trabeculae) in partially dolomitized matrix. Sample is from the Devonian Nisku Formation, Alberta, Canada.

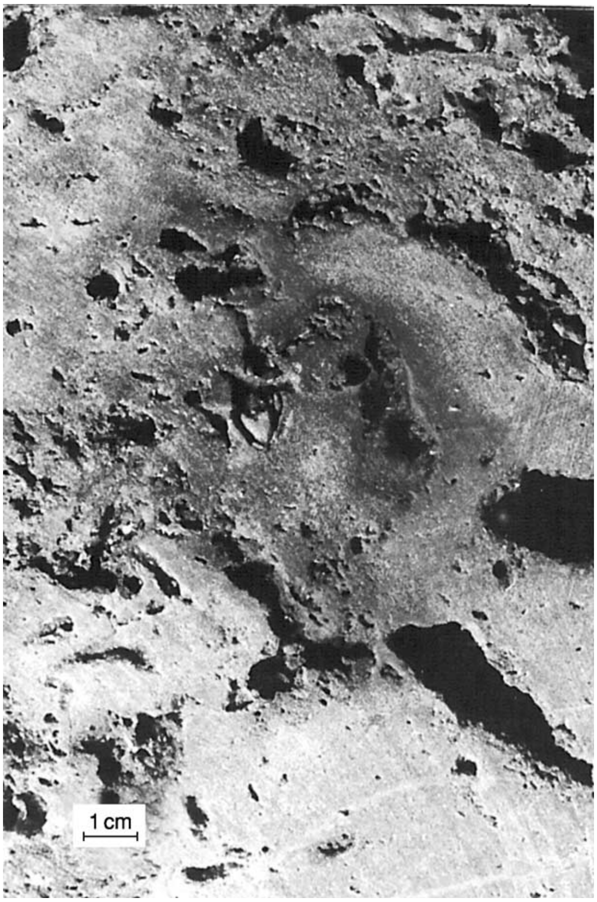


Figure 4 Vuggy dolostone that resulted from (macro-)dissolution of unreplaced calcite matrix and fossils, similar to the sample shown in Figure 3. Connection of pores is intercrystalline pervasive. Sample is from the Devonian Nisku Formation, Alberta, Canada.

by-product of thermochemical sulphate reduction, and from hydrothermal fluids. Saddle dolomite commonly occurs as gangue in MVT-type metal sulphide deposits.

Porosity and Permeability

Comparison of the molar volumes of calcite and dolomite reveals that about 13% of porosity is generated in the so-called 'mole-per-mole' replacement of calcite by dolomite according to reaction [1] (whereby two moles of calcite are replaced by one mole of dolomite). However, several other processes are involved. As a generalization, dolostones can have higher, the same, or lower porosity and permeability than their precursor limestones, and the poroperm evolution has to be investigated on a case-by-case basis. Many/most dolostones have higher porosities than limestones, and this fact may be the result of one or several of six processes (Figure 13): (1) mole-per-mole replacement; (2) dissolution of unreplaced calcite (solution undersaturated for calcite after all Mg in excess of dolomite saturation is exhausted); (3) dissolution of dolomite (without externally controlled acidification); (4) acidification of the pore waters (via decarboxylation, clay mineral diagenesis, etc.); (5) fluid mixing ('*Mischungskorrosion*'); and (6) thermochemical sulphate reduction, which may generate porosity under certain circumstances.

Dolomitization almost invariably involves the reorganization of permeability pathways. Commonly, permeability increases along with porosity, and vice versa, such as in the Upper Devonian Grosmont Formation in eastern Alberta, which hosts a giant heavy-oil reservoir, and in the Cambrian–Ordovician Bonneterre Formation of Missouri, USA, which hosts one of the world's largest MVT-type sulphide deposits. Planar-e dolomites tend to have the highest porosities and permeabilities, the latter caused by well-connected pore systems with low pore to throat size ratios (as indicated by mercury injection curves);

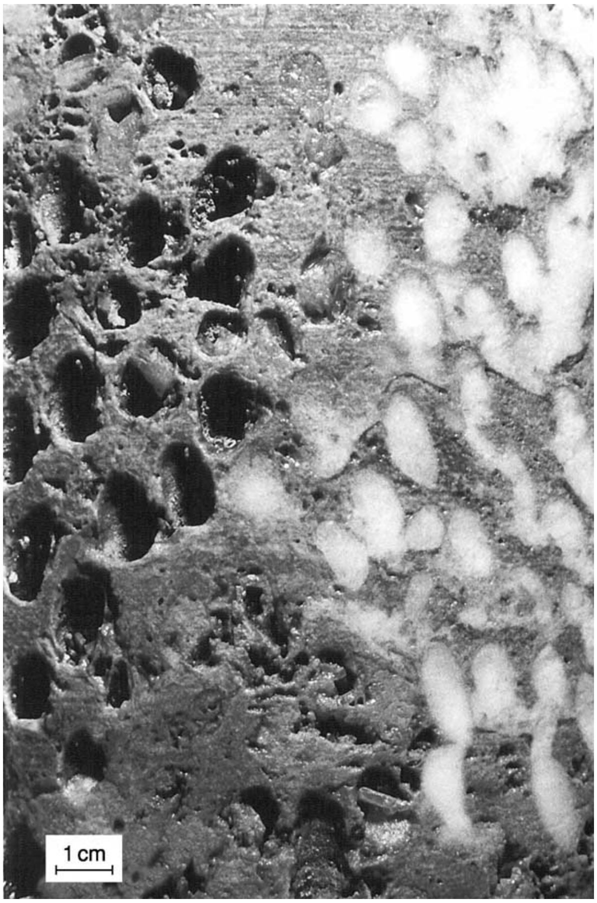


Figure 5 Coral mouldic porosity in tight matrix dolomite. White anhydrite occurs partially as a replacement (top right) and partially as a cement in coral mouldic porosity (centre and bottom right). From the Devonian Nisku Formation, Alberta, Canada.

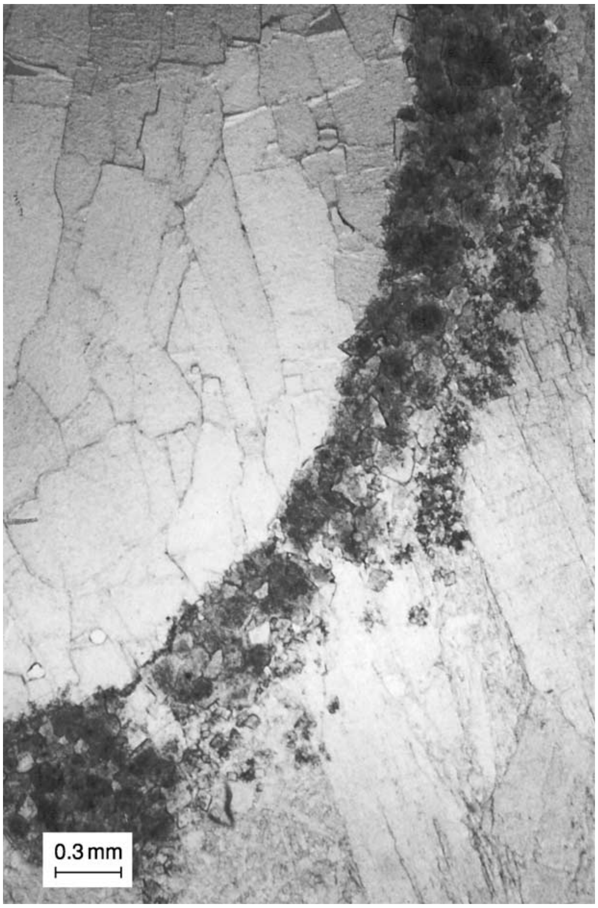


Figure 6 Thin section photomicrograph of the sample shown in Figure 5.

in planar-s dolomite, the permeabilities do not increase as rapidly with increasing porosity, corresponding to relatively large pore to throat size ratios; nonplanar dolomites have a statistically insignificant porosity–permeability relationship, whereby the pore systems have a high tortuosity and large pore to throat size ratios. Some authors have disputed that there is a systematic correlation between porosity and permeability in dolostones, or that these two petrophysical parameters are enhanced in dolostones relative to limestones. The Grosmont and the Bonneterre clearly show, however, that there is a relationship between porosity and permeability in at least some major and economically important dolostone sequences.

In cases of mole-per-mole replacement, the fabrics of the original limestone must be at least partially obliterated in order to account for the volume change during the replacement process. On the other hand, limestones dolomitized in a volume-per-volume replacement should not contain secondary intercrystal pores or dolomite cements, and the primary textures may be partially or largely, even mimetically (if the crystal size is very small), preserved. Partial or complete obliteration of primary textures can occur even in a volume-per-volume replacement, however, if there is a marked change in crystal size (usually an increase, due to Ostwaldt ripening) and/or porosity redistribution.

Dolomite Geochemistry

A wide range of geochemical methods may be used to characterize dolomites and dolostones, and to decipher their origin. One aspect of particular interest is the determination of the type of the dolomitizing fluid(s), i.e., marine, evaporitic, subsurface brine, etc., and the identification of the direction of fluid flow during dolomitization. The latter can often be ascertained by mapping a gradient in dolomite abundance, i.e., complete dolomitization near the upflow direction and decreasing abundance downflow. However, this approach necessarily fails where

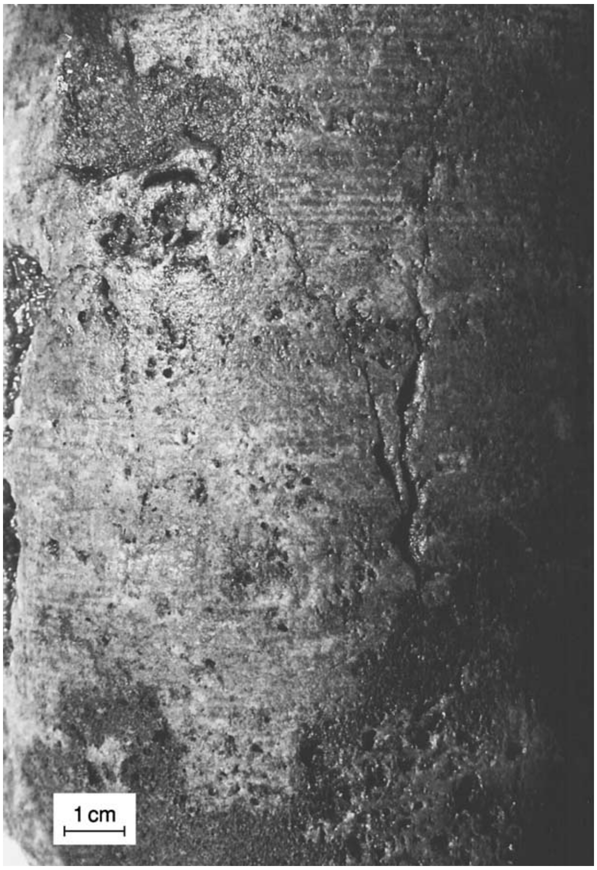


Figure 7 Dolostones consisting of domains of relatively tight, light to medium grey dolomite intergrown with domains of highly porous, brownish dolomite. From the Devonian Nisku Formation, Alberta, Canada.

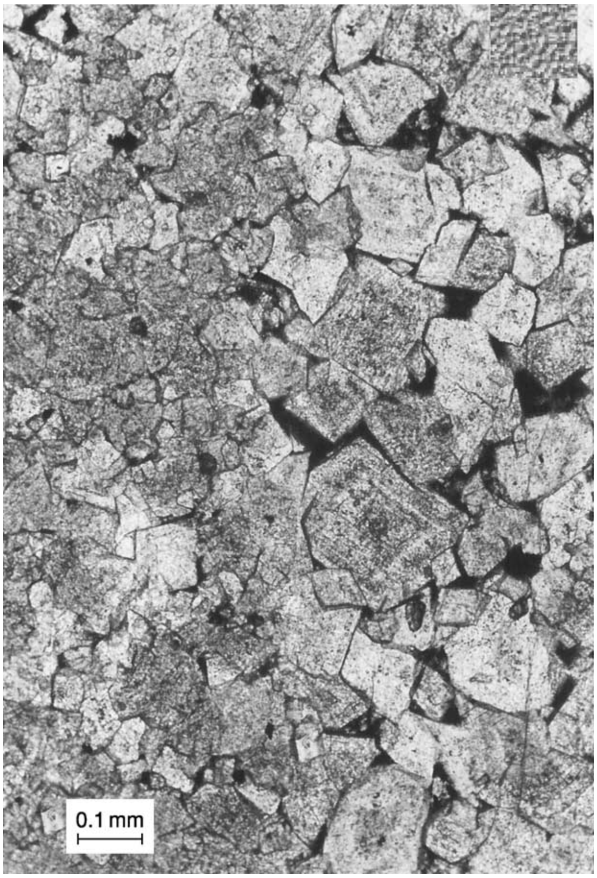


Figure 8 Thin section photomicrograph of the sample shown in Figure 7 from the boundary region between the two dolomite types.

dolomitization is 'complete' or where exposure and/or core are insufficient. In such cases, the geochemical composition of dolomites can be used, within limits, to determine the fluid composition and/or the flow direction during dolomitization.

Oxygen and carbon isotope ratios (δ^{18}O and δ^{13}C) are the most widely applied and probably best understood geochemical methods in dolomite research. δ^{18}O values of carbonates can be used, within limits, to determine the δ^{18}O value and/or temperature of the fluid present during crystallization, including a possible distinction between meteoric, marine, and/or evaporitic waters.

Fluid inclusion homogenization temperatures arguably are the best method to determine the temperature of formation of dolomites (and other minerals), in addition to the highly desirable information on fluid compositions that can be gained from freezing experiments. Unfortunately, the vast majority of fluid inclusions in dolomites are too small for standard heating–freezing runs, as phase transitions within the inclusions are not observable. This is especially true of matrix-selective, replacive dolomites. On the other hand, the sparry saddle dolomite cements found in late-diagenetic dissolution vugs, but also as a replacement, commonly yield excellent fluid inclusion data.

Where possible, fluid inclusion homogenization temperatures are used in conjunction with δ^{18}O values to further characterize the conditions of dolomite formation. This type of analysis can reveal the direction(s) and temperature gradient(s) of the dolomitizing fluid flow on a local scale (a few kilometres) or on a regional scale (over several hundred kilometres). Mapping and contouring of the oxygen isotope and/or fluid inclusion homogenization temperatures have shown clear, spatially resolved gradients in some locations. Unfortunately, such gradients do not appear to be common.

The δ^{13}C values of the carbonates can be used to identify whether meteoric water (carrying soil CO_2) was involved, whether thermogenic or biogenic CH_4 was oxidized, whether CO_2 from microbial processes or organic matter maturation was available,

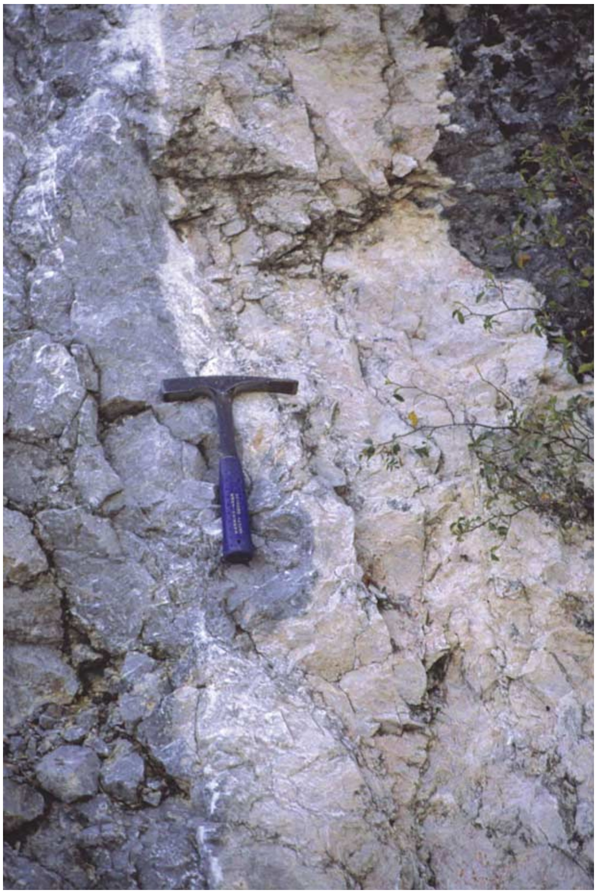


Figure 9 Outcrop photograph of Upper Carboniferous carbonates from the southwestern Cantabrian Zone, Spain: high-temperature dolomitization of limestones. The dolostone appears dark where covered with lichen (upper right corner), yet light beige where cleaned of lichen (centre). The limestone (left) has a medium grey colour. Note the sharp contacts between limestones and dolostones, and that sedimentary and diagenetic textures visible in the limestones are obliterated in the dolostones. Hammer for scale.

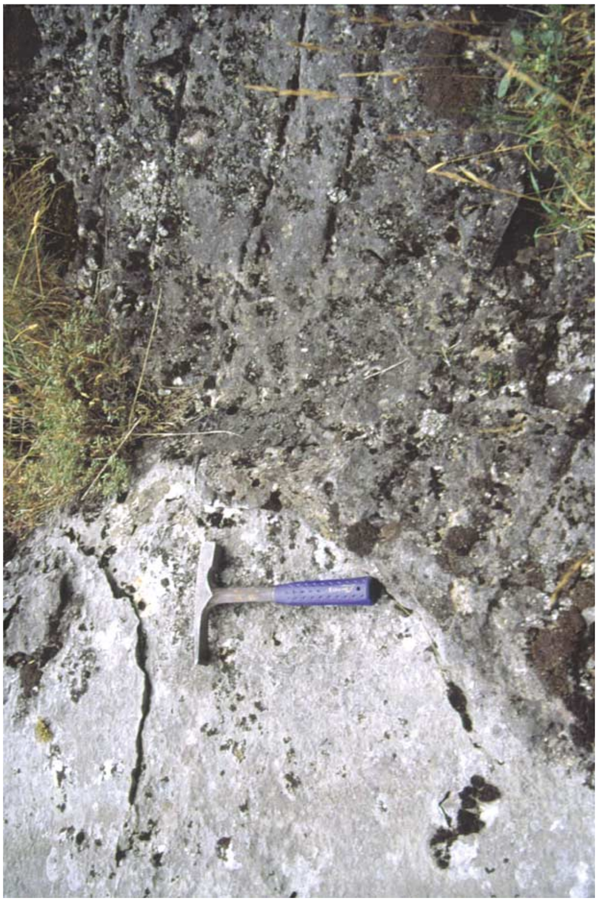


Figure 10 Outcrop photograph of Upper Carboniferous carbonates from the southwestern Cantabrian Zone, Spain: high-temperature dolomitization of limestones. The dolostone is light coloured and at the bottom. Note the sharp contacts between limestones and dolostones, and that sedimentary and diagenetic textures visible in the limestones are obliterated in the dolostones. Hammer for scale.

or whether thermochemical sulphate reduction contributed carbon to the carbonates. Also, there is a secular carbon isotope trend that may be used in the dating of marine dolostones, but only under very favourable circumstances.

Radiogenic isotopes are less commonly used in studies of carbonate diagenesis, mainly because they are analytically much more expensive. Yet, strontium isotopic compositions (usually quoted as $^{87}Sr/^{86}Sr$ ratios) are an excellent parameter to deduce compositional changes and, especially, flow directions of the fluids from which the diagenetic carbonates have formed. This is because strontium isotopes, as opposed to the more commonly used stable isotopes of oxygen and carbon, are not fractionated by pressure, temperature, and (as in the case of carbon) microbial processes.

The direction of fluid flow can also be determined using trace elements, which is especially attractive because trace element analysis is the cheapest of all the common geochemical methods. Trace element trends have been documented in several Phanerozoic dolostone sequences.

For all practical applications, i.e., the determination of fluid composition and/or fluid flow direction, the absence, presence, and/or degree of recrystallization is important. If changes via recrystallization in texture, structure, composition, and/or palaeomagnetic properties are so small that the total data range after recrystallization is the same as when the dolomite first formed, a dolomite/dolostone is said to be ‘insignificantly recrystallized’ (Figure 14, top), and its properties are still representative of the fluid and environment of dolomitization. On the other hand, if these changes result in data ranges that are larger than

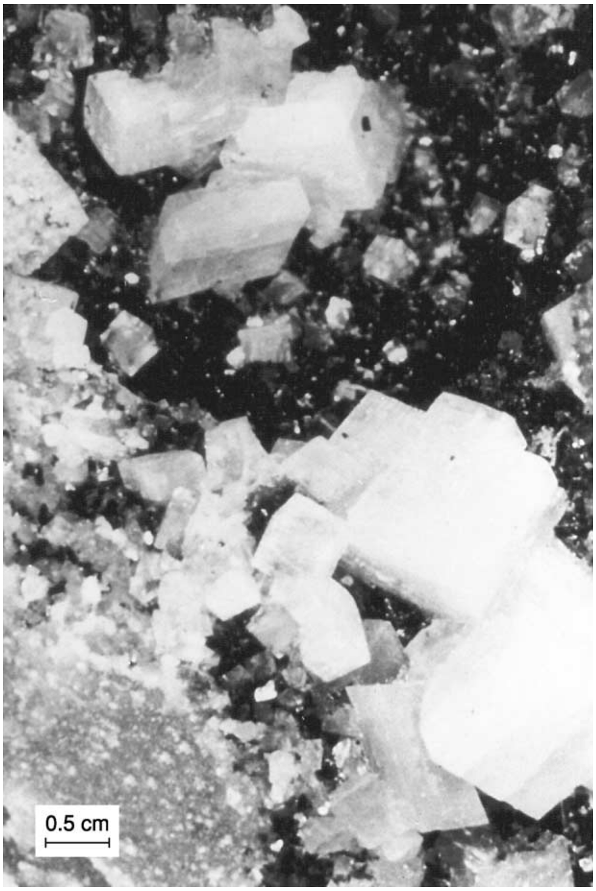


Figure 11 Core specimen of milky-white saddle dolomite cement in a vug that is coated with solid bitumen (‘dead oil’). Host rock is grey matrix dolomite. Saddle dolomite appears as a large crystal in the centre and lower right, with undulous extinction. From the Devonian Nisku Formation, Alberta, Canada.

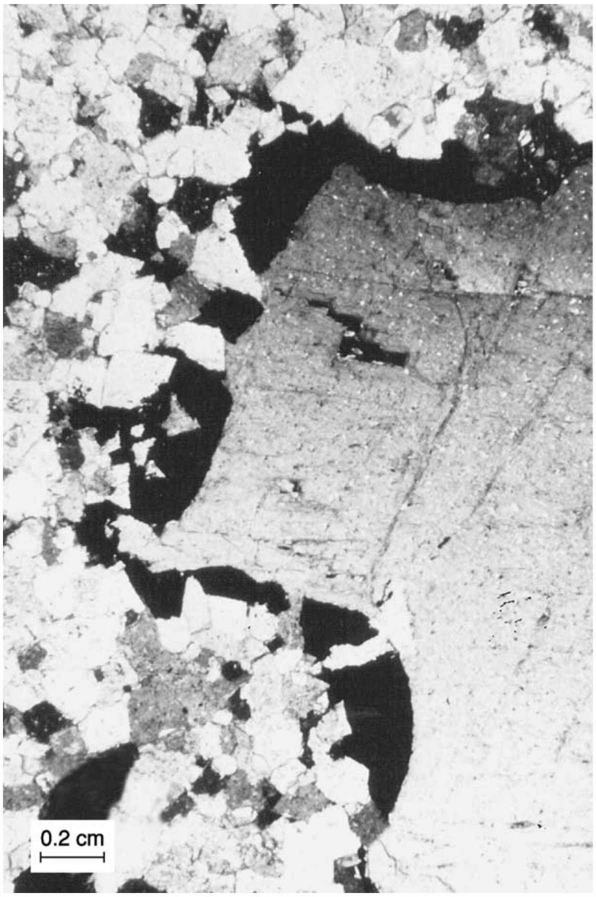


Figure 12 Thin section photomicrograph (transmitted light with crossed polarizers) of the sample shown in Figure 11.

the original ones, a dolomite/dolostone is said to be ‘significantly recrystallized’ (Figure 14, bottom), and its properties are no longer representative of the fluid and environment of dolomitization. In this case, the measured properties are reset and they characterize the last event of recrystallization. Furthermore, not all measurable properties must be reset during recrystallization. For a dolomite to be recognized as significantly recrystallized, only one of the measurable properties has to be modified to a range larger than the original one. In this case, inherited properties may still represent the event of dolomitization, whereas reset properties represent recrystallization.

Most dolomites that originally form very close to the surface and/or from evaporitic brines tend to recrystallize with time and during burial, because they form as metastable protodolomite phases that become thermodynamically highly unstable as a result of increasing temperature, increasing pressure, and changing fluid composition. On the other hand, dolomites that form at several hundred to a few thousand metres of depth are not or hardly prone to recrystallization, because these dolomites tend to form as rather stable (nearly stoichiometric, well-ordered) phases, whose stability does not change much during further burial and with increasing time.

Another important aspect of dolomite research requiring the application of geochemical methods is the recognition of hydrothermal activity. In many studies, the presence of saddle dolomite has been taken as an indication of elevated heat flow and/or increased temperatures during dolomite formation. However, the presence of saddle dolomite merely indicates a temperature of formation that is relatively high in the context of diagenetic studies. Its presence in uplifted dolomites, or in structurally inverted subsurface systems currently at lower temperatures, may merely reflect processes formerly operating at depths (and temperatures) at or around maximum burial and with normal geothermal gradients. Saddle dolomite may be hydrothermal, geothermal, or hydrofrigid (Figure 15). A distinction between these

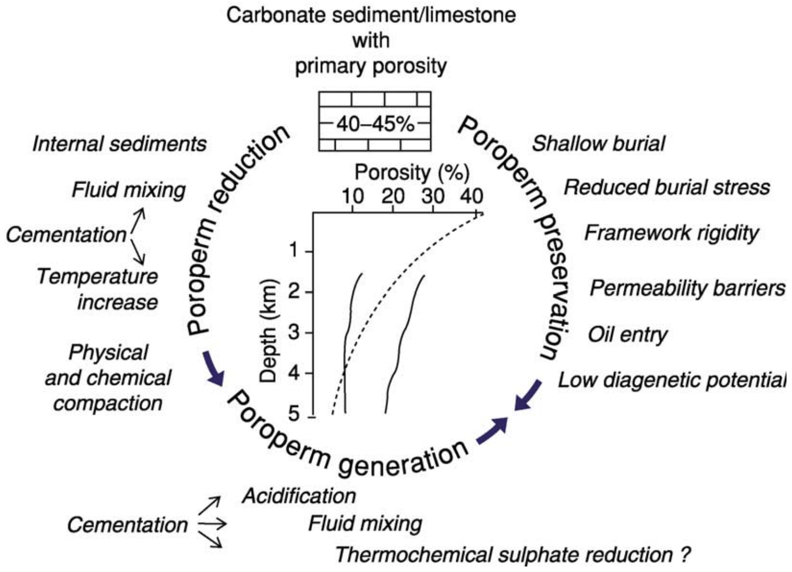


Figure 13 Major processes of porosity and permeability ('poroperm') generation, preservation, and reduction in carbonates. The inset contains averaged porosity/depth data from Mesozoic and Cenozoic limestones and dolostones in south Florida (broken curve; from Schmoker and Halley (1992)) and from the Jurassic Smackover oolite carbonate reservoirs in the southern USA (full curves, which envelop the measured maximum and minimum values below depths of about 1.5 km; from Scholle and Halley (1985) and Heydari (1997)). The Florida trend can be considered typical for most carbonates elsewhere. The large variations in Smackover carbonates at any given depth reflect highly variable degrees of porosity generation, preservation, and reduction due to various competing diagenetic processes. Reproduced with permission from Machel HG (1999) Effects of groundwater flow on mineral diagenesis, with emphasis on carbonate aquifers. *Hydrogeology Journal* 7: 94–107, ©1999 Springer-Verlag.

alternatives can only be made if the temperature of formation of saddle dolomite is considered relative to the temperature of the surrounding rocks at the time of saddle dolomite formation, e.g., via fluid inclusion data in silicates or other carbonates, vitrinite reflectance data, reconstruction of maximum burial and geothermal gradient, etc.

Environments and Models of Dolomitization

In near-surface and shallow diagenetic settings, dolomitization models are defined and/or based on water chemistry, but on hydrology in deeper burial diagenetic settings. This poses an obvious dilemma when some type of near-surface diagenetic fluid moves into the deeper subsurface, and when deeper subsurface fluids (commonly brines) ascend into shallow diagenetic settings. Research over the last 15–20 years has revealed several such 'cross-overs' or 'overlaps' between models, which has resulted in unnecessary ambiguities in semantics and classification.

Penecontemporaneous Dolomites and the Microbial/Organogenic Model

In shallow marine to supratidal environments, penecontemporaneous dolomites commonly form in quantities of <5 vol.%, mostly as Ca-rich and poorly ordered, microcrystalline to fine crystalline cements and/or directly from aqueous solution. These occurrences include: lithified supratidal crusts (e.g., Andros Island, Sugarloaf Key, Ambergris Cay); thin layers in salinas (e.g., Bonaire, West Caicos Island) and evaporative lagoons/lakes (e.g., Coorong); and fine crystalline cements and replacements in peritidal sediments (e.g., Florida Bay, Andros Island). The dolomite-forming fluids are normal seawater and/or evaporated seawater, in some cases with admixtures of evaporated groundwater. There are also cases of penecontemporaneous dolomite formation in association with volcanics or volcanic activity, dolomite as fine crystalline supratidal weathering products of basic rocks, and hydrothermal dolomite forming at submarine vents. One especially important type in this group, commonly classified under hypersaline dolomites, forms lenses and layers of up to 100 vol.% dolomite in sabkhas.

Penecontemporaneous dolomites in hemipelagic to pelagic settings commonly form in very small quantities as microcrystalline protodolomites, generally with less than 1 wt.%. However, under favourable circumstances, the amount of dolomite can reach up to 100 vol.% locally. For example, Miocene hemipelagic carbonate sediments from the margin of the

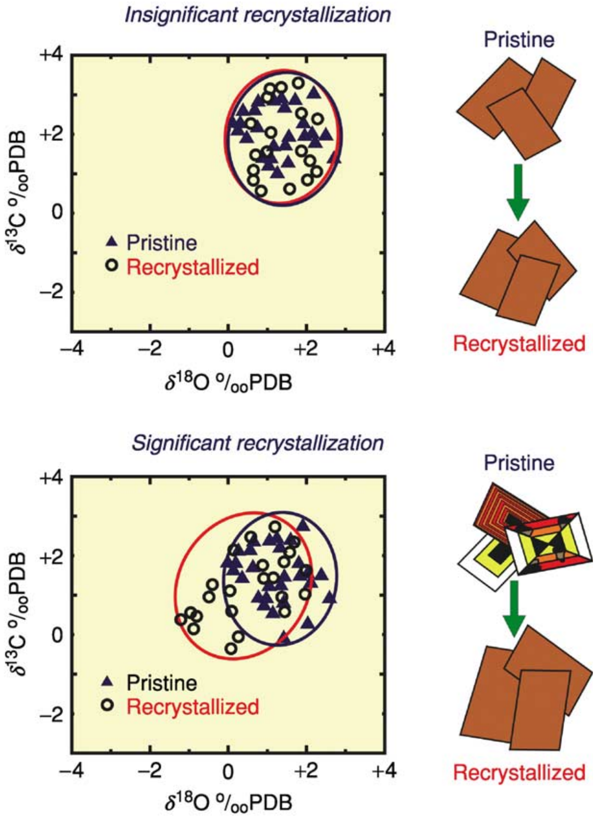


Figure 14 Schematic illustration of insignificant and significant recrystallization. For the properties shown, i.e., δ^{13}C and δ^{18}O values, crystal sizes, and luminescence, the pristine and recrystallized samples are the same in the case of insignificant recrystallization, but different, despite some overlap, in the case of significant recrystallization, i.e., at least some isotope values of the recrystallized samples fall outside the range of the pristine samples. The crystals also have increased in size and lost their respective zonations (concentric, sector, oscillatory). Reproduced with permission from Machel HG (1997) Recrystallization versus neomorphism, and the concept of 'significant recrystallization' in dolomite research. *Sedimentary Geology* 113: 161–168.

Great Bahama Bank are partially to completely dolomitized over a depth range of about 50–500 m subsea. In this setting, dolomite forms as a primary void-filling cement and by replacing micritic sediments, red algae, and echinoderm grains.

Both settings of penecontemporaneous dolomite formation mentioned above appear to be linked to the 'microbial model' or 'organogenic model' of dolomitization. According to this model, dolomite may be formed syndepositionally or early postdepositionally, i.e., at depths of a few centimetres to a few hundred metres, under the influence of, or promoted by, bacterial sulphate reduction and/or methanogenesis. The latter is indicated by organogenic δ^{13}C values. The exact role of microbial activity in reducing the notorious kinetic barriers to dolomitization is unknown, although it seems likely that the reduction of Mg and Ca hydration barriers, an increase in alkalinity, and/or changes in pH are involved. Most microbial/organogenic dolomites are cements; some are replacive, typically fine crystalline to microcrystalline (less than 10 μm), calcic and poorly ordered protodolomites. The chief modes of Mg supply are diffusion from the overlying seawater and/or release from Mg calcites and clay minerals, which obviously places severe limits on the amounts of dolomite that can be formed. Microbial/organogenic dolomites may act as nuclei for later, more pervasive dolomitization during burial.

Hyposaline Environments and the Mixing Zone Model

Hyposaline environments are those with salinities below that of normal seawater (35 g l^{-1}). These environments include coastal and inland freshwater/seawater mixing zones, marshes, rivers, lakes, and caves. Virtually all hyposaline environments are near-surface to shallow burial diagenetic settings at depths of less than about 600–1000 m.

One hyposaline environment, the coastal freshwater/seawater mixing zone (often simply called mixing zone), has given rise to one of the oldest and most popular models, i.e., the 'mixing zone model' (also called the Dorag model) for dolomitization. However, the mixing model has been overrated with regard to its potential to form massive dolostones. Not a single location in the world has been shown to be extensively dolomitized in a freshwater/seawater mixing zone, in recent or in ancient carbonates, and several lines of evidence indicate that massive dolomitization in mixing zones is so unlikely as to be virtually impossible. Rather, mixing zones tend to form only very small amounts of dolomite, commonly along with substantial dissolution porosity, up to the scale of caves. A striking example is the vast cave system with essentially no dolomite generated by mixing zone diagenesis along the coastline of the Yucatan Peninsula, Mexico. The main role of coastal mixing zones in dolomitization may be as a hydrological pump for seawater dolomitization, rather than a geochemical environment favourable for dolomitization.

Most true mixing zone dolomites are petrologically and geochemically distinct. The crystals tend to be relatively clear, planar-e or planar-s, stoichiometric, well-ordered rhombs. However, some mixing zone dolomite is protodolomite. Crystal sizes commonly range from 1 to 100 μm, but may reach several millimetres in some cases. Most mixing zone dolomite occurs as cements in microscopic interstices and macroscopic voids, moulds, vugs, and caverns, and

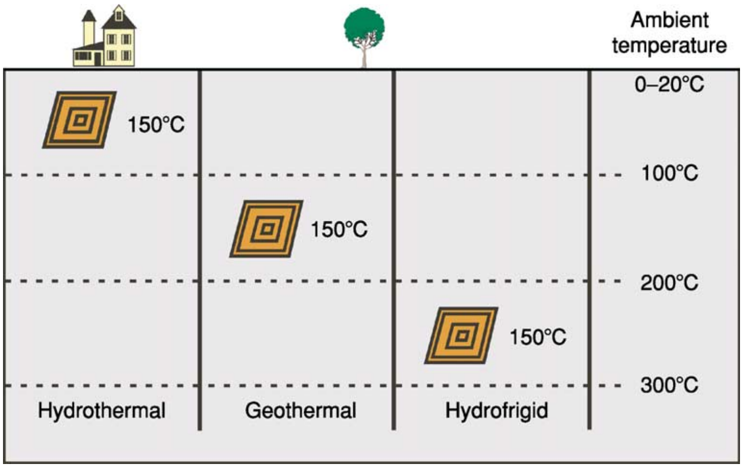


Figure 15 Hydrothermal, geothermal (formed in thermal equilibrium with the surrounding rocks), and hydrofrigid mineral formation. Reproduced with permission from Machel HG and Lonnee J (2002) Hydrothermal dolomite – a product of poor definition and imagination. *Sedimentary Geology* 152: 163–171.

subordinately as a replacement. Alternating generations or growth zones of calcite/dolomite are common in coastal mixing zones with rapid and cyclical changes of salinity.

Hypersaline Environments and the Reflux and Sabkha Models

Hypersaline environments have salinities greater than that of normal seawater and are widespread at latitudes of less than about 30°, while some occur at even higher latitudes. Hypersaline environments thus defined include the so-called mesohaline (also called penehaline) environments, which are mildly hypersaline, i.e., between normal seawater salinity ($35\,\mathrm{g\,l^{-1}}$) and that of gypsum saturation (about $120\,\mathrm{g\,l^{-1}}$). In all these types of environment, dolomite is formed from water whose salinity is controlled by surface evaporation, that is, in near-surface and shallow burial diagenetic settings.

The (evaporative) reflux model of dolomitization was originally proposed for seawater evaporated beyond gypsum saturation in lagoonal and shallow marine settings on a carbonate platform behind a barrier, such as a reef. Surficial water circulation on the platform is severely restricted because of the barrier, which leads to evaporation and a landward salinity gradient behind the barrier. The evaporated seawater flows downwards into and seaward through the platform sediments because of its increased density (active reflux), thereby dolomitizing the penetrated sediments. Platforms can be penetrated by mesohaline reflux to depths of several hundred metres. Furthermore, numerical modelling predicts that 'latent reflux' flows the a platform after the evaporative generation of brine ceases at the platform top, which can be expected after a significant rise in sea-level with concomitant freshening of the waters on the platform. However, a platform can be dolomitized completely only if it has very high permeability and does not contain aquitards (such as shale or evaporite layers), and if reflux is permitted to persist for a very long time. The published literature provides several examples of localities that probably were dolomitized by evaporative reflux, including the Permian carbonates of west Texas and New Mexico, and the peritidal Jurassic carbonates of Gibraltar. Whether active or latent, all refluxing brines exit or even landward of the platform margin, which confines reflux dolomitization to the platform interior.

The sabkha model of dolomitization is hydrologically and hydrochemically related to the reflux model. Sabkhas are intertidal to supratidal deflation surfaces that are flooded episodically. In the sabkha of the Trucial Coast of Abu Dhabi, the type location of the sabkha model for dolomitization, Mg for dolomitization is supplied synsedimentarily/penecontemporaneously by seawater that is propelled periodically onto the lower supratidal zone and along remnant tidal channels by strong onshore winds. The seawater has normal to slightly elevated salinity (up to about $38\,\mathrm{g\,l^{-1}}$), but becomes significantly evaporated beyond gypsum saturation on/within the supratidal flats, through which it refluxes via its increased density, similar to flow in the reflux model. Sabkha dolomite appears to form via evaporative pumping in a narrow (1–1.5 km) fringe next to the strandline, and in the flooded tidal channels that extend more landward. In the Abu Dhabi sabkha, the best dolomitized parts contain 5 to about 65 wt.% protodolomite. Dolomite forms as a cement and aragonite is replaced, but lithification does not occur, or only partially. Furthermore, dolomitization is restricted to the upper 1–2 m of the sediments, and appears to be most intense where the pore waters become chemically reducing, which leads to enhanced carbonate alkalinity via sulphate reduction and/or microbial methanogenesis. Therefore, sabkha dolomites are texturally and geochemically similar to organogenic dolomites in some respects, i.e., they tend to form as protodolomite and may have reduced carbon isotope ratios. In most respects, the Abu Dhabi sabkha appears to be a good recent analogue for dolomitization in many ancient intertidal to supratidal flats, such as landward of the famous Permian carbonates of Texas and New Mexico. Rather than forming reservoir rocks, these dolostones – including the associated evaporites – generally form tight seals for underlying hydrocarbon reservoirs. More generally, sabkhas and similar intertidal to supratidal environments in more humid climates typically form small quantities of fine crystalline protodolomite in thin beds, crusts, or nodules, either within the upper 1–2 m of sediment, or at the sediment surface. Repeated transgressions and regressions may stack such sequences upon one another to cumulative thicknesses of several tens of metres.

Seawater Dolomitization

Most postdepositional formation of massive dolostones probably results from 'seawater dolomitization'. There are a group of models of seawater dolomitization, whose common denominator is seawater as the principal dolomitizing fluid, and which differ in hydrology and/or depth and timing of dolomitization. All dolomites belonging to this group are postdepositional.

The Cenozoic dolostones of the Bahama platform, often used as an analogue for older dolomitized carbonate platforms elsewhere, can be considered the type location for seawater dolomitization. Extensive petrographical and geochemical data indicate that seawater and/or chemically slightly modified seawater was the principal agent of dolomitization in the Bahama platform at shallow to intermediate depths and commensurate temperatures of less than about 60°C.

The hydrology of and during seawater dolomitization is still very much contested. Various hydrological systems have been invoked to drive the large amounts of seawater needed for pervasive dolomitization through the Bahama platform, i.e., thermal convection, a combination of thermal seawater convection and reflux of slightly evaporated seawater derived from above, or seawater driven by an overlying freshwater/seawater mixing zone during partial platform exposure, possibly layer by layer in several episodes.

The Bahamas' dolostones actually represent a hybrid with regard to the traditional, conventional classifications of models. Petrographical and geochemical data indicate that seawater was the principal dolomitizing agent, yet thermal convection, as a hydrological system and drive for dolomitization, is better classified under the burial (subsurface) models discussed below. Analogously, the regionally extensive Devonian dolostones in Alberta, western Canada, are also a hybrid with regard to the conventional dolomitization models. These Devonian dolostones probably formed at depths of 300–1500 m at temperatures of about 50–80°C from chemically slightly modified seawater, and have been classified as burial dolostones. Another example is represented by the regionally extensive dolostones of the Carboniferous of Ireland, which are petrographically and geochemically very similar to the Devonian dolostones of Alberta, and whose genesis has been interpreted in an analogous manner. In all cases, the hydrology that facilitated dolomitization is unclear, with thermal convection, reflux, compaction, tectonic expulsion, or a combination thereof, as alternatives. The regionally extensive dolostones of the Cretaceous Soreq Formation in Israel represent a Mesozoic example of this type of dolomitization. These Palaeozoic and Mesozoic dolostones can be (re-)classified along with the Cenozoic Bahama dolostones as 'seawater dolomites'. This classification dilemma arises from the historical evolution of our understanding of these dolostones, rather than invalidating the earlier 'burial' interpretations.

Intermediate to Deep Burial (Subsurface) Environments and Models

Burial (subsurface) environments are those removed from active sedimentation by burial, and in which the pore fluid chemistry is no longer entirely governed by surface processes, i.e., where water–rock interaction has modified the original pore waters to a significant degree, or where the fluid chemistry is dominated by subsurface diagenetic processes. The textures, porosities, and permeabilities of dolostones formed in intermediate and deep burial settings are variable. These textures in themselves are not indicative of the depth of burial. However, three specific textural characteristics may be used to indicate considerable burial: dolomites cross-cut by stylolites suggest burial of at least 600 m (stylolites in dolostones appear to require at least 600 m of burial); the absence of planar crystals suggests temperatures of formation or recrystallization in excess of the critical roughening temperature of about 50°C; and the presence of saddle dolomite suggests temperatures of formation in excess of about 80°C.

All burial (subsurface) models for dolomitization essentially are hydrological models. They differ mainly in the hydrological drives and direction(s) of fluid flow. Four main types of fluid flow take place in subsurface diagenetic settings: (1) compaction flow; (2) thermal convection; (3) topography driven flow; and (4) tectonically driven flow. Combinations of these flow regimes and associated fluid compositions are possible under certain circumstances.

The oldest burial model of dolomitization is the compaction model, according to which seawater and/or its subsurface derivative(s), that were buried along with the sediments, are pumped through the rocks at several tens to several hundreds of metres as a result of compaction dewatering.

The compaction model in its original form was never especially popular because burial compaction can only generate fairly limited amounts of dolostone due to the limited amounts of expelled water. However, despite this mass balance constraint, the compaction model remains a viable alternative for burial/subsurface dolomitization where funnelling of the compaction waters is/was possible.

Thermal convection is driven by the temperature gradient prevailing across sedimentary strata, which is vertical in most geological situations, except in cases of vigorous advection, igneous intrusions, or in the proximity to plate boundaries and/or orogenic fronts, all of which can 'distort' the normal subvertical temperature gradient. Where the temperature gradient and average rock permeability are high enough, convection cells may become established. In principle, there are two types of convection, i.e., open and closed, although mixed cases are possible. Open convection cells (also called half-cells) may form in carbonate platforms that are open to seawater recharge and discharge laterally and at the top, respectively. Numerical modelling has shown that the magnitude and distribution of permeability are the most important parameters governing flow and dolomitization, and that this type of convection can be active to a depth of about 2–3 km, provided that the sequence does not contain effective aquitards, such as (overpressured) shales or evaporites. The amounts of dolomite that can be formed are theoretically very large, i.e., dolomite can be formed as long as convection is sustained, because Mg is constantly (re-)supplied from the surrounding seawater. However, even at a moderate width of only 40 km, complete dolomitization in a 2 km thick sequence takes about 30–60 million years, which is much longer than the time during which most carbonate platforms remain laterally open to seawater recharge. Hence, most carbonate platforms, even if subjected to thermal convection by seawater, would at best become only partially dolomitized during the time that they were open to seawater recharge.

Closed convection can occur, in principle, in any sedimentary basin over tens to hundreds of metres of thickness, provided that the temperature gradient is high enough relative to the permeability of the strata. As a rule of thumb, however, such convection cells will only be established, and capable of dolomitizing a carbonate sequence of interest, if a sequence is highly permeable and not interbedded with aquitards. Such conditions are rarely fulfilled in typical sedimentary basins, most of which do contain effective aquitards. Furthermore, even if closed thermal convection cells are established, the amounts of dolomite that can be formed are limited by the pre-convection Mg content of the fluids, even more so than in the case of compaction flow, as no new Mg is supplied to the system and 'compaction funnelling' is not possible. Therefore, extensive, pervasive dolomitization by closed cell thermal convection is highly unlikely.

Convection cells invariably have rising limbs that penetrate the overlying and cooler strata, linking thermal convection to hydrothermal dolomitization. There are well-documented examples of hydrothermal dolomite on a local and regional scale. Most cases of hydrothermal dolomitization are rather small and restricted to the vicinity of faults and fractures and/or localized heat sources. One striking case of this type is the Pb–Zn mineralized Navan dolomite plume in Ireland, and another is a dolomitized plume in the Latemar in the Italian Alps. There are also some cases of larger scale, even regionally extensive hydrothermal dolomitization, such as the Middle Devonian Presqu'ile barrier, which forms an aquifer in northwestern Canada that contains abundant saddle dolomite as a replacement and as a cement, including MVT-type mineralization near the discharge area at Pine Point. Texturally, most true hydrothermal dolomite is saddle dolomite.

Topography driven flow takes place in all uplifted sedimentary basins that are exposed to meteoric recharge. With time, topography can drive enormous quantities of meteoric water through a basin, often concentrated by water–rock interaction (especially salt dissolution), and preferentially funnelled through aquifers. Volumetrically significant dolomitization can only take place, however, where the meteoric water dissolves enough Mg *en route* before encountering limestones. This does not appear to be common.

Another type of flow that has been suggested to result in pervasive dolomitization is tectonically driven squeegee-type flow. In this type of flow system, metamorphic fluids are expelled from crustal sections affected by tectonic loading so that basinal fluids are driven towards the basin margin. Metamorphic fluids could be injected into compaction and/or topography driven flow, with attendant fluid mixing. However, it is doubtful whether tectonically induced flow, or the related fluid mixing, can form massive dolostones. Diagenetic studies on this type of flow system suggest that the fluxes are low and short lived.

Secular Distribution of Dolostones

The relative abundance of dolostones that originated from the replacement of marine limestones appears to have varied cyclically through geological time, commonly referred to as secular variation. Early data suggested that dolomite was most abundant in rocks of the Early Palaeozoic systems and decreased in abundance with time. Relatively recent reassessments of the dolomite distribution throughout time revealed two discrete maxima of 'significant early' dolomite formation (massive early diagenetic replacement of marine limestones) during the Phanerozoic, i.e., the Early Ordovician/Middle Silurian and the Early Cretaceous.

Various explanations have been proposed for this phenomenon, i.e., that periods of enhanced early dolomite formation were related to or controlled by plate tectonics and related changes in the compositions of the atmosphere and seawater, such as an increased atmospheric CO_2 level, high eustatic sea-level, low saturation state of seawater with respect to calcite, changes in the marine Mg/Ca ratio, or low atmospheric O_2 levels that coincided with enhanced rates of bacterial sulphate reduction. It appears possible that a combination of two or more of these factors was involved. Perhaps the most elegant explanation is that the secular dolomite variations are the result of the lengthy induction period for dolomite formation that was observed in laboratory experiments. Marine carbonates in prolonged contact with seawater may be dolomitized because they remained in contact with the dolomitizing solution (seawater) long enough to exceed the induction period. On the other hand, undolomitized carbonates were not in contact with seawater for long enough, and metastable precursors to dolomite that may have formed were destroyed by freshwater diagenesis during intervening periods of exposure. The secular variations in marine dolomitization may thus reflect periods of seawater contact longer or shorter than the induction period.

See Also

Analytical Methods: Geochemical Analysis (Including X-Ray). **Diagenesis, Overview**. **Minerals:** Carbonates. **Petroleum Geology:** Production. **Sedimentary Environments:** Carbonate Shorelines and Shelves. **Sedimentary Rocks:** Mineralogy and Classification; Evaporites; Limestones.

Further Reading

Allen JR and Wiggins WD (1993) *Dolomite Reservoirs – Geochemical Techniques for Evaluating Origin and Distribution. AAPG Continuing Education Course Note Series, 36*. Town: American Association of Petroleum Geologists.

Braithwaite C, Rizzi G, and Darke G (eds.) (2004) *The Geometry and Petrogenesis of Dolomite Hydrocarbon Reservoirs. Special Publication of the Geological Society* (in press).

Budd DA (1997) Cenozoic dolomites of carbonate islands: their attributes and origin. *Earth Science Reviews* 42: 1–47.

de Dolomieu D (1791) Sur un genre de Pierres calcaires très-peu effervescentes avec les Acides, & phosphorescentes par la collision. *Journal de Physique* 39: 3–10. Translation with notes of Dolomieu's paper reporting his discovery of dolomite by: Carozzi AV and Zenger DH (1981) *Journal of Geological Education* 29: 4–10.

Gregg JM and Sibley DF (1984) Epigenetic dolomitization and the origin of xenotopic dolomite texture. *Journal of Sedimentary Petrology* 54: 908–931.

Hardie LA (1987) Dolomitization: a critical view of some current views. *Journal of Sedimentary Petrology* 57: 166–183.

Land LS (1985) The origin of massive dolomite. *Journal of Geological Education* 33: 112–125.

Luo P and Machel HG (1995) Pore size and pore-throat types in a heterogeneous Dolostone reservoir, Devonian Grosmont Formation, Western Canada Sedimentary Basin. *American Association of Petroleum Geologists Bulletin* 79: 1698–1720.

Machel HG (1997) Recrystallization versus neomorphism, and the concept of 'significant recrystallization' in dolomite research. *Sedimentary Geology* 113: 161–168.

Machel HG (2004) Concepts and models of dolomitization – a critical reappraisal. In: Braithwaite C, Rizzi G, and Darke G (eds.) *The Geometry and Petrogenesis of Dolomite Hydrocarbon Reservoirs. Special Publication of the Geological Society* (in press).

Machel HG and Mountjoy EW (1986) Chemistry and environments of dolomitization – a reappraisal. *Earth Science Reviews* 23: 175–222.

Mazzullo SJ (1992) Geochemical and neomorphic alteration of dolomite: a review. *Carbonates and Evaporites* 6: 21–37.

Morrow DW (1982a) Diagenesis 1. Dolomite – Part 1: The chemistry of dolomitization and dolomite precipitation. *Geoscience Canada* 9: 5–13.

Morrow DW (1982b) Diagenesis 2. Dolomite – Part 2: Dolomitization models and ancient dolostones. *Geoscience Canada* 9: 95–107.

Morrow DW (1999) Regional subsurface dolomitization: models and constraints. *Geoscience Canada* 25: 57–70.

Nordeng SH and Sibley DF (1994) Dolomite stoichiometry and Ostwald's step rule. *Geochimica et Cosmochimica Acta* 58: 191–196.

Pray LC and Murray RC (eds.) (1965) *Dolomitization and Limestone Diagenesis – A Symposium. Society of Economic Paleontologists and Mineralogists Special Publication 13.* Town: Society of Economic Paleontologists and Mineralogists.

Purser B, Tucker M, and Zenger D (eds.) (1994) *Dolomites – A Volume in Honour of Dolomieu: Special Publication 21.* Town: International Association of Sedimentologists.

Sibley DF and Gregg JM (1987) Classification of dolomite rock textures. *Journal of Sedimentary Petrology* 57: 967–975.

Van Tuyl FM (1914) *The Origin of Dolomite. Annual Report 1914,* vol. XXV, pp. 257–421. Town: Iowa Geological Survey.

Wright WR (2001) Dolomitization, fluid-flow and minerlization of the Lower Carboniferous rocks of the Irish Midlands and Dublin Basin. Unpub. Ph.D. thesis, University College Dublin, Belfield, Ireland, 407 p.

Zenger DH, Dunham JB, and Ethington RL (eds.) (1980) *Concepts and Models of Dolomitization. Society of Economic Paleontologists and Mineralogists Special Publication 28.* Town: Society of Economic Paleontologists and Mineralogists.

Evaporites

A C Kendall, University of East Anglia, Norwich, UK

Deposits Produced by the Evaporation of Seawater

Seawater is considered to be the major or the only feedstock capable of generating large bodies of evaporite. All deposits of potash salts are associated with large basin-central evaporites and, consequently, are believed by most to have been formed by the evaporation of seawater. The problem with this marine origin is that the chemical and mineralogical characters of most potash deposits depart significantly from those that would be expected from simple seawater evaporation. If the marine origin is correct, then other processes must have been involved to cause the differences.

Seawater becomes progressively more concentrated as it evaporates until it is supersaturated with respect to a particular mineral phase, which then precipitates. Precipitation of a salt preferentially extracts chemical components from the seawater-derived brine, altering its overall composition. Initially, calcium combines with bicarbonate, but, after seawater has been concentrated approximately 3.5 times, gypsum ($CaSO_4 \cdot 2H_2O$) saturation is reached and calcium and sulphate are extracted from the brine. Seawater contains abundant sulphate, and, after the greater part of the calcium has been extracted (as carbonates and as gypsum), fully two-thirds of the original sulphate remains in the brine. At this stage, and in all subsequent stages of evaporation, a marine-derived brine will be impoverished in calcium and should contain abundant sulphate.

The next mineral to precipitate by continued seawater evaporation is halite (NaCl), and this extracts sodium and chloride from the brine. At about 60 times seawater concentration, the sulphate remaining in the brine should begin to be removed in the form of various magnesium sulphate minerals. Only after considerable amounts of sulphate have been eliminated from the brine will the next mineral – carnallite (hydrated magnesium and potassium chloride) – precipitate. Finally, at the last stages of concentration, the mineral bischoffite ($MgCl_2$) is precipitated.

Typical Composition of Evaporite Deposits

Only about 10% of all potash-bearing evaporites contain significant quantities of magnesium sulphate, which would be expected from simple seawater evaporation. Of these 10%, all differ from direct seawater precipitation sequences in having different magnesium sulphate minerals in different amounts from those expected. These differences can be explained in two ways. First, most magnesium sulphates precipitated during experimental seawater evaporation are highly unstable hydrous phases. These alter to less hydrous minerals upon burial. Second, during evaporation, the concentrated brines may react with previously precipitated calcium sulphate to form the mineral polyhalite. This reaction removes sulphate and potassium from the brine, so changing its composition. Further evaporation of this modified brine is capable of generating the mineral sequences found in 10% of potash salt deposits.

The majority of potash salts, however, differ substantially from those expected to result from seawater evaporation. After halite has precipitated, the next mineral to appear is carnallite (hydrous magnesium and potassium chloride) or sylvite (potassium chloride); usually there are no magnesium sulphates. The uppermost parts of some deposits, especially from the Atlantic-marginal Cretaceous salt basins of Brazil and Gabon, also contain the mineral tachyhydrite (hydrous calcium chloride). This highly soluble salt (which dissolves in atmospheric moisture) must represent the final evaporative stages of a brine. However, if this brine were concentrated seawater, then all the available calcium should have been extracted at a much earlier evaporative stage, during gypsum and early-stage halite precipitation.

Sylvite should not precipitate during simple seawater evaporation. Some sylvite can be explained as a later alteration product of carnallite, but in other cases textural evidence indicates that sylvite is a primary mineral. The absence of magnesium sulphates cannot be explained by later diagenetic changes (for instance, converting them to carnallite) because textural evidence also suggests that much carnallite is primary.

The Missing Sulphate

The brines that precipitated potash deposits low in magnesium sulphate did not lack magnesium (carnallite and bischoffite contain this element), and so the problem is to explain why marine-derived evaporites are impoverished in sulphate in their later evaporative stages.

Explanations of this missing marine-derived sulphate are unsatisfactory. One hypothesis suggests that sulphate is removed from the brine by the addition of river water containing additional calcium. This calcium strips the brine of its remaining sulphate by the precipitation of additional gypsum. Simple calculations indicate that the amount of river water needed would be enormous and more than enough to dilute the brines so that no evaporites would form in the first place.

A commonly proposed explanation is that sulphate is removed by the activity of sulphate-reducing bacteria. The sulphate is reduced to hydrogen sulphide, which is then lost to the atmosphere. Hite convincingly argued, by analogy with Holocene environments, that sulphate reduction would be confined to the uppermost metre of sediment, and so the sulphate-reducing capabilities of evaporite basins would be limited. His calculations showed that bacteria would be unable to remove all the seawater sulphate and, furthermore, that the amount of organic carbon required to reduce the marine sulphate could not be supplied, even if evaporite basins were extraordinarily productive.

The problem of the missing sulphate exists only if seawater was the original feedstock. It is clear from the presence of tachyhydrite in some sequences that either the seawater was substantially modified or some other water was evaporated to generate the potash salts. In essence, no brine containing more sulphate than calcium (including all modern seawater-derived brines) can generate tachyhydrite and evaporites deficient in magnesium sulphate, whereas these sequences can be generated if waters are used where more calcium than sulphate is present. By definition, any concentrated water that has excess calcium (calcium > sulphate + bicarbonate) is a calcium chloride brine.

Hardie believed that evaporites that are impoverished in magnesium sulphate formed by the evaporation of calcium chloride brines that were generated in rift or transtensional basins – basins with high heat flows and active hydrothermal groundwater circulations. Calcium chloride brines are being expelled today in these types of basin. Groundwater circulates deep in the crust and reacts with hot host rocks. Hydration of host-rock minerals removes water, concentrating the groundwater into a brine. The hot brine reacts with calcium-bearing minerals, commonly feldspars (exchanging sodium for calcium), and some of the expelled calcium reacts with any groundwater sulphate to precipitate anhydrite, thus stripping the brine of its contained sulphate. The final product is a brine that is depleted in sulphate and enriched in calcium. When heated, these brines become buoyant and may be expelled to the surface, where they evaporate.

Evaporites as Hydrothermal Deposits in Rift Basins

Rift and transtensional basins are ideal locations for evaporites to form. Commonly they are isolated or have very restricted access to the ocean. Even if they are not located in arid climatic zones, they may develop evaporites by virtue of uplifted rift shoulders or transpression ridges causing orographic aridity. Calcium chloride waters entering such basins as hot springs can evaporate and generate evaporite sequences without magnesium sulphate minerals. Many evaporites deficient in magnesium sulphate are located in present-day and ancient rift basins.

When hydrothermal waters are the only feedstock, the resultant evaporites may be entirely deficient in calcium and other sulphates. Evaporite sequences should have little, if any, gypsum and anhydrite. It is also possible, however, that the main water in the basin is seawater but that this is substantially modified by the addition of relatively small volumes of hydrothermal brines, which nevertheless carry large amounts of dissolved materials. Hardie calculated that modern seawater could be modified by the addition of only just over 3% Salton Sea brine into water that would not precipitate evaporites containing magnesium sulphate.

Evaporites in Non-Rift Basins

Hardie's hydrothermal brine explanation is not convincing for some evaporite sequences because they occur in non-rift basins. Kendall provided an alternative explanation for some of these evaporites. Desiccation of large evaporite basins produces large and deep depressions. This induces a hydrodynamic drive, which causes subsurface waters to migrate into the basins, where they evaporate. Dolomitization of limestones by migrating formation waters with seawater-like compositions would release calcium and generate calcium chloride waters. Basin desiccation thus provides both the drive that allows formation waters to enter the basin and a mechanism to generate waters with more calcium than sulphate. Middle Devonian evaporites in western Canada provide evidence to support this desiccation–drive model. Where calcium-rich formation waters entered the Devonian evaporite basin, spring-associated carbonates were precipitated, and there was mixing with seawater-derived brines that had already precipitated gypsum. The addition of spring-water calcium stripped the remaining sulphate from the basin brine, precipitated anomalous concentrations of calcium sulphate far into the evaporite basin, and led to the formation of a sulphate-depleted brine, which may have caused later potash salts deficient in magnesium sulphate to form in the basin.

Halite-saturated brines, refluxing beneath large evaporite basins, react with all types of sediments (not just limestones) by exchanging sodium for calcium, to generate calcium chloride brines. The main problem in understanding how these deep dense brines could form potash salts is to explain how the brines are transported to the surface. This could occur at times of basin inversion, by heating (creating buoyant hydrothermal brines) or by evaporative draw into a later evaporite basin.

Past Composition of Seawater

A more exciting alternative that explains the seemingly anomalous compositions of most ancient potash salts is that seawater compositions were substantially different in the past. Secular variations in the distributions of magnesium sulphate and evaporites deficient in magnesium sulphate are in phase with better-known variations in the mineralogy of ancient marine carbonates. These secular variations can be attributed to changes in the major-ion composition of seawater over time.

A model that explains how seawater can change over time was used to predict periods when aragonite and evaporites containing magnesium sulphate are dominant, and episodes when calcite and evaporites deficient in magnesium sulphate are favoured. Seawater chemistry is controlled by steady-state mixing of river water and mid-ocean ridge hydrothermal brines (coupled with calcium carbonate and silica precipitation). Relatively small changes in mid-ocean ridge fluxes cause significant changes in magnesium : calcium, sodium : potassium, and chloride : sulphate ratios in seawater. Such changes are believed to be responsible for variations in the primary mineralogies of marine evaporites and carbonates. Variations in mid-ocean ridge flux correspond to variations in the production rate of oceanic crust (seafloor spreading rate), and this can be estimated using various proxies, such as areas of ocean floor of different ages, the global sea-level curve, and granite-pluton emplacement rates.

Predictions of the variation in past seawater chemistry produced by variation in mid-ocean-ridge flux rates are in agreement with the known age distribution of primary marine carbonate and evaporite mineralogies. The coherence of the datasets strongly suggests that past variations in evaporite and carbonate mineralogy were largely caused by secular variations in seawater chemistry.

The idea that varying seawater chemistry can explain potash salt composition has been challenged. Holland *et al.* predict similar changes of seawater composition but of much smaller magnitude. They argue that an apparent near constancy of the level of potassium in seawater during the Phanerozoic (demonstrated by the compositions of brines trapped in ancient halites) supports this view: Hardie's model predicts significant changes in the sodium : potassium ratio. Instead, Holland *et al.* suggest that changes in past evaporite mineralogy are due to differences in the extent to which dolomitization of carbonate sediments occurred before or during seawater evaporation. During times of rapid seafloor spreading, sea-levels are higher and large carbonate platforms are more abundant. Changes in seawater chemistry (caused by increased mid-ocean ridge flux) coupled with increases in the extent of dolomitization of carbonate platforms are believed to be responsible for the formation of potash deposits that are impoverished in magnesium sulphate. This explanation resembles, in part, that suggested earlier by Kendall.

Variations in the chemistry of primary fluid inclusions from ancient halite deposits are significant. They also imply that seawater chemistry has changed significantly. Variations are in phase with inferred seafloor spreading rates, global changes in sea-level, and the primary mineralogies of ancient marine carbonates and evaporites. Of particular significance is the fact that inclusions in halites of the same age from different geological basins exhibit similar compositions. This suggests that the association with dolomitization (proposed by Holland *et al.*) is incorrect: more interbasin variation in the amount of dolomitization would be expected, resulting in a greater variation in the chemistry of fluid inclusions than that observed. It is surprising, however, that the question of whether or not variations in sodium : potassium ratios match model predictions was not addressed. More recent, unpublished, work suggests that Cretaceous and Permian seawaters were enriched in potassium and relatively depleted in sodium, as would be expected from the Hardie hypothesis.

See Also

Minerals: Sulphates. **Sedimentary Environments:** Lake Processes and Deposits. **Sedimentary Rocks:** Mineralogy and Classification; Dolomites. **Tectonics:** Hydrothermal Activity; Hydrothermal Vents At Mid-Ocean Ridges; Rift Valleys.

Further Reading

Hanor JS (1996) Variations in chloride as a driving force in siliciclastic diagenesis. In: Crossey LJ, Loucks R, and Totten MW (eds.) *Siliciclastic Diagenesis and Fluid Flow: Concepts and Applications,* pp. 3–12. Special Publication 55. Tulsa: Society for Sedimentary Geology.

Hardie LA (1990) The roles of rifting and hydrothermal $CaCl_2$ brines in the origin of potash evaporites: a hypothesis. *American Journal of Science* 290: 43–106.

Hardie LA and Spencer RJ (1990) Control of seawater composition by mixing of river waters and mid-ocean ridge hydrothermal brines. In: Spencer RJ and Chou I-M (eds.) *Fluid–Mineral Interactions: A tribute to H-P Eugster,* pp. 409–419. Special Publication 2. San Antonio: Geochemical Society.

Hite RJ (1983) The sulphate problem in marine evaporites. In: Schreiber BC (ed.) *Proceedings of the 6th International Salt Symposium, Toronto*, pp. 217–230. Alexandria, VA: Salt Institute.

Kendall AC (1989) Brine mixing in the Middle Devonian of western Canada and its possible significance to regional dolomitization. *Sedimentary Geology* 64: 271–285.

Lowenstein TK and Spencer RJ (1990) Syndepositional origin of potash evaporites: petrographic and fluid inclusion evidence. *American Journal of Science* 290: 1–42.

Ironstones

W E G Taylor, University of Lancaster, Lancaster, UK

Introduction

Ironstones have been critical to industry and industrial revolutions. They have been essential raw materials since the dawn of the Iron Age (about 700 BC). Without iron-rich deposits many of the manmade structures and utensils that we take for granted today – tall urban buildings, power pylons, bridges, ships, cutlery, hammers, saws, and the seemingly indispensable motor car – could not exist.

The point at which an ironstone deposit is considered to be an ore has changed considerably over the years and depends upon the particular economic, technological, social, and political circumstances at the time. Nowadays deposits need to have an iron content in excess of 60% by weight to be worked, whilst in the mid-twentieth century ironstones with 28% iron by weight were regularly extracted as ores. The quality of the potential ore, and in particular the proportion of phosphatic material, is an important factor that has to be considered in the mining of iron.

Initially, the availability of water power and the proximity of coal were the factors controlling production. The middle of the nineteenth century saw a change from coal-fired furnaces producing cast iron to the Bessemer process, which produced steel. Later in the same century, the open-hearth process and various refinements were developed. Each of these new production processes demanded ores of a particular quality.

Records of global production are scarce before the latter half of the twentieth century, and certainly in Europe much of the exploitation predates that century. In Great Britain the maximum annual output was in the order of 18 Mt (Milliontonnes) and occurred during two main periods – 1870–1890 and 1940–1945. In the former period the main type of ore extracted was from the blackband and claystone ironstones (see below) of the Carboniferous rocks of various coalfields, whilst in the latter period the

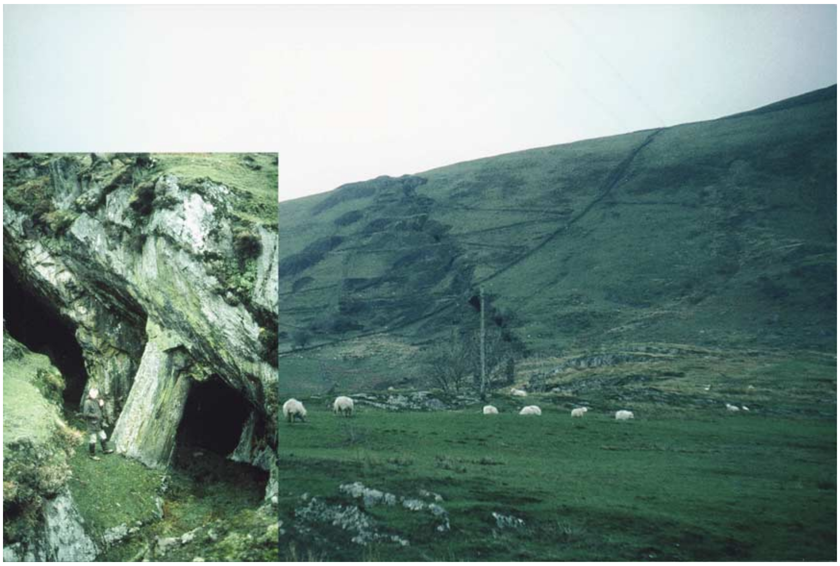


Figure 1 Ordovician ironstones, Betws Garmon, North Wales, UK (inset showing the steeply inclined stoping method of underground mining for the deposit and a residual pillar of magnetite-rich material).

ooidal ironstones of Northamptonshire and Lincolnshire were the dominant ores. Although Ordovician ooidal ironstones from North Wales were extracted until early part of the twentieth century (Figure 1).

The terms used to describe both the processes of ironstone formation and the ironstones themselves have been many and varied. Attempts to simplify and standardize the terminology have recently met with some success, mainly through the International Geological Correlation Programme (IGCP). For example, the terms ‘Clinton’ (from the Silurian Clinton Group of New York State, USA) and ‘Minette’ (from the Jurassic Minette oolitic ironstones of northeastern France and adjacent areas) as descriptions of ironstone types have proved to be unsatisfactory and have now fallen into disuse.

Definition

Largely because iron may invade and impregnate a wide range of rocks, defining what constitutes an ironstone has proved difficult. Exhortations to merge the nomenclatures of the various iron-rich deposits, such as the banded iron formations and ironstones, have been resisted on the basis that the mineralogy, petrology, and genesis of these deposits are distinct and separate. A precise definition of ‘ironstone’ was agreed only in the last decade of the twentieth century and stems from the work of the IGCP 277 (Phanerozoic Ooidal Ironstones). Ironstones are sedimentary rocks consisting of at least 15% iron by weight, which may be quoted as 19% FeO or 21% Fe_2O_3 or an equivalent admixture in a chemical analysis. They occur almost exclusively in the Phanerozoic Era and are distinguished from the mainly Precambrian banded iron formations (*see* **Sedimentary Rocks:** Banded Iron Formations) by their lack of both regular banding and chert and by their age: banded iron formations were produced when there was a deficiency of oxygen in the Earth’s atmosphere. The ferruginization (iron enrichment) may be the result of either direct deposition or subsequent chemical changes.

Ironstone Mineralogy

The iron-ore minerals may be oxides (including goethite, haematite, and magnetite), carbonates (usually siderite), or silicates (normally berthierine or chamosite). They may be associated with other carbonate minerals, sulphides and/or phosphatic minerals.

Goethite – FeO(OH) – is commonly formed by oxidation during weathering. Also, in many ooidal ironstones, it can result from the oxidation of berthierine; the two minerals may be found intermixed, often in alternate concentric layers, in ooids. Limonite was formerly thought to be a distinct mineral with the composition $2Fe_2O_3 \cdot 3H_2O$ but is now considered to have a variable composition (and properties) and to consist of several iron hydroxides (commonly goethite) or a mixture of iron minerals. Generally, it occurs as a secondary alteration product. Haematite – Fe_2O_3 – can be an important mineral in some ironstones, where it is usually formed as a late-stage diagenetic product of the alteration of goethite. Experimental synthesis indicates that this transformation occurs at a temperature above 80°C and at a depth of about 2 km. Magnetite – Fe_3O_4 – normally forms during the low-grade metamorphism of ironstones, although the mineral has been reported from unmetamorphosed deposits in Libya.

Siderite – $(Fe,Mg,Ca,Mn)CO_3$ – is a very important mineral in ironstones. It is the only iron-bearing mineral in many claystone and blackband ironstones. Substitution of magnesium, calcium, or manganese for iron in the structure of siderite has been hypothesized to be related to the environment of formation. Substitution appears to have been greatest in marine sediments and in those ironstones formed during the later diagenetic stages of non-marine sediments.

Berthierine – $(Fe^{2+},Fe^{3+},Al,Mg,Mn)_2(Si,Al)_2O_5(OH)_4$ – is a 0.7 nm repeat serpentine. Reported variations in the chemical composition may reflect the analytical difficulties of dealing with very fine-grained samples. The formation of berthierine in ironstones is the subject of some debate and will be considered later. Chamosite – $(Fe_5^{2+},Al)(Si_3Al)O_{10}(OH)_8$ – is a 1.4 nm repeat chlorite with a very similar chemical composition to berthierine. Experimental work has shown that berthierine may be transformed into chamosite at a temperature of about 150°C and a depth of about 3 km. The phyllosilicate glauconite – $(K,Na)(Al,Fe^{3+},Mg)_2(Al,Si)_4O_{10}(OH)_2$ – is generally thought to be restricted to marine environments and occurs in some ironstones.

Other carbonate minerals, such as calcite, aragonite, dolomite, and ankerite – $Ca(Mg,Fe)(CO_3)_2$ – may be particularly common in ironstones both as a constituent of the cement and as discrete bioclasts. Phosphate minerals, such as francolite (carbonate fluorapatite) and vivianite – $Fe_3(PO_4)_2 \cdot 8H_2O$ – can be major components of ooidal ironstones. They can be a detrimental contributory factor to the viability of a deposit as an ore, particularly since, in most cases, the mineral grains are very small and difficult to separate.

Types of Ironstones

Extensive use of high-precision microscopy and analytical techniques has allowed detailed insights into the composition and formation of ironstones. A formal classification of ironstones has not yet been universally accepted, but three distinctive categories have been recognized (Figure 2).

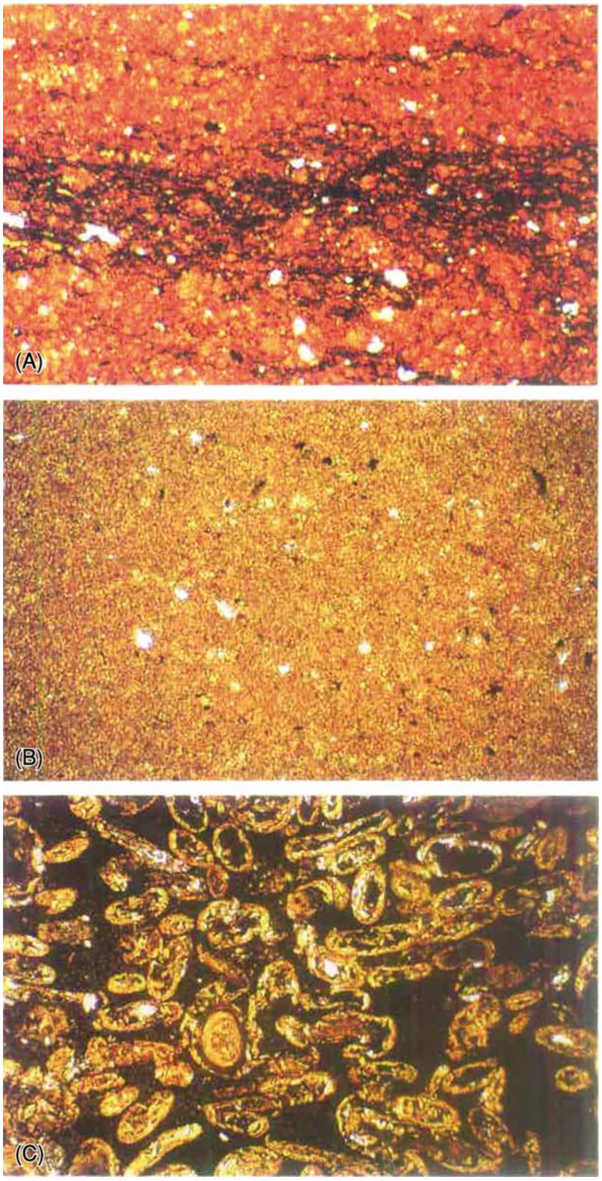


Figure 2 Typical photomicrographs of the three categories of ironstone. (A) Blackband ironstone. Note the dark organic-rich lamination in a mainly sideritic matrix. Plane-polarized light; horizontal dimension is 1.3 mm. (B) Claystone ironstone. Note the lack of organic material in the mainly sideritic matrix. Plane-polarized light; horizontal dimension is 1.3 mm. (C) Ooidal ironstone. Ooids, showing selective replacement by siderite and phosphatic minerals, are set in a berthierine-rich mudstone matrix. Plane-polarized light; horizontal dimension is 5.2 mm (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446–489. London: Chapman & Hall, plates 14b, 14a and 14e).

Blackband Ironstones

Blackband ironstones are, typically, fossiliferous sapropel-rich (usually with an organic content in excess of 10%) finely laminated sideritic ironstones. Although non-laminated types are known, more frequently they are formed of alternating siderite- and organic-rich laminae. They are found almost exclusively above coal seams in a lacustrine parasequence with mudstone and seat earth deposits (Figure 3). Palaeontological and mineralogical evidence indicates that these ironstones were formed during freshwater inundation. Unlike non-marine clayband ironstones, there is an absence of early diagenetic pyrite, and the occurrence of coal balls (calcite–pyrite concretions) is an indication of marine incursion. The ironstones typically form thin (less than 10 cm) sheets of less than 10 km^2 extent and often change laterally into limestones with similar textural characteristics. Bog iron ores, which occur as lenses of ferruginous concretions within peat deposits, are thought to be the modern analogues of blackband ironstones.

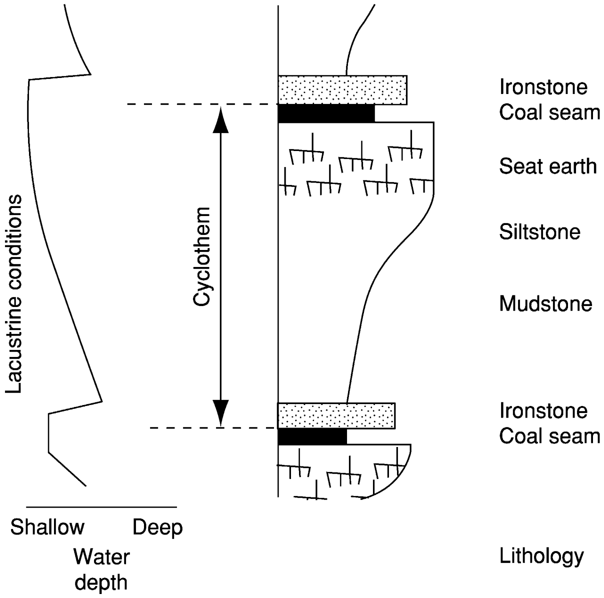


Figure 3 Idealized stratigraphical column for blackband ironstones showing relative water depths of sedimentation, not to scale.

Claystone Ironstones

Claystone or clayband ironstones have been the basis of the steel industry in many industrialized countries, largely because of their association with coalfields. Essentially, they are accumulations of iron carbonates (usually siderite) that have replaced the non-marine shales of coal-measure cyclothems (parasequences) and occur as either thin sheets or, more commonly, layers of concretions (Figure 4). Occasionally these sheets may extend over several hundred square kilometres. Normally, each concretion is unlaminated and does not contain high amounts of organic material, and the siderite grains are usually microscopic or sub-microscopic in size (less than 10 μm). Marine claystone ironstones are predominately rich in ankerite with pyrite, and production of siderite is suppressed. Irregularly shaped sphaerosiderites (ball ironstones), which usually occur at the base of palaeosols, are composed of siderite cement in the form of distributed spherulites (0.5–1 mm in diameter).

Ooidal Ironstones

Ooidal ironstones are characterized by the presence of ooids and/or pisoids and are very diverse, with a

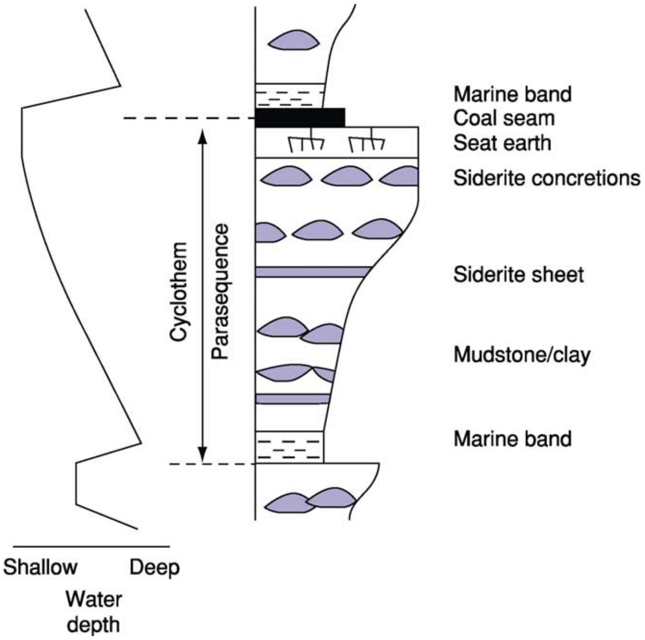


Figure 4 Idealized stratigraphical column for claystone ironstones showing relative water depths of sedimentation, not to scale.

wide range of mineralogy, textures, and chemical compositions. Because they possess oolites, shell fragments, and mud matrices in various admixtures similar to limestones (*see* **Sedimentary Rocks:** Limestones), most researchers in the field use the petrographic terminology advocated by Young to describe and classify ooidal ironstones. Most ooidal ironstones are less than a metre in thickness and are laterally persistent over approximately 100 km^2. A few deposits are in excess of 15 m thick (e.g. the Gara Djebilet Ironstone in Algeria). Although an idealized stratigraphical model for this type of ironstone consists of an upward shoaling sequence from black shales at the base, through progressively coarser deposits, to the ironstone at the top (Figure 5), in practice there are many deviations from this standard. Ironstones develop during periods of reduced sediment input (starvation), with abundant bioturbation, and often exhibit signs of storm reworking to form tempestites. The earliest-formed minerals are usually iron oxides and silicates. Iron-rich carbonates may be generated subsequently, often during early diagenesis.

Modern Examples of Ironstone Development

Bog iron ores are found associated with peat deposits in swampy conditions. Typically they contain hydrated ferric-oxide and manganese-oxide cements but, below the water table, they may be cemented by siderite. It has been suggested that microbial activity in tropical climates particularly promotes the direct precipitation of siderite.

A possible present-day analogue of ancient ooidal ironstones appears to be the verdine facies. In this facies, iron-rich aluminous green clay minerals replace bioclasts and pellets. Ferruginous peloids, in many cases altered faecal pellets, are known to be forming today in sediments deposited in front of equatorial deltas, such as those on the continental shelves off Senegal, Guinea, Nigeria, Gabon, Sarawak, and east Kalimantan. Present-day examples of ferruginous ooid accumulation are rare. In the interior of Africa, along the southernmost parts of Lake Malawi, amorphous ferric-oxide ooids have been found associated with geothermal springs, and, in the brackish open water of southern Lake Chad, goethitic brown ooids are being formed. In the shallow seas of northern Venezuela, berthierine-rich green-brown muddy ooidal sediments with peloids have been discovered.

Environment of Deposition and Subsequent Alteration during Lithification

Very few generalizations can be made about the sedimentary environment of ironstones. Ironstones may be deposited in shallow-marine, interdeltaic, nonmarine lacustrine, and alluvial environments and may interfinger or replace sandy and shelly marine deposits laterally. They are frequently associated with

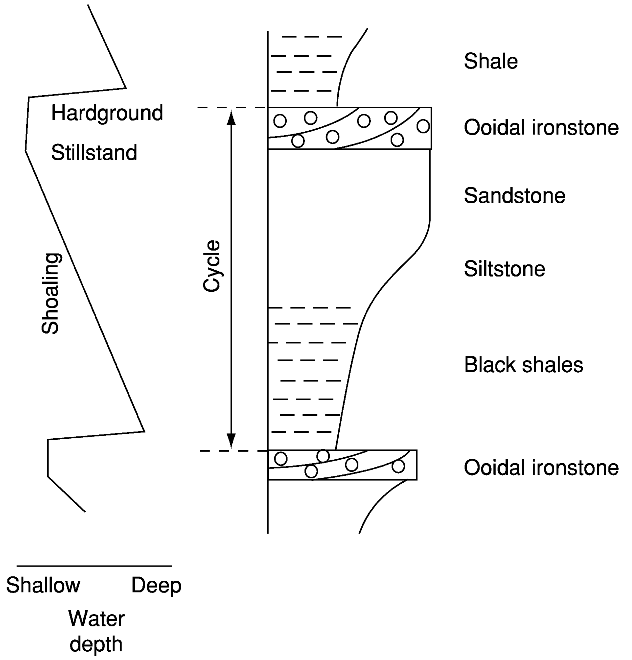


Figure 5 Idealized stratigraphical column for ooidal ironstones showing relative water depths of sedimentation, not to scale.

phosphates, coals, evaporites or laterites, and most have no direct relation to volcanism.

Blackband ironstones have many of the characteristics of bog iron ores, which are developed *in situ*, soon after deposition, by a reaction between organic material and underground colloidal iron-rich solutions under a thick vegetative cover. Progression of the process could yield siderite by reduction. Alternative evidence has been put forward suggesting that these deposits could form by direct sideritic precipitation from tropical swamp waters that are already rich in iron. Blackband ironstones are always developed in close proximity to coal seams, so either process could be feasible.

The diagenesis of the fine-grained claystone ironstones has been studied in great detail (Figure 6). Most became enriched in iron during very early diagenesis along or near the sediment–water interface. Based upon distinct chemical reactions involving the oxidation of organic matter buried within the sediment, diagenetic zones have been established. Although the zones can be considered as due to burial, their development is especially dependent upon the availability of oxidizing agents and organic matter, the sedimentary environment, the nature and amount of organic material, the composition of the inorganic sediment, the hydrological regime of the sedimentary pile, and the composition of the overlying water. The reactions below the oxic zone may be complicated by kinetic controls, which could explain the occasional appearance of residual ferric iron in an anoxic environment. Because some siderite concretions are developed early and are associated with many nonsequences, the sedimentation rate must have been relatively low (less than 40 m Ma^{-1}). Whilst claystone ironstones are formed during diagenesis by the growth of siderite in the pore spaces of argillaceous materials, sphaerosiderites form by the direct precipitation of siderite from pore fluids, and their size and shape probably reflect a higher growth rate. They can occur in a variety of environments, including the deep sea, but are usually products of a waterlogged zone below a leached soil profile.

The exact genesis of ooidal ironstones remains controversial. Particularly, the origin of the ooids is the subject of a long-lasting debate. The original constituents of ooids and how they vary from deposit to deposit are not known with any certainty. It is debatable whether the ooids grew from solutes, colloidal particles in solutions, or gels. The ferrous ion in bicarbonate form survives only in an anoxic or reducing environment, so this would place a severe constraint on its presence in solution. Ferruginous ooids are commonly built of alternating ferric oxide and berthierine sheaths of submicroscopic thickness. Whether the initial crystalline phase was berthierine

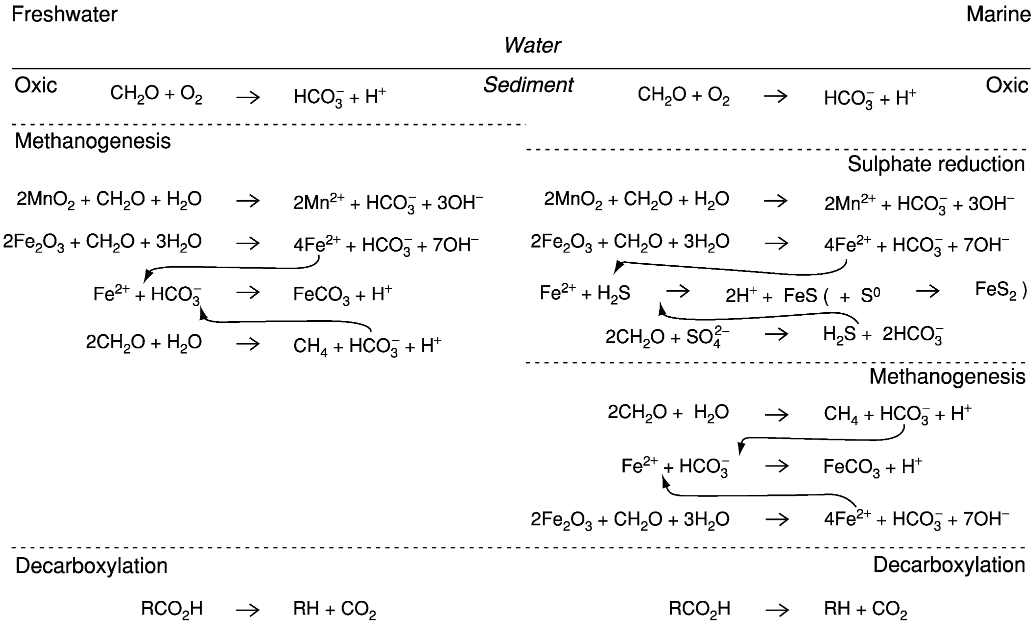

Freshwater
Marine
Water
Sediment
Oxic
$CH_2O + O_2 \rightarrow HCO_3^- + H^+$
Oxic
$CH_2O + O_2 \rightarrow HCO_3^- + H^+$
Methanogenesis
Sulphate reduction
$2MnO_2 + CH_2O + H_2O \rightarrow 2Mn^{2+} + HCO_3^- + 3OH^-$
$2Fe_2O_3 + CH_2O + 3H_2O \rightarrow 4Fe^{2+} + HCO_3^- + 7OH^-$
$Fe^{2+} + HCO_3^- \rightarrow FeCO_3 + H^+$
$2CH_2O + H_2O \rightarrow CH_4 + HCO_3^- + H^+$
$2MnO_2 + CH_2O + H_2O \rightarrow 2Mn^{2+} + HCO_3^- + 3OH^-$
$2Fe_2O_3 + CH_2O + 3H_2O \rightarrow 4Fe^{2+} + HCO_3^- + 7OH^-$
$Fe^{2+} + H_2S \rightarrow 2H^+ + FeS\ (+\ S^0 \rightarrow FeS_2)$
$2CH_2O + SO_4^{2-} \rightarrow H_2S + 2HCO_3^-$
Methanogenesis
$2CH_2O + H_2O \rightarrow CH_4 + HCO_3^- + H^+$
$Fe^{2+} + HCO_3^- \rightarrow FeCO_3 + H^+$
$2Fe_2O_3 + CH_2O + 3H_2O \rightarrow 4Fe^{2+} + HCO_3^- + 7OH^-$
Decarboxylation
$RCO_2H \rightarrow RH + CO_2$
Decarboxylation
$RCO_2H \rightarrow RH + CO_2$

Figure 6 Summary of reactions and zonation that may occur during the diagenesis of sediments in marine and non-marine conditions (after Curtis and Coleman 1986, Spears 1989 and reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446–489. London: Chapman & Hall, Figure 9.5 after Curtis and Coleman 1986 and Spears 1989).

or a precursor ferric mineral (e.g. odinite) is uncertain. Some feel that it was crystallized at the earliest stage, probably from a gel; others have suggested that it formed during the early stages of burial diagenesis. Alternatively, it could be a product of the transformation of either a mixture of kaolinite and hydrous ferric oxide or a complex synthesis of silicic, ferric, and aluminous substances. The processes involving micro-organisms (such as bacteria) are not understood, particularly in terms of how they promote the growth of ooids. Reworking of ooidal sediments in shallow-water environments often separates, concentrates, and highly sorts the ooids, forming lenses, which probably accumulated in shallow depressions. Often zonation of ironstones may be observed when the body is less affected by redistribution.

The variable nature of the nuclei of ooids and the trapping of marine microflora during growth indicate that ooids are probably generated within the host sediment. Ferruginous ooids could have grown on the seafloor, at the water–sediment interface, by either concentric growth due to precipitation of mineral matter, frequently around heterogeneous nuclei, or mechanical accretion by rolling (like a snowball). Alternatively, they could have grown inside the sediment at shallow depths below the water–sediment interface either as early diagenetic microconcretions or by replacement or addition of iron to peloids. Fluviatile examples do exist (e.g. the Late Oligocene deposits of Aral Lake, Russia), in which ooids have been developed on land and then moved, but this has not been convincingly demonstrated to be of general application.

The IGCP 277 came to the conclusion that ferruginous deposition must have been due to the interplay of a number of different processes and hence that there is rarely a single genetic explanation. The salinity of seawater, the carbon dioxide and oxygen contents of the atmosphere, the action of organisms, the sources and availability of iron compounds, seasonal or long-lasting climatic conditions, specific physicochemical conditions, the marine water depth, diagenetic processes, and tectonism are all potential factors. However, the dominant influences seem to be the local hydrodynamic conditions and the topographical relief of the land and seafloor, which may help to protect the ooidal deposits from excessive dilution by clastics. Paradoxically, the rates of deposition of the stratigraphical equivalents of many ooidal ironstones do not always correspond to the periods of lowest detrital input. Changes during burial are numerous and complex and include the formation of phosphatic minerals, iron oxides, siderite, pyrite, and quartz. In most cases, these are followed by alterations due to the effects of meteoric waters.

The Ferruginization Process

Although ironstones are generally considered to be the products of ferruginization during diagenesis, the physical sedimentary environment is thought to control the style of diagenesis in ironstones. Blackband ironstones are geochemically and isotopically homogeneous, suggesting stability of conditions during growth. They were probably formed close to the sediment surface, with precipitation of siderite, and not during progressive burial (Figure 7). The high manganese content of siderite, the relatively low calcium and magnesium contents, and the high carbonate content support this. Studies of carbon isotopes show that calcareous shells from limestones and ironstones have similar $\delta^{13}C_{PDB}$ values (from +4‰ to −6‰), indicating that the siderites replaced original calcite or aragonite and precluding the domination of methanogenesis. As has been previously noted, very early siderite could be precipitated directly from swamp waters, but could precipitation have occurred from primitive freshwater too?

Claystone ironstones usually have lower manganese and ^{13}C enrichment than blackband ironstones, which can be related to slightly later ferruginization, which takes place below the oxic zone in non-marine waters by diagenetic distribution of iron within the sediment (Figure 8). The relationships leading to the precipitation of iron minerals are complex and are susceptible to slight shifts in the concentrations and availability of reactants especially S and organic C. The thermodynamics of the reactions predict the observation of manganese enrichment within the concretion cores. The iron and manganese would have been present in the silicate minerals of the sediment. Sulphate reduction would be inhibited, as the sediments were isolated from potential sources of sulphate (e.g. seawater), and changes in organic matter would be methanogenic, giving rise to bicarbonates rich in ^{13}C ($\delta^{13}C_{PDB}$ values in the order of +10‰). Also precipitation would be enhanced by an increase in alkalinity resulting from the combination of changes in organic matter and the reduction of Fe^{3+} and Mn^{4+}. Any growth at deeper levels of burial would be slow under decarboxylation conditions with bicarbonate depleted in ^{13}C. In marine claystone ironstones, sulphate and iron reduction would proceed broadly simultaneously, leading to the production of iron pyrites. Siderite is normally rare in marine sediments because iron can become incorporated in carbonates only below the zone of sulphate reduction.

Over the past two decades there have been significant developments in research into the environmental conditions under which ooidal ironstones are formed.

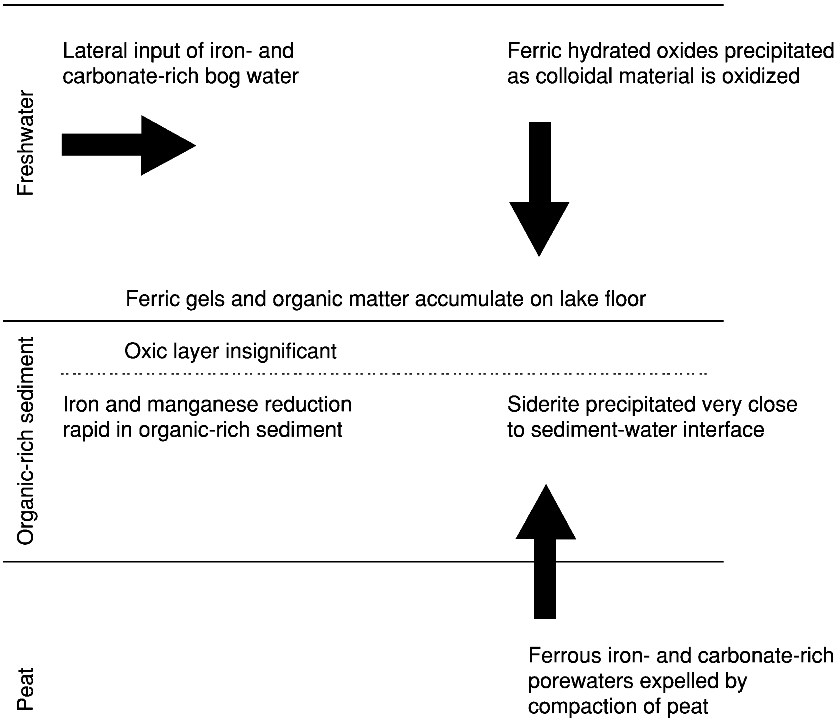

Freshwater
Lateral input of iron- and carbonate-rich bog water
Ferric hydrated oxides precipitated as colloidal material is oxidized
Ferric gels and organic matter accumulate on lake floor
Organic-rich sediment
Oxic layer insignificant
Iron and manganese reduction rapid in organic-rich sediment
Siderite precipitated very close to sediment-water interface
Peat
Ferrous iron- and carbonate-rich porewaters expelled by compaction of peat

Figure 7 Model of mineralization for blackband ironstones (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446–489. London: Chapman & Hall, Figure 9.9).

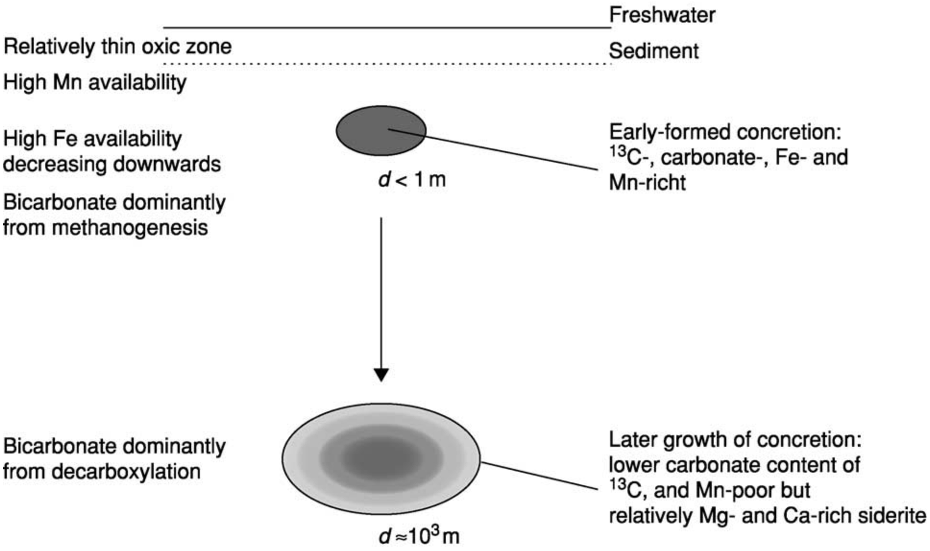

Freshwater
Sediment
Relatively thin oxic zone
High Mn availability
High Fe availability decreasing downwards
Bicarbonate dominantly from methanogenesis
d < 1 m
Early-formed concretion: 13C-, carbonate-, Fe- and Mn-richt
Bicarbonate dominantly from decarboxylation
d ≈ 10³ m
Later growth of concretion: lower carbonate content of 13C, and Mn-poor but relatively Mg- and Ca-rich siderite

Figure 8 Model of mineralization for claystone ironstones (*d*, diameter) (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446–489. London: Chapman & Hall, Figure 9.7).

Sea-level change may be the most significant genetic control since it can generate very low accumulation rates within basins with low overall sedimentation rates (Figure 9). Widespread sediment starvation could be produced by relative sea-level rise in shallow epeiric seas with a topographically low hinterland. There is a dispute as to whether these conditions appertain to the end of regression or to the beginning

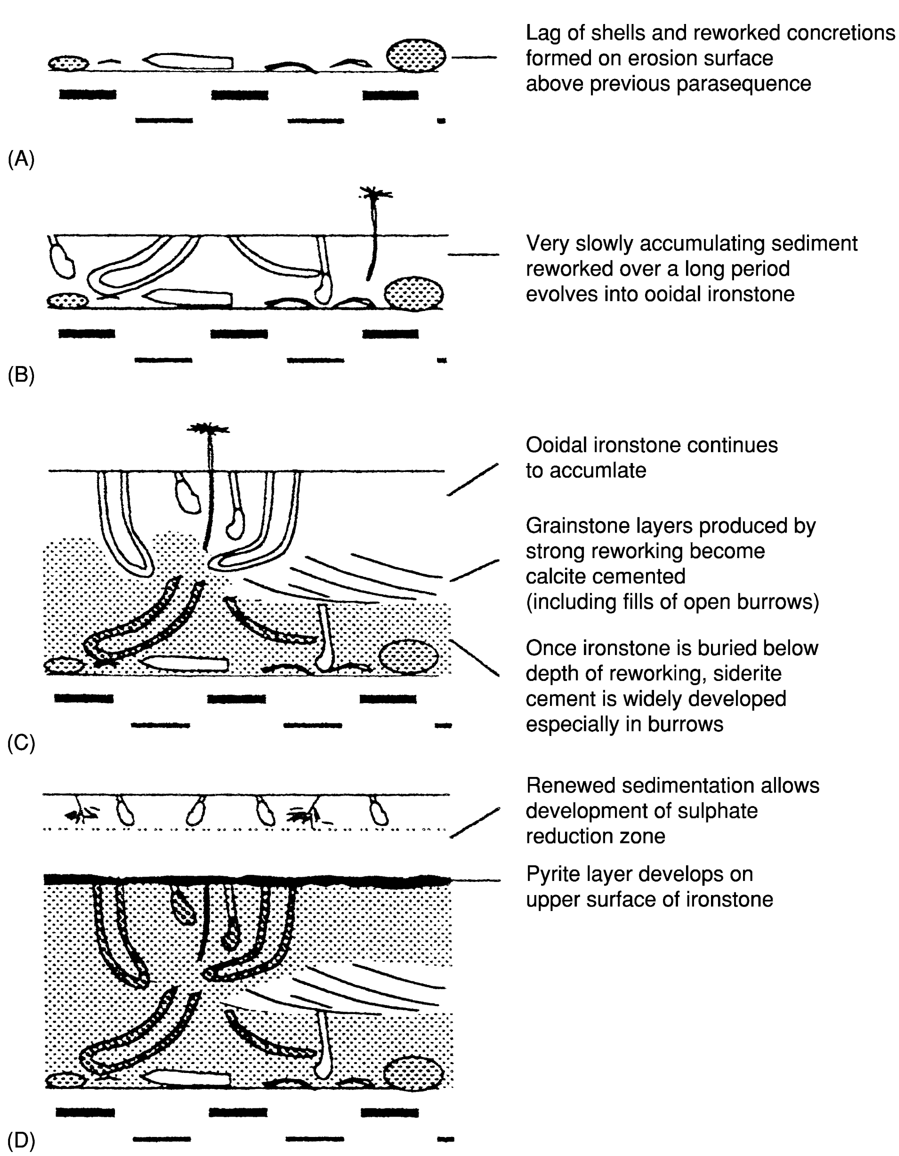

Lag of shells and reworked concretions formed on erosion surface above previous parasequence
(A)
Very slowly accumulating sediment reworked over a long period evolves into ooidal ironstone
(B)
Ooidal ironstone continues to accumlate
Grainstone layers produced by strong reworking become calcite cemented (including fills of open burrows)
Once ironstone is buried below depth of reworking, siderite cement is widely developed especially in burrows
(C)
Renewed sedimentation allows development of sulphate reduction zone
Pyrite layer develops on upper surface of ironstone
(D)

Figure 9 Model of mineralization for ooidal ironstones; (A), earliest phase; (B) and (C), middle phases; (D), latest phase (reproduced with kind permission of Kluwer Academic Publishers from Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles*, pp. 446–489. London: Chapman & Hall, Figure 9.13).

of transgression. Similarly, the origin of early sideritic units can be related to a decrease in the role of sulphate reduction, a low sedimentation rate (less than 40 m Ma^{-1}), and oxygenated and carbon-poor sediments (values of $\delta^{13}C_{PDB}$ for various cements vary from −3 to −22‰).

The source and influx of iron is a subject of much controversy. In 1856, Sorby proposed that extensive ferruginization occurred during later diagenesis, but this idea is no longer accepted, since most ferruginous ooids were formed within the depositional environment. There are three proposals for the origin of iron enrichment:

1. iron-rich exhalative fluids, supplying the sediment–water interface (some examples do seem to be related to the episodic reactivation of faults involving exhalative hydrothermal or seep sources);
2. mechanical accretion of lateritic terrestrial weathering products (e.g. kaolinite and iron oxides) or lateritic soils to form ooids in a marine environment, with subsequent transformation to berthierine (this does not seem to be generally applicable since unaltered primary ooids with mixed iron oxide–kaolinite composition have not been found in marine ooidal ironstones); and

3. leaching from underlying sediments, especially organic-rich shales, during very early marine alteration of detrital material (the diagenetic redistribution of iron within sediments is difficult to demonstrate in ancient ooidal ironstones since the process would probably require considerable time).

The role of clay minerals in effecting ferruginization is unknown, particularly in respect of the transformation of non-iron-bearing phyllosilicates into iron-bearing ones and the role of iron-rich green trioctohedral clay minerals of warm seas as possible precursors of later ooidal minerals.

Stratigraphical Record (Temporal Occurrences) and Tectonic Settings

It has puzzled geologists that some geological periods have significant numbers of ironstones, whilst other periods are devoid of them. Ironstones are almost completely restricted to the Phanerozoic. Blackband and claystone ironstones are particularly prevalent in the Carboniferous, when the depositional basins occupied near-tropical locations. Ooidal ironstones are particularly common in the Ordovician, Devonian, Jurassic, and Cretaceous periods. Most ironstones were formed in warm climates, although some were deposited in cooler climates (e.g. the Late Ordovician and Late Permian ironstones). Palaeolatitude data has shown that the Ordovician and Devonian ironstones formed in a zone of the Gondwanan shelf seas ranging from 45°N to 65°S of the palaeoequator. In the Jurassic and Cretaceous, ironstones formed between 70°N and 10°S. For this reason climate cannot be the major contributory factor in their formation.

Ironstones are largely confined to three types of cratonic setting.

1. Many developed in anorogenic basins dominated by prolonged stability and sometimes with complex extensional faulting that involved the formation of marine basins and swells in areas of subdued relief.
2. Some developed along the margins of cratons during initial convergence or divergence of plates.
3. Other ironstones accumulated on the inner sides of mobile belts at times of diminished deformation.

Glossary

Condensed deposit A relatively thin but uninterrupted sedimentary sequence representing a significant period of time during which the deposits have accumulated very slowly. It is generally correlated with a thicker time-equivalent succession elsewhere.

Ferruginization A synonym of ferrification and the preferred term by IGCP 277 to describe the processes of iron-enrichment of various Earth materials.

Hardground A zone at the seafloor a few centimetres thick, where the sediment is lithified to form a hardened surface, which is often encrusted and bored.

IGCP The International Geological Correlation Programme.

Neoformation A synonym of neogenesis, the formation of new minerals.

Ooid A synonym of oolith and the preferred term to describe a spherical or ellipsoidal accretionary sand-sized (diameter of 0.25–2 mm) particle in a sedimentary rock (mainly limestones and ironstones). Ooids usually consist of successive concentric layers (often carbonates) around a central nucleus.

Pellet A small, usually ellipsoidal, aggregate of accretionary material (mainly micrite) that has, in most cases, formed from the faeces of molluscs and worms.

Peloid An allochem composed of micrite, irrespective of size or origin, without internal structure. Includes both pellets and intraclasts.

Pisoid A synonym of pisolith and the preferred term to describe small round or ellipsoidal particles (diameter of 2–10 mm) in a sedimentary rock (mainly limestones and ironstones). Pisoids are larger than ooids and usually consist of concentric layers around a central nucleus.

Stillstand A period of time when an area of land is stable relative to mean sea-level (or some other global measure), leading to a relatively unvarying base level of erosion.

Verdine facies Green marine clay characterized by the authigenesis (neoformation *in situ*) of iron-rich aluminous clay minerals, especially 0.7 nm repeat odinite, but not berthierine or glauconite.

See Also

Economic Geology. **Palaeozoic:** Carboniferous. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Mineralogy and Classification; Banded Iron Formations; Clays and Their Diagenesis; Limestones.

Further Reading

Boardman EL (1989) Coal measures (Namurian and Westphalian) blackband iron formations: fossil bog iron ores. *Sedimentology* 36: 621–633.

Curtis CD and Coleman ML (1986) Controls on the precipitation of early diagenetic calcite, dolomite and siderite concretions in complex depositional sequences. In: Gautier DL (ed.) *Roles of Organic Matter in Sediment Diagenesis,* pp. 23–33. Special Publication 38. Denver: Society of Economic Palaeontologists and Mineralogists.

Curtis CD and Spears DA (1968) The formation of sedimentary iron minerals. *Economic Geology* 63: 257–270.

Kearsley AT (1989) Iron-rich ooids, their mineralogy and microfabric: clues to their origin and evolution. In: Young TP and Taylor WEG (eds.) *Phanerozoic Ironstones,* pp. 141–164. Special Publication 46. London: Geological Society of London.

Kimberley MM (1994) Debate about ironstone: has solute supply been surficial weathering, hydrothermal convection, or exhalation of deep fluids? *Terra Nova* 8: 116–132.

Odin GS (ed.) (1988) *Green Marine Clays, Oolitic Ironstone Facies, Verdine Facies, Glaucony Facies and Celadonite-Bearing Facies – A Comparative Study.* Developments in Sedimentology 45. Amsterdam: Elsevier.

Petranek J and Van Houten F (1997) *Phanerozoic Ooidal Ironstones.* Special Papers 7. Prague: Czech Geological Survey.

Spears DA (1989) Aspects of iron incorporation into sediments with special reference to the Yorkshire Ironstones. In: Young TP and Taylor WEG (eds.) *Phanerozoic Ironstones,* pp. 19–30. Special Publication 46. London: Geological Society of London.

Taylor JH (1949) *The Mesozoic Ironstones of Britain: Petrology of the Northampton Sand Ironstone.* Memoir of the Geological Survey of Great Britain. London: Geological Survey of Great Britain.

Van Houten FB and Arthur MA (1989) Temporal patterns among Phanerozoic oolitic ironstones and oceanic anoxia. In: Young TP and Taylor WEG (eds.) *Phanerozoic Ironstones,* pp. 33–49. Special Publication 46. London: Geological Society of London.

Young TP (1993) Sedimentary iron ores. In: Pattrick RAD and Polya DA (eds.) *Mineralization in the British Isles,* pp. 446–489. London: Chapman & Hall.

Young TP and Taylor WEG (eds.) (1989) *Phanerozoic Ironstones.* Special Publication 46. London: Geological Society of London.

Limestones

R C Selley, Imperial College London, London, UK

Introduction

Limestones are one of the most important of all the sedimentary rocks introduced in (*see* **Sedimentary Rocks:** Mineralogy and Classification). Limestones are composed largely of calcium carbonate ($CaCO_3$) in the mineral form calcite, but there are several other important carbonate minerals with which limestones are associated. This article opens by discussing important differences between limestones and sandstones, and continues by outlining the mineralogy, classification, and rock names of limestones. This is followed by a brief account of limestone depositional environments, and, logically, by their postdepositional diagenesis. The article concludes with a description of the economic importance of limestones, which is considerable, and a selected reading list.

Differences between Limestones and Sandstones

Limestones and sandstones are the two most important groups of sedimentary rocks. However, limestones pose a completely different set of problems to those of sandstones, the solutions of which require the application of different concepts and techniques.

First, limestones, unlike sandstones, are intrabasinal in origin. That is to say they form in the environment in which they are deposited. The source material of sandstones, by contrast, has been weathered, eroded, transported, and may finally be deposited hundreds of kilometres from its point of origin. Sandstones (or siliciclastic rocks) therefore often contain many different minerals. Limestones, by contrast, have a much simpler mineralogy, generally consisting of only calcite and two or three others (which will be mentioned shortly). Siliciclastic sand grains may hold clues to their source, but tell little of their depositional environment. Limestone grains, by contrast, although largely monomineralic, occur in a wide range of sizes and shapes, reflecting their multiple origins. These grains form in specific environments from which they are seldom transported. Limestone grains thus give important clues about their environment of deposition.

When studying sandstones, vertical profiles of grain size and analysis of sedimentary structures are the keys to environmental diagnosis. With limestones, however, it is the analysis of grain type and texture that aids environmental diagnosis.

The second large difference between sandstones and limestones lies in their chemistry. Sandstones are composed largely of quartz (SiO_2) sand, whilst limestones are composed largely of the mineral calcite ($CaCO_3$). In the subsurface environment, silica is chemically relatively inert, whereas calcite is much more reactive. This means that diagenesis in sandstones is relatively less important. Primary intergranular porosity may be preserved as it was when the sand was first deposited. In limestones, however, primary intergranular porosity is often quickly infilled by cement, even before the sediment has been buried. In the subsurface though, limestones are more vulnerable to the effects of acid solutions which can selectively leach out the rock and generate secondary pores. These may just be where individual fossil shells have leached out (Biomouldic porosity). These pores may enlarge to cross-cut the fabric of the rock (vuggy porosity), or even form caves, described by geologists as cavernous porosity (cavernous pores are defined as those that are large enough to contain a crouched geologist, or for the drill string to drop by 1 m or more). Fenestral porosity is a less common but significant type of pore system found in intertidal lagoonal muds. It forms from the buckling of laminae or the trapping of gas bubbles in carbonate mud when exposed to hot sunshine. It is characterized by thin, horizontally elongated pores, thus giving good horizontal and poor vertical permeability. The replacement of limestone by dolomite (dolomitization) may create intercrystalline porosity.

Through the creation of pore systems of diverse shapes and sizes, diagenesis may completely destroy the original depositional fabric of the sediment, a feature unknown in sandstones whose diagenetic overprint has little impact on the primary features.

These differences between limestones and sandstones will become clearer as this article unfolds.

Limestone Mineralogy, Grains, and Rock Names

Limestone Mineralogy

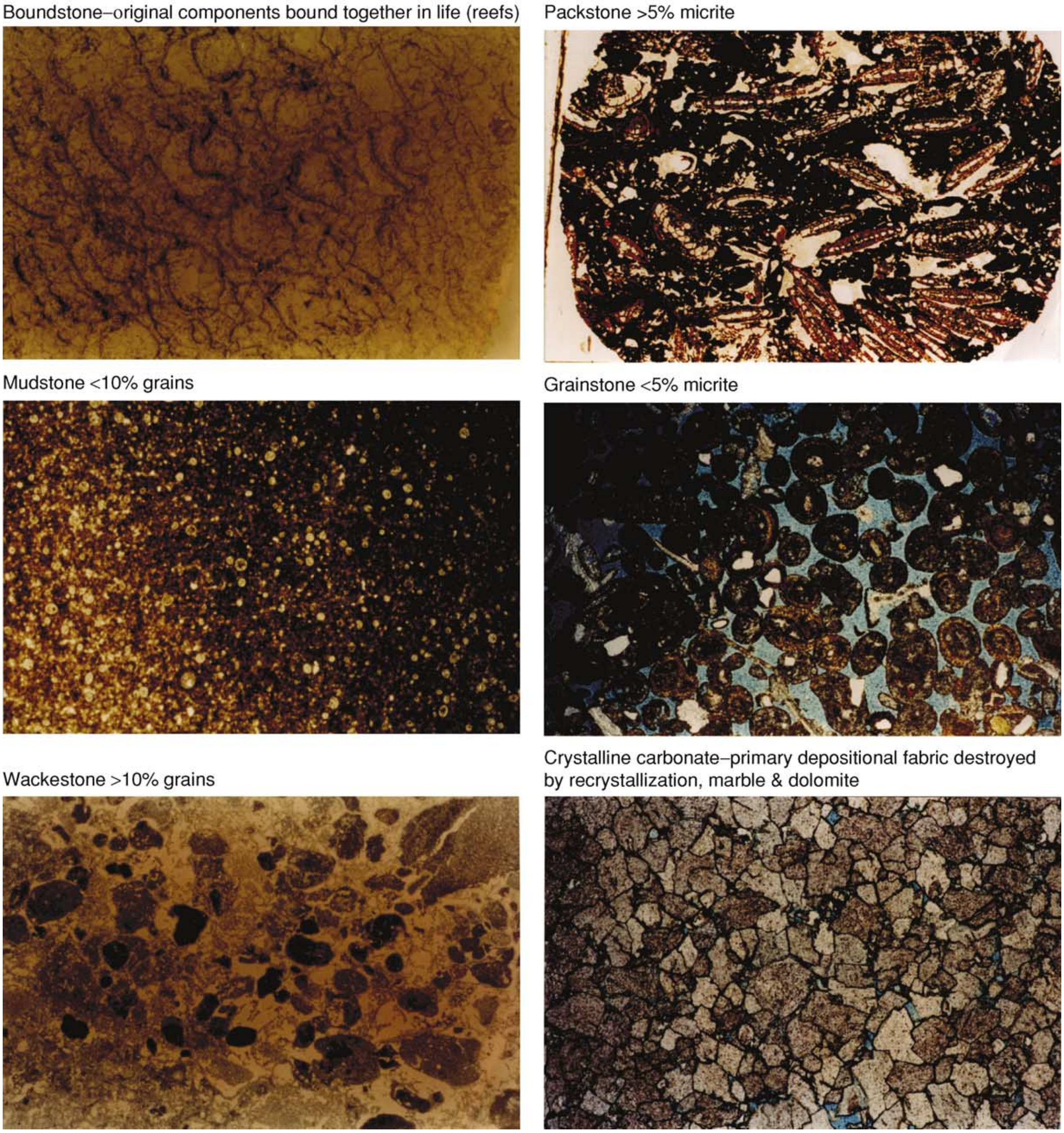
Limestones are principally composed of calcium carbonate in the form of calcite ($CaCO_3$). They may also contain several other carbonate minerals, listed in Table 1, and several non-carbonate impurities. There are two varieties of calcium carbonate ($CaCO_3$): 'aragonite', which has an orthorhombic crystal system, and 'calcite', which has a hexagonal crystal system. Aragonite is an important component of carbonate mud and of many shells. It is, however, relatively unstable in the subsurface, and soon goes into solution, often generating mouldic porosity, either at the surface or during shallow burial. It is very rarely preserved in old and/or deeply buried limestones. Calcite also occurs in many shells and other carbonate grains. It is more stable than aragonite. 'Dolomite' ($CaMg(CO_3)_2$) is the third and most important mineral associated with limestones (described in detail in (*see* **Sedimentary Rocks:** Dolomites)). It rarely forms on the Earth's surface, but commonly does in the subsurface. With increasing abundance of dolomite, limestones grade, via dolomitic limestones, into limey dolomites, and finally dolomite rock. Geopedants restrict the name dolomite to the mineral, and dolostone to the rock. Not everyone is so particular. 'Magnesite', 'ankerite', and 'siderite' are rare constituents of limestones.

Limestone Grains and Matrix

Just like sandstones, limestones consist of framework grains, matrix (syndepositional), cement (postdepositional), and, sometimes, pores. There are many types of carbonate grain. They are briefly described here, and are illustrated in Figure 1. Probably the most common grain type in limestones is shell debris. Indeed, many limestones are made up of nothing but fossils, whole or fragmented. These are termed bioclastic or biogenic limestones. Because of their origin, palaeoecology is an important tool in the diagnosis of the depositional environment. Not only whole fossils but even fragmented bioclasts may be identifiable, and hence of diagnostic value. Some limestones are composed of rounded grains termed ooids, or ooliths, and the rock oolite (named from '*oos*', the Greek for egg, the rock having the appearance of the roe of a fish). Internally, ooids show a concentric growth ring structure around a nucleus of a quartz grain or shell

Table 1 Summary of the minerals commonly associated with limestones

Mineral	*Formula*	*Crystal system*	*Occurrence*
Aragonite	$CaCO_3$	Orthorhombic	Some shells and mud, unstable during burial
Calcite	$CaCO_3$	Hexagonal	Some shells and mud, relatively stable during burial
Magnesite	$MgCO_3$	Hexagonal	Rare surface mineral
Dolomite	$CaMg(CO_3)_2$	Hexagonal	Rarely at the surface, more common as a subsurface replacement
Ankerite	$Ca(MgFe)(CO_3)_2$	Hexagonal	A rare cement
Siderite	$FeCO_3$	Hexagonal	As ooliths and cement



Boundstone–original components bound together in life (reefs)

Packstone >5% micrite

Mudstone <10% grains

Grainstone <5% micrite

Wackestone >10% grains

Crystalline carbonate–primary depositional fabric destroyed by recrystallization, marble & dolomite

Figure 1 Illustrations of carbonate grain types, rock types, and names set within the Dunham classification of limestones. Boundstone: Colonial '*Halysites*' from Wenlock (Silurian) reef, Welsh Marches. Mudstone: Chalk (Upper Cretaceous), Beer, Devon. Wackestone: Rounded pellets and angular intraclasts in a micrite matrix, Marada Formation (Miocene), Jebel Zelten, Libya. Packstone: Foraminifera in a modest micrite matrix (Oligocene), west of Marada Oasis, Libya. Grainstone: Ooids with quartz and shell fragment nuclei. Blue is preserved primary intergranular porosity. Portland Limestone (Upper Jurassic), Dorset. Crystalline carbonate: Dolomite with minor intercrystalline porosity (blue), Zechstein (Upper Permian), UK North Sea. Once upon a time, this was probably a bryozoan reef. All illustrations from Selley RC (2000) *Applied Sedimentology*, 2nd edn. San Diego: Academic Press.

fragment. Ooids form in shallow, high-energy marine environments with elevated temperatures and salinity, where carbonate precipitates episodically around an agitated nucleus. Larger sized concentric carbonate grains are known as pisoliths and oncolites; these are algally coated clasts. Some limestones are composed of structureless, bullet-shaped, sand-sized grains of lime mud and comminuted shell fragments. These grains are faecal pellets, the excreta of diverse burrowing aquatic creatures. Bizarre as it may seem, whole rock formations are composed of such material. Faecal pellets are the characteristic grain type of inner shelves, sheltered bays, and lagoons. Intraclasts are irregular, generally platy-shaped, carbonate grains of various lengths and size. They are formed by the penecontemporaneous erosion of lithified carbonate sediment. Intraclasts are typically found in continental shelf and slope environments.

Between the framework grains briefly described above, there may be a finer grained syndepositional matrix. In sandstones, this is generally composed of clay minerals. In limestones, the matrix is more usually composed of lime mud, termed micrite. Micrite is sometimes aragonitic, sometimes calcitic. Micrite has several origins. It forms when calcareous algae decompose to liberate skeletal aragonite needles into the water. Waves and tidal currents, together with shell-munching predators, also play a part in disaggregating structured shells into comminuted lime mud. There is some evidence in modern warm shallow seas for the direct precipitation of aragonite mud in seawater. Limestone is commonly cemented by calcite, referred to as 'spar' or 'sparite' in this context. Several other carbonate and evaporite minerals precipitate out in limestone pore spaces as postdepositional cement.

Limestone Classification and Nomenclature

There are several different classifications of carbonate rocks. The one most widely used was proposed by Dunham in 1962 (Figure 1) and is briefly described below. 'Boundstone' is the term applied to limestone formed from organic skeletal material that grew bound together at the Earth's surface: in other words, reef rock. 'Mudstone' is composed of micrite with less than 10% grains; 'wackestone' is composed of micrite with over 10% grains. Both mudstone and wackestone are mud supported. That is to say the grains appear to 'float' within the micrite. In contrast, 'packstone' is grain supported, and the space between the grains is partly or completely filled with micrite matrix. 'Grainstone' is grain supported with negligible micrite matrix. The sequence mudstone–wackestone–packstone–grainstone reflects increasing depositional turbulence and energy, and is therefore useful in palaeoenvironmental reconstruction. Dunham's rock names can be qualified by grain type. For example, faecal wackestone, bioclastic packstone, ooidal grainstone, and so forth. The last rock name in Dunham's classification is 'crystalline carbonate', which would normally include dolomite and marble.

Limestone Depositional Environments

All carbonate sediment is precipitated by organic processes, either directly, as animals and plants secrete lime skeletons, or indirectly, as biochemical changes in water cause carbonate to precipitate as individual crystals. Except in a few deep marine environments, all ecosystems are based on plants, and all plants require sunlight to photosynthesize and grow. Plants provide the food for higher life forms to develop. Thus carbonate precipitation, caused or aided by plants, occurs in shallow water, and most of it takes place on the seafloor. Carbonate skeletal development decreases with increasing water depth, as darkness inhibits photosynthesis. Over time, therefore, a carbonate shelf will develop on a gently sloping seafloor (Figure 2). If sea-level remains constant, this shelf will gradually build out or prograde into deeper water. In certain situations, this gently sloping ramp may have an abrupt break in slope. This may occur in one of two ways. A fault may downthrow the seabed into deeper water. Rapid deepening may also occur if sea-level drops, erodes a sea cliff, and rises again, whereupon the rim will be oversteepened by rapid carbonate growth on the crest of the drowned sea cliff. These processes give rise to two types of carbonate setting: the gently sloping accretionary ramp, and the rimmed carbonate platform (Figure 3). This figure also shows the grain types and textures of the carbonate lithologies in these settings, and elegantly illustrates how carbonate rock type correlates with depositional environment. Considered in more detail, the following range of carbonate sediments may be found in sequence from the deep basin across the shelf towards the land. Basinal lime mud may form from the settling of organic detritus of plant and animal plankton that drifted near the surface. In this manner, many lime mudstones, including chalk, formed. These basinal muds may be interbedded with shallow-water carbonate sediment that was transported downslope as turbidity flows, submarine debris flows, and slides. Such transported carbonates, referred to sometimes as 're-deposited' or 'allodapic' limestones, are particularly common on the steep flanks of rimmed platforms and reefs. In warm, clear, shallow water, organic reefs may form by the *in situ* growth of corals,

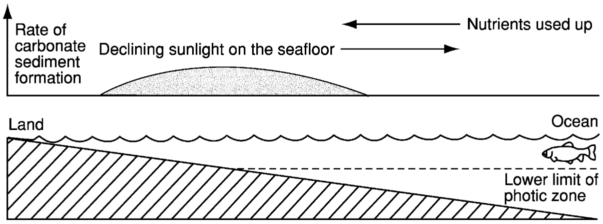

Rate of carbonate sediment formation
Nutrients used up
Declining sunlight on the seafloor
Land
Ocean
Lower limit of photic zone

Figure 2 Diagram to show how carbonate sediment forms optimally in a zone between deep water, where the seafloor is too dark for photosynthesis to occur, and shallow water, where all the nutrients from the open sea have been used up. If sea-level remains constant, the carbonate factory will gradually accrete seaward across the shelf into deeper water.

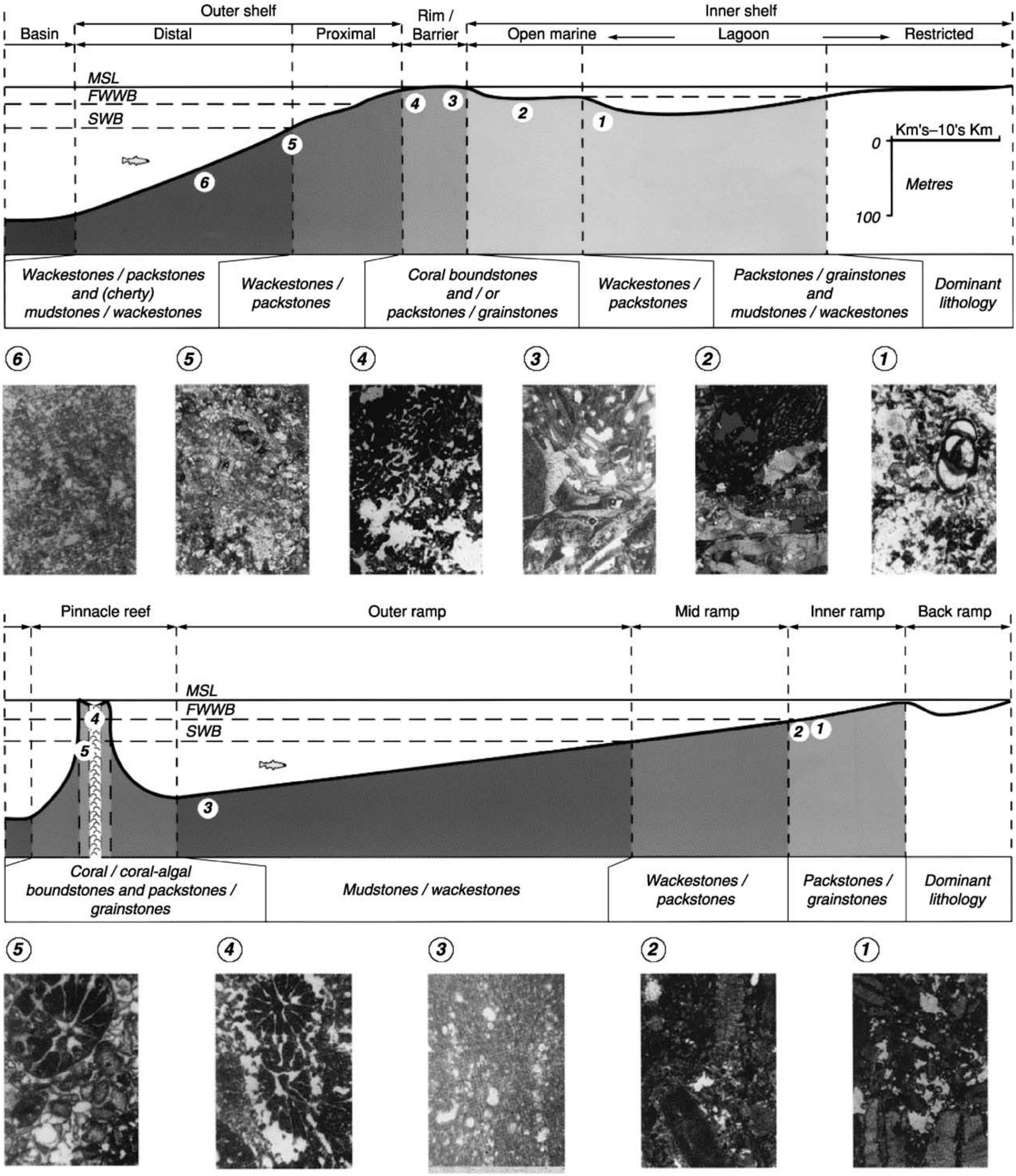


Figure 3 Cross-sections to illustrate the correlation between depositional environments and carbonate rock types (grains and textures) for a rimmed carbonate platform (top) and a carbonate ramp (bottom). Reproduced with permission from Spring D and Hansen OP (1998) The influence of platform morphology and sea level on a carbonate sequence: the Harash Formation, Eastern Sirte Basin, Libya. In: McGregor DS, Moody RTJ, and Clark-Lowes DD (eds.). *Special Publication of the Geological Society of London 132*, pp. 335–353. London: Geological Society of London.

bryozoa, algae, and many other sedentary biota. In turbulent conditions, shoals of oolitic and skeletal grainstone may form, as seen in the modern carbonate banks of the Bahamas, as described in more detail in (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves). In sheltered lagoons behind the high-energy environments of reefs and shoals, burrowing marine animals may excrete faecal pellets to deposit thick formations of peloidal packstones and wackestones. In arid climates, these sediments may, in turn, pass into 'sabkha' ('*sabkha*' is Arabic for salt marsh) where dolomite and evaporite minerals may form. In humid climates, where terrigenous sediment runs off from the land, the carbonate lagoons may interfinger with siliciclastic sand and mud.

The depositional environments of carbonate shorelines and shelves in general, and reefs in particular, are described in greater detail in (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups')), respectively.

Limestone Diagenesis

As noted earlier, the minerals that form limestones are far less stable in the subsurface than are those that form sandstones. Recent carbonate sediment at the Earth's surface is composed of the two isomorphs of calcium carbonate: aragonite and calcite. Recent lime mud is largely aragonitic, but skeletal material is composed of both varieties, which vary in importance between different animal and plant groups.

The change of unconsolidated lime sediment into limestone happens very quickly, and with negligible burial. The 'fossilized' beer bottles and other anthropogenic detritus found in modern 'beach rock' prove this. These early cements are of both calcite and aragonite. In skeletal sands, one of the first diagenetic reactions is the dissolution of aragonite shells. This generates biomoldic porosity.

During burial, aragonitic muds undergo a reordering of the crystal lattice to form calcite. This change is concomitant with a volumetric increase of 8%, and a corresponding loss of porosity. This is why most ancient lime mudstones, certainly those of pre-Mesozoic age, are normally hard, tight, splintery rocks. By contrast, many Cretaceous and younger lime mudstones are light, porous, and chalky. Chalk consists mainly of the fossils of planktonic algae, termed the Coccolithophoridae, together with their disaggregated skeletal plates, termed coccoliths, coccolith-rich faecal pellets, calcispheres, and unicellular planktonic foraminifers. Coccoliths are not composed of unstable aragonite, but of the stabler calcite. Thus, during burial, these lime muds do not undergo expansive diagenesis like aragonitic muds. They maintain their chalky texture, being highly porous, but normally impermeable unless fractured. Chalks are described in greater detail in (*see* **Sedimentary Rocks:** Chalk).

Returning to the diagenesis of carbonate sands, during shallow burial, early cementation may destroy some porosity, but aragonite dissolution may enhance it. With continued burial, calcite cement may infill both biomolds and any remaining intergranular porosity. There are, however, several other diagenetic processes to which a cemented limestone may be subjected.

Limestones may undergo recrystallization, during which some or all of the primary fabric may be destroyed. Individual carbonate grains, generally bioclasts or ooids, may undergo pressure solution. This is a process whereby dissolution occurs at grain contacts due to overburden pressure. Concomitantly, the dissolved mineral matter may be precipitated as cement in adjacent pores. Additional evidence of dissolution is provided by stylolites. These are sutured surfaces, generally subparallel to bedding, where extensive dissolution has left an insoluble residue of clay, kerogen, and other matter along the suture. Stylolites occur in both pure limestones and quartzose sandstones.

Limestone diagenesis must not be thought of as a 'one-way street' that leads to the total loss of porosity and permeability. Limestones may be flushed through with acidic pore fluids, whose leaching properties may generate secondary porosity and permeability. The acidic fluids may come from adjacent compacting clay beds, conveniently generating secondary porosity ahead of petroleum invasion. More usually, however, secondary solution porosity is the result of uplift and erosion, and the flushing of limestone by acidic meteoric water (there is nothing new in acid rain). Solution may form moldic and vuggy pores. It may enlarge fractures and, in extreme cases, develop karstic caverns with concomitant collapse breccias (*see* **Sedimentary Processes:** Karst and Palaeokarst). Many of the best carbonate petroleum reservoirs occur where solution porosity has been developed and preserved beneath unconformities. The best way of preserving porosity in a limestone is for petroleum invasion to occur and expel cementing connate fluids. Renewed burial, without the benefit of petroleum invasion, may, of course, result in total recementation of the limestone as it makes its way to a completely cemented and recrystallized rock, termed marble.

The last important diagenetic process to which limestones are subjected is dolomitization, a process of such complexity and importance that it merits an article to itself (*see* **Sedimentary Rocks:** Dolomites).

Economic Importance of Limestones

Limestones are of great economic importance for many reasons. First, limestones contain lime, an essential ingredient for plant growth, and so limestone quarries are ubiquitous adjacent to farmland with lime-poor acid soil. Hard cemented limestones make excellent building stone and aggregate. Porous and permeable limestones, by contrast, serve as aquifers.

Limestone is used in the manufacture of cement and as a flux in the smelting of iron. Limestones are the host of several metallic minerals, including the eponymous Mississippi Valley telethermal Pb–Zn sulphide ores described in (*see* **Mineral Deposits and Their Genesis**). About 45% of the known petroleum reserves in the world occur in carbonate reservoirs (limestones and dolomites). Six main settings are recognized that preserve large volumes of porous and permeable limestone which have the potential to serve as petroleum reservoirs. These are: oolite grainstone shoals; reefs (often dolomitized); fore-reef talus; grainstone shoals sealed up-dip by evaporites; subunconformity traps, with extensive secondary porosity; and chalk, uplifted and fractured over salt diapirs. Small wonder, then, that limestones, their depositional environments, and diagenesis have been so intensively studied by geologists.

See Also

Building Stone. **Diagenesis, Overview**. **Mineral Deposits and Their Genesis**. **Minerals:** Carbonates. **Sedimentary Environments:** Carbonate Shorelines and Shelves; Reefs ('Build-Ups'). **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Mineralogy and Classification; Chalk; Dolomites.

Further Reading

Dunham RJ (1962) Classification of carbonate rocks according to depositional texture. In: Ham WE (ed.) *Classification of Carbonate Rocks,* American Association of Petroleum Geologists. Tulsa. Ok: pp. 108–121.

Jordan CF and Wilson JL (1994) Carbonate reservoir rocks. *Memoir of the American Association of Petroleum Geologists* No. 60.

Leeder MR (1999) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics.* Oxford: Blackwell Science.

Lucia FJ (1999) *Carbonate Reservoir Characterization.* Berlin: Springer-Verlag.

Reading HG (ed.) (1996) *Sedimentary Environments, Processes, Facies and Stratigraphy,* 3rd edn. Oxford: Blackwell Science.

Selley RC (1996) *Ancient Sedimentary Environments and Their Subsurface Diagnosis,* 4th edn. London: Chapman & Hall.

Selley RC (2000) *Applied Sedimentology,* 2nd edn. San Diego: Academic Press.

Spring D and Hansen OP (1998) The influence of platform morphology and sea level on a carbonate sequence: the Harash Formation, Eastern Sirt Basin, Libya. In: McGregor DS, Moody RTJ, and Clark-Lowes DD (eds.) *Special Publication of the Geological Society of London 132,* pp. 335–353. London: Geological Society of London.

Tucker ME and Wright VP (1990) *Carbonate Sedimentology.* Oxford: Blackwell Scientific Publications.

Oceanic Manganese Deposits

D S Cronan, Imperial College London, London, UK

Introduction

Manganese nodules and encrustations (crusts) together with micronodules are ferromanganese oxide deposits which contain variable amounts of other elements (Table 1). They occur throughout the oceans, although the economically interesting varieties have a much more restricted distribution. Manganese nodules are spherical to oblate in shape and range in size from less than 1 cm in diameter up to 10 cm or more. Most accrete around a nucleus of some sort, usually a volcanic fragment but sometimes biological remains. Crusts are usually tabular.

The deposits were first described in detail in the Challenger Reports. This work was co-authored by J. Murray and A. Renard, who between them initiated the first great ferromanganese oxide controversy. Murray believed the deposits to have been formed by submarine volcanic processes whereas Renard believed that they had precipitated from continental runoff products in seawater. This controversy remained unresolved until it was realized that they could obtain their metals from either or both sources. The evidence for this included the finding of abundant nodules in the Baltic Sea where there are no volcanic influences, and the finding of rapidly grown ferromanganese oxide crusts associated with submarine hydrothermal activity of volcanic origin on the Mid-Atlantic Ridge. Subsequently, a third source of metals to the deposits was discovered, diagenetic remobilization from underlying sediments. Thus, marine ferromanganese oxides can be represented on a triangular diagram (Figure 1), the corners being occupied by hydrothermal (volcanically derived), hydrogenous (seawater derived), and diagenetic (sediment interstitial water derived) constituents.

There appears to be a continuous compositional transition between hydrogenous and diagenetic deposits, all of which are formed relatively slowly at normal deep seafloor temperatures. By contrast,

Table 1 Average abundances of elements in ferromanganese oxide deposits

	Pacific Ocean	*Atlantic Ocean*	*Indian Ocean*	*Southern Ocean*	*World Ocean average*	*Crustal abundance*	*Enrichment factor*	*Shallow marine*	*Lakes*
B	0.0277	—	—	—	—	0.0010	27.7		
Na	2.054	1.88	—	—	1.9409	2.36	0.822	0.81	0.22
Mg	1.710	1.89	—	—	1.8234	2.33	0.782	0.55	0.26
Al	3.060	3.27	2.49	—	2.82	8.23	0.342	1.80	1.16
Si	8.320	9.58	11.40	—	8.624	28.15	0.306	8.76	5.38
P	0.235	0.098	—	—	0.2244	0.105	2.13	0.91	0.15
K	0.753	0.567	—	—	0.6427	2.09	0.307	1.30	0.40
Ca	1.960	2.96	2.37	—	2.47	4.15	0.595	2.40	1.14
Sc	0.00097	—	—	—	—	0.0022	0.441		
Ti	0.674	0.421	0.662	0.640	0.647	0.570	1.14	0.212	0.338
V	0.053	0.053	0.044	0.060	0.0558	0.0135	4.13	0.012	0.001
Cr	0.0013	0.007	0.0029	—	0.0035	0.01	0.35	0.002	0.006
Mn	19.78	15.78	15.10	11.69	16.02	0.095	168.6	11.88	12.61
Fe	11.96	20.78	14.74	15.78	15.55	5.63	2.76	21.67	21.59
Co	0.335	0.318	0.230	0.240	0.284	0.0025	113.6	0.008	0.013
Ni	0.634	0.328	0.464	0.450	0.480	0.0075	64.0	0.014	0.022
Cu	0.392	0.116	0.294	0.210	0.259	0.0055	47.01	0.002	0.003
Zn	0.068	0.084	0.069	0.060	0.078	0.007	11.15	0.011	0.051
Ga	0.001	—	—	—	—	0.0015	0.666		
Sr	0.085	0.093	0.086	0.080	0.0825	0.0375	2.20		
Y	0.031	—	—	—	—	0.0033	9.39	0.002	0.002
Zr	0.052	—	—	0.070	0.0648	0.0165	3.92	0.004	0.045
Mo	0.044	0.049	0.029	0.040	0.0412	0.00015	274.66	0.004	0.003
Pd	0.602^{-6}	0.574^{-6}	0.391^{-6}	—	0.553^{-6}	0.665^{-6}	0.832		
Ag	0.0006	—	—	—	—	0.000007	85.71		
Cd	0.0007	0.0011	—	—	0.00079	0.00002	39.50		
Sn	0.00027	—	—	—	—	0.00002	13.50		
Te	0.0050	—	—	—	—	—	—		
Ba	0.276	0.498	0.182	0.100	0.2012	0.0425	4.73	0.287	0.910
La	0.016	—	—	—	—	0.0030	5.33		0.027
Yb	0.0031	—	—	—	—	0.0003	10.33		
W	0.006	—	—	—	—	0.00015	40.00		
Ir	0.939^{-6}	0.932^{-6}	—	—	0.935^{-6}	0.132^{-7}	70.83		
Au	0.266^{-6}	0.302^{-6}	0.811^{-7}	—	0.248^{-6}	0.400^{-6}	0.62		
Hg	0.82^{-4}	0.16^{-4}	0.15^{-6}	—	0.50^{-4}	0.80^{-5}	6.25		
Tl	0.017	0.0077	0.010	—	0.0129	0.000045	286.66		
Pb	0.0846	0.127	0.093	—	0.090	0.00125	72.72	0.002	0.063
Bi	0.0006	0.0005	0.0014	—	0.0008	0.000017	47.05		

Note: Superscript numbers denote powers of ten, e.g., $^{-6} = \times 10^{-6}$.
(Reproduced with permission from Cronan (1980).)

although theoretically possible, no continuous compositional gradation has been reported between hydrogenous and hydrothermal deposits, although mixtures of the two do occur. This may be partly because: (i) the growth rates of hydrogenous and hydrothermal deposits are very different with the latter accumulating much more rapidly than the former, leading to the incorporation of only limited amounts of the more slowly accumulating hydrogenous material in them; and (ii) the temperatures of formation of the deposits are different, leading to mineralogical differences between them which can affect their chemical composition. Similarly, a continuous compositional gradation between hydrothermal and diagenetic ferromanganese oxide deposits has not been found, although again this is theoretically possible. However, the depositional conditions with which the respective deposits are associated i.e., high temperature hydrothermal activity in mainly sediment-free elevated volcanic areas on the one hand, and low-temperature accumulation of organic rich sediments in basin areas on the other, would preclude much mixing between the two. Possibly they may occur in sedimented active submarine volcanic areas.

Internal Structure

The main feature of the internal structure of nodules and crusts is concentric or tabular banding which is

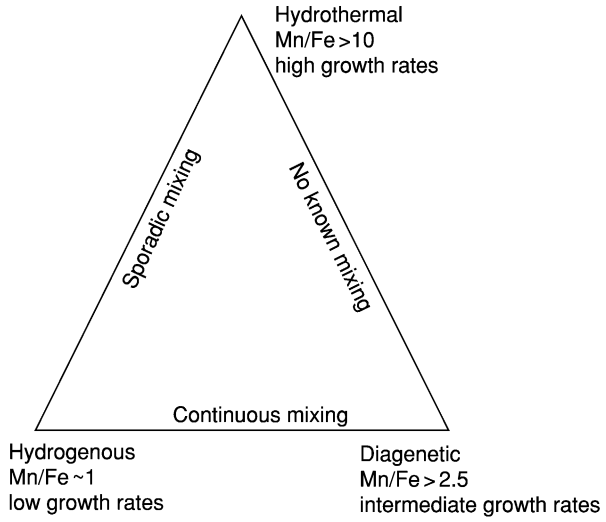


Figure 1 Triangular representation of marine ferromanganese oxide deposits.

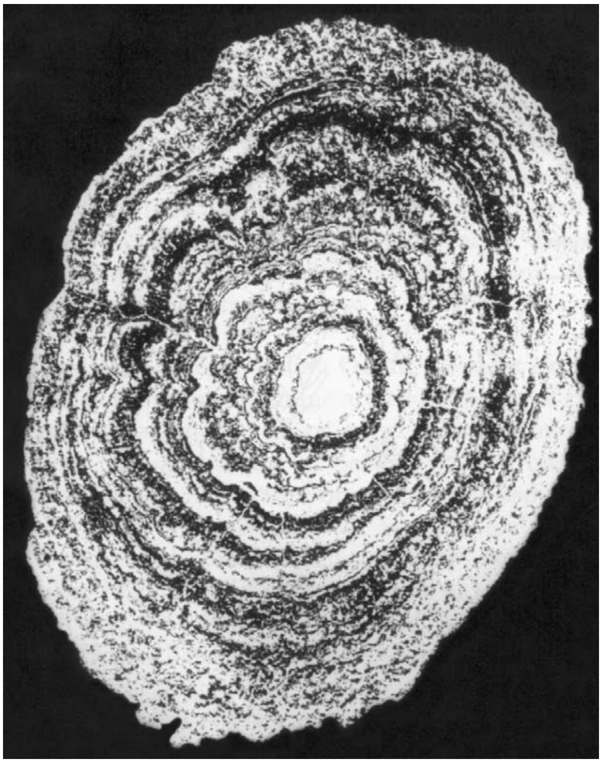


Figure 2 Concentric banding in a manganese nodule. (Reproduced by kind permission of CNEXO, France.)

developed to a greater or lesser extent in most of them (Figure 2). The bands represent thin layers of varying reflectivity in polished section, the more highly reflective layers being generally richer in manganese than the more poorly reflective ones. They are thought to possibly represent varying growth conditions.

On a microscopic scale, a great variety of structures and textures are apparent, some of them indicative of postdepositional alteration of nodule and crust interiors. One of the most commonly observed and most easily recognizable is that of collomorphic globular segregations of ferromanganese oxides on a scale of tenths of a millimetre or less, which often persist throughout much of the nodule or crust interior. Often the segregations become linked into polygons or cusps elongated radially in the direction of growth of the deposits. Several workers have also recognized organic structures within manganese nodules. Furthermore, cracks and fissures of various sorts are a common feature of nodule and crust interiors. Fracturing of nodules is a process which can lead to their breakup on the seafloor, in some cases as a result of the activity of benthic organisms, or of bottom currents. Fracturing is an important process in limiting the overall size of nodules growing under any particular set of conditions.

Growth Rates

It is possible to assess the rate of growth of nodules and crusts either by dating their nuclei, which gives a minimum rate of growth, or by measuring age differences between their different layers. Most radiometric dating techniques indicate a slow growth rate, from a few to a few tens of millimeters per million years. Existing radiometric and other techniques for dating include uranium series disequilibrium methods utilizing ^{230}Th ^{231}Pa, the ^{10}Be method, the K-Ar method, fission track dating of nodule nuclei, and hydration rind dating.

In spite of the overwhelming evidence for slow growth, data have been accumulating from a number of sources which indicate that the growth of nodules may be variable with periods of rapid accumulation being separated by periods of slower, or little or no growth. In general, the most important factor influencing growth rate is likely to be the rate at which elements are supplied to the deposits, diagenetic sources generally supplying elements at a faster rate than hydrogenous sources (Figure 1). Furthermore, the tops, bottoms, and sides of nodules do not necessarily accumulate elements at the same rate, leading to the formation of asymmetric nodules in certain circumstances (Figure 3). Differences in the surface morphology between the tops, bottoms, and sides of nodules *in situ* may also be partly related to growth rate differences. The tops receive slowly accumulating elements hydrogenously supplied from seawater and are smooth, whereas the bottoms receive more rapidly accumulating elements diagenetically supplied from the interstitial waters of the sediments

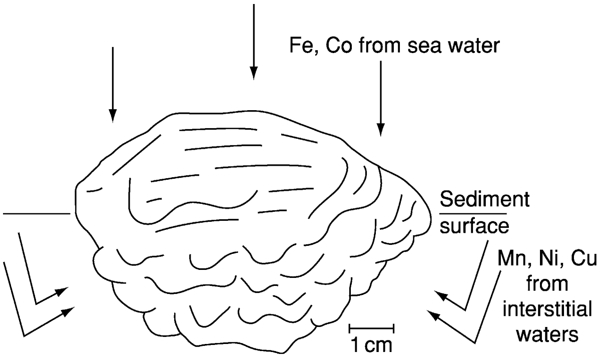


Figure 3 Morphological and compositional differences between the top and bottom of a Pacific nodule. (Reproduced with permission from Cronan, 1980.)

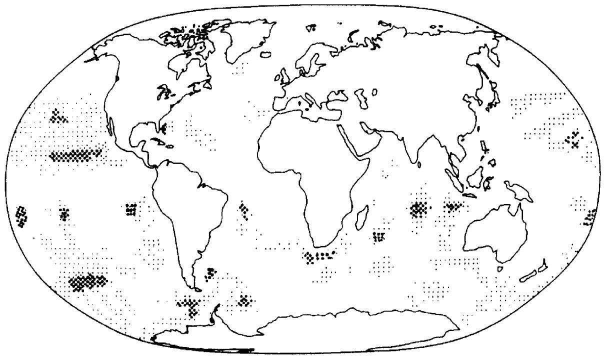


Figure 4 Distribution of manganese nodules in the oceans (updated from Cronan, 1980 after various sources.) ⁚⁚⁚, Areas of nodule coverage; ●●●, areas where nodules are locally abundant.

and are rough (Figure 3). The 'equatorial bulges' at the sediment-water interface on some nodules have a greater abundance of organisms on them than elsewhere on the nodule surface, suggesting that the bulges may be due to rapid growth promoted by the organisms.

It is evident, therefore, that growth cannot be regarded as being continuous or regular. Nodules and crusts may accrete material at different rates at different times and on different surfaces. They may also be completely buried for periods of time during which it is possible that they may grow from interstitial waters at rates different from those while on the surface, or possibly not grow at all for some periods. Some even undergo dissolution, as occurs in the Peru Basin where some nodules get buried in suboxic to reducing sediments.

Distribution of Manganese Nodules

The distribution and abundance of manganese nodules is very variable on an oceanwide basis, and can also be highly variable on a scale of a kilometre or less. Nevertheless, there are certain regional regularities in average nodule abundance that permit some broad areas of the oceans to be categorized as containing abundant nodules, and others containing few nodules (Figure 4), although it should always be borne in mind that within these regions local variations in nodule abundance do occur.

The distribution of nodules on the seafloor is a function of a variety of factors which include the presence of nucleating agents and/or the nature and age of the substrate, the proximity of sources of elements, sedimentation rates, and the influence of organisms. The presence of potential nuclei on the seafloor is of prime importance in determining nodule distribution. As most nodule nuclei are volcanic in origin, patterns of volcanic activity and the subsequent dispersal of volcanic materials have an important influence on where and in what amounts nodules occur. Other materials can also be important as nodule nuclei. Biogenic debris, such as sharks' teeth, can be locally abundant in areas of slow sedimentation and their distribution will in time influence the abundance of nodules in such areas.

As most nuclei are subject to replacement with time, old nodules have sometimes completely replaced their nuclei and have fractured, thus providing abundant nodule fragments to serve as fresh nuclei for ferromanganese oxide deposition. In this way, given sufficient time, areas which initially contained only limited nuclei may become covered with nodules.

One of the most important factors affecting nodule abundance on the seafloor is the rate of accumulation of their associated sediments, low sedimentation rates favouring high nodule abundances. Areas of the seafloor where sedimentation is rapid are generally only sparsely covered with nodules. For example, most continental margin areas have sedimentation rates that are too rapid for appreciable nodule development, as do turbidite-floored deep-sea abyssal plains. Low rates of sedimentation can result either from a minimal sediment supply to the seafloor or currents inhibiting its deposition. Large areas in the centres of ocean basins receive minimal sediment input. Under these conditions, substantial accumulation of nodules at the sediment surface is favoured.

Worldwide Nodule Distribution Patterns

Pacific Ocean As shown in Figure 4, nodules are abundant in the Pacific Ocean in a broad area, called the Clarion–Clipperton Zone, between about 6° N and 20° N, extending from approximately 120° W to 160° W. The limits of the area are largely determined by sedimentation rates. Nodules are also locally abundant further west in the Central Pacific Basin.

Sediments in the northern part of the areas of abundant nodules in the North Pacific are red clays with accumulation rates of around 1 mm per thousand years, whereas in the south they are siliceous oozes with accumulation rates of 3 mm per thousand years, or more.

Nodule distribution appears to be more irregular in the South Pacific than in the North Pacific, possibly as a result of the greater topographic and sedimentological diversity of the South Pacific. The nodules are most abundant in basin environments, such as those of the south-western Pacific Basin, Peru Basin, Tiki Basin, Penrhyn Basin, and the Circum-Antarctic area.

Indian Ocean In the Indian Ocean the most extensive areas of nodule coverage are to the south of the equator. Few nodules have been recorded in the Arabian Sea or the Bay of Bengal, most probably because of the high rates of terrigenous sediment input in these regions from the south Asian rivers. The equatorial zone is also largely devoid of nodules. High nodule concentrations have been recorded in parts of the Crozet Basin, in the Central Indian Ocean Basin, and in the Wharton Basin.

Atlantic Ocean Nodule abundance in the Atlantic Ocean appears to be more limited than in the Pacific or Indian Oceans, probably as a result of its relatively high sedimentation rates. Another feature which inhibits nodule abundance in the Atlantic is that much of the seafloor is above the calcium carbonate compensation depth (CCD). The areas of the Atlantic where nodules do occur in appreciable amounts are those where sedimentation is low. The deep water basins on either side of the Mid-Atlantic Ridge which are below the CCD and which accumulate only limited sediment, contain nodules in reasonable abundance, particularly in the western Atlantic. Similarly, there is a widespread occurrence of nodules and encrustations in the Drake Passage-Scotia Sea area, probably due to the strong bottom currents under the Circum-Antarctic Current inhibiting sediment deposition in this region. Abundant nodule deposits on the Blake Plateau can also be related to strong bottom currents.

Buried nodules Most workers on the subject agree that the preferential concentration of nodules at the sediment surface is due to the activity of benthic organisms which can slightly move the nodules. Buried nodules have, however, been found in all the oceans of the world. Their abundance is highly variable, but it is possible that it may not be entirely random. Buried nodules recovered in large diameter cores are sometimes concentrated in distinct layers. These layers may represent ancient erosion surfaces or surfaces of non-deposition on which manganese nodules were concentrated in the past. By contrast, in the Peru Basin, large asymmetrical nodules get buried when their bottoms get stuck in tenacious sediment just below the surface layer.

Compositional Variability of Manganese Nodules

Manganese nodules exhibit a continuous mixing from diagenetic end-members which contain the mineral 10Å manganite (todorokite) and are enriched in Mn, Ni, and Cu, to hydrogenous end-members which contain the mineral δMnO_2 (vernadite) and are enriched in Fe and Co. The diagenetic deposits derive their metals at least in part from the recycling through the sediment interstitial waters of elements originally contained in organic phases on their decay and dissolution in the sediments, whereas the hydrogenous deposits receive their metals from normal seawater or diagenetically unenriched interstitial waters. Potentially ore-grade manganese nodules of resource interest fall near the diagenetic end-member in composition. These are nodules that are variably enriched in Ni and Cu, up to a maximum of about 3.0% combined.

One of the most striking features shown by chemical data on nodules are enrichments of many elements over and above their normal crustal abundances (Table 1). Some elements such as Mn, Co, Mo, and Tl are concentrated about 100-fold or more: Ni, Ag, Ir, and Pb are concentrated from about 50- to 100-fold; B, Cu, Zn, Cd, Yb, W, and Bi from about 10- to 50-fold; and P, V, Fe, Sr, Y, Zr, Ba, La, and Hg up to about 10-fold, above crustal abundances.

Regional Compositional Variability

Pacific Ocean In the Pacific, potentially ore-grade nodules are generally confined to two zones running roughly east–west in the tropical regions, which are well separated in the eastern Pacific but which converge at about 170°–180° W (Figure 5). They follow the isolines of intermediate biological productivity, strongly suggestive of a biological control on their distribution. Within these zones, the nodules preferentially occupy basin areas near or below the CCD. Thus, they are found in the Peru Basin, Tiki Basin, Penrhyn Basin, Nova Canton Trough area, Central Pacific Basin, and Clarion–Clipperton Zone (Figure 5). Nodules in all these areas have features in common and are thought to have attained their distinctive composition by similar processes.

The potentially ore-grade manganese nodule field in the Peru Basin, centred at about 7°–8° S and 90° W

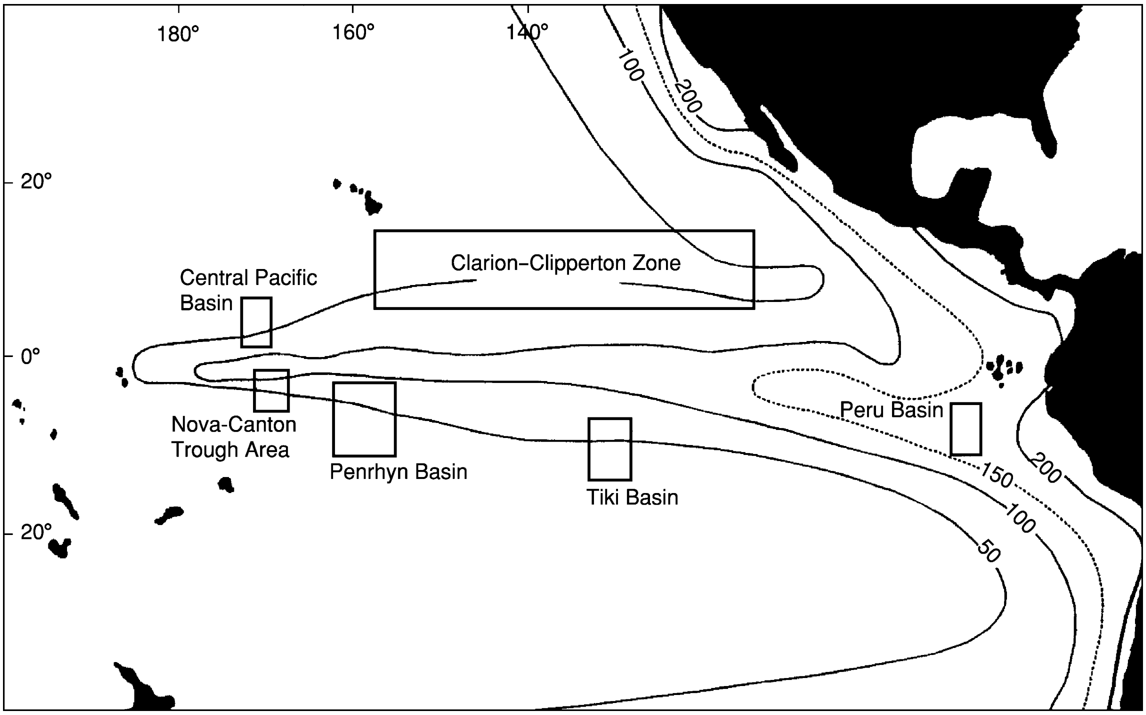


Figure 5 Approximate limits of areas of nickel- and copper-rich nodules in the subequatorial Pacific referred to in the text (productivity isolines in g Cm^{-2} y^{-1}).

(Figure 5), is situated under the southern flank of the equatorial zone of high biological productivity on a seafloor composed of pelagic brown mud with variable amounts of siliceous and calcareous remains. Nodules from near the CCD at around 4250 m are characterized by diagenetic growth and are enriched in Mn, Ni, and Cu, whereas those from shallower depth are characterized mainly by hydrogenous growth. The Mn/Fe ratio increases from south to north as productivity increases, whereas the Ni and Cu contents reach maximum values in the middle of the area where Mn/Fe ratios are about 5.

In the Tiki Basin there is also an increase in the Mn/Fe ratio of the nodules from south to north. All Ni + Cu values are above the lower limit expected in diagenetically supplied material.

The Penrhyn Basin nodules fall compositionally within the lower and middle parts of the Mn/Fe range for Pacific nodules as a whole. However, nodules from the northern part of the Basin have the highest Mn/Fe ratios and highest Mn, Ni, and Cu concentrations, reflecting diagenetic supply of metals to them, although Ni and Cu decrease slightly near the equator. Superimposed on this trend are variations in nodule composition with their distance above or below the CCD. In the Mn-, Ni-, and Cu-rich nodule area, maximum values of these metals in nodules occur within about 200 m above and below the CCD. The latititudinal variation in Mn, Ni, and Cu in Penrhyn Basin nodules may be due to there being a hydrogenous source of these metals throughout the Basin, superimposed on which is a diagenetic source of them between about 2° and 6° S at depths near the CCD, but less so in the very north of the Basin (0–2° S) where siliceous sedimentation prevails under highest productivity waters.

In the Nova Canton Trough area, manganese concentrations in the nodules are at a maximum between the equator and 2.5° S, where the Mn/Fe ratio is also highest. Manganese shows a tendency to decrease towards the south. Nickel and copper show similar trends to Mn, with maximum values of these elements being centred just south of the equator at depths of 5300–5500 m, just below the CCD.

In the central part of the Central Pacific Basin, between the Magellan Trough and the Nova Canton Trough, diagenetic nodules are found associated with siliceous ooze and clay sedimentation below the CCD. Their Ni and Cu contents increase south-eastwards, reaching a maximum at about 2.5°–3° N and then decrease again towards the equator where productivity is highest.

The Clarion–Clipperton Zone deposits rest largely on slowly accumulated siliceous ooze and pelagic clay below the CCD. The axis of highest average Mn/Fe ratio and Mn, Ni, and Cu concentrations runs roughly south-west–north-east, with values of these elements decreasing both to the north and south as productivity declines respectively to the north and increases towards the equatorial maximum in the south.

Indian Ocean In the Indian Ocean, Mn-, Ni-, and Cu-rich nodules are present in the Central Indian Ocean Basin between about 5° and 15° S. They are largely diagenetic in origin and rest on siliceous sediments below the CCD under high productivity waters. The deposits show north–south compositional variability with the highest grades occurring in the north.

Atlantic Ocean In the Atlantic Ocean, diagenetic Mn-, Ni-, and Cu-rich nodules occur most notably in the Angola Basin and to a lesser extent in the Cape/Agulhas Basin and the East Georgia Basin. These three areas have in common elevated biological productivity and elevated organic carbon contents in their sediments which, coupled with their depth near or below the CCD, would help to explain the composition of their nodules. However, Ni and Cu contents are lower in them than in areas of diagenetic nodules in the Pacific and Indian Oceans.

Distribution and Compositional Variability of Ferromanganese Oxide Crusts

Crusts generally accumulate on sediment free hard rock substrates, and thus their regional distribution is related to that of seamounts, plateaux, and other sediment free areas. In a major study on crusts by Hein *et al.* (2000) it is pointed out that the main substrata on which crusts form include basalt, phosphorite, and limestone. However, other than serving as nucleating surfaces for precipitation to occur, the substrata do not contribute to the formation or the composition of the crusts to any significant degree.

Ferromanganese oxide crusts (excluding hydrothermal ones) are generally less variable in composition than manganese nodules. In a large-scale study on crusts in the South Pacific, Verlaan *et al.* (in press) have shown that over the depth range from which the analysed crusts were sampled (650–5853 m), Co, Mn, and Ni increase as depth decreases, while Fe and Cu increase as depth increases. However, the relationship between crust composition and depth may be more complex than this, as analysis of crust composition versus depth in 500 m depth intervals, shows that in certain intervals the correlations between individual elements and depth differ from their overall correlations with depth. These differences are mainly found between three depth segments, above 1500 m (shallow), 1500–3000 m (middle), and below 3000 m (deep). Particularly noteable are the relationships (or lack of them) between elements in crusts and depth in the shallow segment in comparison with those in the deeper segments. There is an absence of any correlation with depth in the shallow segment for Co and Cu, and there is an opposite correlation with depth in the shallow segment for Mn compared with that in the middle and deep segments. Also noteable is the disappearance in the deep segment of any depth correlation for Ni and Fe, and the weakening of the correlation between Cu and depth.

Investigations on the regional variability in crust composition in the South Pacific by Verlaan *et al.* (in press), show that Co increases overall towards the equator. Manganese also increases from south to north and is generally low south of the 12th parallel. Nickel likewise increases northwards towards the equator, while Fe increases to the south-west, away from the equator. Copper shows little regional variation in crusts in the South Pacific. Regionally, Co, Mn, and Ni maintain an opposite behaviour to that of Fe throughout the South Pacific, over the full depth range of the samples collected. Furthermore, the overall equator-ward increase in Co, Mn, and Ni remains evident in each depth segment. The opposite trends in Co, Mn, and Ni enrichment, on the one hand, and Fe enrichment on the other, start from about the 10th parallel, which is the approximate latitude dividing the the high from the low biological productivity regions in the area studied, suggesting that the latitudinal compositional variations in crusts are at least partly productivity influenced. Longitudinally, Co, Mn, and Ni show a tendency to increase to the north-west and Fe towards the south-west, but these variations are much less pronounced than the latitudinal variations.

Economic Potential

Interest in manganese nodules commenced around the mid-1960s and developed during the 1970s, at the same time as the Third United Nations Law of the Sea Conference. However, the outcome of that Conference, in 1982, was widely regarded as unfavourable for the mining industry. This, coupled with a general downturn in metal prices, resulted in a lessening of mining company interest in nodules. About this time, however, several government-backed consortia became interested in them and this work expanded as evaluation of the deposits by mining companies declined. Part 11 of the 1982 Law of the Sea Convention, that part dealing with deep-sea mining, was substantially amended in an agreement on 28 July 1994, which ameliorated some of the provisions relating to deep-sea mining. The Convention entered into force in November 1994.

During the 1980s, interest in manganese nodules and crusts in exclusive economic zones (EEZs) started to increase. An important result of the Third Law of the Sea Conference, was the acceptance of a

200-nautical-mile EEZ in which the adjacent coastal state could claim any mineral deposits as their own. The nodules and crusts found in EEZs are similar to those found in adjacent parts of the International Seabed Area, and are of greatest economic potential in the EEZs of the South Pacific.

At the beginning of the twenty-first century, the outlook for deep sea mining remains rather unclear. It is likely to commence some time in this century, although it is not possible to give a precise estimate as to when. The year 2015 has been suggested as the earliest possible date for nodule mining outside of the EEZs. It is possible, however, that EEZ mining for nodules might commence earlier if conditions were favourable. It would depend upon many factors; economic, technological, and political.

Conclusions

Manganese nodules and crusts, although not being mined today, are a considerable resource for the future. They consist of ferromanganese oxides variably enriched in Ni, Cu, Co, and other metals. They generally accumulate on or around a nucleus and exhibit internal layering on both a macro- and microscale. Growth rates are generally slow. The most potentially economic varieties of the deposits occur in the subequatorial Pacific.

See Also

Mineral Deposits and Their Genesis. **Mining Geology:** Exploration. **Sedimentary Processes:** Deep Water Processes and Deposits. **Sedimentary Rocks:** Deep Ocean Pelagic Oozes.

Further Reading

Cronan DS (1980) *Underwater Minerals.* London: Academic Press.

Cronan DS (1992) *Marine Minerals in Exclusive Economic Zones.* London: Chapman and Hall.

Cronan DS (ed.) (2000) *Handbook of Marine Mineral Deposits.* Boca Raton: CRC Press.

Cronan DS (2000) Origin of manganese nodule ‘ore provinces’. *Proceedings of the 31st International Geological Congress,* Rio de Janero, Brazil, August 2000.

Earney FC (1990) *Marine Mineral Resources.* London: Routledge.

Glasby GP (ed.) (1977) *Marine Manganese Deposits.* Amsterdam: Elsevier.

Halbach P, Friedrich G, and von Stackelberg U (eds.) (1988) *The Manganese Nodule Belt of the Pacific Ocean.* Stuttgart: Enke.

Hein JR, Koschinsky A, Ban M, Manhein FT, Kang J-K, and Roberts L (2000) Cobalt-rich ferromanganese crusts in the Pacific. In: Cronan DS (ed.) *Handbook of Marine Mineral Deposits,* pp. 239–279. Boca Raton: CRC Press.

Nicholson K, Hein J, Buhn B, and Dasgupta S (eds.) (1997) *Manganese Mineralisation: Geochemistry and Mineralogy of Terrestrial and Marine Deposits.* Geological Society Special Publication 119, London.

Roy S (1981) *Manganese Deposits.* London: Academic Press.

Teleki PG, Dobson MR, Moore JR, and von Stackelberg U (eds.) (1987) *Marine Minerals: Advances in Research and Resource Assessment.* Dordrecht: D. Riedel.

Verlaan P, Cronan DS, and Morgan C (in press) A comparative analysis of compositional variations in and between marine ferromanganese nodules and crusts and their environmental controls. *Progress in Oceanography.*

Phosphates

W D Birch, Museum Victoria, Melbourne, VIC, Australia

Introduction

Phosphorus is the tenth most abundant element on Earth and plays a key role in geological and biological processes. In the mineral kingdom, phosphates are amongst the most complex and diverse, with approximately 460 recognized species. Over the past five years about twenty new phosphate minerals have been recognized.

Phosphates are found in diverse geological environments and in many associations or assemblages. In igneous and metamorphic rocks, members of the apatite group, in particular fluorapatite, are the dominant phosphates. Because the solubility of phosphorus is generally low in silicate minerals, fluorapatite and other ubiquitous phosphates such as monazite and xenotime usually occur as accessory minerals. An exception to this general rule is the wide diversity of late-crystallising phosphate minerals found in some granite pegmatites (*see* **Igneous Rocks:** Granite). Other important environments include sedimentary rocks, in which phosphates reach their maximum abundance in the form of phosphorite, and the oxidised zones of sulphide-bearing ore deposits. About ten phosphates, including four not found on Earth, have been recorded from meteorites.

Studies of phosphate minerals are important for scientific, environmental, agricultural, and health reasons. For mineralogists and crystallographers, the crystal chemistry of phosphate minerals embraces novel and diverse structures. To mineral collectors, the better crystallized and more colourful phosphates provide an unlimited source of intriguing and attractive specimens. The geochemistry of many phosphate minerals is important, as they commonly include trace amounts of uranium, thorium, and the rare earth elements. Apatite, monazite, and xenotime are significant for geochronology and thermochronology, utilising the trace amounts of radioactive elements uranium and thorium contained in their crystal structure. Phosphate enrichment of soils and discharge of phosphorus-bearing waste are human processes that have significant environmental impact requiring monitoring and control. As well, calcium phosphates are important constituents of human tissue.

There are thousands of references on phosphate minerals, which means that this review cannot be exhaustive. It deals mainly with the classification of phosphate minerals (*see* **Minerals:** Definition and Classification) and the major geological environments in which they occur.

Classification of Phosphate Minerals

The structures of phosphate minerals are almost exclusively built on the tetrahedral anionic unit $(PO_4)^{3-}$, in which the P atom is central to the four O atoms. P–O bond lengths may vary, leading to distortions in the tetrahedra, but the average distance is 1.537 Å. From bond valence considerations, (PO_4) groups link easily with a range of non-tetrahedrally coordinated cations, such as Al^{3+}, Fe^{3+}, Mg^{2+}, Fe^{2+}, Mn^{2+}, Ca^{2+}, Sr^{2+}, Na^+, and K^+. Many phosphates are also hydrated and/or hydroxyl-bearing. Even though the anionic radius of P^{5+} (0.25 Å) is smaller than that of As^{5+} (0.42 Å) and V^{5+} (0.44 Å), a number of near-ideal solid solutions series are observed between phosphates and arsenates and to a more limited extent between phosphates and vanadates.

The crystal structures of most phosphate minerals have been well characterized, thereby facilitating classification. The simplest schemes divide the phosphates into classes based on whether they are anhydrous or hydrated, contain hydroxyl and/or halogen, or contain another anion such as $(SO_4)^{2-}$, $(CO_3)^{2-}$, $(CrO_4)^{2-}$, $(AsO_4)^{3-}$, and $(VO_4)^{3-}$. The well-known Dana system adopts such an approach, as do James Ferraiolo and Hugo Strunz. Alexander Povarennykh combined this approach within divisions based on crystal chemical features. Perhaps because of the great diversity and complexity shown by phosphate structures, overall classification schemes based on crystal chemistry alone have been attempted by only a few researchers, notably Paul Moore, Frank Hawthorne, and Ivan Kostov. These schemes are generally based on the recognition that the PO_4 tetrahedra can polymerize in a number of ways, leading to a broad three-fold subdivision:

i. Polymerization of TO_4 tetrahedra, where T may be P, Be, Zn, B, Al, and Si
ii. Polymerization of PO_4 tetrahedra and MO_6 octahedra. This grouping covers a very large number of species and, within it phosphates can be further subdivided on the basis of whether polyhedra are unconnected, in finite clusters, or in infinite chains, sheets, or frameworks
iii. Polymerization of PO_4 tetrahedra and polyhedra that contain large cations coordinated by more than 6 oxygen atoms.

In many of these structures, OH and H_2O may provide one or more of the oxygen atoms in the PO_4 and MO_6 groups (for simplicity, this is not always indicated in the terminology used in the following review).

Structures with Polymerized TO_4 Tetrahedra

There are about 30 minerals in this category that have structures based mainly on polymerization of two or three PO_4 tetrahedra, or of PO_4 groups with BeO_4, ZnO_4, and AlO_4 tetrahedra. The structures can be based on finite clusters, such as in the rare isostructural zirconium-bearing species gainesite, $Na_2Zr_2[Be(PO_4)_4]$, mccrillisite, $NaCsZr_2[Be(PO_4)_4]$, and selwynite, $NaKZr_2[Be(PO_4)_4]$, or on infinite chains, sheets, and frameworks of tetrahedra. Examples of minerals with structures based on chains of PO_4–BeO_4 linkages included moraesite, $Be_2(PO_4)(OH)$ and roscherite, $CaMn_3[Be_2(PO_4)_3(OH)_3]$, while spencerite, $Zn_4(PO_4)_2(OH)_2$ involves chains of alternating $ZnO_2(OH)(H_2O)$ and PO_4 tetrahedra. Both PO_4–ZnO_4 and PO_4–BeO_4 linkages are present in the sheet-like structures, which include hopeite, $Zn_3(PO_4)_2 \cdot 4H_2O$, scholzite, $CaZn_2(PO_4)_2 \cdot 2H_2O$,

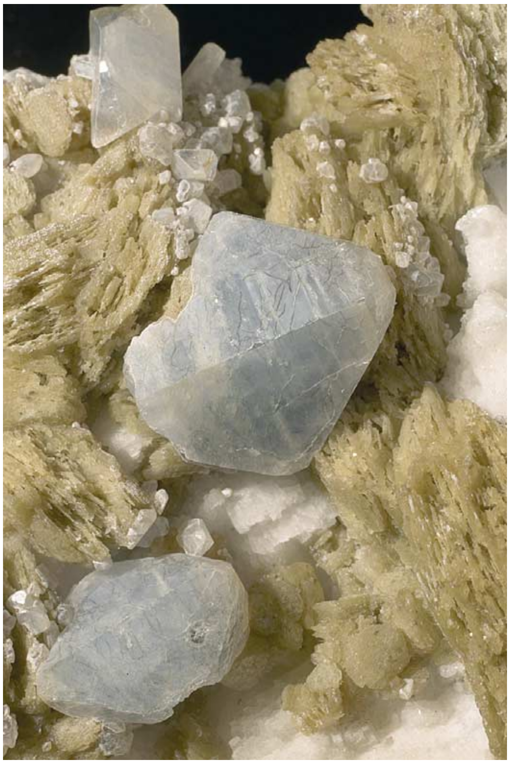


Figure 1 Hydroxylherderite crystals (largest crystal 18 mm long) on muscovite from the Xanda mine, Minãs Gerais, Brazil. Museum Victoria specimen M43389, photography by J Broomfield. Reproduced with permission from Museum Victoria.

the isostructural herderite, $CaBe(PO_4)F$, (Figure 1) and hydroxylherderite, $CaBe(PO_4)OH$. In the infinite framework structures, all except berlinite, $AlPO_4$, which is isostructural with quartz, are based on PO_4–BeO_4 linkages, including beryllonite, $NaBePO_4$ and pahasapaite, $Li_8Ca_8Be_{24}(PO_4)_{24} \cdot 38H_2O$, which has a complex, zeolite-like framework structure.

Structures with Linked TO_4 Tetrahedra and MO_6 Groups

About 180 minerals are known to be represented in this structural grouping, with most being infinite frameworks. Only five minerals are known that have structures based on isolated or finite clusters of tetrahedra and octahedra, linked together by hydrogen bonding. They include struvite, $NH_4Mg(PO_4) \cdot 6H_2O$, anapaite, $Ca_2Fe^{2+}(PO_4)_2 \cdot 4H_2O$, and morinite, $NaCa_2Al_2(PO_4)_2(F,OH)_5 \cdot 2H_2O$. There are twenty-two minerals with structures based on infinite chains of tetrahedra and octahedra. These can be subdivided further into five topologically distinct types, depending on how the TO_4 and MO_6 groups are linked. A group of minerals that includes collinsite, $Ca_2(Mg,Fe)(PO_4)_2 \cdot 2H_2O$, and fairfieldite, $Ca_2(Mn,Fe)(PO_4)_2 \cdot 2H_2O$, consists of chains of alternating $(M^{2+}O_4\{H_2O\}_2)$ octahedra and pairs of (PO_4) tetrahedra, with the chains linked by 7-coordinated Ca atoms and by hydrogen bonding. In childrenite, $(Fe,Mn)Al(PO_4)(OH) \cdot H_2O$, and members of the jahnsite group, $CaMn(Fe,Mn,Mg)_2Fe_2(PO_4)_4(OH)_2 \cdot 8H_2O$, chains are based on corner-sharing octahedra with bridging (PO_4) groups, while in bearthite, $Ca_2Al(PO_4)_2(OH)$, adjacent octahedra share an edge to build chains linked by $(PO_4)_4$ groups and Ca cations.

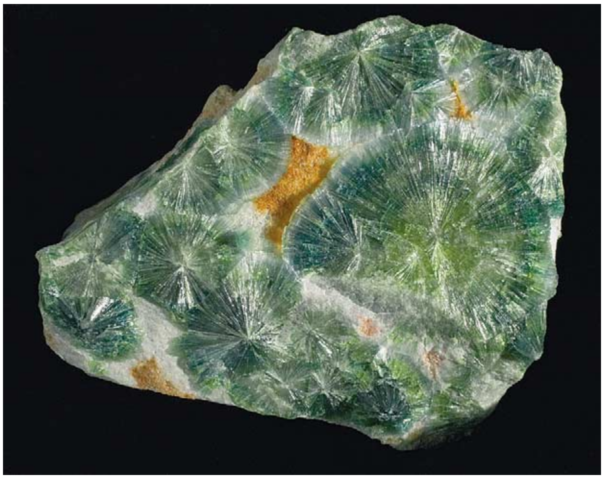


Figure 2 Wavellite sprays (up to 30 mm across) from Montgomery County, Arkansas, USA. Museum Victoria specimen M27840, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

Nearly 50 minerals are known to have structures consisting of infinite sheets of (PO_4) tetrahedra and (MO_6) octahedra. These structures can also be grouped, depending on how the octahedra and tetrahedra are linked. However, due to their complexity, it is not feasible to describe or summarise them here. Some of the more important phosphates with sheet-like structures include members of the crandallite group, based on $[Al_3(PO_4)(PO_3\{OH\})(OH)_6]$, and vivianite, $Fe_3^{2+}(PO_4)_2 \cdot 8H_2O$. About 110 phosphate minerals are known to have framework structures, by far the largest group. Important minerals in this category include wavellite, $Al_3(PO_4)_2(OH,F)_3 \cdot 5H_2O$ (Figure 2), variscite, $AlPO_4 \cdot 2H_2O$, pseudomalachite, libethenite, $Cu_2(PO_4)(OH)$ $Cu_5(PO_4)_2(OH)_4$, members of the turquoise $[CuAl_6(PO_4)_4(OH)_8 \cdot 4H_2O]$ group, the iron phosphates dufrenite, $Fe^{2+}Fe_4^{3+}(PO_4)_3(OH)_5 \cdot 2H_2O$, and rockbridgeite, $(Fe^{2+},Mn)Fe_4^{3+}(PO_4)_3$ $(OH)_5$.

Structures with TO_4 Groups and Large Cations

In the phosphate minerals within this broad grouping, the main cations, which may be either monovalent, divalent, or trivalent, are coordinated to varying numbers of oxygen atoms to form polyhedra, which are then linked in various ways, commonly through (PO_4) tetrahedra, to form chains. Chains are then linked to form sheets, which stack in various crystallographic directions. It is not possible here to describe or summarize individual structures, but some important minerals represented include xenotime, $(Y,Yb)(PO_4)$, and members of the monazite $(REE,Ce,Ca,Th)(PO_4)$ group, as well as a suite of hydroxyl-bearing and hydrated ammonium, sodium, and potassium-bearing species typically found in cave environments. However, the apatite group is the most significant in this structural category and is briefly outlined below. There is also an important and widespread suite of phosphate minerals whose structures are dominated by the uranyl $(U^{6+}O_2)^{2+}$ group (Figures 3 and 4). These include about thirty species in two main groups related to torbernite, $Cu(UO_2)(PO_4)_2 \cdot 10H_2O$, and phosphuranylite, $KCa(H_3O)_3(UO_2)_7(PO_4)_4O_4(UO_2)(PO_4)_2 \cdot 8H_2O$, respectively.

Figure 3 Crystals of meta-autunite, $Ca(UO_2)(PO_4)_2 \cdot 6H_2O$ (up to 7 mm wide) from Autun, Burgundy, France. Museum Victoria specimen M27680, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Figure 4 Crystals of saleeite, $Mg(UO_2)(PO_4)_2 \cdot 8H_2O$ (up to 4 mm across) from the Ranger mine, Northern Territory, Australia. Museum Victoria specimen M45060, photograph by F Coffa. Reproduced with permission from Museum Victoria.

The Apatite structural group The apatite group contains ten species, including pyromorphite, $Pb_5(PO_4)_3Cl$, and belovite, $Sr_3Na(Ce,La)(PO_4)_3(OH)$. However, fluorapatite, $Ca_5(PO_4)_3F$, chlorapatite, $Ca_5(PO_4)_3Cl$, and hydroxylapatite, $Ca_5(PO_4)_3OH$, are the most widespread and influential in geological and biological processes. The essential atomic arrangement for these three species consists of (PO_4) tetrahedra and two Ca polyhedra (Figure 5). Ca1 is coordinated to nine oxygen atoms, and Ca_2 bonds to six oxygen atoms and one anion (F, Cl, or OH) situated in channels running parallel to the *c*-axis. The Ca1 polyhedron shows little response to the effect of different channel anions, whereas in the Ca_2 polyhedron there are significant shifts in the positions of the channel anions, arising from their markedly different sizes. The structure permits a very wide range of substitutions in all cation and anion sites in natural and synthetic apatites. For example, the monovalent ions in the *c*-axis channel sites can be replaced by divalent anions such as $(CO_3)^{2-}$ (eg., in carbonite–fluorapatite) and O^{2-}. Vacancies may also occur in the *c*-axis channels. A large number of divalent cations (for example Pb^{2+}, Ba^{2+}, Mn^{2+}, and Sr^{2+}) can substitute for Ca. The (PO_4) group is commonly replaced by other tetrahedral anion groups, such as $(AsO_4)^{3-}$, $(SO_4)^{2-}$, $(SiO_4)^{4-}$, and $(VO_4)^{3-}$. Apatites may also take up rare earth elements but the mechanisms are complex and beyond the scope of this discussion.

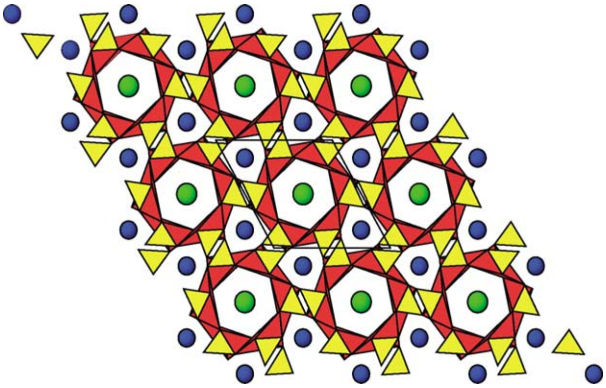


Figure 5 Apatite crystal structural diagram, viewed down the c axes and with the unit cell outlined in black. The PO_4 tetrahedra are yellow, the smaller Ca site is drawn as red polyhedra, the second Ca is shown as blue balls and the (F, Cl, OH) ion is green. Diagram prepared by A Pring.

Geological Environments for Phosphate Mineral Occurrences

Phosphate minerals are found in a wide range of igneous, metamorphic, and sedimentary rocks. In general however, only three minerals—apatite *per se*, monazite, and xenotime—are ubiquitous in typical igneous and metamorphic rocks. With a few exceptions, however, such as granitic pegmatites, alkaline intrusive rocks and some calc-silicate skarns and marbles (Figure 6), these minerals occur as primary accessory phases only. Their abundance is a general indication of the phosphorus content of the rock, as phosphorus has low solubility in most silicate minerals. Extreme phosphorus enrichment in magmas in represented by carbonatites, in which apatite is the most widespread economic mineral, notably in two deposits, Phalaborwa in South Africa and Khibiny on the Kola Peninsula in Russia.

In sedimentary rocks, phosphate minerals are represented throughout the geological time-scale to the present day, and occur in a wide range of host rocks. Many are hosted by metasediments, in which mobilization and recrystallization of original primary phosphate have yielded diverse assemblages of species.

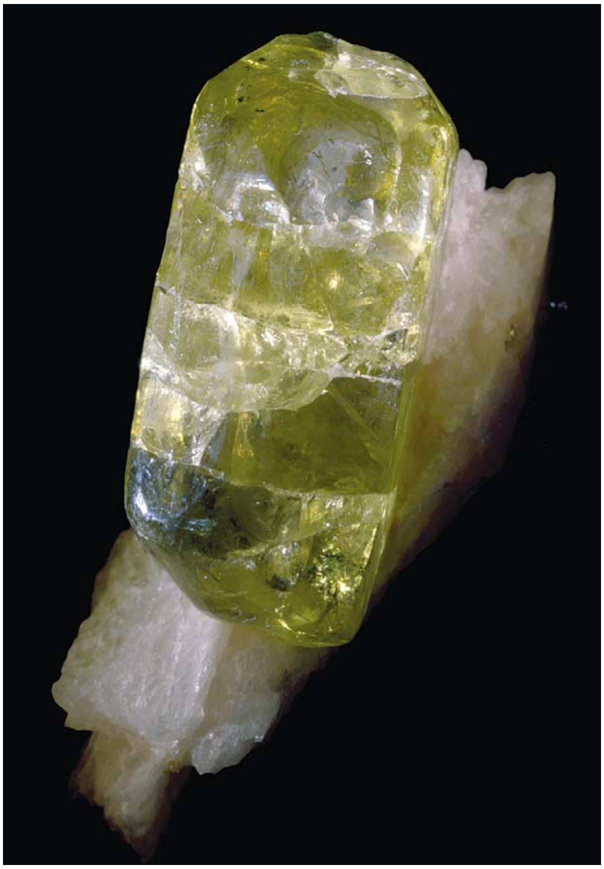


Figure 6 Fluorapatite crystal (2 cm long) in marble from Wilberforce, Ontario, Canada. Museum Victoria specimen M39659, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Granite Pegmatites

Apatite varieties dominate the phosphate suites found in many granite pegmatites and may crystallize at different stages. For example, primary apatite may occur intergrown with feldspar, quartz, and tourmaline, or it may crystallize later as a druse or miarolitic-cavity mineral during the hydrothermal stage (Figure 7). At lower temperatures, crusts of carbonate-bearing apatite may form.

Paul Moore has constructed a ‘paragenetic tree’ of pegmatite phosphates. Primary phosphates are found as giant crystals or lenticular masses which crystallized near the core of the pegmatite, usually embedded in massive quartz. As well as apatite, the triphylite–lithiophilite, triplite–zwieselite, and amblygonite–montebrasite series are significant at the primary stage, with the last series in places constituting an ore of lithium. Metasomatic alteration during the final stages of pegmatite formation may result in these primary phases being partially replaced by such species as alluadite, triploidite–wolfeite, and purpurite–heterosite, which may be nodular and fine-grained. Typical reactions during alteration involve Na, K, and

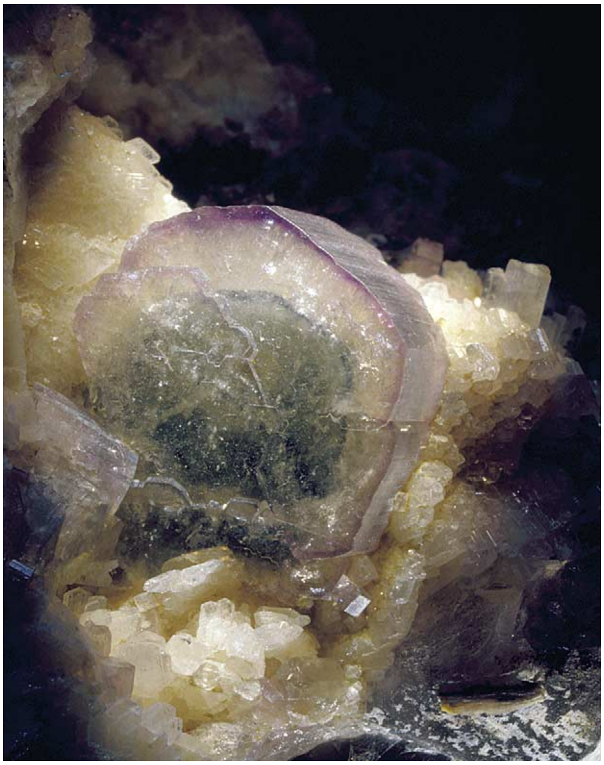


Figure 7 Fluorapatite crystal (12 mm across) showing colour zonation, in granite from Lake Boga, Victoria, Australia. Museum Victoria specimen M29944, photograph by F Coffa. Reproduced with permission from Museum Victoria.

Ca displacing Li, and the addition of OH. The largest diversity of phosphate species in granite pegmatites arises from the oxidation of these primary phases. This may take place under 'hydrothermal' conditions, at temperatures less than 250°C during the cooling history, or much later during atmospheric weathering. The great diversity of these secondary species reflects the mixed valence states of Fe and Mn; the different configurations in which oxygen atoms from H_2O molecules and OH and PO_4^{2-} groups may cluster around these metal cations in octahedral coordination; and the different ways the octahedra can combine (polymerise) to form clusters in solution. Such clusters may be represented in the crystal structures of secondary phosphates. Amongst these are barbosalite, rockbridgeite, beraunite, phosphosiderite, strengite, leucophosphite, bermanite, strunzite, laueite, cacoxenite, cyrilovite, eosphorite–childrenite, and ludlamite (Figure 8), which are derived from Fe- and Mn-bearing primary phosphates. Species such as hurlbutite, herderite, brazilianite, morinite, crandallite, and whitlockite are derived from primary phosphates with low Fe and Mn and/or high Li contents. Many of these secondary species crystallise as small crystals in open cavities, resulting from removal of much of the PO_4^{2-} and most of the alkalis.

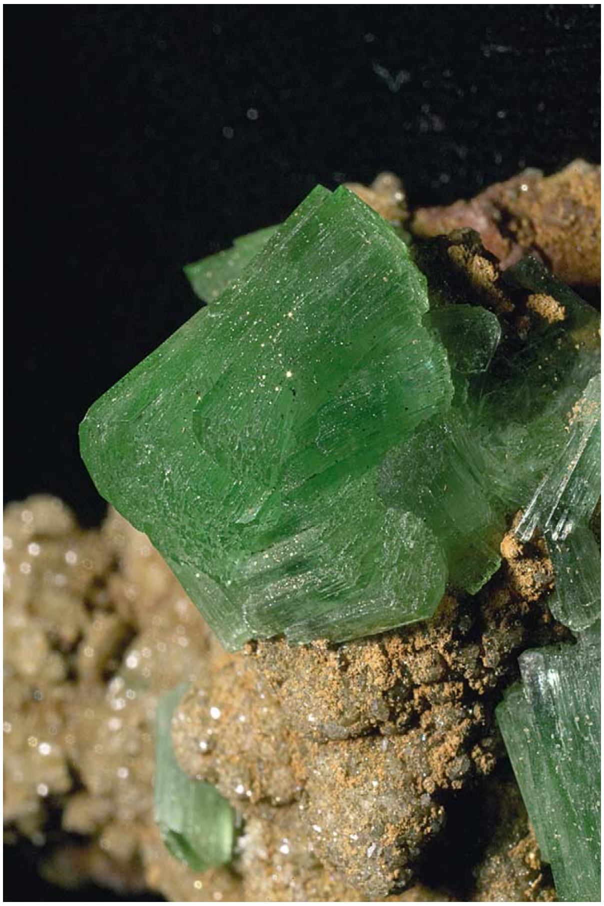


Figure 8 Ludlamite crystal group (12 mm across) from the San Antonio mine, Chihuahua, Mexico. Museum Victoria specimen M37681, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

The best known and most prolific localities for phosphate minerals include the famous pegmatites of the Black Hills, South Dakota; the Palermo mine, New Hampshire; Hagendorf Sud, Bavaria; Tsaobismund, Namibia; Buranga, Rwanda; Viitaniemi, Finland; Sapucaia, Brazil, and occurrences in southern California and Maine in the USA.

Sediment-Hosted Phosphate Deposits

Sedimentary phosphates—phosphorites—are the most important of the world's sources of phosphate rock (*see* **Sedimentary Rocks:** Mineralogy and Classification). They occur on every continent and range in age from Precambrian to Holocene, with nearly all having a marine origin. Modern phosphorites are mineralogically monotonous, consisting of grains of cryptocrystalline or amorphous carbonate-fluorapatite (variously referred to as collophane or francolite), occurring as beds ranging in thickness from a few centimetres up to tens of metres. Other forms of phosphorite include nodules and concretions. Phosphorites are commonly observed in shallow seas, along the edges of continental shelves, and on ocean plateaus. The phosphorus is believed to be derived from faecal matter, bone material, and decaying marine organisms that accumulate locally or are carried into shallow coastal regions by upwelling deep ocean currents. These nutrients encourage a diverse biota to flourish, ultimately producing organic-rich sediments. During early diagenesis, collophane precipitation occurs within the upper layers of these sediments from pore waters rich in phosphorus leached from the organic remains; precipitation is enhanced where phosphatic nuclei are already present. A changing depositional environment with periods of reduced deposition and reworking of sediments in shallow seas favours phosphogenesis. This model is generally applicable to old phosphorites that remain recognizable, such as those of the Cambrian–Ordovician Georgina Basin, in Queensland, Australia. However, settings and methods of deposition (including transport of phosphate grains) vary widely and are subject to debate. Study of the age and global distribution of phosphorites has led to the identification of major phosphogenic episodes as far back as the Proterozoic.

Older phosphorites are more likely to have undergone diagenesis, deformation, and metamorphism, to the extent that the original nature of the deposits may become obscured. Phosphorus may be mobilized in solution and distributed into surrounding rocks, where diverse suites of well-crystallized secondary phosphates may form. Perhaps the most notable such occurrence is in the Cretaceous Rapid Creek Formation in the Canadian Yukon. Here, a marine sequence of ironstone and shale containing unusual Fe- and Mn-bearing phosphates instead of collophane has been deformed and uplifted. This resulted in sets of fractures in which a wealth of well-crystallized phosphate minerals have formed. Four major assemblages characterized by the predominance of specific elements and related to a specific host rock have been identified, with at least five new phosphate species recognized (baricite, garyansellite, gormanite, kulanite, and penikisite). As well, remarkable crystals of arrojadite, augelite, and lazulite, amongst others, occur in this fracture-filling paragenesis.

In south-eastern Australia, a variety of settings for sedimentary phosphate deposits has been recognized, with several producing a range of unusual, in some cases new species. Some deposits have been exploited for phosphate, but all are low grade. The oldest deposits are in South Australia, where there are two main phosphatic horizons, one Late Precambrian, the other Early Cambrian, associated with limestones. There has been local leaching and intermittent concentration of phosphate by replacement. The Moculta deposit has been affected by regional metamorphism, which has recrystallised and brecciated the phosphatic rock. A range of secondary phosphate minerals, such as wavellite, beraunite, cyrilovite, leucophosphite, variscite, crandallite, and aldermanite (for which Moculta is the type locality) has been recorded in veinlets and small cavities and probably formed during near-surface weathering. An intense and prolonged weathering origin can probably be ascribed to a suite of phosphate minerals found in metamorphosed Early Proterozoic iron-rich sediments at Iron Monarch, in the Middleback Ranges. Over thirty phosphate species, including bermanite, collinsite, cyrilovite, fairfieldite, kidwellite, montgomeryite, turquoise, and wavellite, have been identified. A number of vanadates also occur in the assemblage. In central Victoria, small, low-grade phosphate deposits within Ordovician black slate–chert host rocks exhibit a number of mineralization styles, such as phosphorite bands, intraformational breccias, and vein networks. Secondary minerals resulting from weathering of the primary phosphorites include wavellite, turquoise, variscite, cacoxenite, and fluellite.

Amongst the world's largest phosphate deposits are those of Morocco, where Late Cretaceous marine sediments occur on the plains fronting the Atlas Mountains (*see* **Africa:** North African Phanerozoic). These are nodular and sandy deposits riddled with fish teeth and fit well into the upwelling nutrient-rich current model outlined above. Other significant world producers of sedimentary phosphate are the USA, Brazil, and China.

Guano Deposits

Phosphate deposits derived from bird and bat guano represent only a small proportion of the total world reserves of phosphate rock. Insular deposits are common in warm-arid or semi-arid regions with large bird populations either at the present day or in the recent past. The most important deposits, now essentially worked out are on larger islands over 50 metres above sea-level, such as Nauru and Christmas Island, and are thought to be older than about one million years old. In these deposits, solutions derived from overlying bird droppings have percolated into the bedrock, where minerals such as apatite, whitlockite, crandallite, and millisite have crystallized. This phosphatized bedrock forms much of the resource.

Cave phosphate deposits derived from bat droppings are of more interest for the unusual minerals they may contain than for their economic value. Such deposits are mostly in limestone caves, with a minority in lava-tube caves. The chemical reactions involved in forming phosphate minerals are complex, but usually begin with leaching of very soluble nitrogen from the guano. This leaves phosphorus to combine with whatever cations are available from the surrounding rocks. The resulting sequence of minerals may be well stratified within the guano. Typical cave phosphates include brushite, carbonate-hydroxylapatite carbonate-fluorapatite, taranakite, and variscite, generally occurring as powdery nodules within the guano or as coatings on bedrock or cave walls. Distinct crystals of phosphate minerals, such as newberyite and struvite are rare, with a notable occurrence in lava caves at Skipton, Victoria.

Phosphates in Oxidized Metal Sulphide Deposits

Large numbers of phosphate minerals occur in the oxidized zones of base metal orebodies. Solubility phenomena play the most important role in determining which phosphates crystallize in these low-temperature environments, where generally acidic groundwaters dominate. Phosphates of Pb^{2+} are generally the least soluble, so these minerals, particularly pyromorphite (Figure 9), are prominent in oxidized zones above galena-bearing ores. As primary ores commonly contain a mix of lead, copper, and zinc sulphides, as well as arsenopyrite, a diverse suite of secondary phosphates and arsenates can form in oxidized zones. Whether phosphates will be prominent over arsenates depends on the availability of phosphorus, usually as apatite, in either the primary ore or the host rocks – it can vary widely. These differences are illustrated by the two most mineralogically diverse oxidized zones known. At Tsumeb, Namibia, arsenates generally dominate the secondary assemblage, whereas at Broken Hill,

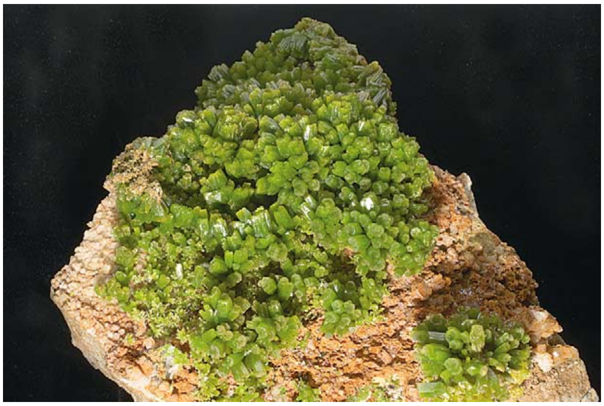


Figure 9 Pyromorphite crystals (up to 5 mm long) from Yang Shao mine, Guangxi Province, China. Museum Victoria specimen M48184, photograph by J Broomfield. Reproduced with permission from Museum Victoria.

in New South Wales, Australia, both phosphates and arsenates occur. The phosphate assemblage libethenite–pseudomalachite is particularly widespread above copper-bearing sulphides in arid regions. Notable occurrences of zinc phosphates, including parahopeite, hopeite, tarbuttite, and scholzite, occur at Broken Hill, Zambia, and at Reaphook Hill, in South Australia. The great chemical diversity shown by secondary phosphates is reflected in their often-spectacular colours and crystal habits, making them much prized by mineral collectors.

Geochronological and Thermochronological Applications of Phosphate Minerals

Apatite, monazite, and xenotime commonly contain between a few tens and hundreds of ppm U and Th in their crystal structures. As a result, several different isotopic dating techniques can be applied to them. While the underlying principles, assumptions, and counting methods for each technique are complex and beyond the scope of this review, a brief summary of each is useful. Fission track dating (*see* **Analytical Methods:** Fission Track Analysis) uses damage tracks in apatite arising from the spontaneous fission of ^{238}U, which occurs at a known rate. Measuring the number of tracks that have accumulated since a crystal formed, along with estimating the amount of uranium it contains, means that a geological age can be calculated. Because fission tracks in apatite are 'healed' or annealed at temperatures over about 120°C, only rocks which have not undergone subsequent heating events can be dated this way. However, the annealing properties of apatite fission tracks have led to a growing number of opportunities to model significant thermal processes in the upper parts of the Earth's crust. These include reconstructing the thermal histories of sedimentary basins (*see* **Sedimentary Environments:** Depositional Systems and Facies) and evaluating their potential for oil and gas resources, and estimating the timing and magnitude of erosional and tectonic denudation of mountain ranges (*see* **Plate Tectonics**).

The U–Th–Pb and (U–Th)/He dating methods applied to apatite, monazite, and xenotime have as their basis the decay series of the long-lived isotopes of uranium, ^{238}U and ^{235}U, and of thorium, ^{232}Th. These decay at a known rate through a series of short-lived radionuclides ultimately to Pb isotopes. Determination of the ratios of ^{206}Pb/^{238}U and ^{207}Pb/^{235}U enables a concordia plot to be drawn, which provides an age for the crystals being analysed. However, there are many complicating factors involved in interpreting these plots and in measurement of the data. As well, different methods of determining isotopic compositions are available and need to be selected, depending on which mineral is involved, the precision required, and other factors. The assumption behind the (U–Th)/He method is that all three phosphates appear to retain He, which is produced during the alpha decay of ^{147}Sm. By measuring U, Th, and He contents, an apparent He age can be calculated on the assumption that the initial He content of the mineral was zero. Both the U–Th–Pb and (U–Th)/He methods are still being developed and refined but offer great scope for accurate dating of Earth processes.

Phosphate Biomineralization

The main inorganic constituent of bones and teeth in vertebrate animals, including human beings, is an apatite-like mineral similar to carbonate-fluorapatite. Small amounts of other elements such as sodium, potassium, magnesium, and zinc are present in the structure. The precipitation of apatite takes place after secretion of certain proteins by specialized cells. Other phosphate minerals such as whitlockite, struvite, and brushite, as well as a number of amorphous calcium and/or magnesium-bearing phosphates, have been found in pathological tissue calcifications, such as dental and urinary calculi. Formation of these and other biophosphates is sensitive to conditions such as temperature and pH, so that transformation by dissolution and recrystallization, especially of apatites, may take place. A range of synthetic apatites, in the form of cements and porous ceramics, is now being developed and trialled in order to repair defects and damaged tissue and to correct deformities. These have the capacity to considerably improve both the quality and span of human life.

Environmental Significance of Phosphate

Phosphorus is essential for all forms of life. Considerable cycling of phosphorus takes place within the biosphere and interchange occurs between ecosystems. While the overwhelming amount of phosphorus fluxing takes place between marine organisms and ocean water, human activities play a significant role at the ecosystem scale. The widespread use of phosphate-based fertilisers and insecticides, the disposal of sewerage sludge and industrial waste, including some derived from nuclear reactors, are examples of larger-scale processes that can have serious environmental impacts. Perhaps the best known involves the overload of phosphorus in streams and lakes, leading to an explosion of plant life, especially algae, which upon decay uses up most of the dissolved oxygen. This process, known as eutrophication, results in fish kills (*see* **Fossil Vertebrates:** Fish) and degradation of water quality. On a more restricted scale, there is some evidence for the formation of lead phosphates such as plumbogummite–crandallite and pyromorphite–apatite in soils verging on roads and highways used by vehicles burning leaded gasoline. Increasing awareness of all these problems has meant that control programmes are in place in many regions. Phasing out of lead-based fuels and phosphate-based detergents, together with possible use of crystalline phosphates and phosphate glasses for nuclear waste immobilization, are also helping to improve environmental outcomes.

See Also

Africa: North African Phanerozoic. **Analytical Methods:** Fission Track Analysis. **Fossil Vertebrates:** Fish. **Igneous Rocks:** Granite. **Minerals:** Definition and Classification. **Plate Tectonics**. **Sedimentary Environments:** Depositional Systems and Facies. **Sedimentary Rocks:** Mineralogy and Classification.

Further Reading

Anthony JW, Bideaux RA, Bladh KW, and Nichols MC (2000) *Handbook of Mineralogy.* Volume IV: arsenates, phosphates, vanadates. USA, Tucson: Mineral Data Publishing.

Birch WD and Henry DA (1993) *Phosphate Minerals of Victoria.* Australia, Melbourne: Mineralogical Society of Victoria Inc.

Cook PJ (1984) Spatial and temporal controls on the formation of phosphate deposits – a review. In: Nriagu JO and Moore PB (eds.) *Phosphate Minerals,* pp. 242–274. Germany, Berlin: Springer-Verlag.

Filippelli GM (2002) The global phosphorus cycle. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 391–425. Washington, DC: Mineralogical Society of America.

Gleadow AJW, Belton DX, Kohn BP, and Brown RW (2002) Fission track dating of phosphate minerals and the thermochronology of apatite. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 579–630. Washington, DC: Mineralogical Society of America.

Hill C and Forti P (1997) *Cave Minerals of the World.* USA, Alabama: National Speleological Society Inc.

Huminicki DMC and Hawthorne FC (2002) The crystal chemistry of the phosphate minerals. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 123–253. Washington, DC: Mineralogical Society of America.

Kostov I and Breskovska V (1989) *Phosphate, Arsenate and Vanadate Minerals. Crystal Chemistry and Classification.* Bulgaria, Sofia: Kliment Ohridski University Press.

Moore PB (1973) Pegmatite phosphates: descriptive mineralogy and crystal chemistry. *The Mineralogical Record* 4: 103–130.

Nash WP (1984) Phosphate minerals in terrestrial igneous and metamorphic rocks. In: Nriagu JO and Moore PB (eds.) *Phosphate Minerals,* pp. 215–241. Germany, Berlin: Springer-Verlag.

Piccoli PM and Candela PA (2002) Apatite in igneous systems. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 255–292. Washington, DC: Mineralogical Society of America.

Robinson GW, Velthuizen J van, Ansell HG, and Sturman BD (1992) Mineralogy of the Rapid Creek and Big Fish River area, Yukon Territory. *The Mineralogical Record* 23(4): 3–47.

Spear FS and Pyle JM (2002) Apatite, monazite and xenotime in metamorphic rocks. In: Kohn KJ, Rakovan J, and Hughes JM (eds.) *Phosphates: Geochemical, Geobiological and Materials Importance,* Reviews in mineralogy and geochemistry 48, pp. 293–255. Washington, DC: Mineralogical Society of America.

Williams PA (1990) *Oxide Zone Geochemistry.* England, Chichester: Ellis Horwood Limited.

Rudaceous Rocks

J McManus, University of St. Andrews,
St. Andrews, UK

Introduction and Terminology

Sedimentary rocks in which coarse particles are dominant are termed 'rudites'. They consist of broken fragments, clasts, of pre-existing rocks, and have formed in a wide range of conditions, such as in scree, in landslides, as tills, on alluvial fans and in many sites along river courses, on beaches, in offshore reef-fringing areas and in deep-water environments. Characteristic inter-relationships between the rudites, and other environmentally significant features of these rocks and their associated sediments provide clues to their modes of origin. The clasts themselves provide additional evidence from their shapes and composition. The particle shapes evolve during transport, and textural sorting by size, shape or form may characterize certain depositional conditions. The composition of the particles often indicates the nature of the source from which they were derived. The rudaceous deposits, therefore, provide a stimulating variety of geological challenges at all levels.

The term conglomerate is applied to rudaceous rocks composed of rounded pebbles, and breccia to those composed of angular clasts. A distinction is made between rudaceous rocks where the clasts are in contact with one another, and those in which the clasts 'float' in a finer matrix of sand and clay. These are termed 'clast-supported' and 'matrix-supported' conglomerates respectively. Intraformational rudaceous rocks are composed of clasts of penecontemporaneously cemented sediment; limestone 'beach rock', for example, or intraformational shale pellet conglomerates. Most rudaceous rocks, however, are composed of 'extraformational' clasts derived from outside the formation in which they occur. A further distinction is made between rudaceous rocks composed of many or one rock type. These are termed 'polymictic' and 'ologomictic' conglomerates respectively.

Rudaceous Rock Textures and Fabrics

The size of the particles is of primary importance. Pebbles are defined as particles between 4 mm and 64 mm in diameter and cobbles up to 256 mm. Coarser materials are boulders or blocks. No natural deposits of clastic materials consist of clasts with a single size of particles and a range of diameters are always present. This is partly a function of the material supplied and partly due to variations in the dynamics of the transporting medium and in the conditions during deposition. As large quantities of sediment (often tens of kg) need to be analyzed to obtain statistically meaningful information on coarse grain size populations, such information is relatively rare. In many cases the coarseness quoted is related to the diameter of the largest clast observed, or the average of the largest clasts.

When clasts are released from their source rocks their shapes are defined initially by the distribution of weaknesses in the parental rocks. Fractures such as bedding planes, joints, or cleavages exert a major influence in both the size and shape of the materials produced. Likewise the composition of the bedrock determines the ease with which the large fragments become broken and rounded in transport. Soft, poorly cemented sandstones and limestones break apart more readily and form better rounded clasts than schists, quartzites or granites.

Three aspects of particle shape need to be considered characterizing pebbly materials, namely roundness, sphericity, and form. The three measures may appear related but they address totally different aspects of the clasts. In a numerical sense the roundness is the relationship between the radius of curvature of the sharpest edge and the length of the longest or intermediate axes or a combination of the two. It is conventionally expressed:

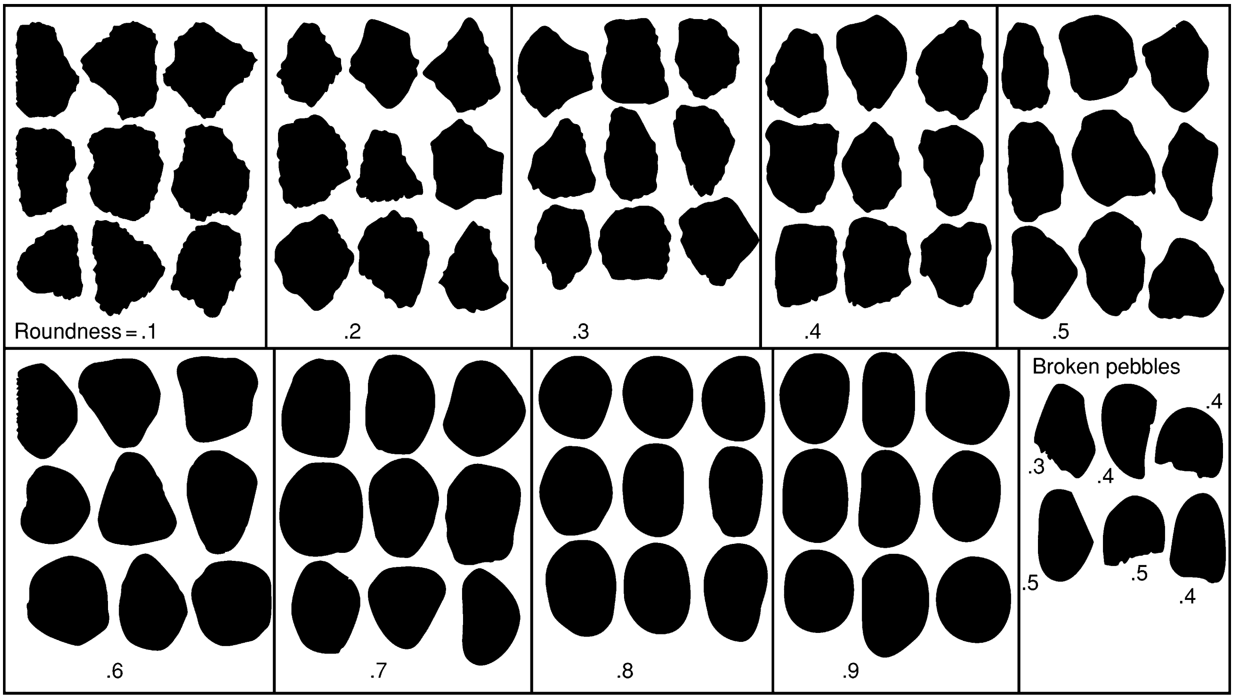
$$\text{Roundness} = \frac{\text{average radius of corners and edges}}{\text{radius of maximum inscribed circle}}$$

For speed of processing, most workers assess roundness with the aid of visual comparator charts, as shown in Figure 1.

The sphericity of a clast is the ratio of the diameter, D, of a sphere having the same volume as the clast to that of a circumscribing sphere (i.e., the longest axis, A). It may also be defined as a triaxial ellipsoid based on the product of the lengths of the three diameters of the particle to the volume of the circumscribed sphere, i.e., BC/A^2, where B and C are the intermediate and short axes respectively. Sphericity is also expressed as:

$$\text{Sphericity} = \frac{\text{surface area of the particle}}{\text{surface area of a sphere of equal volume}}$$

Since particles settle through any transporting medium with their maximum projection area



Pebble images for visual roundness

Figure 1 Roundness chart for particles 16–32 mm diameter (Reproduced from Source; Krumbein WC (1941) Measurement and geological significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Petrology* 11: 64–72).

perpendicular to the direction of settling a more dynamically related measure may be obtained by comparing that value with the projection area of the maximum circumscribing sphere as (C^2/AB) 1/3. The reciprocal of this value is of importance, relating to the ease of transport.

Clast form notation (Figure 2) is based on two ratios (2/3) of the particle axes, B/A and C/B to define four form fields: spheres (equant), discs, blades and rods. Using the same three axial lengths it is possible to create a triangular, ten-field form diagram (Figure 3). Although various other combinations of axial ratios have been suggested none has achieved the lasting impact of the above methods of characterizing pebble form. In any attempt to relate the particle form to particular environments of deposition it is necessary to measure a significant number of pebbles (over 200), so that a definitive spread of values may be obtained.

Plotting the values of A against A/B enabled Moss (1962) to identify three particle populations, which he termed framework, interstitial and contact, in a range of gravel deposits. The framework consists of a pebble population graded in size from small and equant particles to relatively large and elongated clasts over a small size range. The interstitial population is subsidiary to and always associated with the framework. Its coarsest pebbles are the same size as, but more elongated than the finest framework material. The contact population, which may be very minor in proportion or may dominate the deposit, is normally coarser than the coarsest part of the framework and characteristically is more equant in form. The contact population is of materials that are unable to fit into the stable gravel bed and commonly move more rapidly along the river than the bulk of the bed material. The value of 1.5 for A/B provides a separation of blades and rods from the discs and spheres, but as indicated above, in natural systems the rods and spheres generally behave similarly. In order to provide a dynamically meaningful plot, while retaining the use of ratios of the A and B axes, the use of D, the volumetrically determined nominal diameter of the clasts, again enables the four particle forms to be recognized (Figure 4).

Since each of the three properties roundness, sphericity, and form are at least partly defined from the particle diameter they are not entirely independent. Along a sediment transport path, such as a river, the mean clast diameter decreases with distance traveled, an exponential relationship, in which the most rapid changes occur near the source and progressively lesser changes in more distal locations (Figure 5).

Both sphericity and roundness increase as the particles decrease in size, and again the changes are most

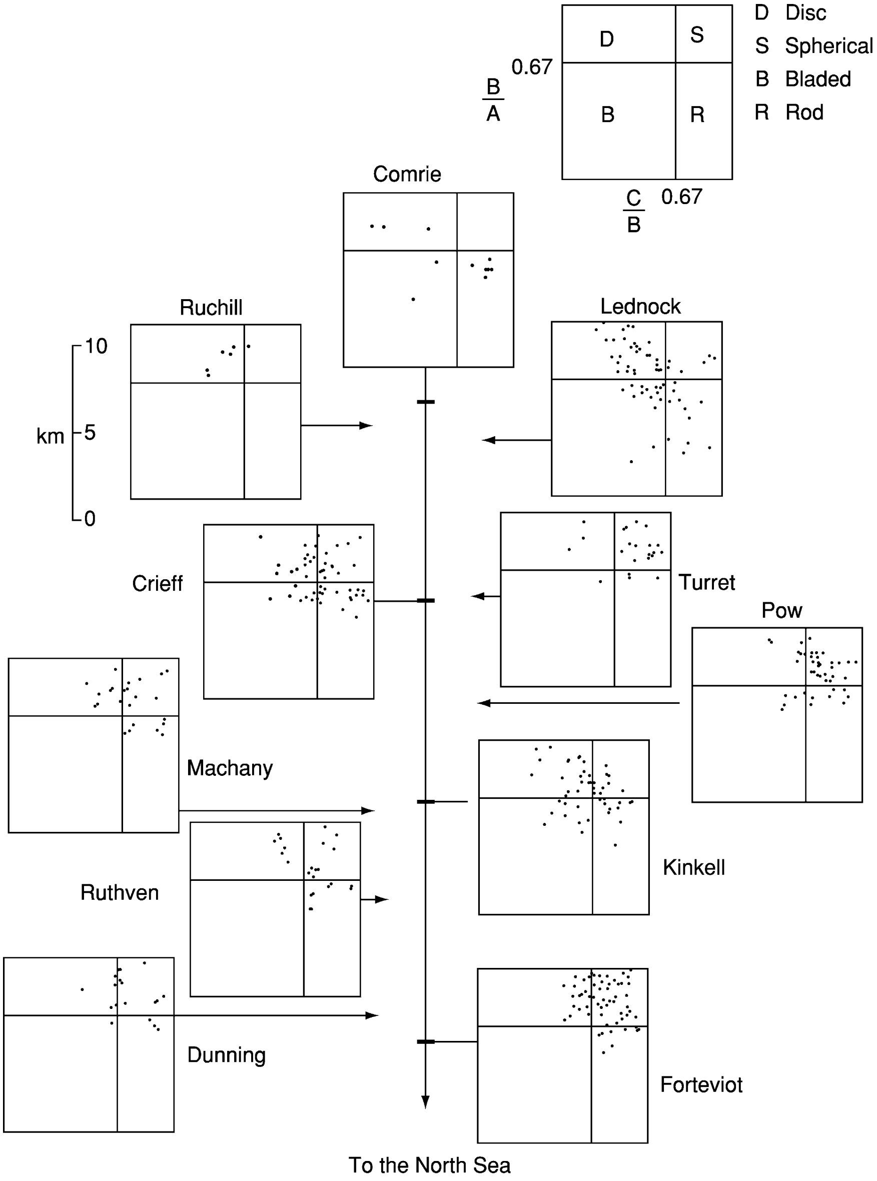


Figure 2 Variation of form of quartz pebbles along the River Earn, Scotland, using the Zingg (1935) plot (after Al-Jabbari, Al-Ansari and McManus (1982) *Journal of Water Resources* 1, 81–110).

rapid near the source. The main controls of roundness are a) distance traveled; b) composition of the pebble; c) clast size; d) initial clast form; e) nature of the bed material, and f) the dynamics of the transporting medium. Whilst these controls are readily isolated in the laboratory they are not so readily assessed in the field, where most gravel-transporting streams receive detritus derived from tributary catchments that includes pebbles having different transport histories that are added along the length of the stream. Furthermore, erosion of stream bank terraces or tills may lead to the addition of clasts at any point along the stream.

Clasts in Natural Environments

When the coarse particles are released from the exposed rock surfaces to form scree move down slopes debris flows or landslides, or enter streams to be carried into lakes or the sea, where they may form beaches, gravity is the prime motive force. Gravity is all-important in the initial stage, whether the clasts are released from the rock face by frost wedging or by a combination of physical and chemical weathering processes. Within scree, debris flow or landslide particle motion is enhanced by the lubricating and hydraulic effects of water, or ice, working in

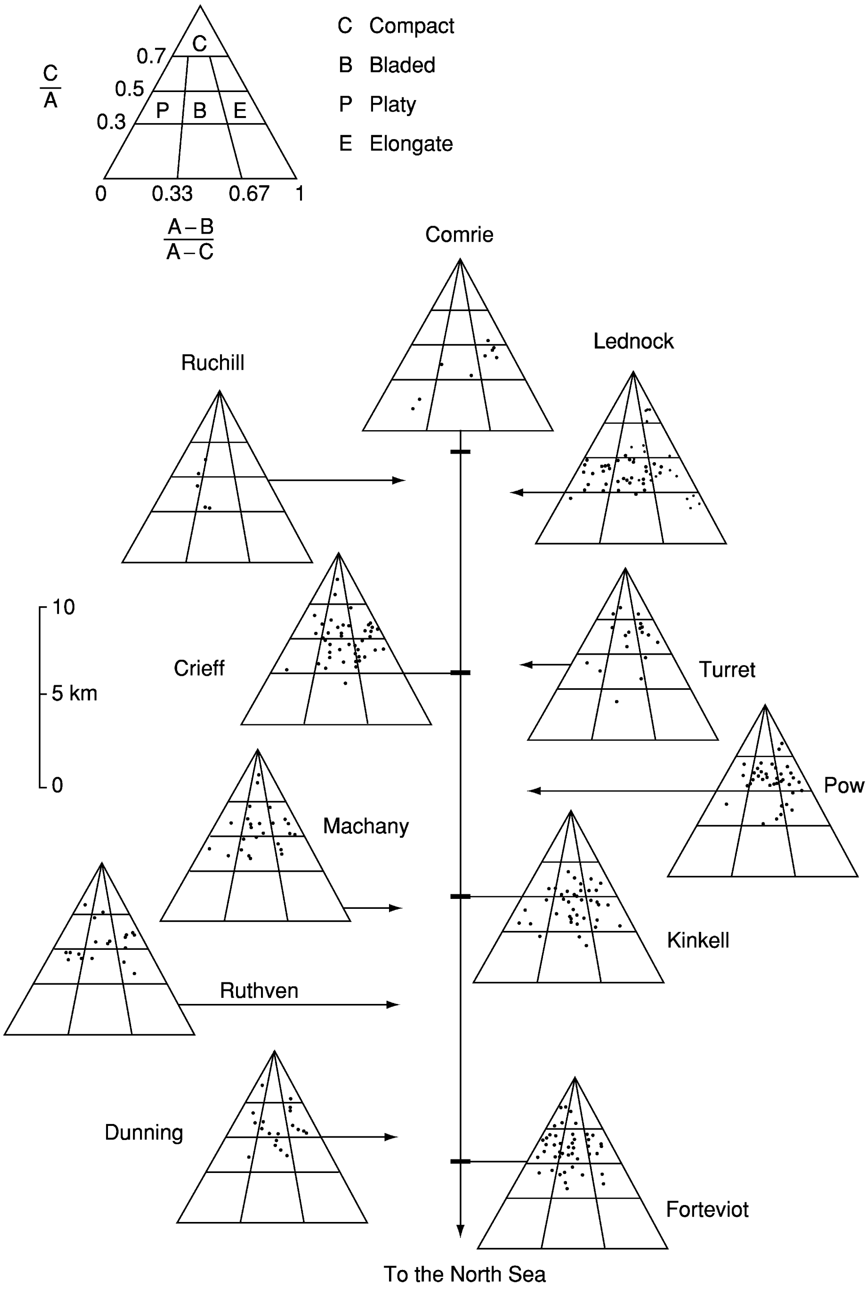


Figure 3 Variation of form of quartz pebbles along the River Earn, Scotland, using the Sneed and Folk (1958) plot. (After Al-Jabbari, Al-Ansari and McManus (1982) *Journal of Water Resources* 1, 81–110).

combination with gravitational downhill pull, although frictional forces generated by neighbouring materials impede free movement.

Measurement of clasts from screes of different rock types in Scotland revealed that very few of the original clasts fall into the form field for rods or spheres. Virtually all were either discs or blades when released into the environment. Both the composition and structure of the materials, as well as physical weathering by frost action, are believed to exert major controls on form.

Clasts in Streams

When a mixed population of particles occurs on a stream bed the finest, silt-sized particles are carried away in suspension, and the sands saltate downstream, bouncing along with the flow. The pebbles normally remain on the bed, often with their upper parts extending through the boundary layer of the flow. Once flow strength exceeds some critical value the smaller pebbles begin to slide or roll along the bed, and as flow strength increases so increasing quantities

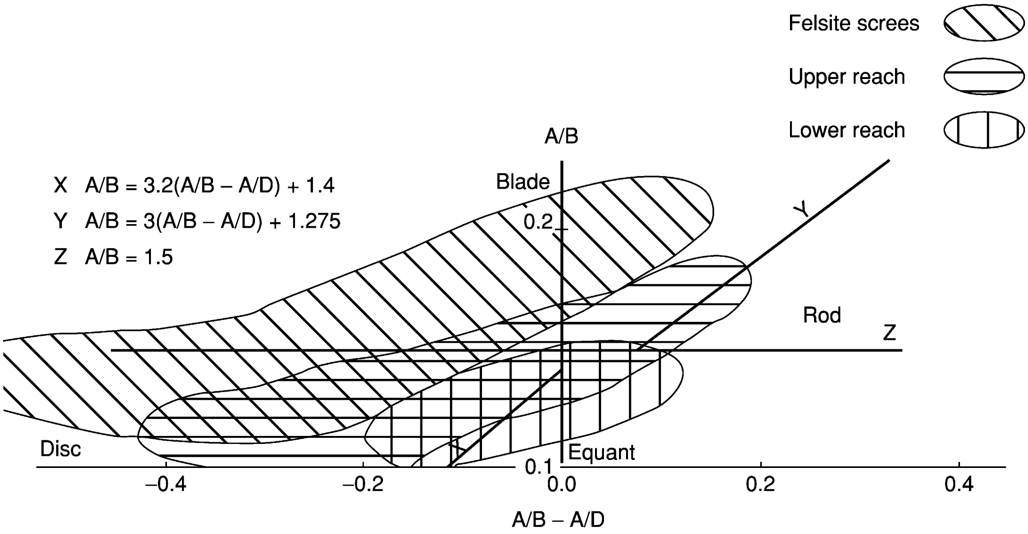


Figure 4 Variation of pebbles form from scree, and the upper and lower reaches of the Shee Water, Scotland, using the A, B and D diameters (after Dhiab IH 1979, unpublished M.Sc. Thesis, University of Dundee).

migrate until, in extreme floods, most of the bed is in motion. As the power of the flow decreases so particles become deposited and the bed aggrades, as material accumulates. During floods the entire bed down to bedrock may become removed from a reach of the river, to be replaced by new material as the peak of flooding passes. There are records of over 8 m of bed removal and replacement during individual storms in the western USA. The various interrelationships between water flow, sediment size, sediment load, and stream slope are summarized in Figure 6.

Under normal conditions the 'vibration' of the turbulent flow induces minor movement within the bed, leading to the upward migration of the coarser particles, while finer materials move downwards into the deposit. In this way the stream bed develops a residual lag of coarse pebbles, and there is a progressive downward fining of the gravels. Frequent inter-pebble contacts induce progressive rounding of edges through abrasion in the bed and surfaces exposed on the streambed become scoured through collisions with the saltating sand grains.

Within the stream bed the clasts may become arranged such that an internal texture, related to both size and form of the particles is developed. The texture results from burial of blades or discs whose intermediate axes slope upstream (up current). This is termed imbrication. During motion the more equant particles roll along the bed once they have been disturbed, whereas the flatter clasts are more likely to slide across the other pebbles, turning over as they encounter immobile clasts. The ability of the clast to find a place into which it can fit in the bed often determines the distance that it travels during any single displacement from the bed. The more nearly spherical the particle the greater the difficulty it experiences and the further it is likely to move once displaced from its resting place.

Attrition of the sliding and turning particles increases the probability of their splitting to form smaller more equant or rod-shaped pebbles. There is commonly a size gradation along the length of a gravel bar within a braided stream reach, with the coarsest clasts at the upstream end and progressively smaller ones downstream. As breaking of the pebbles occurs along the river so there is a downstream increase in the proportion of spherical clasts within the bed load.

Clasts on Beaches

Once the pebbles reach the sea, either from a river, or cliff collapse, they become subject to the motion of waves, which carry them along the coast by longshore drift, and deposit them on beaches. Beaches formed entirely of gravel occur on dynamic coasts. The presence of sand is an indication of quieter conditions. The latter used the 10-fold form diagram to differentiate between beach and river gravels with some success.

The gravel beaches of South Wales develop a fourfold structure, recognized on the basis of pebble size, form and the texture of the deposit. Between an upper zone of coarse discs and a lower zone principally of spheres lies a zone with imbricate discs, which passes down slope into a infill zone comprising particles of many shapes, and often incorporating sand (Figure 7).

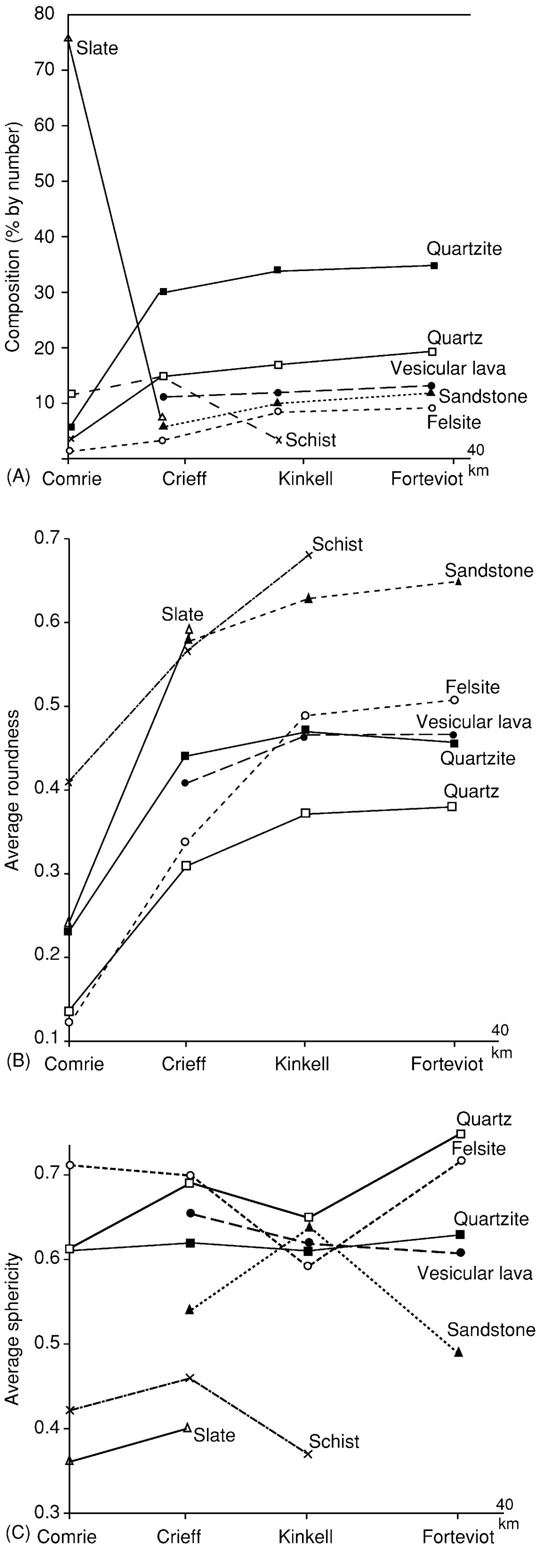


Figure 5 Variation of A) the pebble composition, B) roundness, and C) sphericity along the River Earn, Scotland (after Al-Jabbari, Al-Ansari and McManus (1982) *Journal of Water Resources* 1, 81–110).

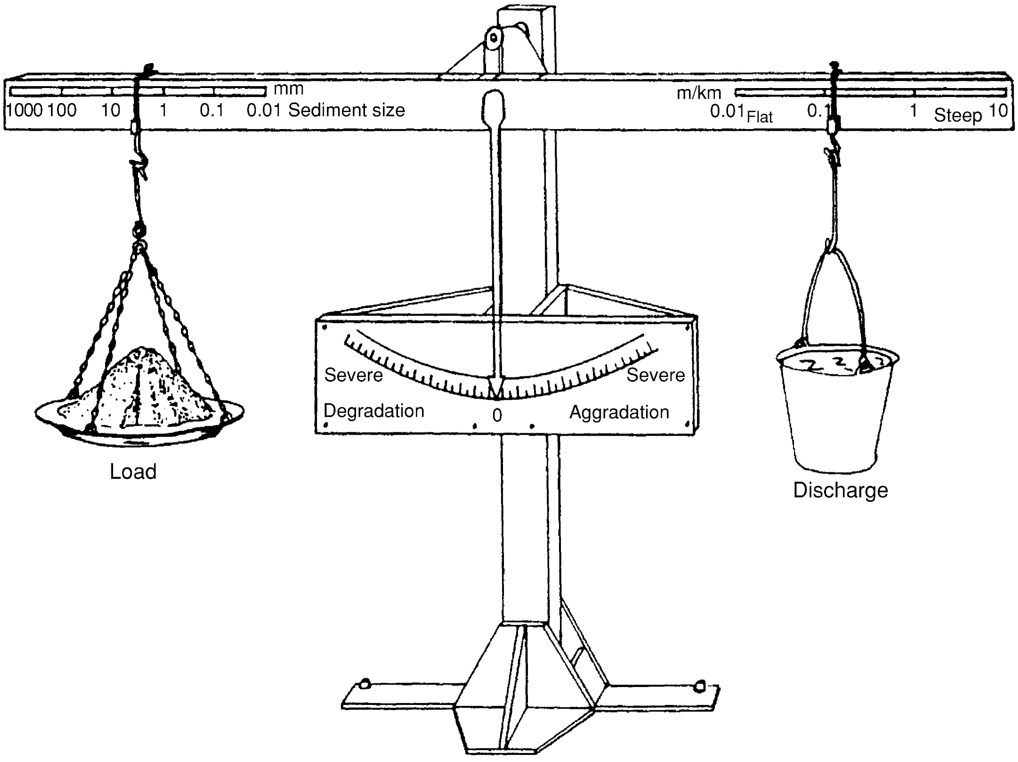
As clasts are brought to a beach they are carried by waves that wash over the surface many times during transport of the particles. The mobile clasts interact with the underlying gravels. If they are larger than the pore spaces between the settled clasts the pebbles bridge the gap and continue to migrate along or up and down the beach. If the clasts are too small they enter the open spaces, but wash through to continue migrating. A clast of similar size will fully occupy the space provided. In this way clasts of similar size and form characteristics gather together usually on the lower parts of beaches and provide 'selection pavements', effectively providing a gauntlet through which clasts must pass if they are to move up or down the beach face. Noting the internal structure of many beaches in western Scotland, Bluck examined gravel beaches forming and migrating through time, drawing attention to the prevalence of imbricate structural zones in almost all beaches examined (Figure 8).

Particle tracing techniques using dyes, paint or creating artificial electronically tagged pebbles have enabled the motion of individual pebbles to be tracked sometimes for periods exceeding ten tides. Disturbance of the sediment often penetrates to 10–20 cm below the beach surface under moderate sea conditions. Large clasts migrate along the beach face more rapidly than do their smaller counterparts, and under identical conditions (on the same days) clasts of ironstone migrate more slowly than similar clasts of sandstone, and they in turn are more sluggish than matched coal fragments. The density of the particles provides the control in this case. On many British beaches rates have been measured of longshore movement of 5–8 cm diameter clasts of flint, chert, sandstone and ironstone of up to 10 m per tide under moderate wave conditions. This suggests that many of the features illustrated by Bluck may be essentially short lived, although regularly regenerated in the same locations.

Where steep rocky cliffs lie behind the coast, direct cliff fall contributes boulders and cobbles to the beach. The large clasts become rounded through attrition but remain on the beach.

Ancient or 'fossilized' cliffed coasts are rarely preserved, but at Enard Bay, north-west Scotland, a Precambrian coastline has been exhumed showing cliffs cut into Lewisian Gneiss, and cut into Lewisian gneisses with Torridon Group marginal fanglomerate and beach deposits banked against it (Figure 9).

Isostatic uplift following deglaciation has allowed many former coasts to rise above current sea level, and the raised beach features preserved around many northern European and north American coasts display most of the internal textures explored by Bluck.



Sediment load × Sediment size ∝ Stream slope × Discharge

Figure 6 Interrelationships between factors controlling a stable channel bed in a river (after Lane 1955).

Lakeshore gravel beaches are common, but the low power in the waves generated on small lakes ensures that the beaches rarely achieve a fully mature condition. The large clasts are commonly essentially static, show limited size and shape sorting, and are often poorly rounded. Rods and spheres are generally more common in the coarser particles than in the smaller clasts.

In shallow tropical or subtropical seas carbonate sediment dominates. These include in situ reef rock, and carbonate sands and muds. All of these may become quickly cemented, and subsequently eroded during heavy storms generating large amounts of rudaceous detritus. An apron of limestone breccia may extend into deep waters at the foot of reefs and other abrupt carbonate shelf margins. Such reef apron breccias have been recognized in the Permian Capitan Reef of Texas and New Mexico, around many Carboniferous structures in Britain and Belgium, and round Cretaceous reefs in the French Juras.

Clasts on Alluvial Fans

Adjacent to newly emergent mountains, or uplifting fault blocks rapid changes of stream bed gradient ensure that deposition of sediment occurs adjacent to the source, creating alluvial fans of coarse detritus. These sediments are particularly well preserved in arid or semi-arid areas providing important sites in which rudaceous materials accumulate. In the geological past, before evolution of extensive land-living plants in the Devonian, fans were widely developed and thick accumulations of pebbly sandstones mark the margins of areas of active uplift.

The evolving alluvial fans, or the coalescent adjacent fans forming a bajada, produce wedges of sediment, commonly but against a fault. Repeated uplift generates successions of wedges of coarse materials stacked above each other, with thicker beds of coarse sediments near the source, grading to thinner beds of finer materials in more distal areas. The uppermost parts of the fans are characterized by the presence of debris- and mud-flow deposits, each of which contain large clasts, but the latter also contain much fine sediment. In essence the water drains from the moving sediment-enriched flow to induce deposition. The fan head is dominated by ribbons of the mass flow deposits, through which the streams erode as they flow towards the lower, outer parts of the fan, where slopes decrease from 5° to less than 1° and the waters become divided into many distributaries. The sediments of the outer part of the fan are dominated by sheets of sand or gravel from the often ephemeral, braided streams. Characteristically, the fan sediments

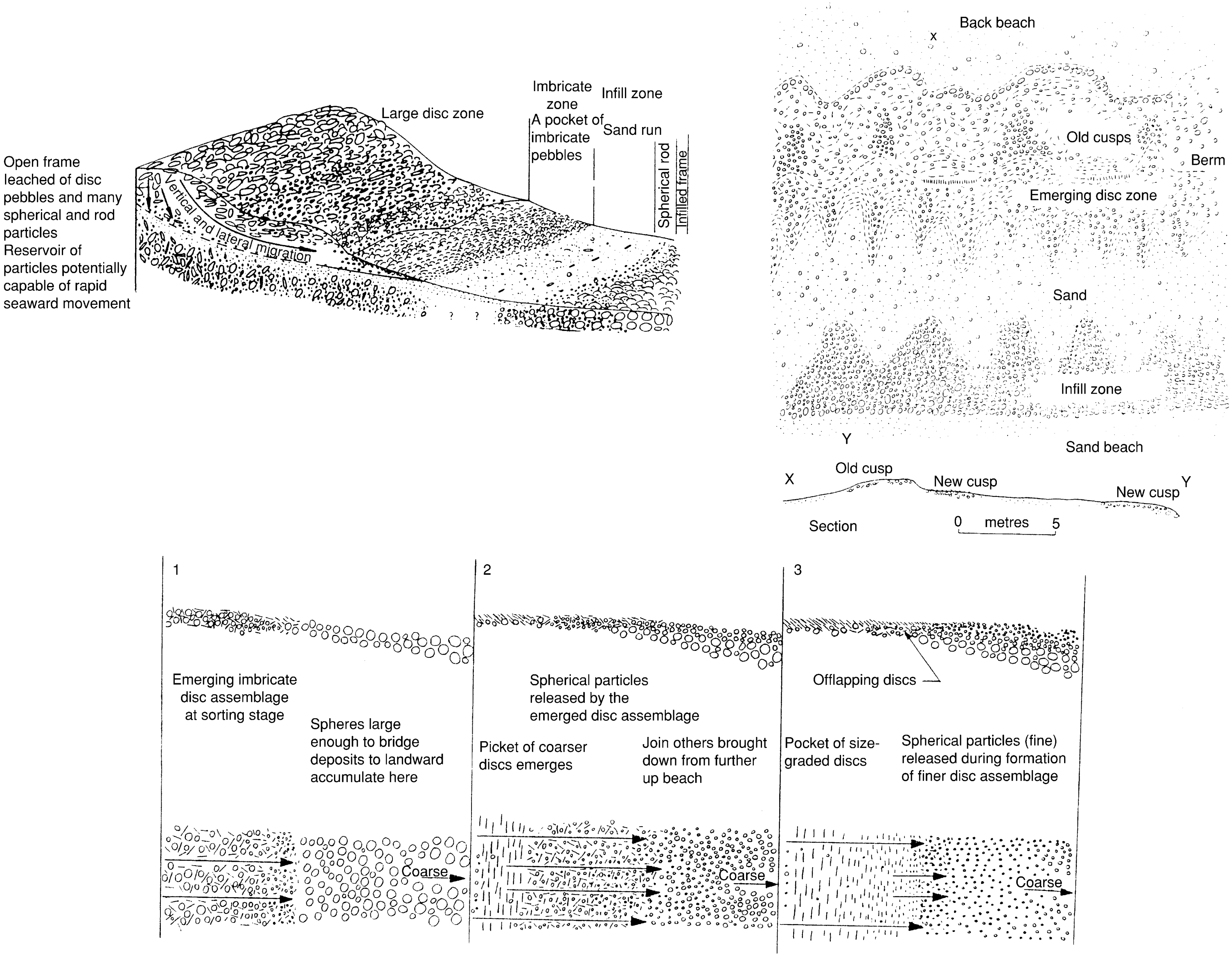

Open frame leached of disc pebbles and many spherical and rod particles
Reservoir of particles potentially capable of rapid seaward movement
Large disc zone
Vertical and lateral migration
Imbricate zone
A pocket of imbricate pebbles
Infill zone
Sand run
Spherical rod
Infilled frame
Back beach
X
Old cusps
Berm
Emerging disc zone
Sand
Infill zone
Y
Sand beach
X
Old cusp
New cusp
New cusp
Y
Section
0 metres 5
1
Emerging imbricate disc assemblage at sorting stage
Spheres large enough to bridge deposits to landward accumulate here
Coarse
2
Spherical particles released by the emerged disc assemblage
Picket of coarser discs emerges
Join others brought down from further up beach
Coarse
3
Offlapping discs
Pocket of size-graded discs
Spherical particles (fine) released during formation of finer disc assemblage
Coarse

Figure 7 Arrangement of clasts on gravel beaches (after Bluck BJ (1999) *Transactions of the Royal Society of Edinburgh* 89, 291–323).

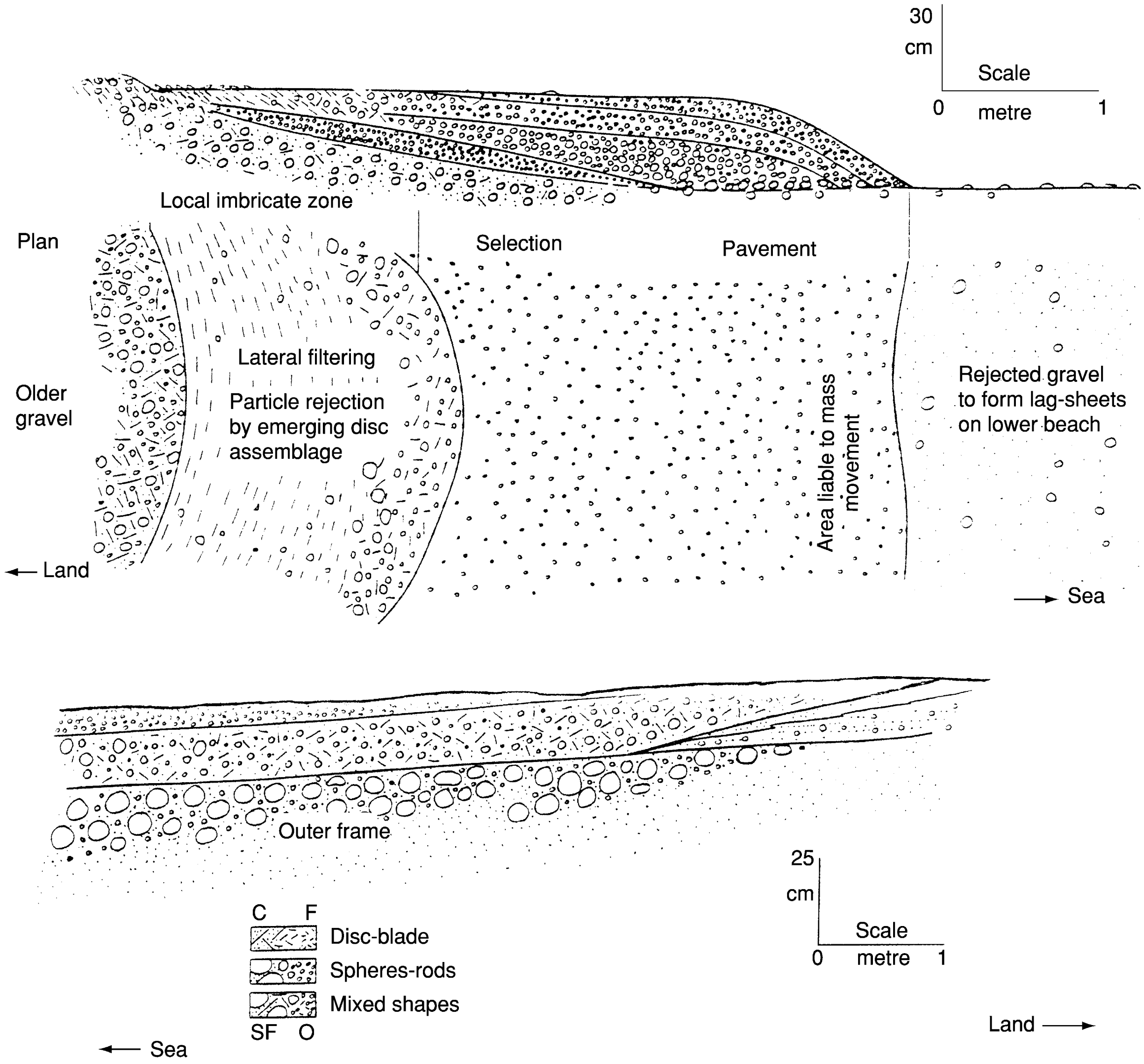


Figure 8 Schematic cross-sections of beaches from south-west Scotland (after Bluck BJ (1999) *Transactions of the Royal Society of Edinburgh* 89, 291–323).

are coarse, and dominated by gravels in the upper reaches, becoming more fine grained in the lower areas.

Sheet-like conglomerates with pebbles of similar sizes result from the formation of temporary selection pavements during sheet flows of flood discharge. Other deposits fill shallow channels scoured into pre-existing sediments. Sometimes their steeply inclined imbricate clasts serving to identify channel margins, while more gently inclined clast axes occur in the central parts of the channels. The maximum clast size decreases exponentially down fan, as does the thickness of the individual conglomeratic beds.

The alluvial fan-bajada systems may extend for 25–30 km from the fan head on to alluvial plains, for tens of kilometres along active mountain fronts, and, where long-lived geologically the deposits may reach several kilometres in thickness. Such dimensions are recorded from modern systems in Nevada, ancient fans in Texas, and from Neogene deposits of Italy and Switzerland, related to the rising Apennines and Alps respectively. In Britain large ancient fans have been identified from the Applecross Formation of the Torridonian (Late Precambrian) of northwest Scotland, with smaller fans in the Devonian successions of the Midland Valley and Orcadian basins of Scotland, and of Permian age in south-west England.

Clasts in Braided Rivers

The outermost parts of some fans are dominated by braided streams, whose normally shallow channels of low sinuosity, become subdivided by mid-channel

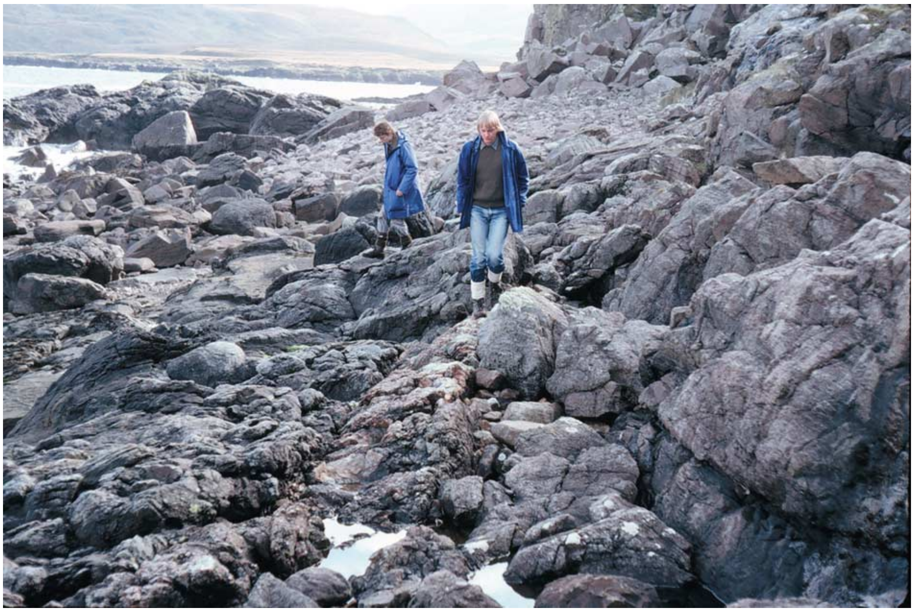


Figure 9 Exhumed cliffline and beach sediments of Early Torridonian age, Enard Bay, north-west Scotland.

bars. The braided rivers beyond the fans typically develop in areas with greatly varying water discharge histories. The wadi-floor streams of desert and semi-desert areas, for example, receive little water for many months before floods sweep through the area carrying sediment-charged waters capable of transporting material up to the size of large boulders. Thus more people drown in deserts than die of thirst. As the waters subside the sediments are rapidly deposited to give a layer of coarse, matrix-supported gravel.

Another important site for the formation of braided systems is in the periglacial sandurs and plains associated with glacial retreat, such as those of Iceland. The rivers carry little water during the winter months, but in summer may carry large quantities of glacial melt water, accompanied by the transport of high sediment loads, which become deposited as the bed gradient and flow velocities fall. These streams drain areas in which sediments of all size are available and movement is minimally restricted by vegetation. In rift valleys rudaceous marginal fault-bounded fanglomerates may pass out into braided river sands and gravels on the floor the central parts of basin (Figure 10).

The deposits of braided rivers, explored by Miall and Bluck, typically show successions of fillings of stacked channels, some of the major stream, and others of second or third-order channels (Figure 11). Typically upward fining sheets of sand and gravel result from the migration of mid-channel and over-bank bars that are the principal sites of deposition. Miall showed that in the upper reaches of some braided systems debris flows occur within the more normal upward fining cycles of the flood deposits. He identified three pebble-rich assemblage types, dominated by channel gravels with intervening debris flows, by superposed channel bars or by channel floor gravels passing upwards into current bedded sands. Pebbles also occurred less frequently in other sand-rich braid deposits. In the British geological column the braided systems have been recognized from the Precambrian Torridonian, in both the upper and lower Old Red Sandstone (Figure 12) of the Devonian and in the New Red Sandstones of the Permo-Trias.

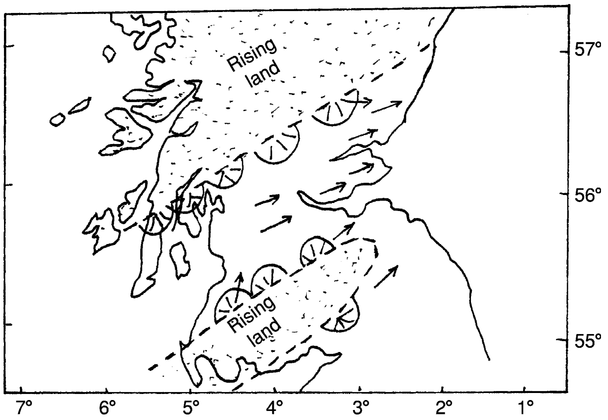


Figure 10 Suggested distribution of Upper Old Red Sandstone alluvial fans in central Scotland (after Trewin NH and Thirlwall MF (2003) Old Red Sandstone. In: Trewin NH (Ed.) *The Geology of Scotland*, 4th edn, pp. 213–251. London: The Geological Society of London).

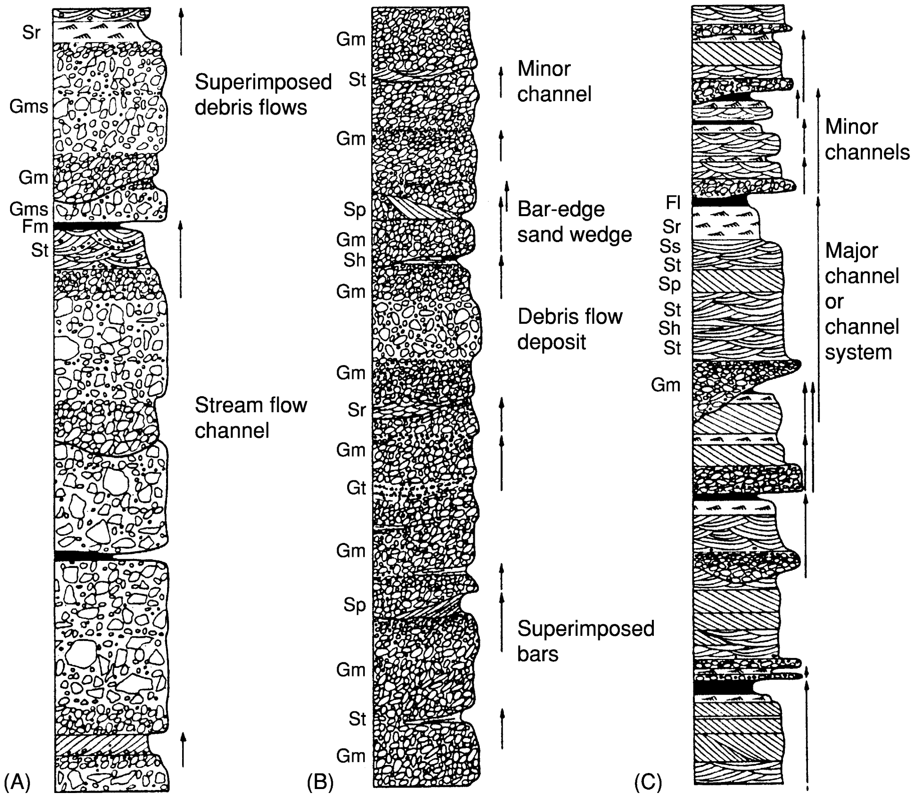


Figure 11 Upper Old Red Sandstone braided river deposits at Whiting Ness, Arbroath, Scotland, showing the flood plain deposits against a buried unconformity, with associated debris flows, the presence of secondary channels and the dominance of bar head deposits.

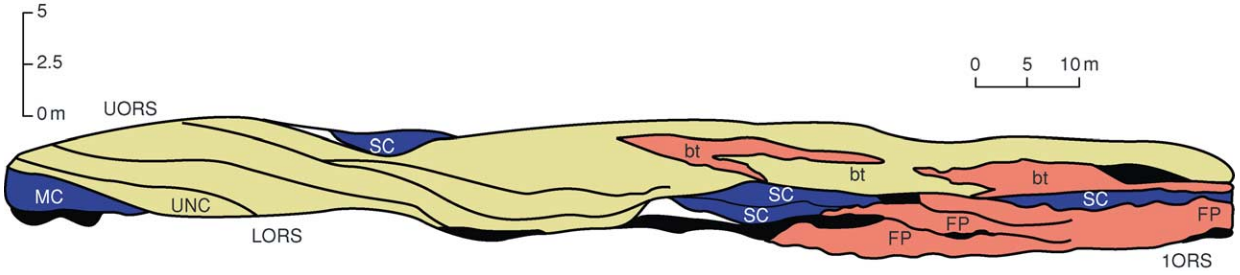


Figure 12 Coarse Lower Old Red Sandstone conglomerates at Stonehaven, Scotland. The larger clasts are over 1 m in diameter.

Clasts in Tills

Rudaceous sediment is found in glacial tills (*see* **Sedimentary Processes:** Glaciers). Tills may be composed of any mix of coarse and fine materials. Whereas in the mountains large boulders are common in clast-supported tills, in more distal areas the more readily transported fine materials dominate. Two forms of till are recognized. Lodgment tills are in direct contact with the underlying rocks, and have usually been deposited beneath the moving glacier, and been partly compressed by the weight of overlying ice. Particles in this deposit may be striated. The overlying ablation till is structurally weaker, having been deposited from down-melting ice. Imbrication, indicating the direction of ice flow, may form in clast-supported or matrix-supported tills. Often in areas away from the mountain sources the large Pleistocene ice sheets became enriched with locally derived materials in addition to those derived from upper catchment areas. Tills containing glacial erratic clasts have been recorded from the Precambrian and Permo-Carboniferous and Cretaceous glaciations in many parts of the world.

Deep Water Rudaceous Deposits

Rudaceous deposits, both terrigenous and carbonate, occur in deep-water environments, ranging from turbidites to mega-boulder complexes. Gravels in turbidites are restricted to the basal part of the graded bed, and increase in size and abundance towards the source. Isolated clasts occur in debris and grain flows. At the other extreme is the 'Wild-Flysch' of Alpine geologists, termed 'olostostroma' by Italians. These are irregularly shaped formations that contain clasts the size of sky-scrapers and jumbo-jets. Such deposits usually occur at the foot of submarine fault scarps, and are associated with tectonic disturbances of violent intensity.

Conclusions

The clasts of rudaceous sediments hold important information about not only the rock types of the hinterland from which they were derived, but also about its geological history. As a terrain is unroofed it will shed progressively older and more lithified clasts into the depositional system. With continuous or discontinuous uplift erosion unmantles progressively older or more changed, often more highly metamorphosed materials, which are transported and deposited in the resultant conglomerates and breccias. Structurally or compositionally weak rocks do not preserve as well as stronger materials and allowance must be made in attempting to reconstruct unroofing histories. Furthermore, it is generally the more chemically stable silica-rich rocks that contribute to rudaceous deposits. Thus, of all of the sediments it is sandstones and cherts that are preserved at the expense of shales and limestones, of metamorphic rocks quartzites are preserved at the expense of slate and schist. Of all of the igneous rocks, pebbles of rhyolite are more usually preserved than those of basalt or gabbro.

Now that our understanding of the processes leading to the formation of rudaceous deposits is fairly advanced, much present research is moving into the field of exploring the geological characteristics of ancient catchments, even to the level of distinguishing separate phases of advance of thrust sheets into an area during orogeny.

The rudaceous rocks have much to offer the sedimentologist and the geological historian. The ability to recognize particular depositional environments in the ancient record and to recreate the conditions at the land surface during mountain-building enables the geologist to postulate the locations of potential metalliferous and hydrocarbon economic resources. Gold and uranium occur in Precambrian rudaceous rocks in Canada, the USA, Brazil and South Africa, wherein those of the Witwatersrand basin are probably the best known. Rudaceous rocks host placer ores in many parts of the world. Because they are composed of clasts, which of their very nature are tough, and therefore of low porosity, Rudaceous rocks are seldom good petroleum reservoirs. But it is as aggregates for road building and construction that unconsolidated rudaceous sediments are economically most important.

See Also

Sedimentary Environments: Alluvial Fans, Alluvial Sediments and Settings; Lake Processes and Deposits; Shoreline and Shoreface Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures; Fluvial Geomorphology; Glaciers; Landslides. **Weathering**.

Further Reading

Bluck BJ (1967) Sedimentation of beach gravels; examples from South Wales. *Journal of Sedimentary Petrology* 37: 128–156.

Bluck BJ (1980) Structure, generation and preservation of upward fining braided stream cycles in the Old Red Sandstones of Scotland. *Transactions of the Royal Society of Edinburgh, Earth Sciences* 71: 29–46.

Bluck BJ (1999) Clast assemblages, bed-forms and structure in beach gravels. *Transactions of the Royal Society of Edinburgh, Earth Sciences* 89: 291–323.

Bluck BJ (2000) Old Red Sandstone basins and alluvial systems of Midland Scotland. In: Friend PF and Williams BPJ (eds.) *New Perspectives on the Old Red Sandstone,* 180, pp. 417–437. London: Geological Society of London.

Bray M, Workman M, Smith J, and Pope DJ (1996) Field measurements of shingle transport using electronic tracers. In: Proceedings of 31st Ministry of Agriculture, Fisheries and Food Conference on River and Coastal Engineering. Keele: University of Keele.

Bull WB (1977) The alluvial fan environment. *Progress in Physical Geography* 1: 222–270.

Cailleux A (1945) Distinction des galets marins et fluviatiles. *Bulletin of the Geological Society of France* 5: 125–138.

Dobkins JE and Folk RL (1970) Shape development on Tahiti-Nui. *Journal of Sedimentary Petrology* 40: 116–203.

Glennie KW (2002) Permian and Triassic. In: Trewin NH (ed.) *Geology of Scotland,* 4th ed, pp. 301–322. London: The Geological Society of London.

Griffiths JC (1967) *Scientific Method in the Analysis of Sediments,* p. 508. New York: McGraw-Hill.

Krumbein WC (1941) Measurement and geological significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Petrology* 11: 64–72.

Lane EW (1955) Design of stable channels. *Transactions of the American Society of Civil Engineers* 120: 1234–1279.

Laming DJC (1966) Imbrication, paleocurrents and other sedimentary features in the lower New Red Sandstone,

Devonshire, England. *Journal of Sedimentary Petrology* 36: 949–959.
Miall AD (1977) A review of the braided-river depositional environment. *Earth Science Reviews* 13: 1–62.
Moss AJ (1962) The physical nature of common sand and pebbly deposits 1. *American Journal of Science* 262: 337–373.
Moss AJ (1963) The physical nature of common sand and pebbly deposits 2. *American Journal of Science* 263: 297–343.
Selley RC (1965) Diagnostic characters of fluviatile deposits of the Torridonian. *Journal of Sedimentary Petrology* 35: 366–380.
Selley RC (2000) *Applied Sedimentology,* 2nd edn. San Diego: Academic Press.
Sneed ED and Folk RL (1958) Pebbles in the lower Colorado River, Texas, a study in particle morphogenesis. *Journal of Geology* 66: 114–150.
Trewin NH and Thirlwall MF (2002) Old Red Sandstone. In: Trewin NH (ed.) *The Geology of Scotland,* 4th edn, pp. 213–251. London: The Geological Society of London.
Wadell H (1935) Volume, shape, and roundness of quartz particles. *Journal of Geology* 27: 507–521.
Williams GE (1968) Neoproterozoic (Torridonian) alluvial fan succession, northwest Scotland, and its tectonic setting and provenance. *Geological Magazine* 138: 161–184.

Sandstones, Diagenesis and Porosity Evolution

J Gluyas, Acorn Oil and Gas Ltd., Staines, UK

Introduction

Sand comprises particles of rock and mineral with a mean grain size between 0.0625 and 2 mm and deposited by sedimentary processes on the Earth's surface (*see* **Sedimentary Rocks:** Mineralogy and Classification).

The composition of sand is highly variable, depending on the source of the sediment and the extent to which weathering and erosion during transport have removed unstable minerals (*see* **Weathering**). As a general rule, sands derived, first cycle, from igneous and metamorphic terrains tend to contain more mineral phases that are unstable under surface and shallow burial conditions than do sands that have been involved in many cycles of erosion, transport, and deposition.

Following deposition, sand may become buried. It may also be lithified (indurated) into sandstone. The process whereby sand becomes sandstone is known as diagenesis (*see* **Diagenesis, Overview**). It includes three distinct components: one mechanical (compaction) and two chemical (cementation and dissolution).

When sands are deposited, they are commonly highly porous and highly permeable. Any given volume of newly deposited sand will contain between 40% and 50% pore space (Figure 1). The permeability of loose sand is enormous, measured in tens to hundreds of darcy. Sandstones are less porous and less permeable, there being a continuous range from the values for sand shown above to sandstones that are non-porous and impermeable.

Grain Size and Sorting

Sand having grain sizes between 0.0625 and 2 mm is further divided into a series of subcategories, from very fine sand at the lower end of the size range to very coarse sand at the upper end of the range. Smaller grains (silt and clay grade) and larger grains (granules to boulders) are defined in **Sedimentary Rocks:** Mineralogy and Classification. Grain size is governed by the grain or crystal size in the provenance area and the degree of abrasion suffered by the sediment *en route* from the source area to deposition. Sorting is a measure of the range of grain sizes in a given sand sample. Well-sorted sand has a narrower range of grain sizes than poorly sorted sand. Sorting within sand is controlled by both provenance and sedimentary process. Surface processes which constantly rework sediment, such as in shallow marine settings (wave

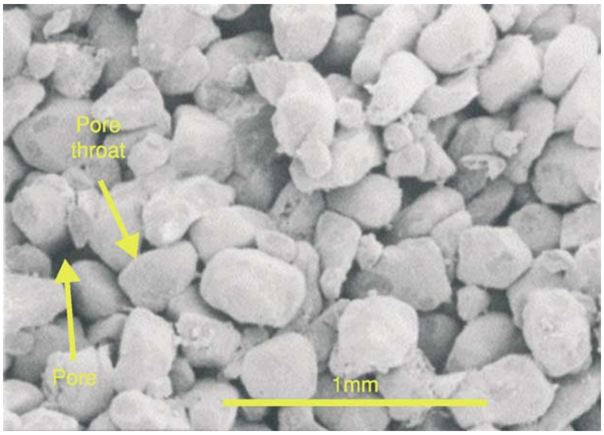


Figure 1 Well-sorted, medium-grained, aeolian, uncemented sandstone from the Cleeton Field, UK southern North Sea, showing well-developed pores and pore throats; scanning electron photomicrograph. From Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience*. Oxford: Blackwell Science.

and tidal action), are likely to produce better sorted sand than, for example, gravity-driven processes such as debris flow. However, if the sediment provenance area comprises well-sorted sand, so too will the deposition area, irrespective of the specific sedimentary process responsible for deposition.

The grain size and sorting of sand control its initial permeability and the sorting of sand controls its initial porosity. Moreover, as compaction and diagenesis proceed, the 'memory' of the depositional characteristics can be retained, such that the sands that were the most permeable at deposition become the most permeable sandstones after compaction and cementation.

Compaction

Loose sand compacts easily. During initial burial, much of the compaction is taken up by the rearrangement of grains. Simple burial combined with seismic shock will turn loose sand into consolidated sand. The amount of porosity lost will depend largely on how well sorted the sand is. In poorly sorted sand, more porosity will be lost than in well-sorted sand – small grains fill in between larger ones. As burial continues, rough edges tend to be knocked off grains, so aiding greater compaction. At deeper levels (about 1–4 km), the sand begins to behave like a linearly deformable solid. Deeper still, plastic deformation is probably more common. The boundaries between the occurrence of these processes will vary from sand to sand and basin to basin and, in some instances, may be gradational.

The net outcome of all the above processes is that sands compact when stressed, but decompact very little when the stress is released. This means that, for compacted but uncemented sandstone at the Earth's surface, it is possible to calculate the maximum stress suffered in any previous burial phase. Such a stress calculation can be used to provide an estimate of the maximum burial depth.

In the geological literature, there are a large number of so-called compaction curves for sandstones. Alas, most of these curves are porosity/depth plots, rather than porosity/stress plots. As such, the great swathes of data on these plots include, but do not differentiate, the effects of fluid overpressure and sandstone cementation. However, experimental data are available on the way in which sands compact and, for clean quartzose or arkosic sandstones, these data have been used to formulate a compaction equation

$$\Phi = 0.5 \exp\left(\frac{-10^{-3}z}{2.4 + 10^{-4}z}\right)$$

where z is in metres.

In this equation, porosity is expressed as a fraction (i.e., <1) and the equation is calibrated to a normal hydrostatic pressure gradient. If the system is overpressured, the pressure borne by the grains is less than in a hydrostatic system and an effective depth must be calculated. As a simple rule of thumb for typical burial depths of 2–4 km, 1 MPa of overpressure is equivalent to about 80 m less burial. The equation is well tested, predicting porosity to within ±3% at 95% confidence limits.

Sands that contain easily squashed grains, such as glauconite or mica, and those rich in matrix clay lose porosity much more readily at a given applied stress. Empirical curves linking porosity to applied stress have also been constructed for sands with various quantities of easily deformed grains.

Detrital Mineralogy

Quartz is the most common mineral found in sands and sandstones (*see* **Minerals:** Quartz). Feldspar and lithic (rock) fragments are also common in most sandstones and, as a consequence, these three components are often used to classify sandstones. The QFL plot sums the three components (quartz, feldspar, lithics) to 100% on a triangular diagram (Figure 2). The triangle is divided into fields: quartz arenite, sublithic arenite, arkosic arenite, etc. There is no strict convention as to whether polycrystalline quartz is included with (monocrystalline) quartz or with lithic fragments, although it is common to label the diagram so as to show where the polycrystalline quartz has been included. The feldspar component includes both alkali and plagioclase, whilst the lithic component can include sedimentary, igneous, and metamorphic

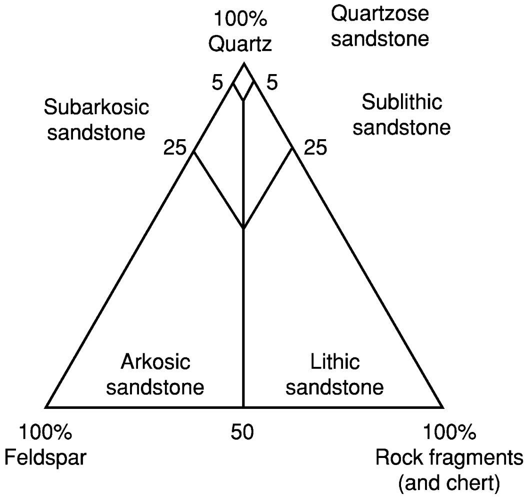


Figure 2 Sandstone classification using the QFL (quartz, feldspar, lithic fragments) system.

Table 1 Summarized mineralogy of the Upper Jurassic Brae Formation (Miller Field) and Middle Jurassic Etive Formation (Columbia Field), UK North Sea

	Brae formation	*Etive formation*
Quartz + polycrystalline quartz	86.3	74.1
Total feldspar	2.0	3.3
Mica	1.0	1.9
Other detrital minerals	1.6	1.1
Matrix clay	0.9	1.2
Organic matter	1.1	0.4
Calcite cement	0.2	1.8
Siderite cement	0.0	0.0
Quartz cement	5.3	3.9
Pyrite cement	0.2	0.4
Kaolinite cement	0.1	6.1
Illite cement	1.3	5.8
Number of samples	56	18

rock fragments. Although the QFL diagram is widely used, it may not be adequate for some sandstones, in which case alternative classifications and descriptions may be employed, e.g., micaceous sandstone, glauconitic sandstone, shelly sandstone, and tuffaceous sandstone. Most sandstones contain between about 10 and 20 distinct mineralogical and rock components (Table 1).

Diagenetic Mineralogy

Minerals that precipitate during diagenesis are commonly referred to as cements. A wide variety of cements have been identified in sandstones. Some are common, others are rare. The most common cements, in decreasing order of abundance, are quartz (Figure 3), carbonates, zeolites, clays, and evaporite minerals (Figure 4). Less common cements include barite, celestite, opal, amorphous silica, albite, haematite, pyrite and other sulphides, apatite, and many more. A systematic study of more than 100 case histories of diagenesis from a range of sandstones worldwide has revealed several recurring patterns, in addition to demonstrating the relative abundance of the five mineral groups (quartz to evaporites) listed above (Figure 5). There appear to be five common styles of diagenesis that can be seen in sandstones of different ages from across the globe. The mineral associations that form these styles are as follows.

- Quartz-dominated diagenesis with lesser quantities of clay minerals and carbonate minerals that precipitated after the quartz.
- Clay mineral-dominated diagenesis with lesser quantities of carbonate minerals and quartz or zeolite that precipitated after the quartz.

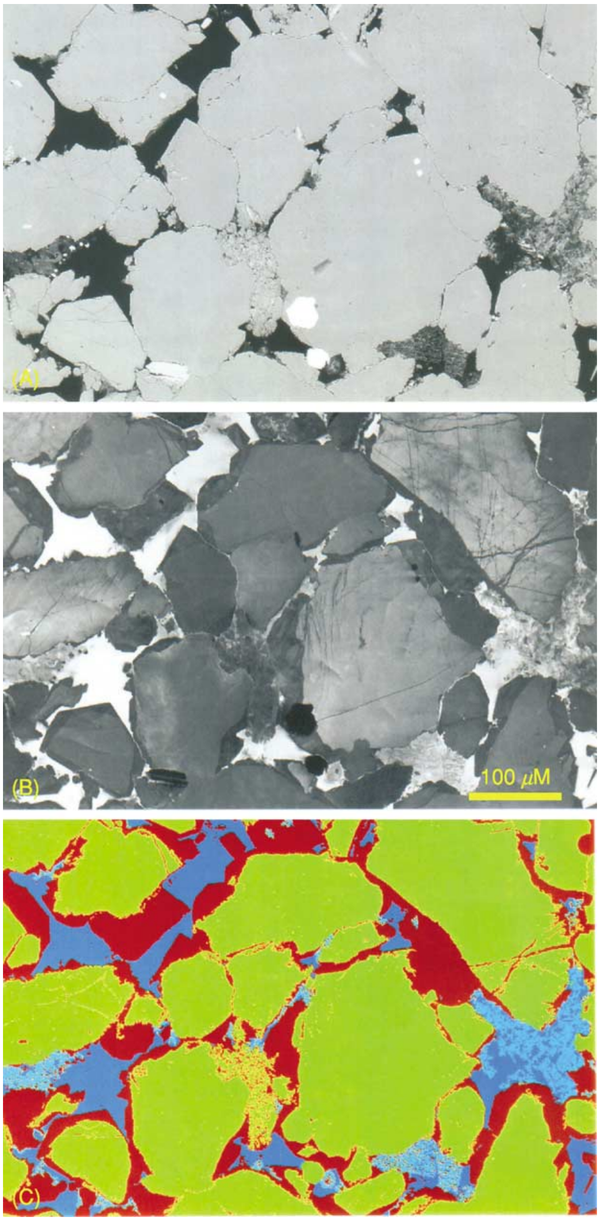


Figure 3 Quartz-cemented quartzose sandstone, Miller Field, North Sea. (A) Backscattered scanning electron microscopy (BSEM) photomicrograph. Minerals with highest mean atomic number appear white and those with lowest mean atomic number appear black. Pore space is filled with a low mean atomic number resin, so appearing black. (B) Scanning electron microscopy (SEM) cathodoluminescence (CL) image of the same field of view as in (A). There are a few impurities and lattice defects in the syntaxial quartz cement and fracture fills, and so these areas appear darker than the detrital grains. (C) Combined BSEM and CL images with false colour added. Green, quartz grains; red, quartz cement; blue, pore space. The areas of pale blue are resin-impregnated kaolinite plates and partially dissolved feldspar grains. From Gluyas JG, Garland CR, Oxtoby NH, and Hogg AJC (2000) Quartz cement; the Miller's tale. In: Worden RH and Morad S (eds.) *Special Publication of the International Association of Sedimentologists 29*, pp. 199–218. Oxford: International Association of Sedimentologists.

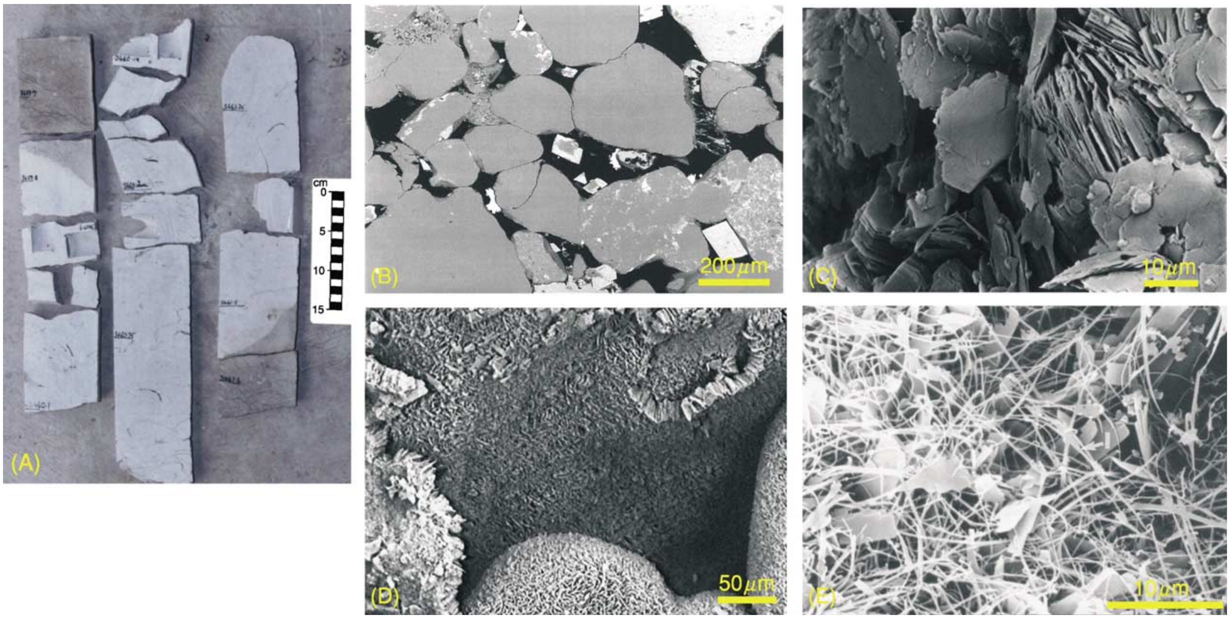


Figure 4 Common mineral cements in sandstones. (A) Spherical calcite concretion in core, Upper Jurassic, Ula Formation, North Sea; scale, 15 cm. (B) Rhombs of dolomite cement (with ferroan dolomite rims), Lower Permian Rotliegend Sandstone, North Sea; backscattered scanning electron microscopy (BSEM) photomicrograph. (C) Pseudohexagonal plates of kaolinite, Upper Jurassic, Magnus Member, North Sea; scanning electron microscopy (SEM) photomicrograph. (D) Grain coating chlorite cement, Cretaceous, Tuscaloosa Sandstone, Louisiana, USA; SEM photomicrograph. (E) Pore bridging illite cement, Triassic Skagerrak Formation, North Sea; SEM photomicrograph. (C)–(E) Secondary electron microscope images. Photographs reproduced courtesy of BP.

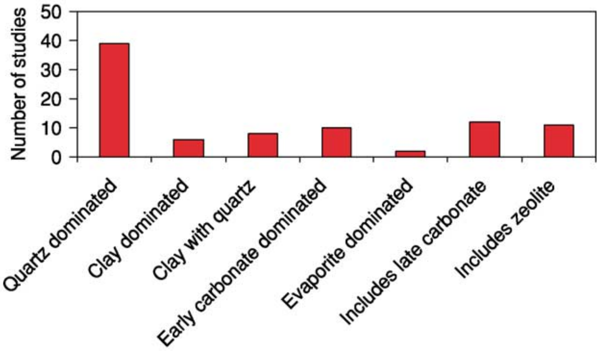


Figure 5 Styles of diagenesis summarized from a worldwide survey. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

- Grain coating clay precipitated soon after deposition and wholly or partially inhibiting subsequent precipitation of quartz and carbonates.
- Carbonate cements precipitated soon after deposition.
- Zeolites precipitated with clays, followed by carbonates and opal or quartz.

The reasons why such associations are common are investigated in the following sections.

Diagenetic Sequence

From the associations listed above, it is clear that diagenesis has a chronology. Observations made under the microscope (optical microscopy, scanning electron microscopy (SEM), transmission electron microscopy (TEM), backscattered scanning electron microscopy (BSEM), cathodoluminescence (CL), **Analytical Methods:** Geochemical Analysis (Including X-Ray)) allow mineral precipitation (and dissolution) events to be arranged in a temporal sequence. It is also possible to include the relative timing of compaction within such sequences. An example of the mineral precipitation sequence for the Middle Jurassic Brent Sandstone from the North Sea is shown in Figure 6. The sequence of diagenetic events was deduced from observations made using thin sections and SEM. Although such diagrams are useful in conveying the sequence of events, they often accidentally convey two other impressions, neither of which is likely to be true. In the absence of quantitative data on when and where cements precipitated, it is common to display the high-abundance cements as having taken the longest to precipitate. This is probably an error. It is also common for the sequence of events to fill up all the available time from the deposition of the sand to the present day. This is certainly an error. A similar diagenetic sequence diagram is

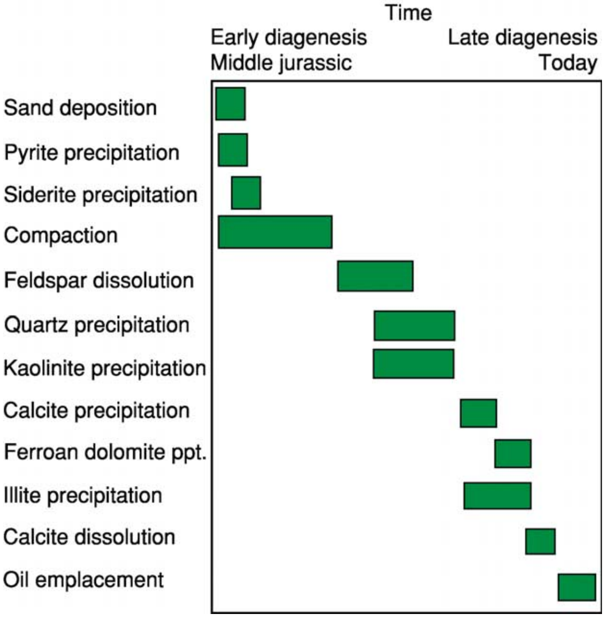


Figure 6 Diagenetic sequence deduced from thin section and scanning electron microscopy (SEM) analysis, Brent Group, North Sea. Adapted from Eglington G, Curtis CD, McKenzie DP, and Murchison DG (1985) Geochemistry of buried sediments. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences* 315.

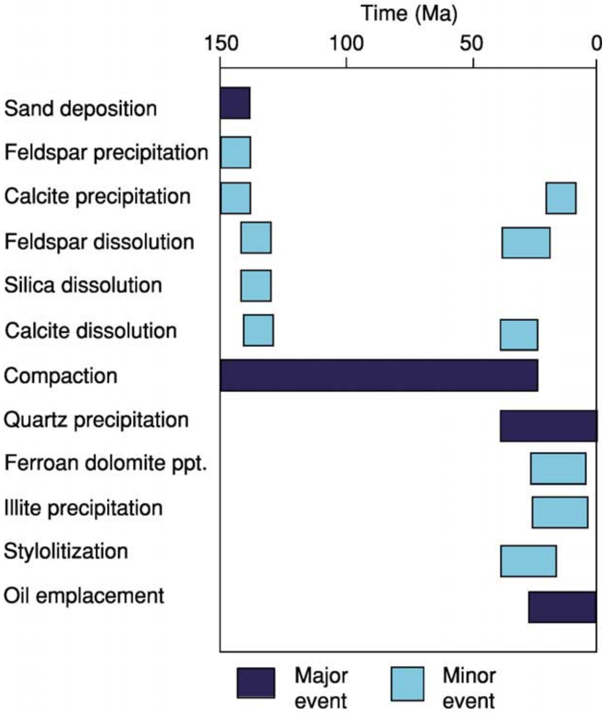


Figure 7 Diagenetic sequence calibrated to absolute time using geochemical and isotopic analysis in addition to conventional petrographical observations, Ula Formation, North Sea. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

shown in Figure 7, although here the duration of events has been constrained using additional data from geochemical, isotopic, and petrographical analyses.

There are many descriptive terms used to refine the qualitative description of diagenesis. Thus, it is possible to read of 'early carbonate', 'mesogenic quartz', 'burial cements', 'late ferroan dolomite', and so on. It is all too easy to be confused by the plethora of terms, particularly when some are contractions of long, although better, descriptive terms. For example, 'early carbonate' is often used to describe calcite or dolomite that precipitated soon after the deposition of a sand, before significant compaction, and whilst the sand was still in contact with surface or near-surface formation water.

Mineral Dissolution

Dissolution of either grains or cements in a sandstone leads to the development of secondary porosity. In the 1970s and 1980s, many publications suggested that mineral dissolution was a key process whereby significant porosity could be created at depth. Such porosity could then be occupied by petroleum. A range of dissolution mechanisms were proposed to explain this. More recently, new work has indicated that none of these mechanisms is likely to be capable of generating significant secondary porosity in the deep subsurface. That is to say, secondary porosity is rarely so extensive as to significantly improve reservoir quality.

Many minerals will dissolve during deposition and subsequent diagenesis. The only requirement is that the connate (formation) water that surrounds the grains is undersaturated with respect to the mineral in question. However, proof that a particular mineral has dissolved during diagenesis is often more difficult to come by. Grains that have partially dissolved are positive proof that secondary porosity has been created, as is mouldic porosity within otherwise tight rock (Figure 8). However, so-called oversized pores are commonly cited as evidence for the complete dissolution of grains and, although such claims are sensible, proof of secondary porosity creation is lacking.

Advocates of secondary porosity often claimed the wholesale dissolution of mineral cements (particularly calcite) during deep burial, rendering once cemented, low-porosity sandstones highly porous and permeable. Popular amongst the various processes invoked for such widespread dissolution was appeal to

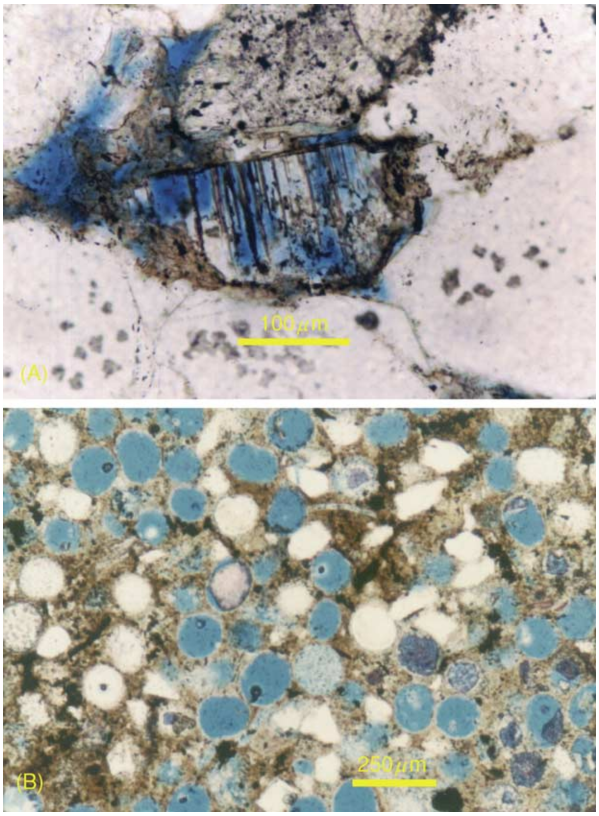


Figure 8 (A) Skeletal feldspar grain (blue, porosity), Upper Jurassic Fulmar Sandstone, North Sea; plane-polarized light photomicrograph. (B) Sponge spiculite sandstone in which many of the spicules have dissolved (blue pore space) and microcrystalline quartz has precipitated in the original pore space (stained brown by oil), Jurassic Alness Spiculite, North Sea; plane-polarized light photomicrograph.

organic acids created during the initial phases of oil source rock maturation. The hypothesis invoked such acids racing ahead of the migrating oil, leaching carbonates as they went. Oil then followed in the newly created porosity. In an anthropomorphic twist, this became known as the 'John the Baptist Hypothesis' – porosity created ahead of the oil coming. Although appealing and superficially elegant, there is scant evidence to support such a hypothesis. Quite apart from the difficulties of creating sufficient acid and getting it to the reservoir where secondary porosity is required, it remains difficult to find convincing evidence of large-scale, large-volume mineral dissolution in the deep subsurface. A partial exception to this rule occurs in association with unconformities. There is commonly ample evidence of porosity creation due to reaction between rock and meteoric water beneath unconformity surfaces. The improved porosity is then commonly (partially) retained during reburial of the sequence (Figure 9).

Diagenesis Quantified

The foregoing text describes diagenesis in terms of minerals that can precipitate and others which dissolve. It also investigates the relative timing of diagenetic events. However, in order to understand how diagenetic processes operate, it is important to determine when and where minerals precipitate and dissolve and the quantities of matter involved in such reactions.

Before about 1990, there were few published examples in which the absolute date of precipitation, temperature of precipitation, and isotopic composition of the host fluid were known. A wide range of analytical techniques are now available which enable quantitative or semi-quantitative data to be gathered on the 'when' and 'where' of diagenesis. The most commonly used techniques for obtaining quantitative data are listed below. A description of the methods is given in **Analytical Methods:** Geochemical Analysis (Including X-Ray).

- Fluid inclusion analysis: homogenization temperature data obtained from aqueous inclusions within diagenetic minerals can be used to estimate trapping temperatures and hence the precipitation temperature of the minerals (Figure 10) (*see* **Fluid Inclusions**).
- Radiogenic dating: potassium–argon, argon–argon, and several other methods can be used to give absolute precipitation dates for a few diagenetic minerals, such as illite (clay) and feldspar (Figure 11).
- Stable isotope analysis: δ^{13}C, δ^{18}O, δ^{24}S, and δD (deuterium) are in common usage for helping to identify the source of elements, together with the temperature and composition of formation water during the precipitation of carbonates, sulphates, clay minerals, quartz, and sulphides (Figures 12 and 13).

The techniques outlined above allow some attempt to be made at quantifying when, at what temperature (and depth), and from what formation water a particular mineral precipitated. None of the methods addresses directly how much material moved and over what scale such movement took place during diagenetic processes. There have been many papers questioning whether sandstone diagenesis occurs in open or closed systems. There is no consensus. It is, however, important to try to answer the question because it has a direct bearing on the prediction of porosity (and permeability) ahead of drilling petroleum exploration wells.

One method which has been used to address the movement of matter during diagenesis is whole rock geochemistry. The basic premise of the method is to compare portions of the sandstone that have and have

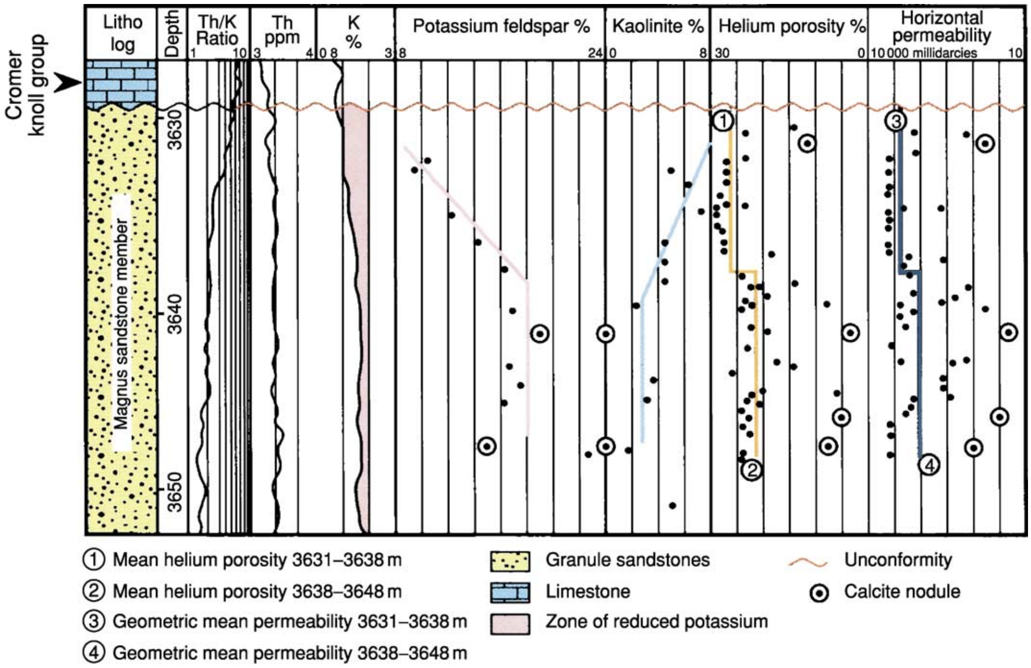


Figure 9 Secondary porosity creation beneath an unconformity during exposure and weathering, Upper Jurassic Magnus Member Sandstone, North Sea. Reproduced courtesy of BP.

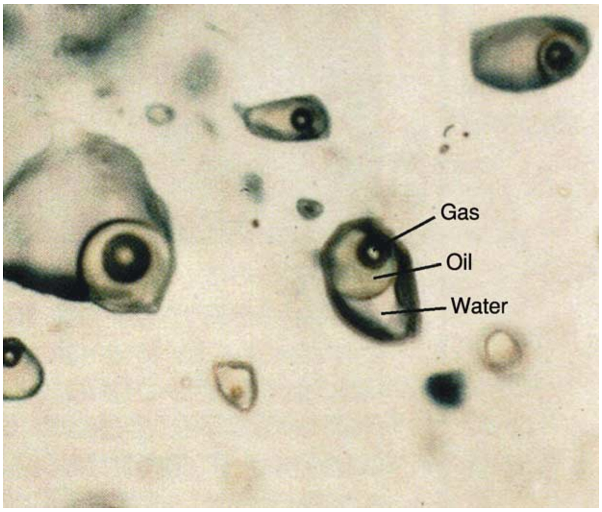


Figure 10 Fluid inclusion within mineral cement. On heating, the fluid phases within the inclusion homogenize. For aqueous inclusions, the homogenization temperature commonly equates to the minimum trapping temperature. Freezing the same inclusions yields a measure of the salinity of the trapped fluid. For those inclusions containing petroleum, ultraviolet fluorescence analysis can give a measure of the maturity of the oil. Moreover, if abundant, oil can be extracted from the inclusions and analysed. Reproduced courtesy of BP.

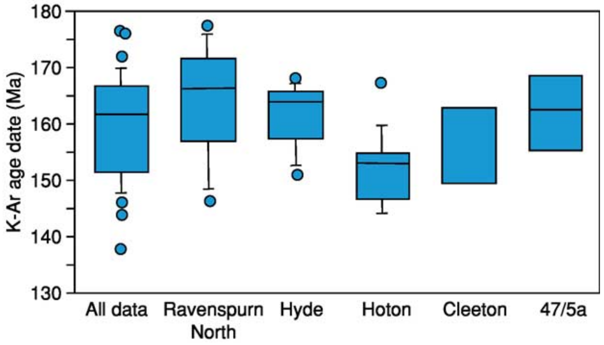


Figure 11 Potassium–argon (K–Ar) age distributions for illite extracted from the Permian Rotliegend Sandstone within the North Sea Gas Fields. The box plots show modes, 10th, 25th, 75th, and 90th percentiles. Modified from Emery D and Robinson AG (1993) *Inorganic Geochemistry: Applications to Petroleum Geology*. Oxford: Blackwell Science.

not undergone diagenesis. There are several possibilities for sand which has been unaffected by deep diagenesis. Part of the formation may never have been significantly buried or, alternatively, part of the formation may have been preserved from the effects of diagenesis. The Garn Formation from the Middle Jurassic of the Norwegian Sea area provides a good example, where little buried sandstone from the Draugen Field at 1.6 km can be compared with deeply buried sandstone in the Smørbukk Field at >4 km. The deep samples are relatively enriched in silica and depleted in potash compared with the shallow samples when normalized to TiO_2 content which is essentially immobile. The same results have been obtained when comparing sand trapped within calcite

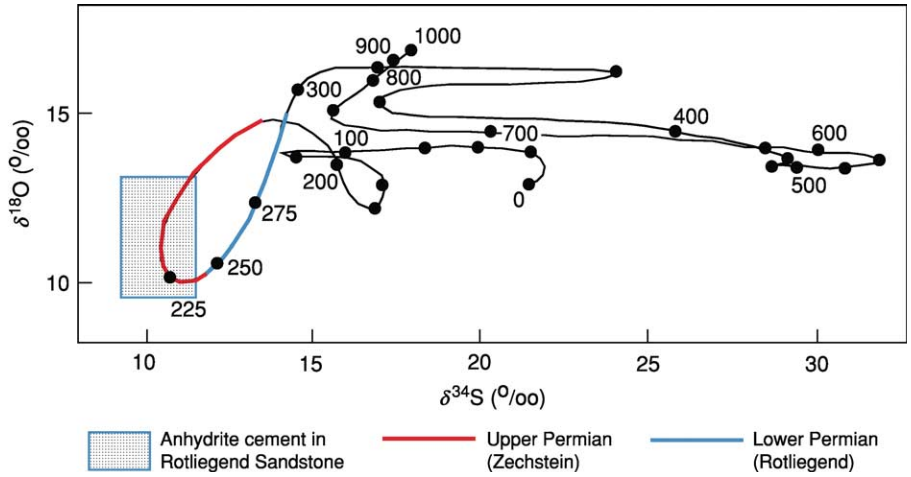


Figure 12 Global secular curve for sulphur and oxygen isotope covariance in marine-derived sulphate minerals. The Permian interval is highlighted, as is the distribution of data obtained from anhydrite and barite cements in the Rotliegend Sandstones of the Amethyst Field (North Sea). The sulphur isotope data clearly indicate derivation from the overlying Zechstein. Modified from Gluyas JG, Jolley EJ, and Primmer TP (1997) Element mobility during diagenesis: sulphate cementation of Rotliegend sandstones, Southern North Sea. *Marine and Petroleum Geology* 14: 1001–1012.

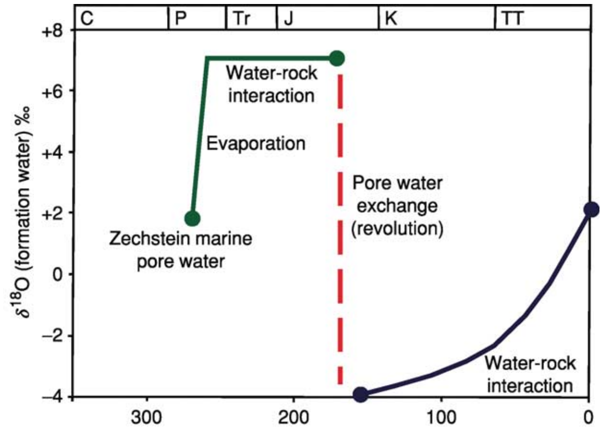


Figure 13 Pore water evolution for the Permian Rotliegend Sandstone of the Village Fields Area (North Sea) deduced from analysis of stable isotope ratios, fluid inclusion homogenization temperatures, and radiometric dating in and of mineral cements.

concretions, which precipitated soon after deposition, with surrounding sandstones. It is tempting to deduce from such studies that (some) sandstones import silica and export potash during diagenesis. Critics of such studies point to the data obtained from formation water isotopic analysis, which have been used to suggest that the water budget is severely limited in the deep subsurface and there is insufficient water to transport the observed cement volumes to the site of precipitation. Others researchers invoke local sources of silica from pressure dissolution along stylolite seams, although this too is not a panacea, as many sandstones are without such pressure dissolution phenomena.

Diagenesis and Petroleum Emplacement

A possible relationship between diagenesis and petroleum emplacement has already been touched upon in the section on 'Mineral Dissolution'. Here, the likelihood of significant porosity creation by organic acids was questioned. Much more controversial is the effect on diagenesis caused by oil emplacement. There are two extreme viewpoints: oil emplacement halts diagenesis by displacing the formation water, and diagenesis continues unaffected by oil emplacement. It is probable that the truth lies somewhere between these two extremes.

Ample evidence exists of continued diagenesis in the presence of (possibly) low oil saturations. Oil-filled fluid inclusions occur in many mineral cements (Figure 10). However, quantitative analysis of these same inclusion distributions often indicates that the presence of petroleum inhibits mineral precipitation. Studies on several sandstones, including those from the Upper Jurassic of the North Sea, have shown that cementation and petroleum migration commonly occur at the same time. In some papers, this has been referred to by the acronym SMAC (synchronous migration and cementation) and in others as the 'Race for Space'. Oilfields so affected have highly porous sandstone at their crest and low-porosity sandstone at the oil–water contact. The rate of porosity decline as a function of depth is perhaps twice that of the regional porosity gradient determined from water-bearing sandstones. In the instance of the North Sea sandstones mentioned above, the regional gradient is

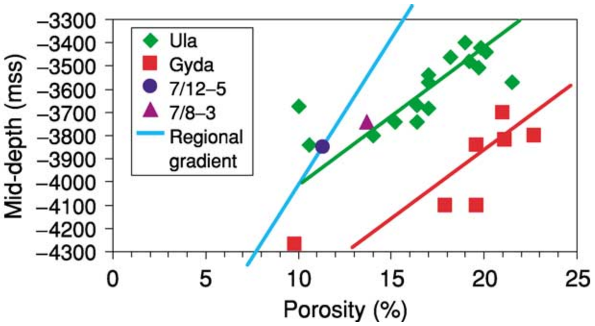


Figure 14 Porosity/depth relationships for fields within the Ula Trend (Norwegian Central Graben). Intrafield porosity gradients are about twice those observed for water-bearing sandstones (regional gradient) in the same area. It is possible that oil emplacement limited cementation within oil-bearing reservoir intervals. From Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69.* Tulsa: American Association of Petroleum Geologists.

8% porosity loss per extra kilometre of burial depth, whilst that seen in the Ula and Gyda Fields is 16% km^{-1} (Figure 14). This same pattern of porosity loss also occurs within individual coarsening-up sequences within the reservoir interval, particularly in the direction of known mature oil source. Detailed observations on the distribution of petroleum-filled fluid inclusions indicate an exponential decline in such inclusions from field crests to field flanks, with the same sort of distribution occurring in the individual retrogradational cycles. The sympathetic patterns of porosity and fluid inclusion distribution are most easily explained by considering that diagenesis was progressively retarded as the fields filled with oil. The coarse, permeable tops of the retrogradational cycles formed the natural migration pathways of oil into the structures and these, too, had retarded diagenesis.

Impact of Diagenesis on Porosity and Permeability

From a physical perspective, sands and sandstones comprise two basic components: solid and void. In the preceding sections, the intrinsic properties of the solid component, its grain size and sorting, and its mineralogy have been examined. The void space is now examined. The void in a sand or sandstone is porosity, an intricate network of pores connected by pore throats (Figure 1). At the Earth's surface, the void space is commonly filled by a combination of water and air (depending on the elevation of the sandstone relative to the local water table). In the subsurface, the void space can, in addition to water, contain petroleum (oil and/or hydrocarbon gas) and possibly non-petroleum gas (CO_2, H_2S, N_2, and He_2).

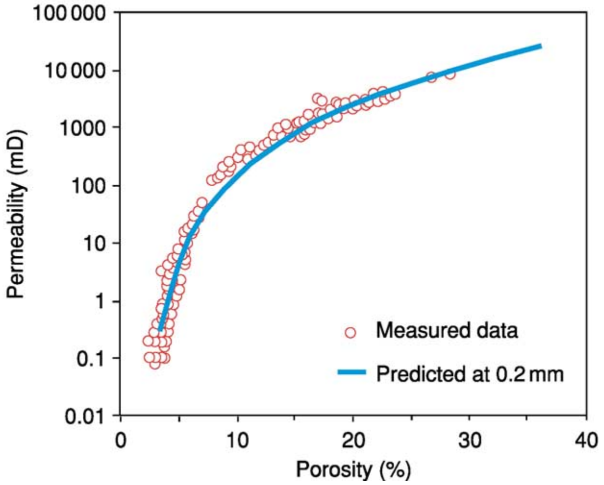


Figure 15 Measured relationship between porosity and permeability for the Fontainbleau Sandstone, and modelled relationship for a monodisperse sphere pack with a grain size of 0.2 mm. Reproduced from Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491–501.

Porosity is measured as a percentage (or fraction) of the rock plus void. For sands, porosity commonly lies in the range 35–50%. Well-sorted sands are more porous than poorly sorted sands, and loosely packed sands are more porous than tightly packed sands.

Sandstones commonly have a lower porosity than sands. This is because compaction and mineral precipitation (diagenesis) reduce the pore space between grains. In extreme instances, the porosity of sandstone can be close to 0%.

The permeability is a measure of the rate at which fluid can be transmitted through a porous medium. It's unit is the Darcy (D), such that a rock has a permeability of 1 D if a potential gradient of $1\text{atm}\times10^{-2}\,\text{m}$ induces a flow rate of $10^{-6}\,\text{m}^3\,\text{s}^{-1}\,10^{-4}\text{m}^{-2}$ and a liquid viscosity of 1cP. For loose sands, the unit of permeability is the darcy, whereas, for sandstones, a more convenient unit is the millidarcy. There is no particular reason why porosity and permeability should be related, other than that, for a rock to have non-zero permeability, it must also have non-zero porosity. However, for individual sands and sandstones, porosity and permeability are commonly positively correlated (more porous sandstones tend to be more permeable than less porous sandstones). Where it does exist in granular porous media, the correlation between porosity and permeability commonly reflects the variation in one or possibly more of the components. For example, the Fontainbleau Sandstone of the Paris Basin is essentially monodisperse (perfectly sorted) and uncompacted. However, the quantity of cement varies between 0% and about 40%. For this

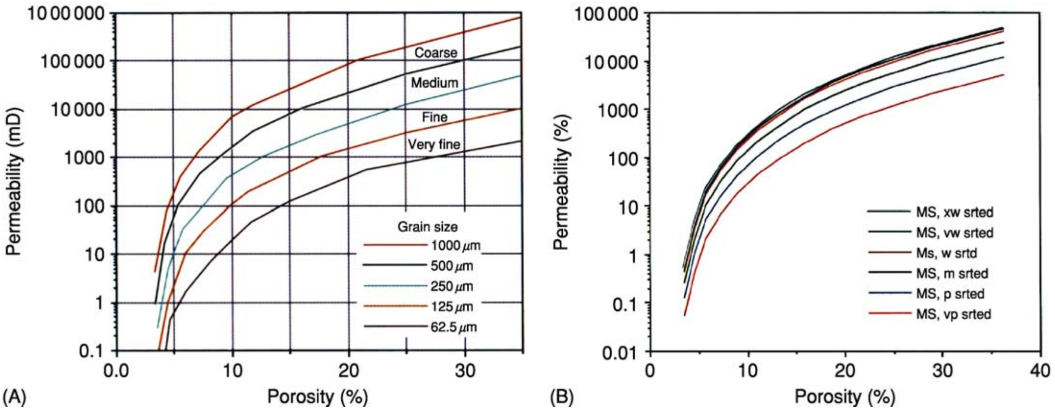


Figure 16 (A) Modelled relationship between porosity and permeability for monodisperse (perfectly sorted) sands of different grain size. (B) Modelled relationship between porosity and permeability for a medium-grained sand with different sorting characteristics (xw, vw, w, m, p, and vp srted denote extra well, very well, well, medium, poorly, and very poorly sorted, respectively). Reproduced from Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491–501.

sandstone, there is a non-linear correlation between porosity and permeability on a semi-logarithmic plot (Figure 15). The increasing rate of decline in permeability at low porosity is due to progressive closure of the pore throats between pores.

The relatively simple relationship between porosity and permeability for the Fontainbleau Sandstone has been used as the foundation for a comprehensive predictive model for permeability based on a real physical model of a porous medium. The model combines data derived from the perfectly sorted porous medium with empirical curves linking porosity and permeability for less well-sorted sands (Figure 16). Cements are then modelled as grain rimming or pore filling, and solid (such as quartz or carbonate) or microporous (clays).

Controls on Diagenetic Processes

In broad terms, near-surface diagenetic processes are much better understood than those occurring at depth. Geochemical and isotopic studies have revealed the importance of bacterial reactions in modifying pore water and inducing the precipitation of carbonates, oxides of iron and manganese, and sulphides.

For the deep subsurface, we know less about what triggers diagenesis, although we can, as shown above, determine when and under what conditions diagenetic reactions occurred. A recurrent observation is that major diagenetic events commonly accompany or follow immediately after significant geological events. This is almost self-evident in the case of mineral dissolution beneath unconformities, but in other situations it is a little more subtle. For example, most of the clay and carbonate minerals in the Permian Rotliegend Sandstone of the southern North Sea precipitated towards the end of the Jurassic, a time of major rifting in the area. At the same time, there was a fundamental change in the connate water from Zechstein (evaporated seawater) derived to meteoric, yet saline, water. There was also a dramatic loss of overpressure from the reservoir system (the reservoirs are normally pressured today). It is tempting to conclude that the rifting led to failure of the salt seals above the Rotliegend, and massive pore water revolution, so causing cementation. In contrast, the Middle Jurassic Brent Sandstone over much of the northern North Sea was cemented at around the Paleocene to Eocene boundary. This, too, may have been associated with the ingress of meteoric water as the rift shoulder became elevated. Although it is possible that such external factors were the cause of cementation in these two sequences, it seems probable that the degree and style of cementation was controlled by the conditions within any particular sandstone (temperature, pressure, mineralogical composition).

See Also

Analytical Methods: Fission Track Analysis; Geochemical Analysis (Including X-Ray); Geochronological Techniques. **Diagenesis, Overview**. **Fluid Inclusions**. **Minerals:** Feldspars; Quartz. **Petroleum Geology:** The Petroleum System. **Sedimentary Rocks:** Mineralogy and Classification. **Sedimentary Processes:** Fluxes and Budgets. **Weathering**.

Further Reading

Burley SD and Worden RH (2003) *Sandstone Diagenesis Recent and Ancient, Reprints Series, International Association of Sedimentologists,* vol. 4. Oxford: Blackwell Science.

Cade CA, Evans IJ, and Bryant SL (1994) Analysis of permeability controls: a new approach. *Clay Minerals* 29: 491–501.

Eglington G, Curtis CD, McKenzie DP, and Murchison DG (1985) Geochemistry of buried sediments. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences* 315.

Emery D and Robinson AG (1993) *Inorganic Geochemistry: Applications to Petroleum Geology.* Oxford: Blackwell Science.

Gluyas JG and Swarbrick RE (2003) *Petroleum Geoscience.* Oxford: Blackwell Science.

Gluyas JG, Garland CR, Oxtoby NH, and Hogg AJC (2000) Quartz cement; the Miller's tale. In: Worden RH and Morad S (eds.) *Special Publication of the International Association of Sedimentologists 29*, pp. 199–218. Oxford: International Association of Sedimentologists.

Gluyas JG, Jolley EJ, and Primmer TP (1997) Element mobility during diagenesis: sulphate cementation of Rotliegend sandstones, Southern North Sea. *Marine and Petroleum Geology* 14: 1001–1012.

Kupecz JA, Gluyas JG, and Bloch S (1997) *Reservoir Quality Prediction in Sandstones and Carbonates, American Association of Petroleum Geologists' Memoir 69*. Tulsa: American Association of Petroleum Geologists.

MacDonald DA and Surdam RA (1984) *Clastic Diagenesis, American Association of Petroleum Geologists' Memoir 37*. Tulsa: American Association of Petroleum Geologists.

Morad S (1998) *Carbonate Cementation in Sandstones, Special Publication of the International Association of Sedimentologists 26*. Tulsa: Blackwell Science.

Selley RC (2000) *Applied Sedimentology*, chap. 8. San Diego: Academic Press.

Worden RH and Morad S (2000) *Quartz Cementation in Sandstones, Special Publication of the International Association of Sedimentologists 29*. Oxford: Blackwell Science.

SEISMIC SURVEYS

M Bacon, Petro-Canada, London, UK

Introduction

Seismic methods study the subsurface by generating seismic waves and observing the way that they propagate through the Earth. Various methods of field acquisition and data processing are used, mainly with the objective of producing cross-sections through the subsurface that can be interpreted in geologically meaningful ways. The methods are particularly widely used in the oil and gas industries.

The type of wave most often used for seismic investigation is a low-frequency sound wave. This is usually called a P wave; during its passage, individual particles oscillate backwards and forwards in the direction that the wave is travelling, so that the wave consists of alternating compressions and rarefactions. The velocity at which the wave travels depends on the rock through which it is passing, and is related to the mineral constituents, the amount and geometry of the porosity, and the type of fluid contained in the pore space. Another type of wave sometimes used is the shear (or S) wave, where the particles vibrate at right angles to the direction in which the wave travels. This type of wave cannot travel through fluids. In rocks, its velocity is affected by similar factors to those that influence P-wave velocity, except that it is relatively insensitive to the type of fluid in the pore space.

Seismic reflection is the method most commonly used. The basic idea is shown in Figure 1. Seismic P waves are generated by a source (such as a small explosive charge) at the ground surface. They travel down through the Earth, are reflected at boundaries between rock layers, and travel back to the surface, where they are detected by a receiver (similar to a microphone, but sensitive to low frequencies down to 5 Hz) and recorded. The time taken for the wave to travel from source to receiver tells us the depth of the reflecting boundary, and, by repeating the measurement at a series of points, it is possible to map the reflecting surface. The principle is similar to the way a

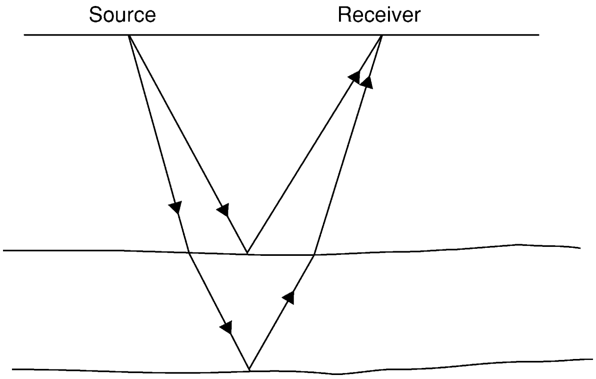


Figure 1 Principle of seismic reflection: seismic waves are generated by a surface source, are reflected at boundaries between rock layers, and are detected and recorded by a receiver at the surface.

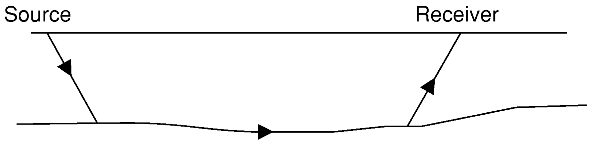


Figure 2 Principle of seismic refraction: a seismic wave travels between a surface source and a receiver along a subsurface boundary across which there is an increase in seismic velocity.

ship's echo sounder is used to measure the distance to the seabed.

Other types of seismic wave can travel along boundaries between layers where there is an increase in wave velocity, and this is the basis of the seismic refraction method (Figure 2). A distance much larger than the depth of investigation separates the source and receiver. The travel time of the seismic signal is used to map the deep interface. This method is usually able to detect only a few such surfaces, across each of which there is a large velocity increase. It is used both for large-scale crustal studies and for shallow engineering investigations. The seismic reflection method is able to detect much more detail, typically allowing us to see many tens or even hundreds of reflecting surfaces. For this reason, it has become the method of choice for the subsurface investigation of sedimentary basins, particularly for petroleum exploration.

Most seismic reflection work uses sources and receivers at or near the surface, and this is what I shall proceed to discuss in detail. It is also possible to acquire data using a surface source and receivers in a borehole. The advantage of such a vertical seismic profile (VSP) is that a particularly detailed image of the subsurface is obtained, which can be closely tied to the drilled succession.

Reflecting Interfaces

What determines how well an interface between two layers will reflect seismic waves? There is a characteristic of a material called its acoustic impedance; this is the product of the seismic velocity and the density of the material. The amplitude of the reflected signal is proportional to the contrast in acoustic impedance across the interface. The formula shown in Figure 3 applies to P waves at an incident angle of 0°, when the seismic wave travels perpendicularly to the reflecting interface; at larger angles of incidence the formula is more complicated and involves a combination of S-wave and P-wave velocities. The acoustic impedance of a rock type depends on many factors. In a sandstone, for example, it will depend on porosity, cementation, and clay content; in a shale it will depend on the degree of

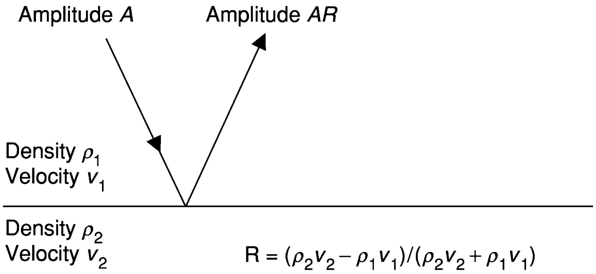
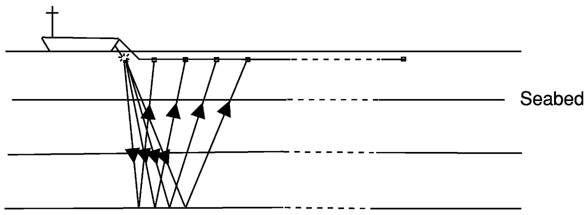


Figure 3 The amplitude of the reflection from an interface depends on the contrasts in density and seismic velocity across it.

compaction. In practice, many interfaces between different lithologies will have enough acoustic impedance contrast to cause appreciable seismic reflection. The interface needs to be sharp rather than gradational. This requirement is not onerous, however; the transition has to take place over a distance very much less than the seismic wavelength. Since the seismic signal typically has a frequency of 30 Hz and the seismic velocity in a sand or shale is typically 3 km s^{-1}, a typical seismic wavelength is about 100 m. A lithological transition over a vertical distance of a few metres will therefore be seen as sharp by the seismic wave. A further requirement is that the reflecting interface should be laterally continuous over distances similar to the seismic wavelength. In sedimentary basins, there are often many interfaces that meet these criteria and are therefore a suitable target for the seismic reflection method. It is sometimes possible to use the method to investigate the internal structures of metamorphic or igneous rocks, but results are often poor owing to a lack of suitable reflecting surfaces or to scattering of seismic energy by internal complexity.

Data Acquisition and Processing

The simple geometry shown in Figure 1 is not usually an adequate approach to acquiring seismic data. The reflections are weak and easily swamped by noise. Increasing the power of the source will help. Modern seismic sources include airguns (which are used at sea and work by releasing a bubble of compressed air into the water) and vibrator trucks (which are used on land and vibrate a metal pad held in contact with the ground). Cost, practicality, and concern about possible environmental damage place limits on the energy that can be put into the ground. The solution is to use an array of receivers to make the best use of the available energy. Figure 4 shows how, in the marine case, a long array of receivers is towed behind a ship, which fires the source at regular intervals along a line. After acquisition is complete, the



Not to scale.
Typical receiver spacing 50 m, typical total length of receiver array 4 km.

Figure 4 Schematic geometry for acquiring marine seismic reflection data.

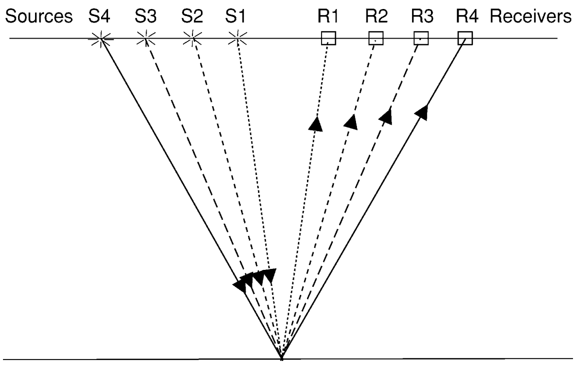


Figure 5 Acquisition geometry of traces sharing a common reflection point.

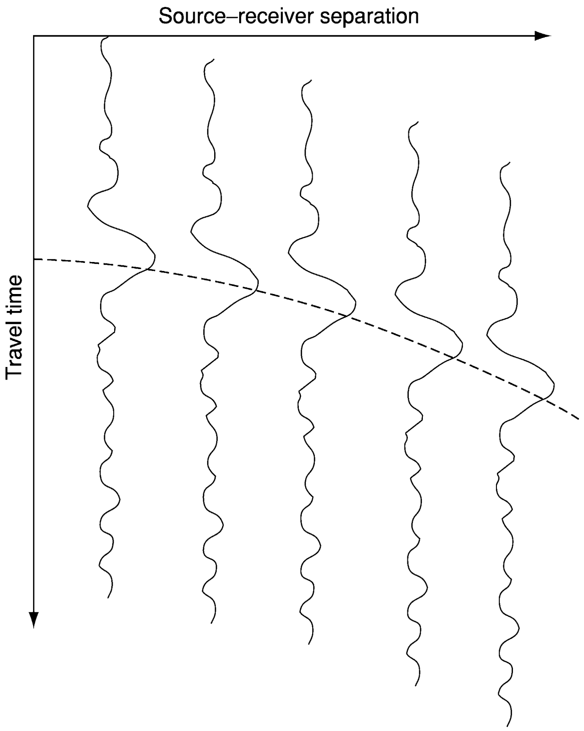


Figure 6 Schematic plot of traces acquired with the geometry shown in Figure 5.

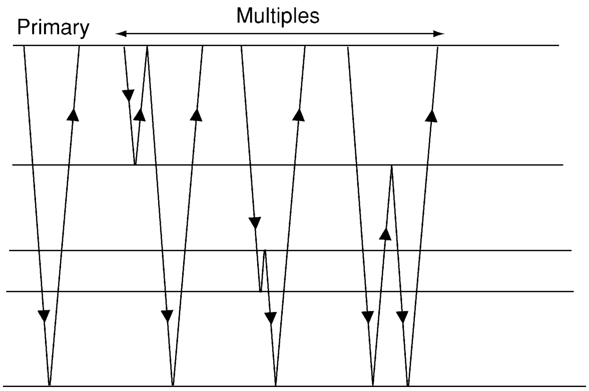


Figure 7 Travel paths for primaries and multiples.

recorded data from different shots can be reordered to bring together traces corresponding to a single reflection point in the subsurface (Figure 5). Of course, the travel time increases as the source–receiver separation becomes larger (Figure 6), but this can be corrected so as to line up all the signal peaks at the same travel time. They can then be added together (stacked) to create a signal with a much higher amplitude. The correction required to align all the traces contains information about the average velocity of the seismic waves, which is useful in later processing.

However, the medium above a target reflector is usually strongly layered. This means that signals can bounce back and forth between these shallow layers, and may perhaps arrive at the receiver at much the same time as a genuine reflection from a deeper layer (Figure 7). There are several ways to remove these ‘multiples’. Many of these methods depend on the difference in average velocity along the travel path between the primary and the multiples, caused by the general increase of velocity with depth due to compaction. The multiples have spent more time at shallow depths, so their average velocity is lower than that of the deeper-penetrating primaries. Correction for variable source–receiver distances will thus line up the primaries but not the multiples, which will be attenuated in the stacked result. Various algorithms exploit this velocity difference to improve the discrimination against multiples further. The same process of multiple bounces, on a smaller layer-thickness scale, acts to blur the crisp initial seismic signal on its passage through the Earth. This combines with the effect of the absorption of seismic energy (which is more pronounced over a given distance for the higher frequencies) to reduce the content of high-frequency energy in deep reflections. Commonly, reflections from a depth of 3000 m will have peak energy at a frequency of 25–30 Hz. As we shall see, this reduces the resolution that can be achieved.

Seismic data are often acquired along a straight line, with the objective of producing a cross-section showing the subsurface reflectors along the line (2D seismic). At first sight, this just requires each stacked trace to be plotted in the correct place along the line; the wiggles corresponding to each reflector will then line up to produce a cross-section through the Earth, though the vertical axis will be travel time rather than depth. There is, however, a complication, which is illustrated in Figure 8. This shows the travel paths of the seismic signals for various source and receiver positions along a line over a schematic buried syncline. Since the data have been corrected for variable source–receiver distances and stacked, we can assume that they are equivalent to the data that would have been recorded with zero source–receiver separation at each point along the line. In that case, the seismic travel paths must hit the reflector at right angles, so that the reflected path is the same as the incoming path. We see from Figure 8A that at some surface locations it is possible to receive reflections from both sides and from the bottom of the syncline, so that the stacked section will present a 'bow-tie' appearance (Figure 8B). For other reflector geometries, the distortion would be less dramatic, but would still be present; it arises wherever the travel paths are not vertical, but we ignore this and plot the reflection traces vertically below the relevant surface point. To correct for the distortion, the reflector segment seen on each trace needs to be moved laterally by the correct amount. This process is known as migration. It requires knowledge of the seismic velocities in the subsurface, which have already been obtained for use in the variable source–receiver distances correction prior to stacking. The effect of migration in transforming the image into a recognizable picture of subsurface structure can be dramatic.

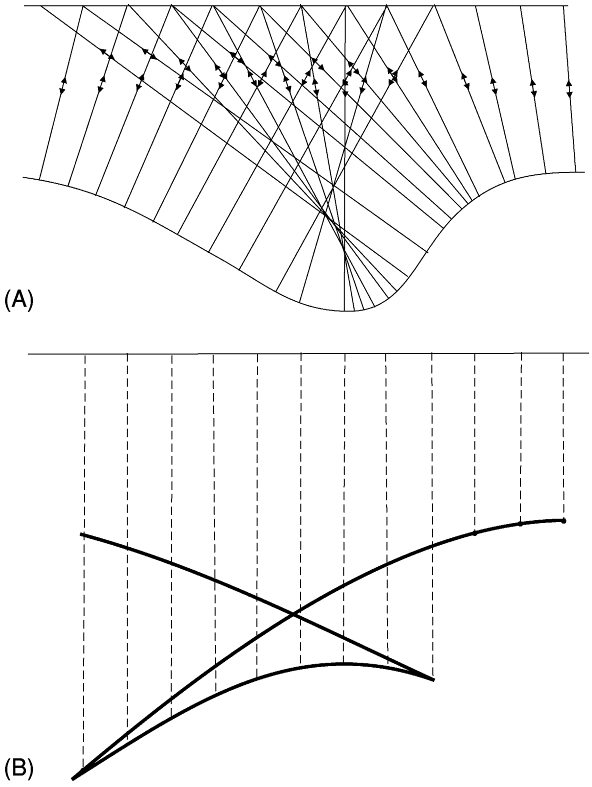


Figure 8 A (B) travel-time section can be more complicated than (A) the real depth section. In (B) time is plotted vertically below the surface point concerned.

3D Seismic

Migration of seismic data along a 2D line does not perfectly position the reflectors in the right place, however. The problem is that, if the line is not exactly along the dip direction, reflection points may be laterally offset from the line (Figure 9). Standard processing has no way of detecting that this is so, and the final migrated section will be plotted vertically below the surface line. To minimize this effect, 2D lines are acquired along the dip direction where possible; however, the dip direction may change with stratigraphic level. A big improvement in subsurface imaging can be gained by the use of 3D seismic. Suppose we acquire a large number of parallel 2D lines, at close spacing (perhaps 50 m). Then information about the structure off to the side of each line is available from other lines in the dataset (except for the lines at the edge of the survey). We can reposition (migrate) the data in 3D, so that, when we plot a vertical section along one of the 2D lines, it contains only information about reflectors that are vertically below it.

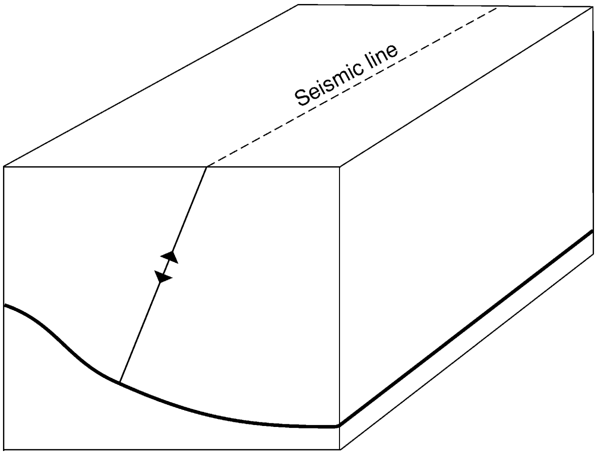


Figure 9 Reflection points for a seismic line may not be vertically below it.

Such 3D surveys are routinely being undertaken, primarily for petroleum exploration, which can support the high cost of acquisition. The output from the processing of such a survey will be a cube of data, made up of traces plotted vertically below points on a square grid. At a typical trace spacing of 25 m × 25 m, a surface area of 200 km^2 would contain 320 000 traces. There are two further benefits from such a survey, besides the 3D imaging.

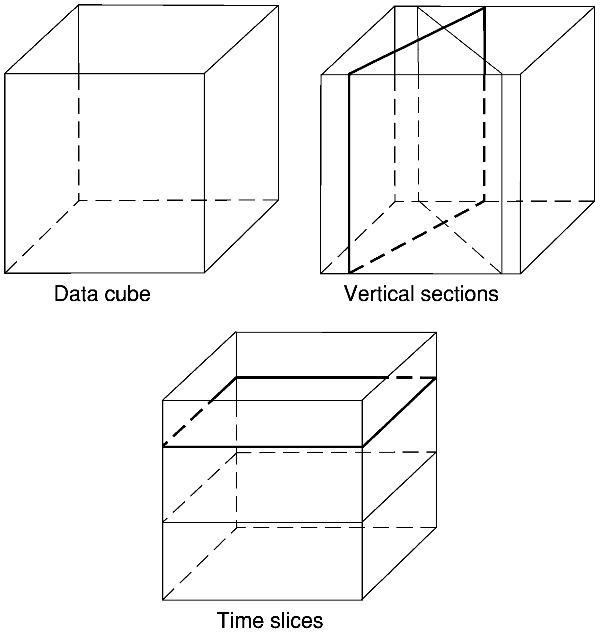


Figure 10 A 3D data cube can be viewed as slices in any direction.

1. The density of the data makes it easy to follow features from line to line across the cube. For example, the development of a fault can be studied across the cube as its throw grows and diminishes, or the detailed geometry of a sedimentary channel can be mapped.
2. Given sufficient computational power, it is easy to construct slices through the data cube in any direction (Figure 10). Vertical sections can be chosen in any direction; for example, sections perpendicular and parallel to a given fault may be helpful in understanding its geometry. The plan view (time slice) can be particularly helpful in understanding depositional systems. It is also possible to view a depositional body in 3D (Figure 11).

These displays extracted from the cube contain a great deal of information. There are, however, two limitations that need to be borne constantly in mind. One is the limited vertical and horizontal resolution. The vertical traces consist of seismic ‘wiggles’; each reflecting interface is marked by a signal that represents the source signal, modified by its passage through the Earth and modified further by data processing. Figure 12 shows a typical response for a thin layer, representing perhaps a sand encased in shale. As the layer thins, there comes a point where the reflections from the top and base of the layer start to coalesce. Beyond this point, the layer is thinner than the separation of the apparent top and base wiggles would suggest. The amplitude response is a maximum when the bed thickness is one-quarter of

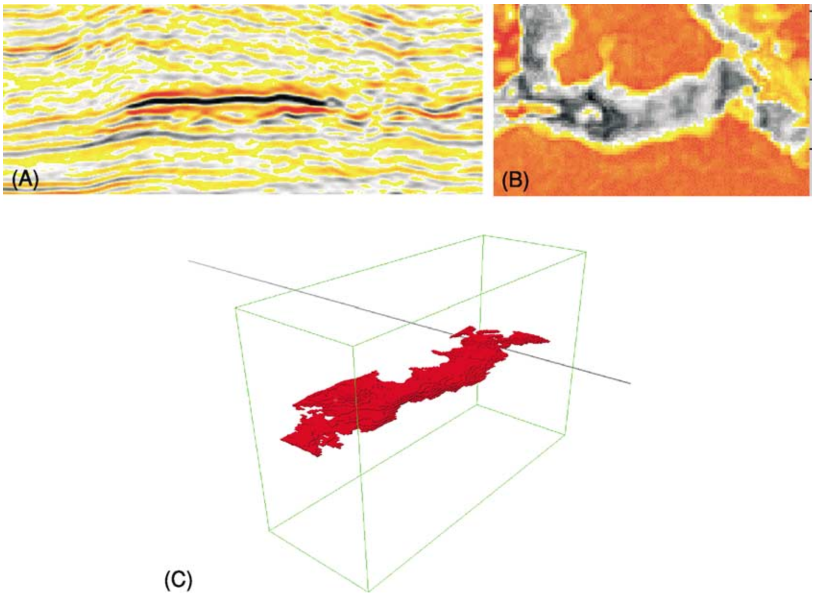


Figure 11 Displays created from a 3D data cube: (A) vertical section, (B) map view and (C) 3D perspective view.

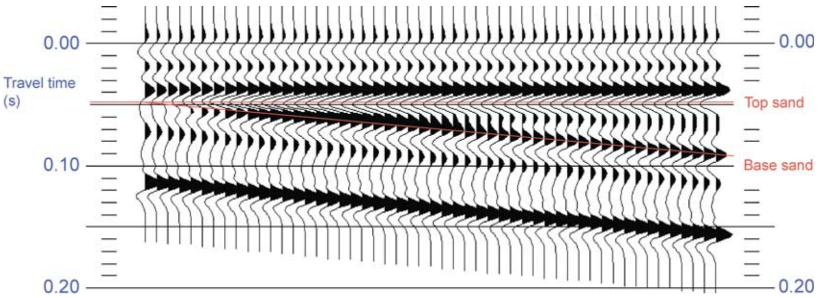


Figure 12 Modelled traces showing the seismic response of a sand of varying thickness.

the seismic wavelength: if a typical wavelength is about 100 m, the amplitude maximum is at a thickness of 25 m. This will be the approximate limit of vertical resolution. Various processes can be applied to the trace data to try to sharpen up the wiggles, but at depths of a few thousand metres it is hard to achieve a better resolution than about 12 m. Horizontal resolution is also limited. The resolution achievable depends on the accuracy with which seismic velocities are known: errors in the velocities degrade the focusing of the migrated seismic image. In practice, the resolution might be 50–100 m at a depth of a few thousand metres.

The second limitation is that the vertical axis of the traces represents travel time, not distance. If we know the seismic velocities, we can of course convert the travel times into depths. However, if there is no well control the velocities may be fairly uncertain. The resulting errors may not be important for mapping on a basin scale, but are often critical in the detailed work of oil and gas prospecting. Even when no detailed depth conversion is intended, it is essential to have a rough idea of the depth scale corresponding to the travel-time scale whenever a seismic section is being interpreted. This is because displays with considerable vertical exaggeration are often used: true-scale seismic sections are usually much wider than they are high, leading to display problems on the typical workstation computer screen with an aspect ratio near to one. If unrecognized, this distortion will hinder the understanding of depositional and tectonic features.

Interpretation

Seismic reflection allows us to see and map layering within the subsurface. We usually need to put some stratigraphic label on the mapped interfaces. Sometimes, distinctive interfaces such as a major angular unconformity are easy to recognize (Figure 13). If some boreholes have been drilled and wireline logs have been run in them to record seismic velocity and density, we can calculate the acoustic impedance of each layer and hence the expected seismic response. We also know the travel time from the surface to the reflecting interfaces, either from direct observation (e.g., in a VSP) or by integration of the sonic log. The interfaces that give rise to the largest reflection amplitudes can thus be related to the sequence drilled by the well. If there are several boreholes, reflectors can be tracked from one to another to establish a consistent identification scheme. Usually, seismic reflectors are time-lines, at least on the broad scale. The overall depositional setting can be inferred from interpretation of seismic sequences. Relative sea-level fall and rise can be inferred from variation in the pattern of onlap, and this provides information about the overall depositional environment.

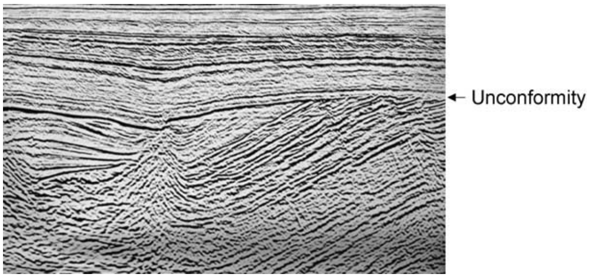


Figure 13 Seismic section showing a prominent angular unconformity.

Structure mapping is often quite easy provided deformation is not extreme. Reflecting interfaces can be followed through a cube of 3D seismic, or around a grid of intersecting 2D lines, and a map constructed. Seismic reflection works best for interfaces with dips of up to 30° or so. Fault planes are therefore seldom imaged directly; they are recognized from the displacement of sedimentary layering across them

(Figure 14). Steeply dipping bedding, for example in an overthrust zone or against the flank of a salt or mud diapir, will often not be imaged.

Depositional environments can often be recognized and mapped from the external geometry of a feature (the shape of its envelope) and from the geometry and character of the reflections within it. For example, within a fluvial system it may be possible to recognize channels by mapping reflection amplitude on a slice through a 3D cube, parallel to the regional dip; the channel fill often has a different acoustic impedance from the rest of the unit. By making a series of such slices, it is possible to follow the evolution of the channel system through time. If seismic resolution permits, it may be possible to see internal depositional geometry, such as the downlap geometry of a laterally accreting point bar. Discrimination between sand and shale infill may be possible: sands often have a mounded appearance due to differential compaction.

In carbonate systems, it is often possible to recognize reefs from seismic reflections. A reflection is usually obtained from the top of a reef, though it may be discontinuous if the topography is complex. The interior is usually quite transparent. Often the reef has separated different depositional environments, so there is a sharp change in the reflection character of the contemporaneous package from one side to the other.

Salt and shale diapirs are often inferred from the deformation of the layered sediments around them. The salt or shale itself is usually acoustically homogeneous and therefore appears as a transparent body on seismic sections. Imaging sedimentary layering below the overhanging top of a mushroom-shaped salt diapir is difficult, because of the complicated paths that seismic waves follow, owing to the much higher seismic velocities in the salt.

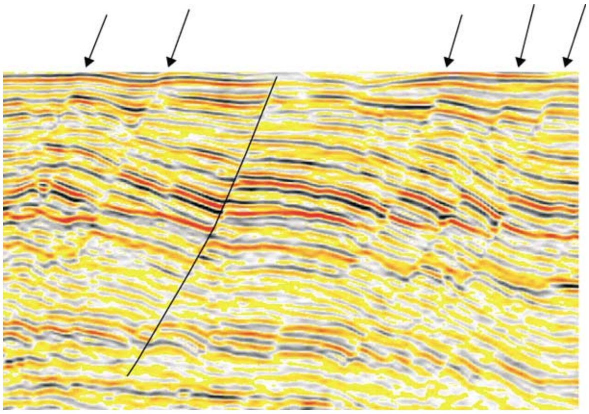


Figure 14 Seismic section showing faults, imaged as discontinuities in reflectors. One fault is shown interpreted; some others are indicated by arrows.

Seismic Reflection in the Oil and Gas Industry

The primary use of seismic reflection in the oil and gas industry is for mapping structure. In exploration **Petroleum Geology:** Exploration, this is mainly a matter of looking for the closed anticlinal features that form potential hydrocarbon traps. These may be either pure (four-way) dip closures or combination fault–dip closures. If a possible trap is fault bounded, it will be necessary to look into the juxtaposition of beds across the fault, to see whether the reservoir is always juxtaposed against a seal (e.g., a shale), or whether it is in places in contact with another reservoir (e.g., a sand). In the latter case, it may still be possible for the trap to work if the fault plane itself provides a seal, but this is inherently more risky. 3D seismic, with its high trace density, is well suited to making a juxtaposition analysis along the whole length of a fault. When a hydrocarbon discovery has been made, seismic reflection can be used to define the internal geometry of the reservoir. Small faults or thin (but laterally extensive) shales may be significant barriers to hydrocarbon flow when the reservoir is put on production, so it is important to plan development wells with this in mind.

Sometimes, it is possible to see directly from the seismic reflections whether hydrocarbons are present in a particular reservoir. When oil or gas replaces brine in the pore space of a reservoir, the acoustic impedance of the material is reduced. This can be a large effect (Figure 15), particularly for gas in high-porosity reservoirs. If the brine sand has an impedance that is less than that of the overlying shale, then the impedance of an oil or gas sand will be much less. The consequence is that the reflection amplitude from the top of the reservoir will be higher where the hydrocarbon is present. Such a 'bright spot' is

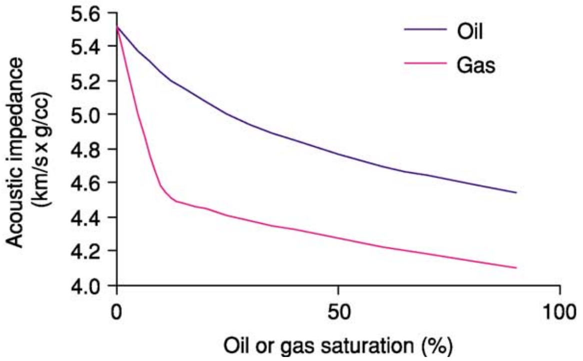


Figure 15 Effect of oil or gas on the acoustic impedance of a high-porosity sandstone.

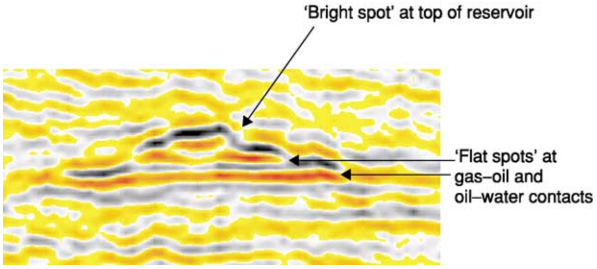


Figure 16 Examples of bright-spot and flat-spot direct hydrocarbon indicators.

a direct hydrocarbon indicator (DHI). Of course, the brightening might be due to a lateral change in lithology, perhaps an increase in the porosity of the reservoir sand. If the brightening is a DHI, then the amplitude increase will follow a particular depth contour, at the hydrocarbon–water contact. Another type of DHI is the 'flat spot', which is a reflection from a fluid (e.g., oil–water) contact. This should be flat and consistent with the amplitude change at the top of the reservoir (Figure 16).

When these effects of fluid fill were first recognized, they were used to reduce the risk of exploration prospects. More recently, it has become possible to use this principle to follow the way that fluids are moving through an oil or gas reservoir as the hydrocarbons are being produced. This is done by acquiring a survey before production starts, and then repeating the survey after some production has taken place. The difference between the two surveys (after careful matching to remove the effects of any differences in acquisition) will be due to fluid movement, for example an oil–water contact rising because some of the oil has been produced. This is useful information if the reservoir has internal barriers to fluid movement, for example due to faulting. In such a case, oil may be produced from some fault blocks and not others; the seismic differences would be confined to the blocks where production is occurring. Such seismic surveys are almost always shot as 3D surveys, because the high trace density is important in identifying small DHI effects, and the repeat survey is often called 4D seismic (elapsed time being the fourth dimension).

Huge amounts of 3D seismic reflection data are acquired by the oil and gas industry. In the year 2000, some 300 000 km^2 were shot offshore, and 30 000 km^2 onshore. This is targeted on hydrocarbon resources, of course, and therefore confined to sedimentary basins with proven or potential hydrocarbon generation and migration. Within such basins, seismic reflection has revealed a great deal about subsurface geology.

See Also

Economic Geology. **Engineering Geology:** Seismology. **Petroleum Geology:** The Petroleum System; Exploration; Production; Reserves. **Sedimentary Processes:** Depositional Sedimentary Structures; Post-Depositional Sedimentary Structures.

Further Reading

Blake BA and Figueroa DE (1999) Interpretation strategy drives acquisition of 2-D seismic in sub-Andean Bolivia. *The Leading Edge* 18: 1360–1365.

Brown AR (1999) *Interpretation of Three-Dimensional Seismic Data*. Tulsa, OK, USA: Society of Exploration Geophysicists.

Gersztenkorn A, Sharp J, and Marfurt K (1999) Delineation of tectonic features offshore Trinidad using 3-D seismic coherence. *The Leading Edge* 18: 1000–1008.

Payton CE (ed.) (1977) *Seismic Stratigraphy – Applications on Hydrocarbon Exploration*. Memoir 26. Tulsa, OK, USA: American Association of Petroleum Geologists.

Sheriff RE and Geldart LP (1995) *Exploration Seismology*. Cambridge: Cambridge University Press.

Story C, Peng P, Heubeck C, Sullivan C, and Lin JD (2000) Liuhua 11-1 Field, South China Sea: a shallow carbonate reservoir developed using ultrahigh-resolution 3-D seismic, inversion, and attribute-based reservoir modeling. *The Leading Edge* 19: 834–844.

Telford WM, Geldart LP, and Sheriff RE (1991) *Applied Geophysics*. Cambridge: Cambridge University Press.

Wescott WA and Boucher PJ (2000) Imaging submarine channels in the western Nile Delta. *The Leading Edge* 19: 580–591.

Yilmaz O (2000) *Seismic Data Analysis*. Tulsa, OK, USA: Society of Exploration Geophysicists.

Zeng H, Tucker FH, and Wood LJ (2001) Stratal slicing of Miocene–Pliocene sediments on Vermilion Block 50 – Tiger Shoal Area, offshore Louisiana. *The Leading Edge* 20: 408–418.

SEQUENCE STRATIGRAPHY

P P McLaughlin Jr, Delaware Geological Society, Newark, DE, USA

Introduction: What is Sequence Stratigraphy?

Sequence stratigraphy is one of the major unifying concepts of the geosciences to arise in the twentieth century. Rooted in the cross-fertilization of regional facies mapping and geophysics, sequence stratigraphy provides an invaluable approach to practical problems in applied geology and fundamental scientific questions in Earth history. It incorporates a variety of disciplines of stratigraphic geology (i.e. lithofacies analysis, biostratigraphy, and chronostratigraphy) and is intrinsically related to a number of other areas of Earth history, notably sea-level change, tectonics, and palaeoclimate.

Sequence stratigraphy is defined as the study of rock relationships within a chronostratigraphic framework of repetitive genetically related strata bounded by surfaces of erosion or deposition or their correlative conformities. The fundamental starting point for sequence stratigraphy is the sedimentary facies, which is a lithostratigraphic body characterized by distinct lithological or fossil characteristics, generally reflecting a certain origin. A group of sedimentary facies genetically linked by common processes and environments comprises a depositional system. These depositional systems can be grouped together within a framework of unconformity-bound relatively conformable stratigraphic packages called sequences.

Early publications on sequence stratigraphy emphasized the relationships between global sea-level change, or eustasy, and large-scale stratigraphic patterns. This work provoked serious debate about the importance of global sea-level change as a genetic control on stratigraphy. Recent work integrating sequence-stratigraphic analysis with isotope data is providing new insights into the relationships between ice-sheets, climate, and sea-level, and is helping to clarify the role of eustasy in the evolution of stratigraphic successions.

However, sequence stratigraphy is more than a record of global sea-level: it is a practical stratigraphic tool. Sequences are a product of the interplay of eustasy, tectonics, and sediment supply. As a result, they can be recognized and correlated regionally, regardless of whether global sea-level change was the dominant control. With an understanding of these controls, the sequence concept provides a framework for understanding the evolution of depositional systems through time, making it a powerful predictive tool for stratigraphic analysis.

Development of the Concept

Sequence stratigraphy has seen major growth and development since the 1970s. However, the roots of the field extend back to the 1940s, when LL Sloss coined the term ‘stratigraphic sequence’ in his regional facies mapping of the Palaeozoic of North America. Sloss defined stratigraphic sequences as “rock stratigraphic units of higher rank than group, megagroup, or supergroup, traceable over major areas of a continent and bounded by unconformities of interregional scope”. He recognized six sequences and gave them Native American names derived from localities where they are well developed: Sauk, Tippecanoe, Kaskaskia, Absaroka, Zuni, and Tejas.

In the 1960s and 1970s, the concept of the stratigraphic sequence was applied to the geophysical data collected by oil companies that were using new seismic-imaging tools to obtain a picture of basin and stratigraphic architecture. Under the leadership of former Sloss student Peter Vail, researchers at Exxon and its predecessors recognized stratigraphic patterns on seismic lines that they believed corresponded to the same types of sequences and unconformities mapped by Sloss. In addition, they identified ‘onlap unconformities’ within marine successions in basins on different continents and inferred global sea-level control and a worldwide extent for ‘onlap cycles’.

These seismic-derived sequence-stratigraphic concepts were brought into the public domain with the 1977 publication of seminal papers by Vail and collaborators in AAPG Memoir 26. The ‘depositional sequence’ was defined as a stratigraphic unit composed of a relatively conformable succession of genetically related strata bounded at its top and base by unconformities or their correlative conformities. Conceptually, depositional sequences resemble Sloss cratonic sequences, but represent much shorter time intervals. The AAPG memoir also provided the first published documentation of the Exxon group’s view of the relationship between inter-regional unconformities and global cycles of sea-level, including documentation of coastal onlap curves established from seismic stratigraphic records on different continental margins. Other papers detailed methods for determining sea-level change from coastal onlap and for interpreting sequence stratigraphy and facies from seismic-reflection patterns. The second major treatise on the subject, SEPM Special Publication 42, was published in 1988. This volume elevated the significance of sequence stratigraphy as a means of understanding Earth history and as a practical tool in petroleum geology. Included articles further defined the key elements of sequence stratigraphy, documented the ages of sequences, examined theoretical aspects, and discussed both outcrop and subsurface examples.

Although these two volumes helped to bring sequence stratigraphy into the mainstream of geological thought, there was notable criticism of the concepts and cycle charts. One of the more contentious issues was the hypothesis that eustasy is the primary control on the timing and patterns of deposition of sequences. Critics argued that the accuracy and precision of chronostratigraphic control are inadequate to demonstrate synchronicity of sequences from around the world and thereby establish the uniqueness of eustatic control.

The late 1980s and 1990s saw the evolution of sequence stratigraphy into a tool for investigating increasingly detailed stratigraphic problems. Studies examined factors controlling sedimentation in specific basins or time intervals and dealt with increasingly finer-scale stratigraphic problems in a variety of depositional environments beyond the marine continental margins. This approach can be used for finer-scale reservoir- and aquifer-scale stratigraphy problems integrating well log, core, and outcrop-based datasets.

Parasequences: The Building Blocks of Sequences

The parasequence is the fundamental building block of a sequence. A parasequence is a relatively conformable succession of genetically related strata bounded by marine flooding surfaces. It is normally a progradational or aggradational package that reflects a shoaling-upwards trend. The succession of facies within a parasequence generally follows Walther's Law, which states that a normal vertical facies succession mirrors the lateral distribution of facies in a sedimentary environment.

The parasequence boundary is the key to correlation in a sequence-stratigraphic framework. It is an approximately planar marine flooding surface commonly characterized by non-deposition or minor erosion. It may be marked by significant burrowing by organisms and may have an associated lag deposit of coarse material such as shells, gravel, authigenic minerals, or rip-up clasts formed by erosion and winnowing in the course of the flooding event.

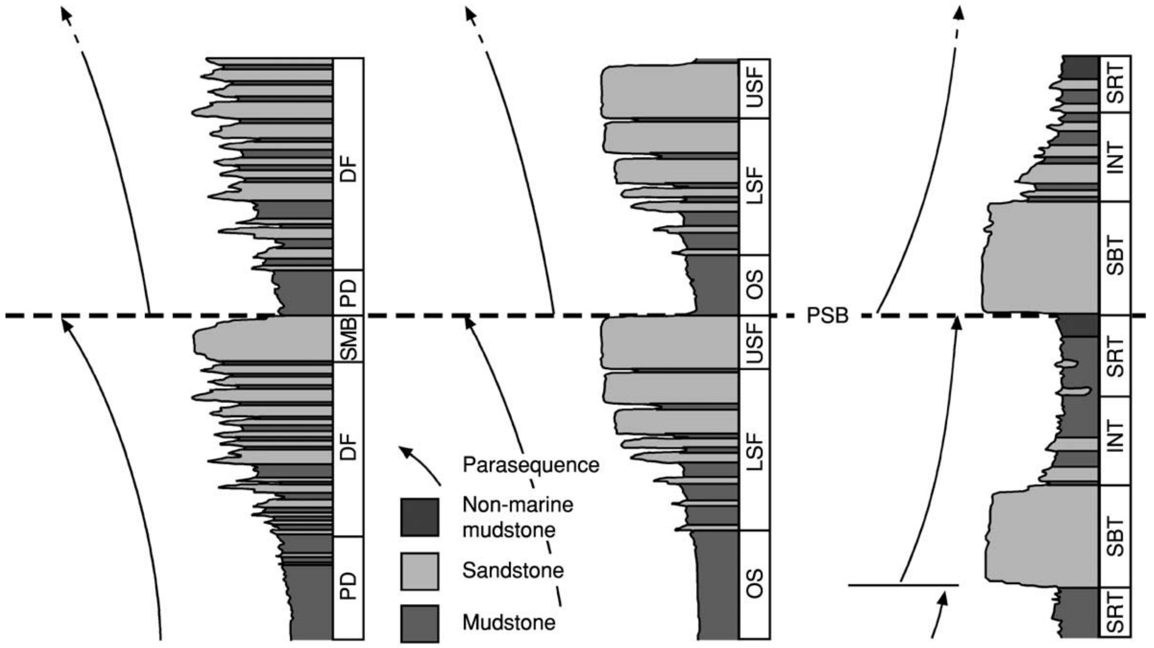
In shallow-marine successions, parasequences typically coarsen upwards with an increase in sand content and a general increase in the thicknesses of the sand beds. Sedimentary facies trace a regular succession of shallower-water sedimentary environments. For example, in a river-dominated deltaic environment, facies could reflect shoaling from prodelta to delta front to stream-mouth bar (Figure 1A); for a wave-dominated shoreline, a parallel succession from offshore to lower-shoreface to upper-shoreface environments might be expected (Figure 1B).

In some cases, fining-upwards parasequences can be recognized. For example, in marginal-marine settings, the base of the parasequence may be marked by the abrupt appearance of marine sand above marginal-marine muds, above which the percentage of sand decreases and the sand beds become thinner. The facies trace a succession of shallower-water environments, in this example shoaling from subtidal to intertidal to supratidal non-marine facies (Figure 1C).

Parasequence-Stacking Patterns

Just as a normal succession of genetically related beds make up a parasequence, so a normal succession of parasequences can be grouped into a unit called a parasequence set. The pattern of changes between successive parasequences in a parasequence set is termed the parasequence-stacking pattern. The concept of accommodation is fundamental to understanding parasequence-stacking patterns. Accommodation is the space available for potential sediment to accumulate and is a function of eustasy and subsidence (Figure 2). Sediment influx controls the rate at which this space is filled. The interplay between accommodation and sedimentation rates controls whether the shoreline advances or retreats and the resulting vertical facies changes.

Three types of parasequence-stacking pattern are progradational, retrogradational, and aggradational. A progradational parasequence set is recognized where parasequence stacking reflects overall shoaling and basinwards advance of a depositional system. Progradation occurs when the rate of deposition exceeds the rate of accommodation: the lack of vertical space for sediment accumulation forces sedimentation basinwards (Figure 3A). A retrogradational parasequence set reflects the opposite case, in which parasequences are stacked in a pattern that reflects overall deepening and a landwards retreat of the depositional system. Retrogradation reflects a sedimentation rate that is lower than the rate of accommodation: the inability of sedimentation to fill the available vertical space shifts sedimentation landwards (Figure 3B). An aggradational stacking



A. River-dominated shoreline	B. Wave-dominated shoreline	C. Tide-dominated shoreline
● Sands thicken upward	● Sands thicken upward	● Sands thin upward
● Percent sand increases upward	● Percent sand increases upward	● Percent sand decreases upward
● Facies shoal upward	● Facies shoal upward	● Facies shoal upward
● Parasequence boundary at abrupt shift to finer-grained facies	● Parasequence boundary at abrupt shift to finer-grained facies	● Parasequence boundary at abrupt shift to coarser-grained facies

Figure 1 Characteristics of parasequences in various coastal environments: (A) deltaic parasequence on river-dominated shoreline; (B) offshore to shoreface parasequence on wave-dominated shoreline; and (C) subtidal to intertidal parasequence on muddy tide-dominated shoreline. PSB, parasequence boundary; SMB, stream-mouth bar; DF, delta front; PD, prodelta; USF, upper shoreface; LSF, lower shoreface; OS, offshore; SRT, supratidal; INT, intertidal; SBT, subtidal. (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1–55.)

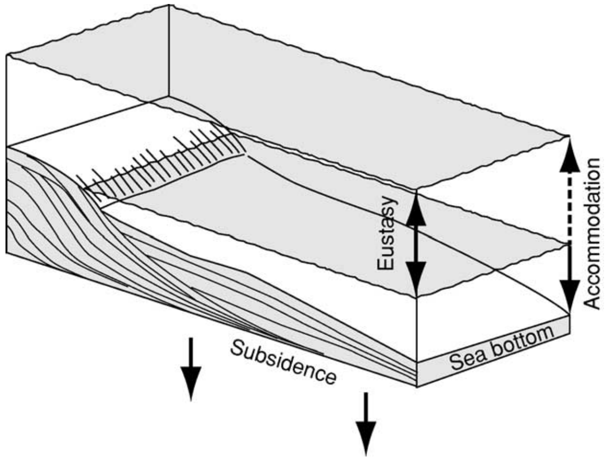


Figure 2 Accommodation as a function of eustasy (sea-level changes) and subsidence (tectonics). Horizontal grey surfaces represent positions of sea-level. Dashed arrow indicates the additional accommodation at the higher sea-level.

pattern is recognized where the facies reflect steady accumulation of sediments without significant shifts basinwards or landwards. Aggradation occurs where the sedimentation rate and accommodation rate are approximately in balance (Figure 3C).

Parasequences are useful tools for chronostratigraphic correlation. Correlations based on sequence stratigraphy may differ significantly from lithostratigraphic correlations. Conventional lithostratigraphic correlations typically emphasize the linkage of similar lithologies, with the underlying philosophy of tracing mappable rock stratigraphic units; in the lithostratigraphic-correlation example in Figure 4A, the sand intervals in each well are correlated. In contrast, correlation of parasequences places the sedimentary section in a chronostratigraphic reference frame. Parasequence boundaries are correlated as local time-lines and provide a basis for connecting genetically linked strata. In the example of chronostratigraphic correlation shown in Figure 4B,

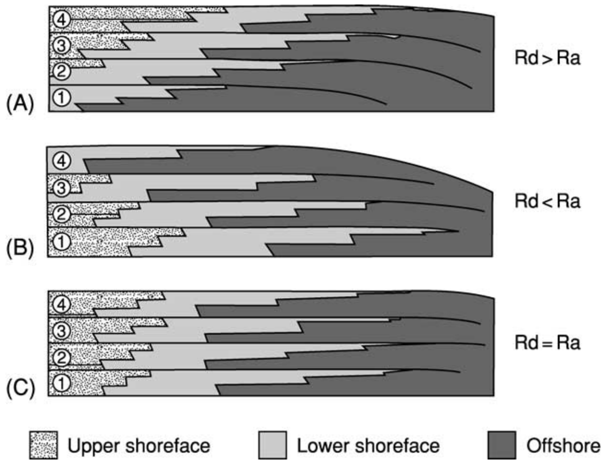


Figure 3 Parasequence-stacking patterns for (A) progradational, (B) retrogradational, and (C) aggradational parasequence sets. Stacking patterns are indicated by the movement of facies in successive parasequences (numbered) in each diagram. Rd, rate of deposition; Ra, rate of accommodation. (Adapted from Van Wagoner JC, Posamentier HW, Mitchum RM Jr, *et al.* (1988) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hasting BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 39–45. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

the resulting stratigraphic interpretation provides an improved understanding of stratigraphic geometry by demonstrating the connection of the thinner shallow-marine sands to thicker sand beds in a landwards direction.

Recognition of Sequences and Systems Tracts

A sequence is a succession of genetically related relatively conformable strata bounded by unconformities or their correlative conformities. Based on parasequence-stacking patterns and facies trends, a number of distinct sequence components called systems tracts can be defined. A systems tract is a linkage of contemporaneous depositional systems, with a depositional system defined as a three-dimensional assemblage of lithofacies. The five systems tracts most commonly recognized are lowstand fan, lowstand wedge, shelf-margin wedge, transgressive, and highstand (Table 1 and Figure 5). These systems tracts are separated by significant stratigraphic surfaces, the most important being the sequence boundary, transgressive surface, and maximum flooding surface.

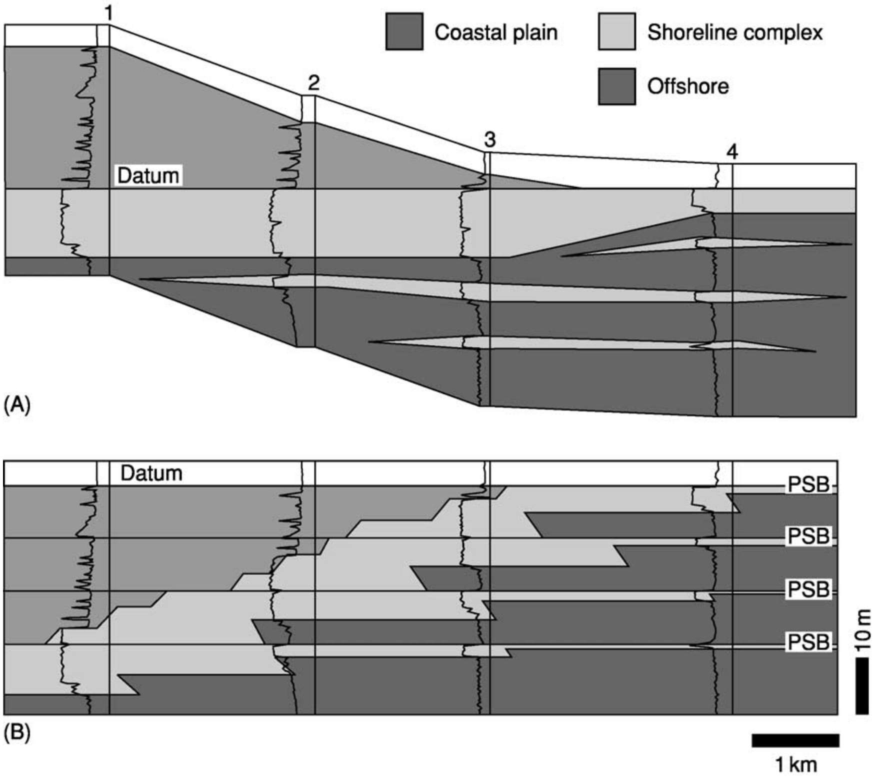


Figure 4 Comparison of (A) lithostratigraphic and (B) chronostratigraphic correlation styles for a prograding parasequence set. The correlation datum for the lithostratigraphic section is the top of the sandstone; the datum for the chronostratigraphic section is the uppermost parasequence boundary. PSB, parasequence boundary. (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1–55.)

Table 1 Characteristics of systems tracts

Systems tract	*Stacking pattern*	*Bounding surfaces*	*Stratal terminations*	*Location of best development*
Highstand	Aggradational to progradational	Base: maximum-flooding surface Top: sequence boundary	Downlap basinwards, toplap or truncation at top landwards	Landwards of the offlap break
Transgressive	Retrogradational	Base: transgressive surface Top: maximum-flooding surface	Downlap basinwards, onlap landwards	Landwards of the offlap break
Shelf-margin wedge	Weakly progradational	Base: sequence boundary Top: transgressive surface	Onlap landwards, downlap basinwards	Near the offlap break
Lowstand wedge	Progradational to aggradational	Base: sequence boundary or downlap surface at top of underlying lowstand fan Top: transgressive surface	Onlap landwards, downlap basinwards	Basinwards of offlap break where one exists; or in incised valleys and basinward end of ramp
Lowstand fan	Aggradational	Base: sequence boundary Top: flooding surface at top of fan	Onlap landwards, or bidirectional downlap	Deep-water basinwards of offlap break

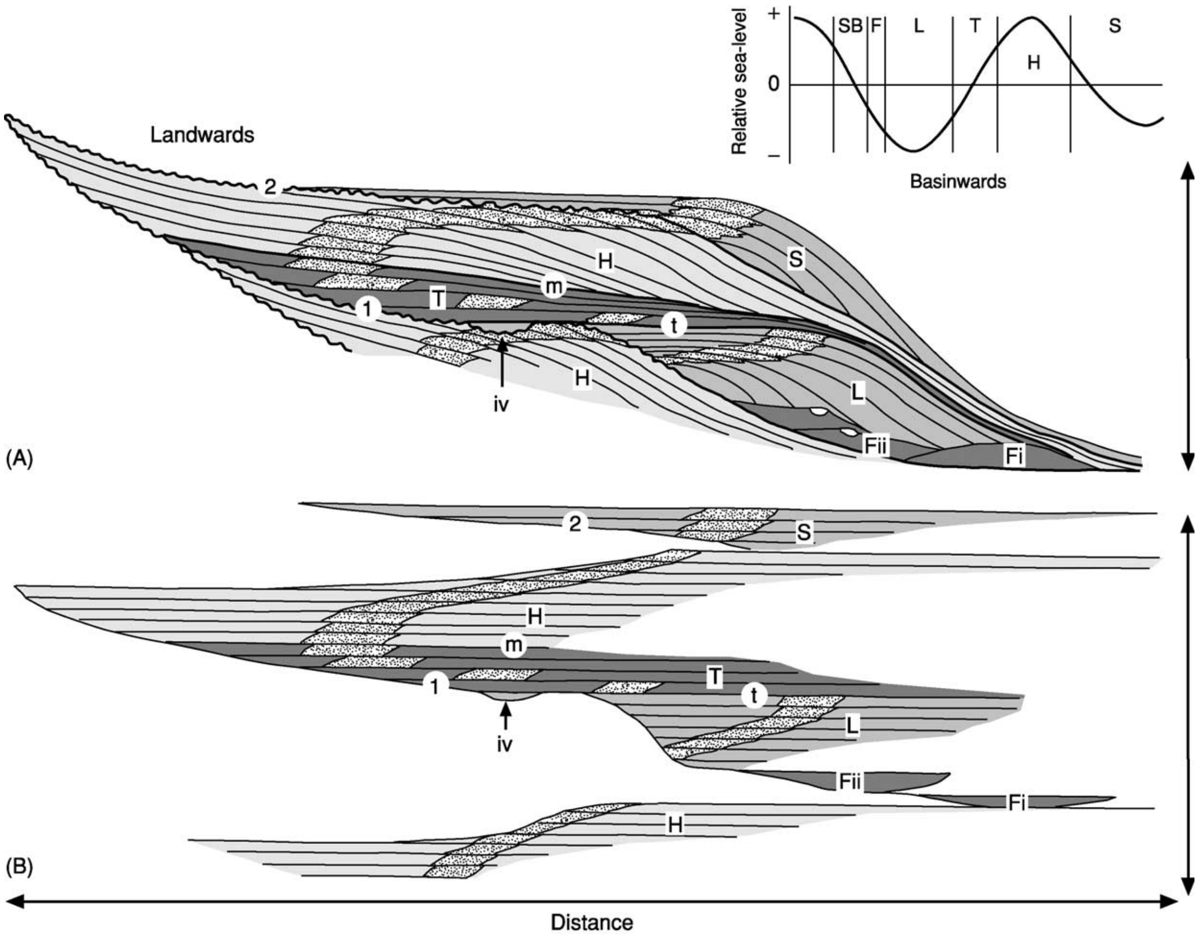


Figure 5 (A) Stratigraphic cross-section and (B) chronostratigraphic section through a conceptual clastic sequence. Systems tracts, in white boxes: Fi, lowstand basin-floor fan; Fii, lowstand slope fan; L, lowstand wedge; T, transgressive; H, highstand; S, shelf-margin wedge. Surfaces, in white circles: 1, type 1 sequence boundary; t, transgressive surface; m, maximum flooding surface; 2, type 2 sequence boundary. Other features: iv, incised valley. Stippled pattern represents sandy shoreline complex. Relative sea-level curve indicates period of deposition for each systems tract. (Adapted from Christie-Blick N and Driscoll NW (1995) Sequence stratigraphy. *Annual Review of Earth and Planetary Sciences* 23: 451–478.)

Descriptive Terminology

Most of the descriptive terms for larger-scale features are derived from seismic stratigraphy (Figure 6). One of the more common geometries for a sequence is a wedge-shaped slug of sediments, with a thin zone of gently dipping strata on the landwards end, a thicker zone of more steeply seawards-dipping strata in the middle, and another thin zone of gently dipping strata on the basinwards end. The term topset is applied to the relatively flat zone of sediments on the proximal part of the basin margin. More steeply inclined strata called clinoforms characterize the thicker zone, and the reflection pattern is termed offlap. The relatively flat thinner zone basinwards of the clinoforms is referred to as the bottomset.

A fundamental principle of sequence stratigraphy is that seismic reflections are produced by contrasts in sonic velocity at chronostratigraphically significant stratal surfaces and unconformities; therefore, they are considered to approximate time-lines in the sedimentary record. Identifying terminations of these reflections is fundamental to the definition of systems tracts and key surfaces (Figure 6). Some reflection types terminate against an underlying surface. Onlap is defined by the termination of a reflection against a more steeply inclined underlying reflection, most commonly in a landwards direction. Downlap is interpreted where an inclined reflection terminates against a less inclined underlying reflection, for example the basinwards termination of prograding clinoforms. Other reflection types terminate against an overlying surface. Toplap is a subtle low-angle termination where a seismic reflection terminates against an overlying reflection without significant erosional truncation. Toplap may reflect the disappearance of an interval in a landwards direction due to sediment bypass or thinning of the bed to below seismic resolution. In contrast, erosional truncation is more abrupt, where a reflection exhibits an angular truncation against a younger surface. This generally signifies an erosional contact.

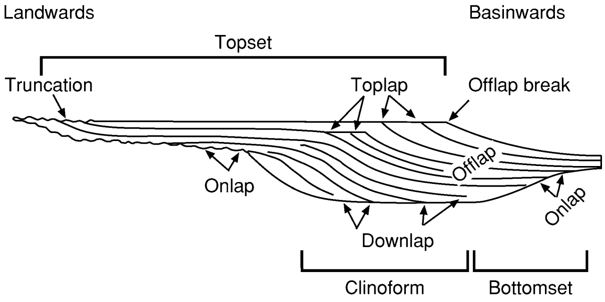


Figure 6 Seismic terminology used in sequence-stratigraphic analysis. Truncation, toplap, offlap, onlap, and downlap are seismic terminations; topset, clinoform, and bottomset are zones of basin margin succession. (Adapted from Mitchum *et al.* (1977) Copyright © 1977 by The American Association of Petroleum Geologists; used by permission of AAPG whose permission is required for further use.)

An important point of reference for the description of sequences is called the offlap break. Offlap is a term sometimes used to describe clinoforms. The offlap break is the main break in slope in the depositional profile and is located at the boundary between the topset and the clinoform. In many sequence-stratigraphy publications, this is referred to as the shelf edge; however, this has created some confusion with the actual topographical break at the edge of a continental shelf, and so the term offlap break is a clearer term for this feature.

Surfaces

Sequence boundary The sequence boundary is the defining surface in sequence stratigraphy. A typical sequence boundary is an areally extensive unconformity above which there is a basinwards shift in facies, a downwards shift in coastal onlap, and onlap of underlying strata (Figure 5).

The facies shift at a sequence boundary commonly does not follow the order predicted by Walther's Law and may have a gap of an environment or two. In such cases, the sequence boundary reflects a significant basinwards shift produced by a rapid decrease in accommodation. It is commonly expressed as an unconformity produced by subaerial erosion that occurs across extensive areas both landwards and basinwards of the offlap break. Such a surface is termed a type 1 sequence boundary (Figure 7A).

In other cases, the characteristics of a sequence boundary are not as distinct, and the area affected by exposure and subaerial erosion is minimal. Although such sequence boundaries generally exhibit some onlap of underlying strata, a downwards shift in coastal onlap, and a change in facies-stacking patterns, the resulting unconformity usually has limited areal extent and the basinwards shift in facies is minor. This is termed a type 2 sequence boundary (Figure 7B). The modest reduction in accommodation permits accumulation of subsequent lowstand sediments landwards of the offlap break, in contrast to the shift of sedimentation off the shelf that occurs at the more significant type 1 sequence boundaries. In early sequence-stratigraphical publications, global sequence charts labelled individual sequence boundaries as either type 1 or type 2 and considered this to be a function of the rate of global sea-level fall; however, it is now understood that a given sequence

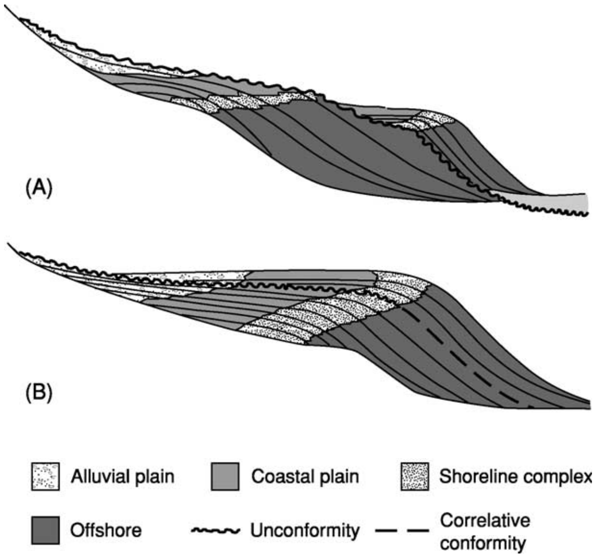


Figure 7 Physical expressions of (A) type 1 and (B) type 2 sequence boundaries. The type 1 boundary exhibits an abrupt downward shift in facies; the type 2 boundary exhibits more a gradual downward shift. (Adapted from Posamentier HW and Vail PR (1988) Eustatic controls on clastic deposition II: – sequence and systems tract models. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 125–154. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

boundary may be of either type depending on local accommodation and sedimentation rates.

The subaerially produced unconformity that occurs at a sequence boundary passes at some basinwards point into a genetically related conformable stratigraphic contact. Some workers contend that a sequence boundary should be recognized only where an unconformity exists. However, many workers correlate this as the same chronostratigraphic surface, potentially allowing a sequence to be recognized over an entire basin.

Transgressive surface The transgressive surface is the first marine flooding surface of significant areal extent landwards of the offlap break (Figure 5). In a complete sequence, it marks a change from progradational or aggradational parasequence stacking at the bottom of the sequence to retrogradation in the middle of the sequence. It may be somewhat diachronous, with the onset of transgression appearing earlier in more basinward areas and reaching more landward areas later.

Maximum-flooding surface The maximum-flooding surface reflects the maximum landward extent of transgression and is marked by a stacking-pattern change from retrogradational in the middle part of a complete sequence to progradational in the upper part (Figure 5). On seismic profiles, it is commonly marked by the downlap of younger horizons onto it and, as such, is sometimes termed the downlap surface. The maximum-flooding surface may have an associated condensed section characterized by strongly burrowed intervals or hardgrounds, in some cases with an associated marine hiatus, as well as enrichment of authigenic minerals such as glauconite or phosphate, a high organic content, and a peak in the abundance of deeper-marine fossils. The condensed section reflects slow sedimentation rates in basinward areas when the peak of the transgression focuses sedimentation in the heads of estuaries and in other landward areas. The elevated concentration of fossils in condensed sections commonly makes them important intervals for the occurrence of age-significant fossils such as ammonites, planktonic foraminifera, and calcareous nannofossils.

Systems Tracts

Lowstand systems tracts Three systems tracts can be deposited during a lowstand of relative sea-level: lowstand fan, lowstand wedge, and shelf-margin wedge (Figure 5). These lowstand systems tracts overlie the sequence boundary and express a basinwards shift in facies produced during periods of relative sea-level fall. The lowstand fan systems tract is the most basinward of the lowstand systems tracts and forms by the accumulation of clastic deposits in a deep-basin setting. During times of relative sea-level fall, large areas of the basin margin are exposed and subjected to erosion. Sedimentation mostly bypasses the basin margin and is fed directly to the basin through incised valleys and submarine canyons. As a result, lowstand fans are commonly detached from the depositional system that built the preceding highstand complex upslope. They may onlap the underlying sequence boundary in the landwards direction; the relief built during fan formation may also produce bidirectional downlap in other directions (Table 1). Aggradational to slightly retrogradational stacking patterns are the most common. The top of the fan may be marked by a shift in deposition to the overlying lowstand wedge systems tract, which produces a downlap surface between the units. In some deep-water systems, the lowstand fan can be divided into two parts: a basin-floor fan, which occurs on the basin floor and may be detached from the depositional system that built the preceding highstand complex, and a slope fan, which develops along the middle or lower part of the slope.

The lowstand wedge systems tract is composed of a prograding wedge basinwards of the offlap break and a thinner unit of incised-valley fill in the landwards direction (Figure 5). The base of the unit is marked by onlap onto the sequence boundary along the landward end of the wedge and in areas of lowstand incised-valley fill (Table 1). In a basinwards direction, it commonly exhibits downlap onto the underlying sequence boundary or the lowstand fan systems tract. The top of the unit is defined by the transgressive surface. The facies-stacking pattern is progradational to aggradational. Because lowstand wedge deposition is generally focused basinwards of the offlap break formed by the preceding sequence, landward areas may be subaerially exposed and subject to fluvial incision, especially during sea-level fall. However, during later parts of the lowstand, the rebound of relative sea-level may result in some sediment accumulation in the incised valleys.

The shelf-margin wedge systems tract represents an accumulation of lowstand deposits near the offlap break of the preceding sequence. Like other lowstand systems tracts, it is produced when a fall in relative sea-level causes a basinwards shift in facies, but the reduction in accommodation is not rapid enough to force sedimentation into the basin. The base of this systems tract is defined by a type II sequence boundary, onto which it onlaps in a landwards direction and downlaps in a basinwards direction (Table 1). The top is marked by the transgressive surface. Facies-stacking patterns are typically weakly progradational.

Recent works have proposed several unique systems tracts for deposits produced during periods of sea-level fall. One of the more commonly cited, the forced regressive wedge systems tract, describes a complex of downstepping shorelines and subaerial erosion overlying a sequence boundary. Successive shorelines partly cannibalize sand through erosion of previous shorelines, producing stranded shoreline sand bodies. Another type, the falling-stage systems tract, is similar but differs in part in the placement of the sequence boundary at the top of the unit. However, these systems tracts are not yet consistently used by sequence stratigraphers.

Transgressive systems tract The transgressive systems tract traces a landward shift in depositional environments that reflects a rise in relative sea-level (Figure 5). The base of this unit is the transgressive surface, and the top is defined by the maximum-flooding surface (Table 1). Because sediment input is overwhelmed by accommodation, a retrogradational facies-stacking pattern is produced. The deposits of transgressive systems tracts are best developed landwards of the underlying offlap break; they onlap the merged sequence boundary–transgressive surface in a landwards direction and downlap onto the transgressive surface in a basinwards direction. They may comprise a thin sheet of deposits that reflect the landwards migration of drowned shoreface complexes; in cases where they directly overlie a sequence boundary with significant erosive relief, they may comprise more laterally variable incised-valley fill.

Highstand systems tract The highstand systems tract traces the basinwards march of deposition over the transgressive systems tract as the rate of sediment input overtakes a slowing rate of relative sea-level rise (Figure 5). It is bounded by the maximum-flooding surface below and the sequence boundary above (Table 1) and is commonly characterized by a prograding topset–clinoform system, the toes of which downlap onto the maximum-flooding surface. The top of the highstand systems tract may be marked by toplap or truncation under the overlying sequence boundary. The facies-stacking patterns change upward from aggradational to progradational, reflecting decreasing accommodation; as a result, accumulation patterns are increasingly driven basinwards rather than vertically, producing a regressive succession (Figure 5).

Variations by Depositional System

The descriptions of sequence-stratigraphic elements in this article have been principally focused on shallow-marine clastic successions. However, sequence concepts are applicable to a variety of other depositional systems. Carbonate systems are very responsive to changes in relative sea-level, but differ from clastic systems in that they generate most of their own sediment from biological sources in the photic zone. When sea-level rises, carbonate systems build upwards to fill the available space, commonly creating thick transgressive systems tracts (*see* **Sedimentary Environments:** Carbonate Shorelines and Shelves); in some cases, very rapid transgression may drown the carbonate system, resulting in a thinner transgressive interval. During highstands, carbonate sedimentation typically continues to be vigorous, but the waning increase in accommodation causes sediment accumulation to prograde basinwards, a phenomenon termed ‘highstand shedding’. When sea-level subsequently falls, little carbonate sediment is produced and little is physically eroded; in wet climates, meteoric diagenesis may produce karst and/or cementation, whereas, in arid climates, evaporites may form in the basin.

Sequence concepts can also be applied to non-marine depositional systems. In alluvial environments, the base level (the surface to which erosion and deposition respond), subsidence (or uplift), sediment supply, and climate all affect the stratigraphic evolution of the system. Sequence interpretation is complicated by the abundance of erosion surfaces, which may be difficult to differentiate from a sequence boundary. Falling base level typically results in sediment bypass and erosion. Lowstand deposits may be characterized by high-gradient stream deposits with sand-rich amalgamated alluvial facies. Where low base level in the early lowstand is accompanied by a low water table, heavily weathered palaeosols may be formed in interfluvial areas; the rise in the water table that accompanies base-level rise in the late lowstand may be conducive to widespread peat formation. Transgressive deposits may reflect decreasing fluvial gradients, which can be expressed as more isolated fining-upwards fluvial sands and crevasse splays upward, with a shift from aggradational to retrogradational facies successions. If accompanied by a rising water table, peat-forming conditions would wane as mires are flooded, and poorly drained gleyed palaeosols would typify interfluvial areas. Highstand deposits may be thick and trace a shift back to aggradational and even progradational facies successions. Fluvial sands are mostly isolated, but amalgamation is increasingly common upwards. Slowing of the water-table rise can provide good conditions for peat formation, but the beginning of base-level (and probably water-table) fall at the end of the highstand would make conditions less favourable.

Sequence stratigraphy of lacustrine deposits has some parallels with marine sequence stratigraphy but is generally independent of changes in sea-level. Instead, tectonically driven accommodation and climate-driven variations in lake level and sediment supply can combine in many different ways to control lacustrine sequence expression (*see* **Sedimentary Environments:** Lake Processes and Deposits). Reduced precipitation reduces both lake level and sediment supply, creating lowstand deposits that are characterized by evaporites in the basin centre surrounded by an exposure surface with little erosion. However, where sediment influx is significant, the lowstand systems tract may be composed of erosive fluvial systems feeding deep-lacustrine turbidite successions. Where lake-level is raised by increased precipitation, increased sediment input may produce a transgressive systems tract of back-stepping lacustrine deltas; where relative lake-level is raised owing to basin subsidence, it may instead be characterized by low clastic input and a thin interval of fine-grained deposits. Highstand deposits may be composed of deltas and associated deep-lacustrine turbidites if high precipitation creates a large influx of clastic sediments; when precipitation is lower, low sediment influx may produce aggradational carbonate packages in shoreline areas and thin fine-grained successions in the lake basin.

Palaeontological Expression of Sequences

Palaeontology is essential to sequence-stratigraphic analysis, and sequence stratigraphy is a useful frame of reference for understanding the fossil record. Fossils provide information on two essential elements of sequence stratigraphy: environment and age. Understanding facies change is also an essential element of sequence-stratigraphic analysis. Biofacies provide palaeoenvironmental constraints that, like lithofacies, trace the facies changes that define sequence-stratigraphic surfaces and systems tracts. Transgression and regression can be readily demonstrated by palaeontological data, providing important information when lithofacies criteria are not definitive. Although coarse clastics reveal palaeoenvironmental information through sedimentary structures that reflect unique hydrodynamic regimes, environmental differences may be more difficult to discern in mudstones. Biofacies analysis can be especially useful in such fine-grained facies, providing critical criteria for differentiating mudstones from different water depths or different depositional systems (e.g. marine versus freshwater).

In addition, some sequence elements have unique biofacies signals. The maximum-flooding surface is one of the most distinctive. Because the maximum-flooding surface commonly has an associated condensed section on basin margins, it typically exhibits an especially high concentration of fossils due to slow sedimentation rates. The fauna and flora typically reflect the culmination of transgression and may indicate the maximum water depth. The condensed section is also commonly marked by a peak in diversity of common marine microfossils (such as foraminifera), the highest abundance of oceanic planktonic microfossils, and the greatest foraminiferal planktonic–benthic ratios. In some cases, a peak of low-oxygen benthic microfossils occurs in the condensed section. Maximum-flooding surfaces can be ideal locations for age control because of the abundance of fossils, particularly of more age-diagnostic oceanic types, such as planktonic foraminifera and ammonites, and of shell material suitable for isotopic analyses. The maximum-flooding surface is especially important in continental margin sequences because it commonly contains the most landward occurrences of biostratigraphically useful open-marine fossils.

The lowstand systems tract may exhibit a distinctly different biofacies signal. Reworked assemblages may be common: falling relative sea-level exposes older sediments higher on the margin and subjects them, and the microfossils in them, to erosion and basinwards redeposition. Lowstand deposits can also have relatively high abundances of terrestrial microfossils, such as pollen, owing to the more direct transport of terrestrial material to the ocean basin.

Chronostratigraphic Aspects of Sequence Stratigraphy

Sequence stratigraphy and chronostratigraphy are intimately entwined. Sequence stratigraphy provides a framework for understanding the relationships between depositional systems in both time and space.

Sequences are chronostratigraphically significant units. Early papers on the subject considered sequence boundaries to be globally synchronous time lines corresponding to times of sea-level fall and provided detailed ties of biostratigraphic schemes to the global sequence record. Although the ages of some sequence boundaries can be established using biostratigraphy, they are often difficult to date because of a paucity of age-significant open-marine fossils in the associated regressive intervals. In contrast, maximum flooding surfaces mark the landwards incursions of open-marine environments and can be well dated because of the abundance of open-marine fossils in the associated condensed sections.

Global and regional sequence records are shown on a type of chronostratigraphic chart referred to as a cycle chart (Figure 8). Cycle charts commonly show the ages of the sequences, the magnitudes of coastal onlap, and the interpreted eustatic changes. Coastal onlap is defined as the progressive landwards onlap of coastal deposits in a depositional sequence; by definition, it excludes marine onlap such as the onlap of lowstand fan deposits. Coastal-onlap curves have their origin in the analysis of seismic-reflection profiles and trace the migration of the point of coastal onlap across a basin margin for each seismic reflection. The migration of this point reflects rises and falls in relative sea-level through time. Sequence boundaries stand out prominently on cycle charts as the horizontal lines at the jagged edge of the saw-tooth curve. The ages of the sequences are indicated by their positions on the time axis, and on some charts the sequences are also tied to biochronostratigraphic zonations. Global and regional cycle charts are derived by comparing coastal-onlap profiles from multiple basin margins. With some understanding of the tectonic history of each basin and a consistent age framework, differences in subsidence can be accounted for in creating a composite global or regional coastal-onlap curve. The first such global cycle chart, produced by Vail *et al.*, assumed that the coastal onlap curves were actually sea-level curves. These workers assumed that the saw-tooth pattern of the curve represented asymmetric rates of sea-level change, with slow sea-level rises and nearly instantaneous sea-level falls. It is now understood that the abrupt basinwards shift of coastal onlap at a sequence boundary reflects the jump of deposition from the topsets of the highstand systems tract, over the offlap break, and into the basin for the lowstand. Later versions of the cycle charts added a more symmetric eustatic curve to represent estimated global sea-level changes.

Our understanding of the chronostratigraphy of global sequences continues to improve. Integration of isotope stratigraphy and biostratigraphy from marine and onshore boreholes drilled as part of the Ocean Drilling Program has resulted in significant refinements to the dating of the sequence record of the Cenozoic and Late Cretaceous. For many Cenozoic sequences, ages of sequence boundaries can be determined at a resolution of 0.5 Ma or better.

Genesis of Sequence-Stratigraphic Units

To understand the genesis of sequence-stratigraphic units, three essential factors need to be considered: sea-level change, tectonics, and sediment supply.

Sea-Level Change

The importance of sea-level as a causal mechanism in the development of sequences is well understood, and it is important at different periodicities and scales (Table 2A). Supercontinent cycles operate at a scale of hundreds of millions of years and exercise a fundamental control on the volume of the ocean basins and hence on sea-level. These drive first-order sea-level cycles. Second-order cycles may be influenced in part by changes in rates of seafloor spreading on a scale of tens of millions of years. Faster seafloor spreading creates a greater volume of hot and more buoyant mid-ocean ridge material, decreasing the volume of the ocean basins and raising sea-level; slower seafloor spreading creates less mid-ocean ridge material, increasing the volume of the ocean basins.

At the period of third-order sequences, generally 1–10 Ma, mechanisms of sea-level change are more problematic. In times of significant continental glaciation, gross patterns of ice volume may provide a mechanism. Growth of continental glaciers decreases the volume of water in the oceans, lowering sea-level; melting of ice-sheets releases water into the oceans, raising sea-level. Continental glaciation was the major factor controlling sea-level in the Pleistocene, and its importance as early as the Miocene is widely accepted. Recent results from the study of

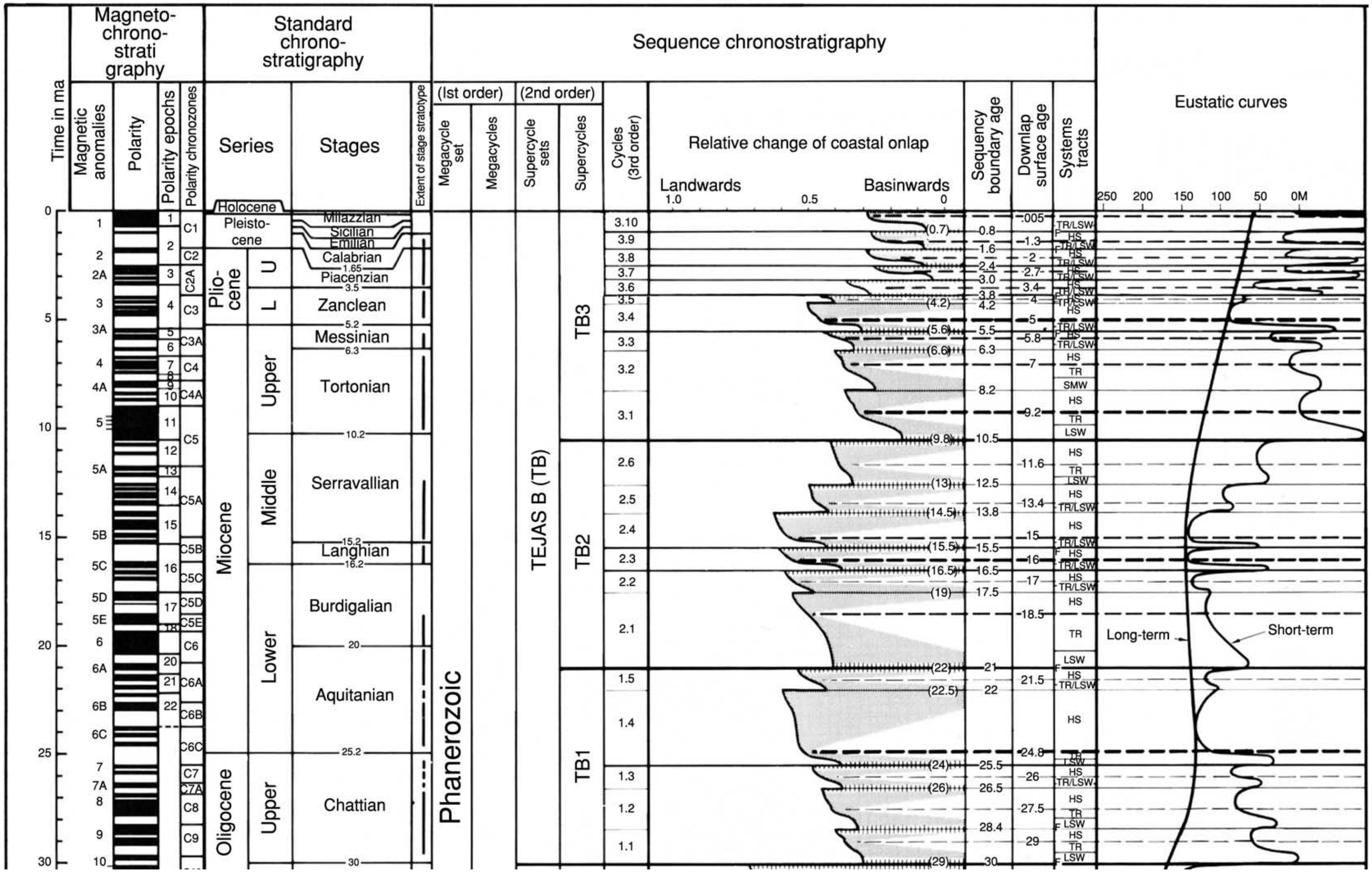


Figure 8 Cenozoic chronostratigraphic- and eustatic-cycle chart. Systems tracts abbreviations: HS, highstand; TR, transgressive; LSW, Lowstand wedge; SMW, shelf-margin wedge. (Reproduced from Haq B, Hardenbol J, and Vail PR (1988) Mesozoic and Cenozoic chronostratigraphy and eustatic cycles. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 71–108. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.)

Table 2 Mechanisms for sea-level change. (A) Orders of sea-level cyclicity. (B) Characteristics of Milankovitch astronomical cycles

(A)

Order	*Duration*	*Mechanism*
First	200–400 Ma	Breakup of continents
Second	10–100 Ma	Volume of mid-ocean ridges
Third	1–10 Ma	Glacioeustasy, possibly tectonics
Fourth	200–500 Ka	Astronomical forcing of glacioeustasy or climate
Fifth	20–200 Ka	Astronomical forcing of glacioeustasy or climate
Sixth	1–10 Ka	Astronomical forcing of glacioeustasy or climate

(B)

Cycle	*Duration*	*Mechanism*
Eccentricity	100 Ka and 400 Ka	Variations in degree of roundness of orbit of Earth around Sun
Obliquity	41 Ka	Variations in angle of tilt of axis of Earth relative to Sun
Precession	19–23 Ka	Variations in wobble of rotation of Earth

Ocean Drilling Program sites suggest that continental ice-sheets may have existed even earlier, with significant volumes as early as the middle Eocene and small- to moderate-sized sheets as far back as the Late Cretaceous. However, during periods of Earth history without significant glaciation, mechanisms for eustatic change are less clear. Variation in intraplate stress has been proposed as a mechanism for inducing apparent sea-level changes of as much as 100 m on the flanks of passive margins. Variation through time in the irregularities of the geoid (equipotential surface of the gravitational field) has been postulated to cause sea-level changes at different times in different parts of the globe. A more exotic mechanism invoked is an asteroid or comet impact that induces the global release of stress at plate boundaries and a resultant isostatic response of continental margins.

Fourth-order and higher cycles have periods of hundreds of thousands of years or less. Cyclic variations in the tilt and wobble of the Earth's axis, called Milankovitch cycles, cause variations in the intensity of solar radiation, which can strongly influence climate on the scale of fourth- and fifth-order cycles (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). Milankovitch cycles include three components with different periods: eccentricity, obliquity, and precession (Table 2B). Milankovitch-related climate changes affect the size of the polar ice-caps and thus global sea-level change. They can also influence monsoonal fluctuations and hence vary the amount of water delivered to and stored in lakes, aquifers, and soils over periods as short as tens of thousands of years. During the opening of the South Atlantic Ocean in the Early Cretaceous, the Parana-Benue basin is estimated to have been able to store enough water to vary sea-level by 3.46 m. It is thought that obliquity is more important at high latitudes, and precession is more important in tropical latitudes.

Small eustatic changes associated with higher-order cycles may be overprinted by larger-scale changes associated with lower-order cycles, resulting in quite different stratigraphic expressions of the same order sequence. Figure 9 shows that the effect of a small (10 m amplitude) fourth-order sea-level fall can be significant if it occurs during a period of overall limited accommodation (e.g. on the falling leg of a third-order sea-level cycle), enhancing the expression of sequence boundaries (Section I). In contrast, during a period of overall high accommodation (e.g. on the rising leg of a third-order sea-level cycle), the fourth-order fall is muted and difficult to detect in the section, while flooding surfaces are enhanced (Section II).

The early cycle charts of Vail *et al.* proposed sea-level falls for third-order sequences from as little as tens of meters to more than 300 m. Second-generation charts by Vail *et al.* indicated smaller variations, but still with some pre-Pleistocene changes of more than 100 m. More recent work arising from the Ocean Drilling Program suggests that these earlier estimates may be too high. Based on analysis of oxygen isotope records and backstripping of continental-margin sites, Cenozoic sea-level changes are estimated to be less than 100 m, approximately one-quarter the older estimates.

The relative importance of self-regulating (autocyclic) and externally forced (allocyclic) cycles in the formation of fourth- and higher-order sequences is an important consideration in sequence-stratigraphic analysis. In a clastic system, a delta lobe normally produces an upward shoaling package as it progrades

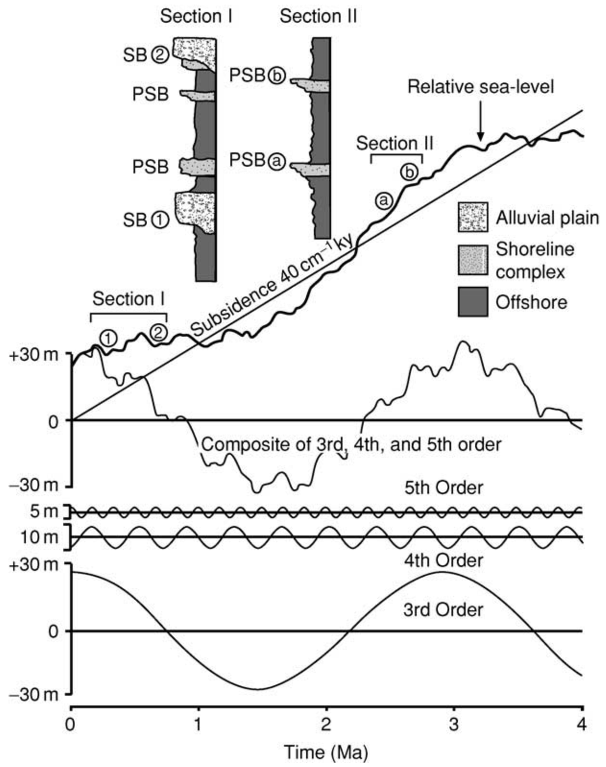


Figure 9 Interaction between subsidence and multiple orders of eustatic cyclicity, and the effect on sequence stratigraphy. Third-, fourth-, and fifth-order eustatic changes combine with subsidence to produce a relative sea-level curve, which controls the character of the sequence expression. Section I reflects deposition during a period of lower accommodation related to the third-order sea-level fall; points 1 and 2 identify fourth-order sea-level falls corresponding to sequence boundaries (SB). Section II reflects deposition during a period of greater accommodation related to the third-order sea-level rise; points a and b designate fourth-order sea-level rises corresponding to parasequence boundaries (PSB). (Adapted from Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1990) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1–55.)

into an area, and is capped by a flooding surface after sedimentation shifts to another lobe and the abandoned lobe sinks through subsidence and compaction. In a carbonate system, upward growth of carbonate sediments normally produces an upward shoaling pattern until all the available accommodation space is used; carbonate production is then shut off until subsidence provides space for carbonate growth to resume. These successions may occur repeatedly at scales of tens of thousands of years, similar to the time scales of higher order sea-level cycles. Thus, differentiating autocyclicity from allocyclicity in high-resolution sequence analysis may be difficult.

Tectonics

Tectonics is another major factor that contributes to accommodation; subsidence creates space for sediment accumulation, and uplift takes it away. Subsidence patterns vary according to tectonic setting. In an extensional basin, subsidence may be a response to lithospheric thinning and cooling following rifting. Subsidence due to extensional faulting is typically greatest in the early phases and decreases with time. In a foreland basin, subsidence may be a response to lithospheric flexure due to loading in the adjacent fold-thrust belt. Subsidence in foreland basins commonly accelerates over time as the load on the lithosphere increases during thrusting, followed by post-orogenic rebound. In cratonic basins, subsidence may be due to a subtler regional warping.

Tectonics is considered to be generally too slow a process to cause third- and higher-order sequences. Certainly, movement on individual faults can be significant at shorter time scales. However, overall, changes in rates of basin subsidence tend to occur on a time scale of tens of millions of years, which is a broader period than third-order sea-level variations. As a result, tectonics exerts a broader-brush influence on the nature of expression of the sequences, upon which higher-frequency sea-level events are superimposed (Figure 9).

Sediment Supply

Sediment supply controls how rapidly the accommodation space is filled, and the balance between accommodation and sediment supply controls sediment stacking patterns and thus the expression of sequences. In a setting with a high input of sediments, such as a delta, sedimentation may fill the available space in the two systems tracts with lower accommodation, the lowstand wedge and the highstand systems tract. As a result, these intervals will be characterized by aggradational and progradational packages. The sequence boundary reflects low accommodation relative to sediment input. Retrogradation of the depositional system will probably occur only at the peak of transgression, and condensed sections will be developed only in offshore locations. In settings with a low influx of sediments, accommodation may exceed the ability of the sedimentary system to fill it. In such cases, only during the part of the sequence when sea-level is falling and accommodation is lowest will sediment advance basinwards. The transgressive and highstand systems tracts will probably be characterized by sediment starvation and represented as a condensed section. Climate changes arising from Milankovitch cycles can exert a strong influence on sediment supply.

Climate variations can affect precipitation rates and vegetation patterns, which in turn influence erosion rates and sediment supply.

Conclusions

Sequence concepts provide a genetic frame of reference that has made stratigraphy a more dynamic process-based field of study compared with the static nomenclatural emphasis of traditional approaches. Sequence stratigraphy encourages an integrative approach, bringing together lithofacies analysis, geophysics, biostratigraphy, and chronostratigraphy. It provides a stratigraphic record of changes in sea-level, tectonics, and climate on local to global scales. The significant volume of jargon common in sequence-stratigraphic literature can be a barrier to understanding, and the key to wider understanding of sequence-stratigraphic concepts is to focus on the processes that create these genetically related rock units rather than on the nomenclature.

The historical association of the sequence concept with a universal explanation of global sea-level change may also hinder the acceptance of the discipline. It is important to recognize that accommodation, rather than sea-level, is the key concept in sequence stratigraphy. Because the interplay of accommodation and sediment supply controls sequence expression, the concepts of sequence-stratigraphic interpretation can be used to correlate and understand genetically related stratigraphic units regardless of whether the units are global or regional in extent.

Research drilling on coastal plains and margins will provide important new data to advance our understanding of Earth history. Continuous cores provide complete records of sequence expression, together with age data, in different geological settings. The isotopic records from such continuous-core material can provide unique insights into ice volume and palaeoceanographic changes, which shed light on the relationship between global sea-level change and sequence stratigraphy.

The nature of lithologic heterogeneities in reservoirs and aquifers and their impact on fluid flow will be better understood through the integration of sequence concepts with process-based geological models and increasingly powerful computer-based 3D-visualization technology. This detailed information allows numerical characterization of vertical and lateral heterogeneities that influence fluid flow. Sequence stratigraphy hold great growth potential as a tool for detailed stratigraphic analysis in petroleum exploration and production and in ground water resource management.

See Also

Earth: Orbital Variation (Including Milankovitch Cycles). **Palaeoclimates**. **Sedimentary Environments:** Depositional Systems and Facies; Carbonate Shorelines and Shelves; Lake Processes and Deposits. **Sedimentary Processes:** Depositional Sedimentary Structures. **Seismic Surveys**. **Stratigraphical Principles**. **Unconformities**.

Further Reading

Bohacs K and Suter J (1997) Sequence stratigraphic distribution of coaly rocks: fundamental controls and paralic examples. *American Association of Petroleum Geologists Bulletin* 81: 1612–1632.

Bohacs KM, Carroll AR, Neal JK, and Mankiewicz PJ (2000) Lake-basin type, source potential, and hydrocarbon character: an integrated sequence-stratigraphic–geochemical framework. In: Gierlowski-Kordesch EH and Kelts KR (eds.) *Lake Basins Through Space and Time,* pp. 3–34. Studies in Geology 46. Tulsa: American Association of Petroleum Geologists.

Christie-Blick N and Driscoll NW (1995) Sequence stratigraphy. *Annual Review of Earth and Planetary Sciences* 23: 451–478.

Christie-Blick N, Mountain GS, and Miller KG (1990) Seismic stratigraphic record of sea-level change. In: National Research Council (ed.) *Studies in Geophysics: Sea-level Change,* pp. 116–140. Washington, DC: National Academy Press.

Coe AL, Bosence DWJ, Church KD, *et al.* (2003) *The Sedimentary Record of Sea-Level Change.* Cambridge: Cambridge University Press.

Emery D and Myers KJ (1996) *Sequence Stratigraphy.* Oxford: Blackwell Science.

Haq B, Hardenbol J, and Vail PR (1988) Mesozoic and Cenozoic chronostratigraphy and eustatic cycles. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach,* pp. 71–108. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Loutit TS, Hardenbol J, Vail PR, and Baum GR (1988) Condensed sections: the key to age dating and correlation of continental margin sequences. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach,* pp. 183–213. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Miall AD (1997) *The Geology of Stratigraphic Sequences.* Berlin: Springer-Verlag.

Miall AD and Miall CE (2001) Sequence stratigraphy as a scientific enterprise: the evolution and persistence of conflicting paradigms. *Earth Science Reviews* 54: 321–348.

Miller KG (2002) The role of ODP in understanding the causes and effects of global sea-level change. *JOIDES Journal* 28: 23–28.

Posamentier HW, Jervey MT, and Vail PR (1988) Eustatic controls on clastic deposition I: conceptual framework. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 109–124. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Posamentier HW, Allen GP, James DP, and Tesson M (1992) Forced regressions in a sequence stratigraphic framework: concepts, examples, and exploration significance. *American Association of Petroleum Geologists Bulletin* 76: 1687–1709.

Sarg JF (1988) Carbonate sequence stratigraphy. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 155–181. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Shanley KW and McCabe PJ (1994) Perspectives on the sequence stratigraphy of continental strata. *American Association of Petroleum Geologists Bulletin* 78: 544–568.

Sloss LL (1963) Sequences in the cratonic interior of North America. *Geological Society of America Bulletin* 74: 93–114.

Vail P, Mitchum RM Jr, Todd RG, *et al.* (1963) Seismic stratigraphy and global changes of sea level. In: Payton CE (ed.) *Stratigraphic Interpretation of Seismic Data*, pp. 49–212. (in 11 parts). Memoir 26. Tulsa: American Association of Petroleum Geologists.

Van Wagoner JC, Posamentier HW, Mitchum RM Jr, *et al.* (1963) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hastings BS, Kendall CStCC, *et al.* (eds.) *Sea-Level Changes: An Integrated Approach*, pp. 39–45. Special Publication 42. Tulsa: Society of Economic Paleontologists and Mineralogists.

Van Wagoner JC, Mitchum RM, Campion KM, and Rahmanian VD (1963) Siliciclastic sequence stratigraphy in well logs, cores, and outcrops. *American Association of Petroleum Geologists Methods in Exploration Series* 7: 1–55.

SHIELDS

K C Condie, New Mexico Tech, Socorro, NM, USA

General Features

Precambrian shields are stable parts of the continents composed of Precambrian rocks with little or no sediment cover (Figure 1). Rocks in shields range in age from 0.5 to >3.5 Ga. Metamorphic and plutonic rock types dominate, and temperature-pressure regimes recorded in rocks now exposed at the surface suggest burial depths ranging from as shallow as 5 km to as deep as 40 km or more. Shield areas, in general, exhibit very little relief and have remained tectonically stable for long periods of time. They comprise about 11% of the total crust by volume, with the largest shields occurring in Africa, Canada, and Antarctica (Figure 1). Platforms are also stable parts of the crust with little relief. They are composed of Precambrian basement similar to that exposed in shields, but overlain by 1 to 3 km of relatively undeformed sedimentary rock. Sedimentary rocks on platforms range in age from Precambrian to Cenozoic and reach thicknesses up to 5 km, as, for instance, in the Williston basin in the north-central United States. Platforms comprise most of the crust in terms of volume (35%) and most of the continental crust in terms of both area and volume. The largest platform is the Eurasian platform (Figure 1). Shields and the Precambrian basement of platforms are collectively referred to as cratons. A craton is an isostatically positive portion of the continent that is tectonically stable relative to adjacent orogens.

Seismic Characteristics

Shields and platforms have similar seismic wave velocities and layers (Figure 2). The difference in their mean thickness reflects primarily the presence of the sediment layer in the platforms with P-wave velocities $\leq 5\,\mathrm{km\,s^{-1}}$. Upper layer thicknesses range from about 10 to 25 km and each of the lower layers ranges from 16 to 30 km. The P-wave velocities in both layers are rather uniform, generally ranging from 6.0 to

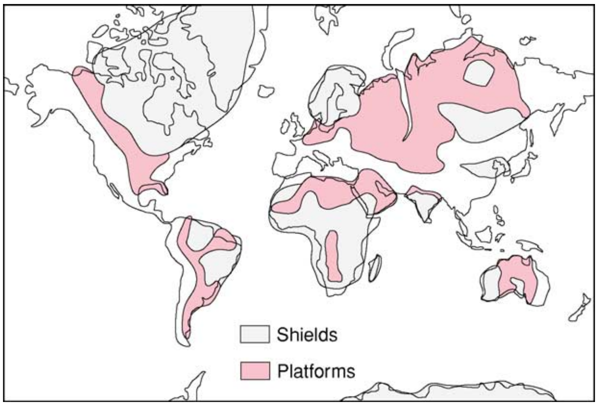


Figure 1 Map showing the distribution of Precambrian shields and platforms.

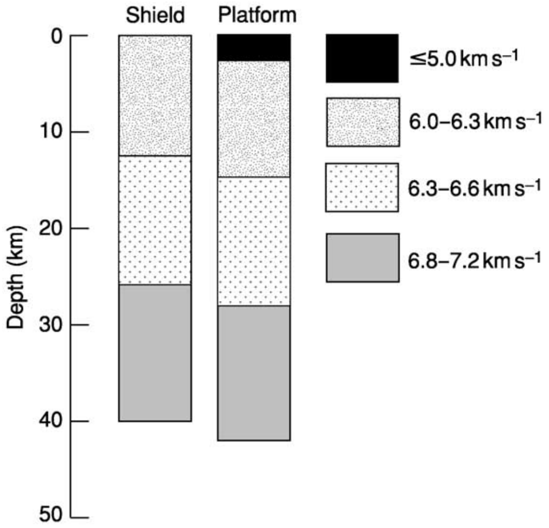


Figure 2 Seismic cross-sections of shield and platform crust with typical P-wave velocity distributions.

6.3 km s^{-1} in the upper layer, 6.3 to 6.6 km s^{-1} in the middle layer, and 6.8 to 7.2 km s^{-1} in the lower layer. Upper mantle velocities beneath shields and platforms are typically in the range of 8.1–8.2 km s^{-1}, rarely reaching 8.6 km s^{-1}. Seismic reflection studies show an increase in the number of reflections with depth, and generally weak, but laterally continuous, reflections at the Moho, the seismic discontinuity defining the base of the crust.

Table 1 Average chemical composition of continental crust

	Crust composition[a]			
Component	*Upper*	*Middle*	*Lower*	*Total*
SiO_2	66.3	60.6	52.3	59.7
TiO_2	0.7	0.8	0.54	0.68
Al_2O_3	14.9	15.5	16.6	15.7
FeOT	4.68	6.4	8.4	6.5
MgO	2.46	3.4	7.1	4.3
MnO	0.07	0.1	0.1	0.09
CaO	3.55	5.1	9.4	6.0
Na_2O	3.43	3.2	2.6	3.1
K_2O	2.85	2.0	0.6	1.8
P_2O_5	0.12	0.1	0.1	0.11
Rb	87	62	11	53
Sr	269	281	348	299
Ba	626	402	259	429
Th	9.1	6.1	1.2	5.5
Pb	18	15.3	4.2	13
U	2.4	1.6	0.2	1.4
Zr	162	125	68	118
Hf	4.4	4.0	1.9	3.4
Nb	10.3	8	5	7.8
Ta	0.82	0.6	0.6	0.7
Y	25	22	16	21
La	29	17	8	18
Ce	59.4	45	20	42
Sm	4.83	4.4	2.8	4.0
Eu	1.05	1.5	1.1	1.2
Yb	2.02	2.3	1.5	1.9
V	86	118	196	133
Cr	112	150	215	159

[a] Major elements in weight percent of the oxide; trace elements in parts per million.

Composition of the Crust in Cratons

Several approaches have been used to estimate the chemical and mineralogical composition of the crust. One of the earliest methods to estimate the composition of the upper continental crust is based on chemical analysis of glacial clays, which were assumed to be representative of the composition of large portions of the upper continental crust. Probably the most accurate estimates of the composition of the upper continental crust come from extensive sampling of rocks exhumed from varying depths in Precambrian shields and from the composition of Phanerozoic shales. Because the lower continental crust is not accessible for sampling, indirect approaches must be used. These include (1) measurement of seismic-wave velocities of crustal rocks in the laboratory at appropriate temperatures and pressures, and comparing these to observed velocity distributions in the crust, (2) sampling and analysing rocks from blocks of continental crust exhumed from middle to lower crustal depths, and (3) analysing xenoliths of lower crustal rocks brought to the surface during volcanic eruptions.

The average chemical composition of the upper continental crust is reasonably well known. An average composition is similar to granodiorite (Table 1), although there are differences related to the age of the crust. The composition of the middle and lower continental crust is much less well constrained. Uplifted crustal blocks, xenolith populations, and seismic velocity data suggest that the middle crust is intermediate (andesitic) in composition and that the lower crust is mafic (basalt-like) in overall composition. The estimate of total continental crust composition in Table 1 is a mixture of upper, middle, and lower crustal averages in equal amounts. The composition is similar to other published total crustal compositions indicating an overall intermediate composition. Incompatible elements, which are elements that are strongly partitioned into the liquid phase on melting, are known to be concentrated chiefly in the continental crust. During melting in the mantle, these elements are enriched in the magma, and thus are transferred upward into the crust as magmas rise. Between 35 and 65% of the most incompatible elements (such as Rb, Th, U, K, and Ba) are contained in the continents, whereas continents contain <10% of the least incompatible elements (such as Y, Yb, and Ti).

Cratonization

Cratonization refers to the process of craton formation. Collisional orogenesis occurs when plates carrying blocks of continental crust collide with each other, producing major mountain chains such as the Himalayas. Cratons are the end product of collisional orogenesis, and thus they are the building blocks of continents. Collisional mountain chains are eroded away in 200–400 Ma, leaving the roots exposed at Earth's surface as 'sutures' between pre-existing crustal blocks. The complex amalgamation of crustal blocks and orogen roots comprises today's cratons.

Using a variety of radiogenic isotopic systems and estimated closure temperatures in various minerals, it is possible to track the cooling and uplift histories of cratons. Results suggest a wide variation in cooling and uplift rates, with most orogens having cooling rates of <2°C My^{-1}, whereas a few (such as southern Brittany) have cooled at rates of >10°C My^{-1} (**Figure 3**). In most cases, it would appear to take a minimum of 300 My to make a craton. Some terranes, such as Enderbyland in Antarctica, have had very long, exceedingly complex cooling histories lasting for more than 2 Ga. Many orogens, such as the Grenville Orogen (*see* **Grenvillian Orogeny**) in eastern Canada, have been exhumed as indicated by unconformably overlying sediments, and then reheated during subsequent burial and then re-exhumed. In some instances, postorogenic thermal events such as plutonism and metamorphism have thermally overprinted earlier segments of the cooling history of an orogen, such that only the very early high-temperature history (>500°C) and, perhaps, the latest exhumation record (<300°C) are preserved. Fission track ages suggest that final uplift and exhumation of some orogens, such as the 1.9-Ga Trans-Hudson Orogen in central Canada, may be related to the early stages of supercontinent fragmentation.

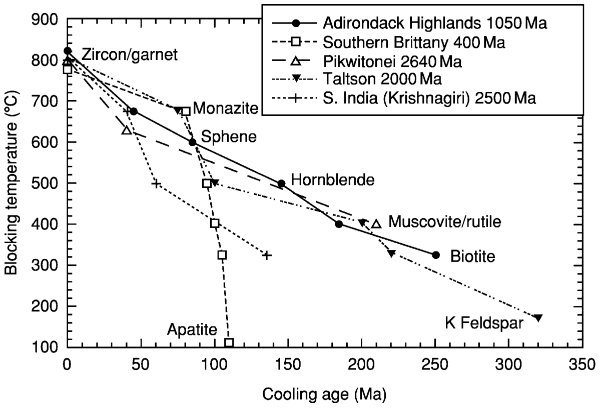


Figure 3 Cooling histories of several orogens leading to the production of stable cratons. Blocking temperature is the temperature at which the daughter isotope is captured by the indicated host mineral.

Crustal Provinces and Terranes

The Canadian Shield can be subdivided into structural provinces based on differences in structural trends and style of folding. Structural trends are defined by foliation, fold axes, and bedding, and sometimes by geophysical anomalies. Boundaries between the provinces are drawn where one trend cuts across another, along either unconformities or structural-metamorphic breaks. Large numbers of isotopic ages from the Canadian Shield indicate that structural provinces are broadly coincident with age provinces. Similar relationships have been described on other continents and lead to the concept of a crustal province (**Figure 4**).

Terranes are fault-bounded crustal blocks that have distinct lithologic and stratigraphic successions and that have geological histories different from neighbouring terranes (*see* **Terranes, Overview**). Most terranes have collided with continental crust, either along transcurrent faults or at subduction zones, and have been sutured to continents. Many terranes contain faunal populations and palaeomagnetic evidence indicating they have been displaced great distances (thousands of kilometres) from their sources prior to continental collision. For instance, Wrangellia, which collided with western North America in the Late Cretaceous, had travelled many thousands of kilometres from what is now the South Pacific and is now represented in Vancouver Island. Results suggest that as much as 30% of North America was formed by terrane accretion in the past 300 Ma and that terrane accretion has been an important process in the growth of continents.

Terranes form in a variety of tectonic settings, including island arcs, oceanic plateaus, volcanic islands, and microcontinents. Continental crust may be fragmented and dispersed by rifting or strike-slip faulting. In western North America, dispersion is occurring along transform faults such as the San Andreas and Fairweather faults, and in New Zealand movement along the Alpine Transform Fault is fragmenting the Campbell Plateau from the Lord Howe Rise. Baja California and California west of the San Andreas Fault were rifted from North America about 4 Ma, and today this region is a potential terrane moving northwards, perhaps on a collision course with Alaska. Terranes may continue to fragment and

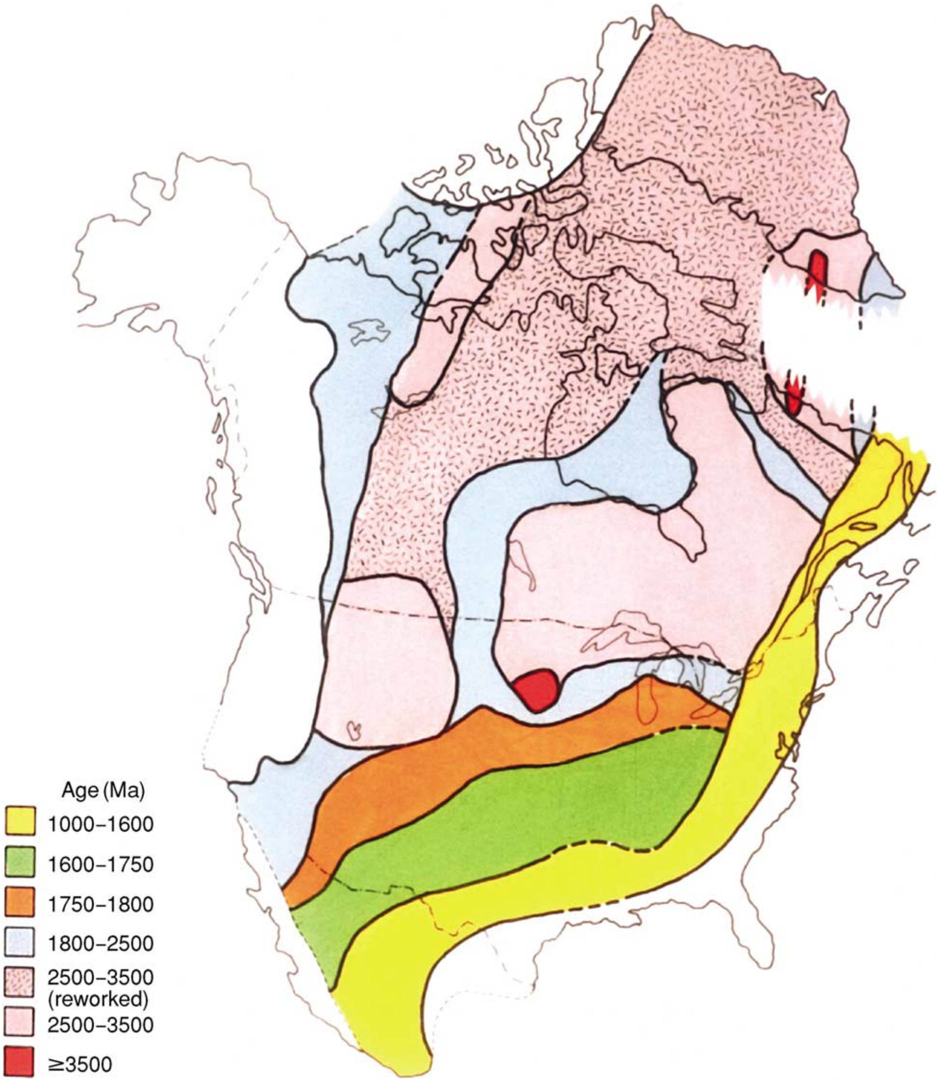


Figure 4 Crustal provinces in North America.

disperse after collision with continents, as did Wrangellia, which is now distributed in pieces from Oregon to Alaska. The 1.9-Ga-old Trans-Hudson Orogen in Canada and the 1.65- to 1.75-Ga-old Yavapai Orogen in the south-west United States are examples of Proterozoic orogens composed of terranes, and the Alps, Himalayas, and American Cordillera are Phanerozoic examples of orogens composed of terranes. Most crustal provinces are composed of terranes, and in turn, cratons are composed of exhumed orogens. In fact, terranes might be considered as the basic building blocks of continents, and terrane collision as a major means by which continents grow in size.

A crustal province is an orogen, active or exhumed, composed of terranes, and it records a similar range of isotopic ages and exhibits a similar postamalgamation deformational history (Figure 4). Shields and platforms are composed of exhumed crustal provinces. Structural trends within provinces range from linear to exceedingly complex swirling patterns reflecting multiple deformation superimposed on differing terrane structural patterns. Exhumed crustal provinces that have undergone numerous episodes of deformation and metamorphism are old orogens. Isotopic dating using several isotopic systems is critical to defining and unravelling the complex, polydeformational histories of crustal provinces.

The definition of ‘crustal province’ is not always unambiguous. Most crustal provinces contain rocks of a wide range in age and record more than one period of deformation, metamorphism, and plutonism. For instance, the Trans-Hudson orogen in North America includes rocks ranging in age from about 1.7 to 3.0 Gy and records several periods of complex deformation and regional metamorphism. Likewise, the Grenville province in eastern North America records a polydeformational history, with rocks ranging in age from 1.0 to 2.7 Ga. Some parts of crustal provinces are new mantle-derived crust, known as juvenile crust, whereas other parts represent reworked older crust. Reworking, also known as overprinting or reactivation, describes crust that has been deformed, metamorphosed, and partially melted more than once. There is increasing evidence that crustal reworking results from continental collisions, and large segments of continental crust appear to have been reactivated by such collisions. For instance, much of central Asia at least as far north as the Baikal Rift, which is in a craton, was affected by the India–Tibet collision beginning about 50 Ma. Widespread faulting and magmatism at present crustal levels suggest that deeper crustal levels may be extensively reactivated. In Phanerozoic collisional orogens where deeper crustal levels are exposed, such as the Appalachian and Variscan orogens, there is isotopic evidence for widespread reactivation.

Sediments Deposited on Cratons

Rock assemblages deposited on cratons are mature clastic sediments, chiefly quartz arenites and shales, and shallow marine carbonates. In Late Archaean and Palaeoproterozoic successions, banded iron formation may also be important. Cratonic sandstones are relatively pure quartz sands, reflecting intense weathering, low relief in source areas, and prolonged transport across subdued continental surfaces. Commonly associated marine carbonates are deposited as blankets and as reefs around the basin margins. Transgression and regression successions in large cratonic basins reflect the rise and fall of sea level, respectively.

Depositional systems in cratonic basins vary depending on the relative roles of fluvial, aeolian, deltaic, wave, storm, and tidal processes. Spatial and temporal distribution of sediments is controlled by regional uplift, the amount of continent covered by shallow seas, and climate. If tectonic uplift is important during deposition, continental shelves are narrow and sedimentation is dominated by wave and storm systems. However, if uplift is confined chiefly to craton margins, sediment yield increases into the craton, and fluvial and deltaic systems may dominate. For transgressive marine clastic sequences, shallow seas are extensive and subtidal, and storm-dominated and wave-dominated environments are important. During regression, fluvial and aeolian depositional systems become dominant.

The rates of subsidence and uplift in cratons are a function of the time interval over which they are measured. Current rates are of the order of a few centimetres per year, whereas data from older successions suggest rates 1–2 orders of magnitude slower. In general, Phanerozoic rates of uplift appear to have been 0.1–1 cm year^{-1} over periods of 10^4–10^5 years and over areas of 10^4–10^6 km^2. Craton subsidence can be considered in terms of two stages: in the first stage, subsidence rate varies greatly, whereas the second stage subsidence is widespread. After about 50 Ma, the depth of subsidence decreases exponentially to a constant value.

Several models have been suggested to explain cratonic subsidence. Sediment loading, lithosphere stretching, and thermal doming followed by contraction are the most widely cited mechanisms. Although the accumulation of sediments in a depression loads the lithosphere and causes further subsidence, calculations indicate that the contribution of sediment loading to subsidence must be minor compared to other effects. Subsidence at passive margins may result from thinning of continental crust by progressive creep of the ductile lower crust towards the suboceanic upper mantle. As the crust thins, sediments accumulate in overlying basins.

Supercontinents and Cratons

Supercontinents are large continents that include several or all of the existing cratons. Matching of continental borders, stratigraphic sections, and fossil assemblages are some of the earliest methods used to reconstruct ancient supercontinents. Today, in addition to these methods, we have polar wandering paths, seafloor spreading directions, hotspot tracks, and correlation of crustal provinces. The use of computers in matching continental borders has resulted in more accurate and objective fits. One of the most definitive matching tools in reconstructing plate positions in a former supercontinent is a piercing point. A piercing point is a distinct geologic feature such as a fault or terrane that strikes at a steep angle to a rifted continental margin, the continuation of which should be found on the continental fragment rifted away.

The youngest supercontinent is Pangaea (*see* **Pangaea**), which formed between 450 and 320 Ma and includes most of the existing continents (Figure 5). Pangaea began to fragment about 160 Ma and is still dispersing today. Gondwana (*see* **Gondwanaland and Gondwana**) was a southern hemisphere supercontinent composed principally of South America, Africa, Arabia, Madagascar, India, Antarctica, and Australia. It formed in the latest Proterozoic and was largely completed by the Early Cambrian (750–550 Ma). Later it became incorporated in Pangaea. Laurentia, which is also part of Pangaea,

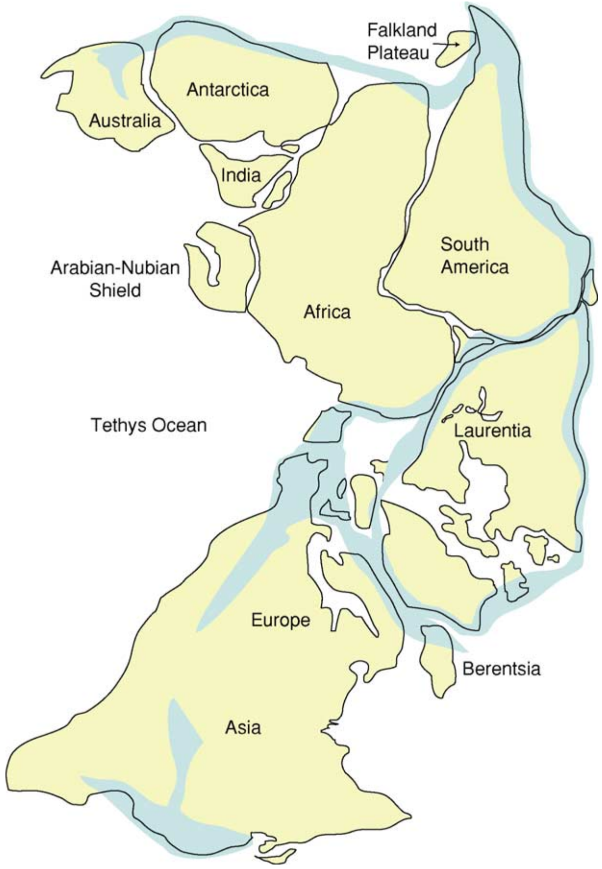


Figure 5 Pangaea, a supercontinent that formed between 450 and 320 Ma and began to fragment about 160 Ma. The major collisional orogens are indicated.

includes most of North America, Scotland, and Ireland north of the Caledonian suture, and Greenland, Spitzbergen, and the Chukotsk Peninsula of eastern Siberia. Although the existence of older supercontinents is likely, their configurations are not well known. Geological data strongly suggest the existence of supercontinents in the Proterozoic and in the Late Archaean. Current thinking is that supercontinents have been episodic, giving rise to the idea of a supercontinent cycle. A supercontinent cycle consists of rifting and break up of one supercontinent, followed by a stage of reassembly in which dispersed cratons collide to form a new supercontinent, with most or all fragments in different configurations, compared to the older supercontinent. The supercontinent cycle provides a record of the processes that control the formation and redistribution of cratons throughout Earth history.

See Also

Analytical Methods: Fission Track Analysis. **Antarctic**. **Earth:** Crust. **Gondwanaland and Gondwana**. **Grenvillian Orogeny**. **Pangaea**. **Precambrian:** Overview. **Tectonics:** Mountain Building and Orogeny. **Terranes, Overview**.

Further Reading

Beardsmore GR and Cull JP (2001) *Crustal Heat Flow.* Cambridge, UK: Cambridge University Press.

Brown M and Rushmer T (eds.) (2003) *Evolution and Differentiation of the Continental Crust*. Cambridge, UK: Cambridge University Press.

Condie KC (ed.) (1992) *Proterozoic Crustal Evolution.* Amsterdam: Elsevier.

Fountain DM, Arculus R, and Kay RW (1992) *Continental Lower Crust*. Amsterdam: Elsevier.

Juteau T and Maury R (1999) *The Oceanic Crust, from Accretion to Mantle Recycling*. New York: Springer-Verlag.

Kleine E (2003) The ocean crust. In: Rudnick RL (ed.) *The Crust, Treatise on Geochemistry,* vol. 3, pp. 433–463. Amsterdam: Elsevier.

Leitch EC and Scheibner E (eds.) (1987) *Terrane Accretion and Orogenic Belts*. Geodynamics Series 19. Washington DC: American Geophysical Union.

Meissner R (1986) *The Continental Crust, A Geophysical Approach*. New York: Academic Press.

Moores EM and Twiss RJ (1995) *Tectonics*. New York: WH Freeman.

Rudnick RL and Fountain DM (1995) Nature and composition of the continental crust: a lower crustal perspective. *Reviews of Geophysics* 33: 267–309.

Taylor SR and McLennan SM (1985) *The Continental Crust: Its Composition and Evolution*. Oxford: Blackwell Scientific Publication.

Windley BF (1995) *The Evolving Continents,* 3rd edn. New York: John Wiley & Sons.

SHOCK METAMORPHISM

P S DeCarli, SRI International, Menlo Park, CA, USA

Introduction

The term 'shock metamorphism', synonymous with 'shock wave metamorphism' or 'impact metamorphism', refers to the range of effects produced by the collision of two bodies, e.g., by the collision of an asteroid with the Earth. These effects include fracturing, the formation of planar deformation features (PDF), the formation of high-pressure phases, melting, and vaporization. Our knowledge of shock metamorphism, currently quite incomplete, is derived from laboratory shock experiments, static high-pressure experiments, studies of naturally impacted materials, theoretical analyses, and numerical computations.

It is generally accepted that the history of the solar system is one of repeated collisions between orbiting bodies. Lunar craters, now widely accepted as impact craters, provide a partial record of that history. Only during the past 50 years has it become evident that the Earth, because of its higher gravity, should have experienced about twice as many large craters per unit area as the Moon. Most of these craters on the continental crust have been deformed, modified by erosion, and buried by sediments or volcanism.

Between 1960 and 2003, about 170 terrestrial impact craters were identified, and three to five newly identified craters are added to the list each year.

The minimum velocity of an encounter between the Earth and a body within our solar system is 11.2 km s^{-1}, the escape velocity of the Earth. Photographic measurements of meteors, the familiar shooting stars, indicate that they enter the atmosphere at velocities in the range 13–30 km s^{-1}; this is an appropriate velocity range for encounters with asteroids. Comets encounter the Earth at velocities in the range of 30 km s^{-1} (short period comets) to 70 km s^{-1} (very long period comets).

The fate of a body entering the Earth's atmosphere at very high velocity depends on such details as the strength and density of the body, its velocity, and its angle of entry. If the body is non-spherical, details of shape and orientation must also be considered, e.g., whether an elongated body enters the atmosphere in point-first or side-first orientation. To simplify further discussion, only vertical impacts of spherical bodies that are strong enough to survive passage through the atmosphere are considered. The interaction of a fast-moving fragile body with the atmosphere can produce an effect equivalent to a large nuclear explosion at an altitude above 20 km. Resultant pressures at the surface may knock down trees, but are too low to produce shock metamorphic effects in minerals.

A very large body, greater than 10 km in diameter, will not be sensibly retarded by the atmosphere. Impact with the Earth will result in the formation of a large crater, greater than 100 km in diameter. Only two craters larger than 100 km in diameter are known to have formed within the past 150 million years. The larger of the two, the 65-million-year-old Chicxulub crater, Yucatan, Mexico, is buried under more than 300 m of carbonate rocks, and was identified in 1981 by the recognition of circular patterns in gravity and magnetic field data. Shock metamorphic features in drill cores have confirmed the identification. The iridium-rich Cretaceous–Teritary K–T boundary layer, which contains shock metamorphosed minerals, coincides with the mass extinction (including dinosaurs) at the end of the Cretaceous, and is believed to be associated with this impact event. Many impact specialists are convinced that the environmental effects of the Chicxulub impact were the primary cause of an abrupt mass extinction. However, many palaeontologists disagree. They argue that the extinction was not abrupt and that there is evidence for other causes. The one matter on which all agree is that the iridium-rich boundary layer serves as an excellent worldwide common time marker that will be essential to further studies of K–T extinctions. Impact specialists continue to search for evidence of large impacts that could be related to other mass extinctions.

Smaller impact events are much more frequent, but the resultant craters are more easily eroded or obscured. Four craters having diameters in the range 7–18 km have been identified as less than 6 million years old. These craters were formed by the impact of bodies in the diameter range of about 300 m to 2 km, sufficiently large to minimize retardation by the atmosphere. Atmospheric retardation becomes significant only for bodies having diameters less than about 10 m, corresponding to masses less than about 1000 tons.

Thus, the velocity of a stony object of 10 m diameter might be reduced from approximately 15 km s^{-1} on atmospheric entry to 10 km s^{-1} on impact with the Earth. The impact would deposit the energy equivalent of approximately 36 000 tons of trinitrotoluene (TNT), and the resultant crater would have a diameter in the range of 100–200 m.

As noted by Melosh in his book on impact cratering (see Further Reading), relationships between crater dimensions and impact parameters are poorly constrained. There are a variety of empirical scaling relations extrapolated from small-scale laboratory impact experiments, high explosive and nuclear experiments, and large-scale computer calculations. Here, we apply the observation that, for a variety of impact conditions, many scaling relations predict D/d, the ratio of the crater diameter to the impactor diameter, to be in the range of 10–20.

Small objects with a mass in kilograms are slowed by atmospheric drag to terminal velocities in the region below about 200 m s^{-1}. The resultant impact pressure of about 1 GPa for an impact on rock is too low to produce shock metamorphic effects other than fracture. Thus, 20 GPa shock metamorphic effects found in some small meteorites may be interpreted as the result of impacts on a meteorite parent body. The exception to this general rule is when there is evidence that a small meteorite is a fragment of a much larger body that impacted the Earth at high velocity. Iron meteorites found in the vicinity of the Meteor Crater (northern Arizona, USA) (1.3 km in diameter) are interpreted as fragments of the rear surface of a 100 000 ton (approximately 30 m in diameter) iron meteorite that is estimated to have impacted the Earth at approximately 20 km s^{-1}.

The bulk of this meteorite was melted or vaporized as a result of very high shock pressures. Intuitively, it might be expected that the entire meteorite would be exposed to the same peak pressure, as predicted by some low-resolution calculations. However, the most recent high-resolution calculations predict that rarefactions originating at free surfaces (the meteorite–air interface) will interact to create low-pressure regions near the rear surface of the meteorite. At least 20 tons of meteorite fragments have been recovered from the vicinity of Meteor Crater. Some of these fragments have shock metamorphic features indicative of peak pressures of less than 10 GPa. Other fragments have shock metamorphic features, including shock-synthesized diamond, indicative of pressures in excess of 100 GPa.

Shock Waves and Large Impacts

Pressure scale definitions: the modern unit of pressure, the pascal, is defined as 1 N m^{-2}. Atmospheric pressure on the Earth at sea-level is approximately 10^5 Pa (100 000 Pa); shock pressures are usually stated as gigapascals (GPa), 10^9 Pa. Earlier literature may refer to bars, kilobars (kb or kbar), atmospheres (atm), dynes per square centimetre (dyn cm^{-2}), kilograms per square centimetre (kg cm^{-2}), and pounds per square inch (psi).

$$1 \text{ bar} = 10^5 \text{ Pa} = 10^6 \text{ dyn cm}^{-2} = 0.9869 \text{ atm}$$
$$= 1.0197 \text{ kg cm}^{-2} = 14.504 \text{ psi}$$

$$1 \text{ GPa} = 10 \text{ kbar} = 10^{10} \text{ dyn cm}^{-2} =\sim 145\,000 \text{ psi}$$

A collision between two bodies produces a high pressure (shock wave) at the point of impact. The shock wave propagates into both bodies and is attenuated by rarefaction waves originating at free surfaces. The magnitude of the peak pressure depends on both the impact velocity and the relative stiffness of the impacting bodies, as shown in Table 1. The pressure calculations are based on material properties extrapolated from much lower pressures.

Table 1 Parameters of typical Asteriod-Earth and Comet-Earth Impacts

Impactor–target	*Velocity (km s^{-1})*	*Peak pressure (GPa)*	*Fate of impactor*	*Fate of target*
Iron–water (ice)[a]	20	~360	Completely molten	Total vaporization of water or ice
Iron–alluvium (1.5 g cm^{-3})	20	~400	Completely molten	Total vaporization
Iron–granite	20	~750	Partial vaporization	Total vaporization
Iron–peridotite	20	~950	Partial vaporization	Total vaporization
Peridotite–water (ice)	20	~280	Partial vaporization	Total vaporization
Peridotite–alluvium	20	~300	Partial vaporization	Total vaporization
Peridotite–alluvium over granite[b]	20	~300, then ~400, shock reflection	Partial vaporization	Total vaporization of alluvium Partial vaporization of granite
Peridotite–granite	20	550	Total vaporization	Total vaporization
Peridotite–peridotite	20	650	Total vaporization	Total vaporization
Snow (0.6 g cm^{-3})–ice[c]	40	~500	Total vaporization	Total vaporization
Ice–alluvium[c]	40	~600	Total vaporization	Total vaporization
Ice–ice	40	~650	Total vaporization	Total vaporization

[a] At these very high pressures, the properties of ice and water are indistinguishable.

[b] Shock interactions occur at interfaces between materials having different properties. The 300 GPa shock in alluvium is reflected at the granite interface as a 400 GPa shock moving back into the alluvium. A 400 GPa shock is transmitted into the granite.

[c] The properties of a comet are approximately bracketed by the properties of snow and ice.

The maximum duration of the pressure peak occurs at the point of impact and depends on such details as the relative sizes and shapes of the two bodies, their properties at very high pressure, the impact velocity, and the angle of impact. One popular approximation is that the duration is equal to the impact velocity divided by the diameter of the smaller body. The peak pressure duration for the 20 km s^{-1} impact on the Earth of an asteroid of 20 km in diameter would thus be about 1 s, i.e., between about 0.5 and 2 s, depending on the material properties and the geometry of the impact.

To put these very high impact pressures and durations into perspective, it should be noted that the pressure at the centre of the Earth is about 350 GPa and the effective high-pressure duration of a very large buried thermonuclear explosion is less than 1 ms.

The very high pressures at the point of impact decay in amplitude as they propagate into the Earth. There are two causes of pressure decay. The first is geometric. If the shock front is considered as an expanding hemispherical shell, the peak pressure would be expected to decay as the radius of the hemisphere increases. The second cause of pressure decay is that rarefaction waves originating from the free surface overtake the shock front and reduce the pressure. Melosh's book cites a number of approximate methods for estimating pressure decay in a homogeneous geological environment. Another approach is to perform large-scale computer calculations with 'hydrocodes', computer programs designed to calculate shock wave propagation in various media. These codes were developed for national defence-related purposes to calculate the attenuation of the shock waves produced by large chemical or nuclear explosions. Numerous comparisons of calculations with experimental measurements have shown that agreement within about ±20% can now be achieved over the pressure range between about 1 and 100 GPa, the range over which geological media have been well characterized by Hugoniot (compressional) and release adiabatic measurements. In order to achieve such agreement, the details of geology (including faults and layering) and accurate material properties must be incorporated into the calculation. It is particularly important to include the details of dynamic phase transitions, including the pressure hysteresis between the forward transitions on loading and the reverse transitions on release. The most detailed calculations require extraordinary computer power.

There are many pitfalls in the use of hydrocodes to calculate large impacts on Earth. One problem is the scale of the event. The 20 km s^{-1} impact of a stony asteroid of 10 km in diameter on the Earth releases an energy of approximately 6.3×10^{28} J, about a million times greater than the sum of all nuclear arsenals. The scope of the calculation must be reduced to bring it within the range of even our fastest of current computers. The details of the geological setting must be ignored and the spatial resolution of the calculation must be very coarse, about 1 km. A second problem is that the calculation requires a knowledge of detailed material properties at very high pressures, well above the range of current experimental data. Data on the compression and release behaviour of geological media are virtually non-existent for the pressure range above 100 GPa. Furthermore, most existing data for the range below 100 GPa were obtained in experiments of submicrosecond duration. A few experiments at longer duration, up to 10 μs, have been performed to explore the effects of pressure duration on dynamic phase transitions; kinetic effects were not observed. However, it may be inferred from static high-pressure studies that kinetic effects could possibly be significant in large impact events where the shock duration may exceed 1 s.

Although the equations governing shock wave propagation are simple, the details of shock wave propagation through a typical polymineralic rock are extraordinarily complex. The shock properties of individual minerals can differ substantially; shock interactions occur at mineral boundaries and in the vicinity of pores and cracks. Recent high-resolution hydrocode calculations have studied the details of pressure and temperature equilibration in a rock-like material. On a time-scale of nanoseconds (10^{-9} s) and a distance scale of micrometres, the shock front appears chaotic. Shock collisions around a pore produce nanosecond duration pressure spikes that may be ten times the amplitude of the pressure within an adjacent millimetre-sized grain. These initial pressure inhomogeneities equilibrate within less than a microsecond (for a mineral grain size of about 1 mm) to a uniform pressure. In general, these initial pressure inhomogeneities are ignored. The term 'peak pressure', as it is commonly used (e.g., in Table 1), refers to the pressure after equilibration of the nanosecond duration spikes. However, a knowledge of the complexity of pressure equilibration may be very useful to the researcher who uses microscopic techniques to study shock metamorphic effects on a micrometre scale.

Although accurate high-resolution calculations of a specific large impact may be presently beyond reach, there is nevertheless an excellent qualitative understanding of generic impacts from the study of numerous low-resolution calculations, from small-scale laboratory experiments, and from theoretical considerations, as summarized by Melosh. Figure 1 shows a

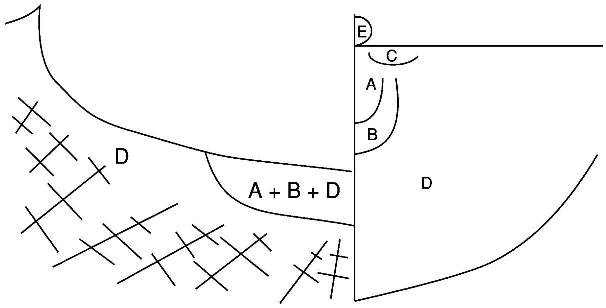


Figure 1 Diagram showing zones of damage by the high-velocity impact of a large meteorite. Right: at moment of impact. Left: after cratering is complete. Zone A: very high-pressure region above approximately 100 GPa; all minerals melted or vaporized. Zone B: high-pressure region, approximately 7–100 GPa; most minerals show distinctive shock metamorphic effects. Zone C: near impactor–near surface region. Together with the atmosphere, material is ejected at high velocity into a high-angle trajectory. The ejected material can range from strongly shocked (even molten) to weakly shocked. This is the probable source region for tektites, natural glasses of crustal composition. Zone D: weakly shocked fractured region, approximately 0.1–7 GPa; radial and circumferential fractures. Most of the material ejected from the crater originates from this zone and is not strongly shocked. However, block motion in this region can produce localized melting caused by frictional heating. The resultant sheet-like formations are called pseudotachylites. Zone E: the impacting body. Melted or vaporized, for the most part, but a small fraction may survive in a relatively unshocked condition. The melted and vaporized material can be ejected into the upper atmosphere and subsequently deposited over a large area. Left: A + B + D, breccia lens; an intimate mixture of shock metamorphosed clasts with melt and weakly shocked material. If melt predominates, the breccia is called tagamite. If the melt is less dominant, the breccia is called suevite.

diagram depicting the effects of an impact, referenced to the pre-impact setting and approximately scaled to the diameter of the impacting body.

Shock Metamorphic Effects

The impact event depicted in Figure 1 can produce numerous shock metamorphic effects. Ejected material, including solid, solidified melt, and condensed vapour, can serve as a stratigraphical marker. The breccia lens provides most of the evidence to distinguish an impact crater from a volcanic crater. As noted above, the breccia lens is a heterogeneous mixture of high-pressure and lower pressure material. This is advantageous in that the lower pressure clasts have been subjected to only minor shock heating and serve to quench the more strongly shock-heated and shock-melted material. Quenching helps to preserve more fragile high-pressure phases, such as stishovite. Some common shock metamorphic effects are presented in Table 2.

Controversial Issues

There are numerous controversies in the literature on shock metamorphism, as should become evident to the person who reads more than one of the items in the Further Reading section at the end of this article. The controversies are, in general, a measure of the incompleteness of our knowledge. Almost without exception, published experimental data and observations are trustworthy and reproducible. The majority of controversies centre on the interpretations of the data, inferences from hydrocode calculations, or the validity of often unstated assumptions.

For example, tektites (distinctive forms of natural glass showered down on the Earth) show unequivocal signs of scuplturing by aerodynamic forces during high-speed entry into the Earth's atmosphere. Although the detailed chemical composition of tektites implies an Earth origin, simple physical arguments indicate that it is impossible to propel tektites into space through the Earth's atmosphere. Those who therefore argued against a terrestrial origin of tektites failed to consider the possibility that a large impact on Earth could melt crustal material and propel it into space, together with a portion of the atmosphere. Large-scale hydrocode calculations indicate that this latter possibility is plausible. The chemical evidence for a terrestrial origin of tektites is so strong that it overwhelms concerns that a calculation having a resolution of 1 km is used to predict the fate of centimetre-sized objects.

There is a related controversy over the minimum shock pressure to which Martian meteorites have been exposed. There is very strong chemical evidence that certain meteorites found on the Earth actually originated on Mars. Some scientists have argued that any meteorite ejected from Mars by a large impact would have melted on release from the requisite high pressure, about 150 GPa. Although various investigators have disagreed about the peak pressure implied by shock metamorphic effects, they have agreed that the meteorites were not melted by their ejection from Mars. Subsequent calculations have indicated that the Martian meteorites could have been accelerated to the Martian escape velocity of 5 km s^{-1} by shock pressures as low as 65 GPa. Some investigators have interpreted these calculations as evidence that all Martian meteorites must necessarily have been exposed to shock pressures of 65 GPa or higher. However, one Martian meteorite shows remanent magnetism that would have been destroyed by shock pressures exceeding about 20 GPa. If the magnetic data and the evidence for a Martian origin are accepted, there must be an even lower pressure mechanism for accelerating a Martian rock to escape

Table 2 Shock metamorphic effects

Effect	*Source material*	*Pressure (GPa) (single shock)*	*Comments*
Melting	Iron	>170	Melts on release of pressure
Melting	Olivine, pyroxene	>100	Melts on release of pressure
Melting	Quartz, granite	>50	Melts on release of pressure
Melting	Sand, soil	>20, possibly as low as 7	Energy increase on shock compression much greater for porous materials
Diaplectic glass	Quartz, feldspars	>15, possibly as low as 7	Diaplectic glass forms by solid-state transformation. It is amorphous, but retains original crystal form and usually has a higher refractive index than melt glass. Lower bound pressure from PDF formation
Stishovite, hollandite	Quartz, feldspars	>15, possibly as low as 7	Stishovite, hollandite, polymorphs of quartz, and feldspar found in impact craters and meteorites. Lower bound pressure from PDF formation
Coesite	Quartz	>15, possibly as low as 3	Coesite found in impact craters in association with diaplectic glass, implying that it formed on release of pressure. Could conceivably be found in a pseudotachylite that solidified under pressure
Ringwoodite, wadsleyite	Olivines	>15	Found in meteorite melt veins; pseudotachylite-like structures that were quenched at high pressure
Akimotite, majorite	Pyroxenes	>17	Found in meteorite melt veins; pseudotachylite-like structures that were quenched at high pressure
Diamond, cubic and hexagonal mixture	Graphite	>25	Found in meteorites; $P \sim 100$ GPa from graphite in iron meteorites. Found in impact craters; $P \sim 30$ GPa from graphite in gneiss. Made in laboratory shock experiments
Cubic diamond	Porous carbon	>15	Made in laboratory shock experiments
Planar deformation features (PDFs)	Quartz, feldspar predominantly. Also other minerals	>7	PDFs in quartz are a primary diagnostic for impact. A PDF is a lamellar feature aligned with a low-index crystallographic plane. A number of different orientations may appear in the same grain. There is evidence that the lamellae contain high-pressure phases that invert to low-pressure forms during electron microscopy
Fractures	All rocks	>~0.1	Laboratory shock experiments show dynamic fracture strength comparable (~1.5 times) to static strength
Pseudotachylite formation	All rocks	>~0.1	Pressure estimate based on observation of pseudotachylites in the fractured zone

Numerous other high-pressure minerals have been observed in meteorites and impact craters. The most common and readily observed are listed. The book by French (see Further Reading) contains numerous micrographs of shock metamorphosed quartz.

velocity. Melosh has suggested one such mechanism: entrainment of the rock in the vapour plume formed by strongly shocked material (see Table 1).

Finally, there are long-standing controversies over the peak pressures associated with various metamorphic effects. Shock metamorphic effects in rocks and minerals have been studied in numerous laboratory shock experiments over the past 45 years. It was initially hoped that a peak shock pressure calibration could be established based on the presence of various metamorphic effects. The assumption that the peak shock pressure is the only significant parameter seems to be incorrect, as may be inferred by analyses of apparent conflicts in research reports. These conflicts can usually be resolved by considerations of experimental differences between the experiments. Samples loaded to the same peak pressure via different loading paths (single shock vs. a sequence of shock reflections) often show marked differences in metamorphic effects. Although the pressure duration of laboratory shock experiments is in the range of a microsecond, the shock pressure duration for a large natural impact may exceed a second. The interpretation of metamorphic effects on the basis of laboratory static high-pressure data may be more appropriate in this regime. Some controversies over the precise peak pressure to which a given natural sample has been exposed may not be resolvable on the basis of present knowledge. However, there is usually no argument about whether a given sample has been shock metamorphosed at all.

See Also

Impact Structures. **Solar System:** Meteorites; Mercury; Moon; Mars. **Tektites**.

Further Reading

Desonie D (1996) *Cosmic Collisions, A Scientific American Focus Book*. New York: Henry Holt and Co.

French BM (1998) *Traces of Catastrophe: A Handbook of Shock Metamorphic Effects in Terrestrial Meteorite Impact Structures*. LPI Contribution 954. Houston, TX: Lunar and Planetary Institute.

Koeberl C and Martinez-Ruiz F (eds.) (2003) *Impact Markers in the Stratigraphic Record*. Berlin, Heidelberg: Springer-Verlag.

McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

Melosh HJ (1989) *Impact Cratering: A Geologic Process*. Oxford: Oxford University Press and Oxford: Clarendon Press.

Rubin AE (2002) *Disturbing the Solar System: Impacts, Close Encounters, and Coming Attractions*. Princeton: Princeton University Press.

SOIL MECHANICS

J Atkinson, City University, London, UK

Soil and Mechanics

Engineering Soils

Soil mechanics describes the mechanical behaviour of granular materials. Mechanical behaviour covers strength, shear stiffness, volumetric compressibility, and seepage of water. Granular materials include powders, grain, and other foodstuffs, mineral ores and concentrates, as well as natural soils.

The simple theories of soil mechanics are intended for collections of grains which are uncemented or only very slightly cemented and which contain fluid, usually water or air, in the pore spaces. This covers dense and loose sands and soft and stiff clays. Rock mechanics describes the behaviour of strongly bonded grains whose overall behaviour is governed by joints and fractures. There is a range of materials between these, including weathered rocks, weak rocks, and cemented soils for which simple theories of soil mechanics have limited application.

The theories of soil mechanics apply equally to sands (coarse-grained soils) and clays (fine-grained soils). Figure 1 shows samples of sand and clay under load in unconfined compression. In each case the strength arises from suctions in the pore water. The clay is stronger than the sand because it can sustain larger suctions: otherwise their behaviour is fundamentally the same.

In describing theories for the behaviour of materials some mathematics is unavoidable. In the following, the mathematics is kept as simple as possible and does not extend beyond simple algebra. Only the most basic and fundamental equations and parameters are included.

Mechanics: Strength, Stiffness, Compressibility, and Permeability

Soils are highly compressible. The volume decreases significantly as it is compressed under an isotropic stress state. This is illustrated in Figure 2C. Soils also change volume when they are sheared and distorted.

Strength is basically the maximum shear stress a soil can sustain before it fails. Stiffness is the distortion which occurs as the soil is loaded before it fails. These are illustrated in Figure 2(D). G is the shear modulus and describes stiffness: τ_f is the shear stress after large distortion and it is the strength. In soils both strength and stiffness increase with increasing mean stress.

The frictional nature and the coupling between shear and volume change are the two main differences between the mechanical behaviour of granular materials and the mechanical behaviour of metals and other similar materials.

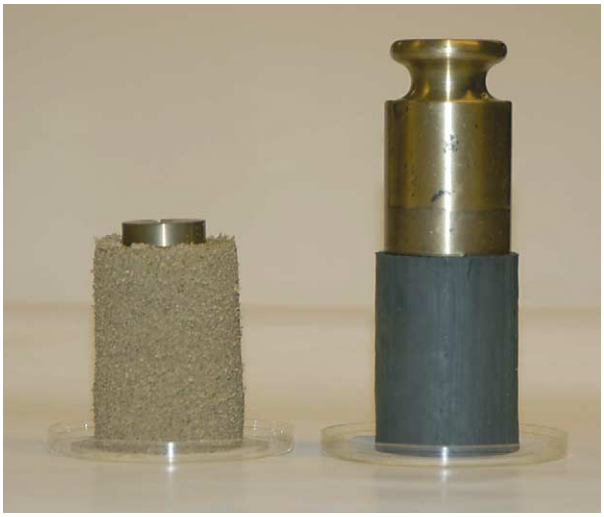


Figure 1 Unconfined compression of sand and clay.

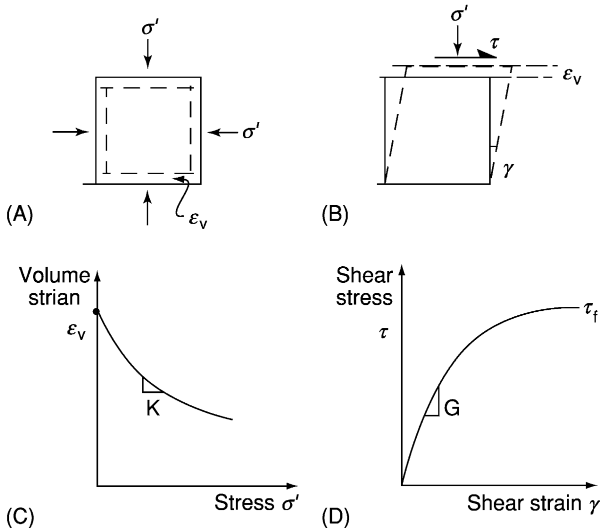


Figure 2 Compression and distortion.

A Brief History of Soil Mechanics

Coulomb and Soil Strength

Theories for soil mechanics originated around the middle of the eighteenth century. Coulomb was a military engineer and he was concerned with calculating soil loads on masonry retaining walls. He carried out experiments on the strength of soils and he found that the resistance of soil to shear loading had two components, one cohesive and the other frictional. Coulomb tested unsaturated samples and his analyses were in terms of forces, not stresses. Methods for analysis of stress discovered later by Mohr were incorporated into Coulomb's results and this is the basis of the well-known Mohr–Coulomb strength equation:

$$\tau_f = c + \sigma \tan \phi \qquad [1]$$

Neither Coulomb not Mohr had a clear understanding of the importance of pore pressures and effective stresses and the original Mohr–Coulomb equation is in terms of total stress. It is now known that the Mohr–Coulomb equation for soil strength is limited but it is still widely used.

Terzaghi and Effective Stress

Karl Terzaghi was an Austrian civil engineer. The major contribution which he made to soil mechanics was to set out a clear theory in the 1920s for accounting for the influence of pore pressure on soil strength and deformation. He proposed an effective stress σ′ which controls all soil behaviour and he discovered that for saturated soil this is related to total stress σ and pore pressure u by:

$$\sigma' = \sigma - u \qquad [2]$$

The Terzaghi effective stress equation has been found to apply for a very wide range of loadings and soils, and it is used universally for geotechnical analysis of saturated soils.

Plasticity and Cam Clay

In the 1960s Andrew Schofield and Peter Wroth were lecturers at Cambridge University. They applied the then relatively new theories of plastic flow to frictional materials and created a complete stress-strain theory for soils. The model they developed they called Cam Clay and this remains the basis for many of the current constitutive equations for soils.

These theories of frictional strength, effective stress, and plastic flow are the basic buiding blocks for modern soil mechanics.

Effective Stress and Drainage

Principle of Effective Stress

The Principle of effective stress first proposed by Terzaghi in 1923, states that the stress which is effective in determining strength, stiffness, and compressibility, the effective stress, σ′ is given by eqn [2].

Total stresses arise from external loads due to foundations and walls and loads from self weight. Pore pressures are the pressures in the fluid in the pore spaces. For dry soils the pore pressure is the pressure in the air in the pores. For saturated soil it is the pore water pressure. For soils which are not fully saturated and which contain both air (or gas) and water in the pores the equivalent pore pressure is some combination of the air and water pressure. At present there is no simple and robust theory for determining the equivalent pore pressure and effective stress in unsaturated soils.

So far as is known, the principle of effective stress and the effective stress equation (eqn [2]) holds for all dry or saturated soils over a very wide range of stress and pore pressure up to several tens of MPa. The strength and stiffness of soil 1 m below the bed of the deep ocean, where the depth of water may exceed 5 km, will be the same as that of soil 1 m below the bed of a duck pond.

Drainage and Consolidation

Because water is relatively incompressible in comparison with soil, volume changes in soil can occur only if water can flow into or out from the pore spaces. Whether or not this happens depends on the rate of drainage and the rate of loading.

If water cannot drain from the soil it is said to be undrained: its volume must remain constant but pore pressures will change in response to the loading. If water has time to drain freely from the soil it is said to be drained: pore pressures remain constant and volume changes occur. Hence:

$$\text{Undrained loading}: \quad \delta V = 0 \text{ and u changes}$$
$$\text{Drained loading}: \quad \delta u = 0 \text{ and volume changes}$$

where the symbol δ means 'a change of.' In many cases soil in the ground is neither fully drained not fully undrained but simple soil mechanics theories are applicable only for fully drained or fully undrained cases.

If soil is loaded undrained the resulting pore pressures will not be in equilibrium with the long-term groundwater pressures. As the excess (out of balance) pore pressures dissipate under constant total stress there will be changes of effective stress and volume changes. This process is known as consolidation. Because the rate of drainage during consolidation depends on hydraulic gradient, which decreases as excess pore pressures dissipate, the rate of consolidation diminishes with time.

Description and Classification of Engineering Soils

There are standard schemes for description and classification of soils for engineering purposes. These essentially classify soils under the two main headings: the nature of the grains and how they are packed together. For natural soils, descriptions are added for structure including bonding, bedding, and discontinuities.

The Nature of the Soil: Characteristics of the Grains

The most important characteristic is the grain size or grading. Figure 3 shows the range of grain sizes commonly found in natural soils and their descriptions (e.g., sand is 0.06 m to 2 mm). The range is very large. Clay grains are of the order of 1000 times smaller than coarse sand grains. Since permeability is related to the square of the size; sands are of the order of 1 million times more permeable than clays. If a soil is essentially single sized it is poorly graded (or well sorted). If it contains a range of sizes it is well graded (or poorly sorted). In a well graded soil it is usually the 10% smaller than (D_{10}) size which governs drainage.

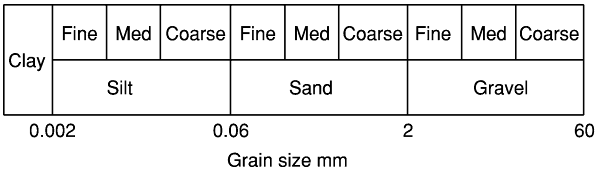


Figure 3 Grain size descriptions.

Grains of silt size and larger normally consist of rock fragments. They may be rounded or angular, rough or smooth. Grains of clay size are normally made of a clay mineral belonging to one of the major families which are kaolinite, illite, and montmorillonite (smectite). These may be distinguished by their Atterberg Limits and Activity (see below). The characteristics of the grains do not effect the fundamental behaviour of soils but they do influence numerical values of strength and stiffness parameters.

Rates of Loading and Drainage

In soil the rate of drainage depends primarily on the permeability which itself depends on the grading of the soil. The Hazen formula for coefficient of permeability k is

$$k \propto D_{10}^2 \qquad [3]$$

where D_{10} is the size of the grains with 10% smaller. Typical values for coefficient of permeability range from greater than $10^{-2}\,\text{ms}^{-1}$ (approx. 1.5 m in a minute) for coarse-grained soils to less than $10^{-8}\,\text{ms}^{-1}$ (approx 1 m in 3 years) for fine-grained soils. This very large difference (more than 1 million times) in rate of drainage between coarse-grained and fine-grained soils accounts for many of the differences in observed behaviour of sands and clays.

The rate of loading also varies widely. Some natural processes, such as deposition and erosion, occur relatively slowly (over decades) while others, such as earthquakes, occur relatively rapidly (over a few seconds). In construction, a shallow trench might be dug in a few hours and a large dam built in a few years.

In determining whether a certain event applied to a certain soil is drained or undrained, it is necessary to consider both the rate of drainage and the rate of loading. During earthquakes, coarse-grained sandy soils may be undrained causing liquefaction failure. In construction it is usual to take fine-grained clay soils as undrained and coarse-grained sand soils as drained.

Atterberg Limits

If a clay soil has a very high water content it will flow like a liquid; if it has a low water content it will become brittle and crumbly. For intermediate water contents it will be plastic. The Atterberg Limits, the Liquid Limit, W_l, and the Plastic Limit, W_p, define the range of water content over which a clay soil is plastic. The Plasticity Index, I_P, is the difference between the Liquid and Plastic Limits:

Table 1 Typical values for some characteristic soil parameters

Parameter	*Symbol and units*	*Kaolinite clay (China clay)*	*London clay*	*Alluvial sand (Thames)*	*Carbonate sand*	*Decomposed granite (Dartmoor)*
Liquid Limit	W_l	65	75			
Plastic Limit	W_p	35	30			
Plasticity Index	I_p	30	45			
Activity	A	0.4	1			
Maximum specific volume	V_{max}	2.72	2.98	2.2	3.2	
Minimum specific volume	V_{min}	1.92	1.80	1.5	2.0	
Coefft of compressibility	Cc	0.44	0.37	0.37	0.78	0.21
Coefficient of swelling	Cs	0.11	0.14	0.03	0.01	0.01
Specific volume on NCL at $\sigma' = 1$ kPa	Vn	3.26	2.68	3.17	4.8	2.17
Specific volume on CSL at $\sigma' = 1$ kPa	Vc	3.14	2.45	2.99	4.35	2.04
Critical state friction angle	ϕ'c degrees	25	23	32	40	39
Very small strain shear modulus at $\sigma' = 100$ kPa on the NCL	G′o MP_a	40	15	60	60	60

(Data from research at City University, London.)

$$I_P = W_l - W_p \qquad [4]$$

For a natural clay soil, which may contain silt and sand sized grains, the Activity is

$$A = \frac{I_P}{\%\text{clay}} \qquad [5]$$

and this is related to the mineralogy of the clay, as shown in Table 1. Many numerical values for soil parameters are related to the clay mineralogy and to the Atterberg Limits.

State: Liquidity Index and Relative Density

Grains in a soil may be densely packed or loosely packed or in an intermediate state of packing. The packing influences strength and stiffness, as shown in Figure 4. In a clay soil, the loosest packing corresponds to the Liquid Limit and the densest to the Plastic Limit. Intermediate states are described by the Liquidity Index:

$$I_l = \frac{w - W_p}{I_P} \qquad [6]$$

where w is the water content. At the Liquid Limit, the Liquidity Index is 1.0 and at the Plastic Limit it is 0, as shown in Figure 4.

In a coarse-grained soil the loosest packing corresponds to the maximum water content, w_{max}, and the densest to the minimum water content, w_{min}. Intermediate states are described by the relative density:

$$I_d = \frac{w_{max} - w}{w_{max} - w_{min}} \qquad [7]$$

At the loosest state the Relative Density is 0 and at the densest state it is 1.0, as shown in Figure 4.

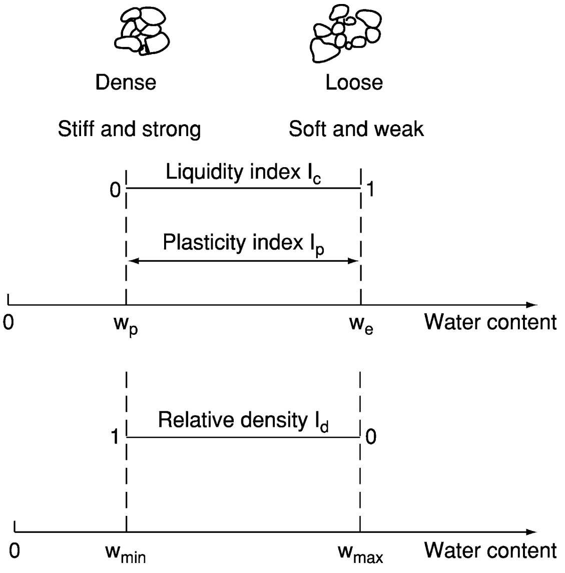


Figure 4 Packing: plasticity index and relative density.

Because soil strength and stiffness are essentially frictional they depend on the current effective stress. Packing, described by Liquidity Index or Relative Density, is not sufficient itself to describe soil behaviour. Soil state will be defined by a combination of packing and stress, as discussed later.

Behaviour in Compression: Change of Size

Isotropic Compression and Swelling

As saturated soil is loaded and unloaded under drained conditions water flows from and into the soil as it compresses and swells, rather like a sponge. The change in volume with changing effective stress is

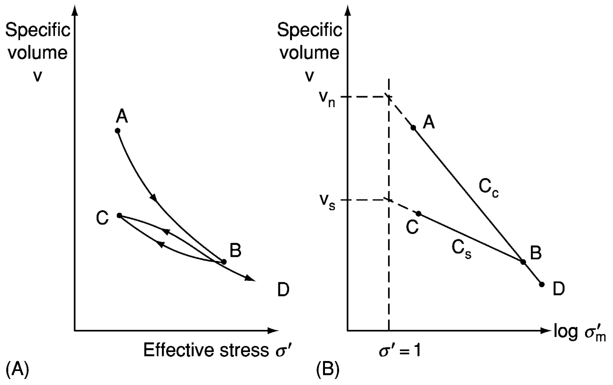


Figure 5 Isotropic compression and swelling.

illustrated in Figure 5(A). The soil is first loaded from A to B and it compresses. The compression of the soil skeleton is due mostly to particles rearranging and also to weak coarse grains fracturing or clay grains bending. The soil is unloaded from B to C and reloaded back to B. Some strains are recovered and there is a hysteresis loop. Volume changes in coarse-grained soils will be small because grains do not 'un-rearrange' or 'unfracture' but may be significant in clay soils as the grains can unbend.

In Figure 5B, effective stresses are plotted to a logarithmic scale and the compression and swelling curves have been idealised. The volume axis is the Specific Volume defined as

$$v = \frac{V}{V_s} \qquad [8]$$

where V_s is the volume of soil grains in a volume V of soil. For many soils v will range from about 1.2 if the soil is dense to over 2 if it is loose. The linear normal compression line ABD is given by

$$v = v_n - C_c \log \sigma' \qquad [9]$$

and the linear swelling and recompression line CB is given by

$$v = v_s - C_s \log \sigma' \qquad [10]$$

The Compression Index, C_c, the Swelling Index, C_s, and the Specific Volume, v_n, at unit stress are material parameters and are related to the characteristics of the grains. Typical values are given in Table 1. The location of a swelling line is given either by the maximum stress, σ'_m, or the specific volume, v_s, at unit stress.

A soil whose state lies on the line ABD is said to be normally compressed and ABC is the normal compression line (NCL). Soil whose state is on the NCL has not experienced a larger stress. A soil whose state is on a swelling line, such as CB, is said to be overconsolidated; it has experienced a greater stress, σ'_m. The overconsolidation ratio is

$$R_0 = \frac{\sigma'_m}{\sigma'} \qquad [11]$$

Equations [9] and [10] relate volume to stress for isotropic loading and unloading. Since the stress scale is logarithmic, the stress-strain behaviour is non-linear; the bulk modulus is not a constant but varies with both stress and overconsolidation.

The idealization of the hysteresis loop in Figure 5A to the line CB in Figure 5(B), common in simple soil mechanics theories, is unrealistic for many soils. Soil stiffness will be discussed later.

One-dimensional Compression in the Ground

Below level ground, the state of stress is not isotropic but one-dimensional, with zero horizontal strain during deposition and erosion; the vertical and horizontal effective stresses σ'_v and σ'_h are related by the coefficient of Earth pressure, K_o, given by

$$K_o = \frac{\sigma'_h}{\sigma'_v} \qquad [12]$$

For normally consolidated soil ($R_o = 1$) K_{onc} is given by

$$K_{onc} = (1 - \sin \phi'_c) \qquad [13]$$

where ϕ'_{cs} is the critical state friction angle. K_o increases with overconsolidation ratio. Horizontal effective stresses given by eqn [12] are for level ground with zero horizontal strain. Near slopes, foundations, and other underground construction stresses will be modified by the stresses imposed by the slope and the structure.

Calculations of settlement in the ground are often carried out in terms of a coefficient of compressibility, m_v, or a one-dimensional modulus, M, given by

$$M = \frac{1}{m_v} = \frac{\Delta\sigma'_v}{\Delta v / v} \qquad [14]$$

where Δv is the change of specific volume observed in a laboratory test on a soil sample with initial specific volume, v, when subjected to an increment of vertical stress, $\Delta\sigma'_v$. Since soil stiffness is non-linear, M is not a soil constant and the increment of stress applied in the test should correspond to the expected change of stress in the ground.

State: Stress and Packing

The behaviour of a particular soil depends on both the current effective stress and on the Relative

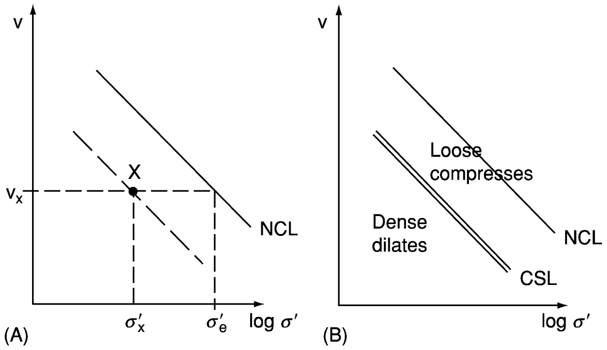


Figure 6 States and state parameters.

Density or Liquidity Index. These may be combined into a state parameter.

Figure 6A, which is similar to Figure 5B, shows the state of an overconsolidated sample at X and the normal compression line. All samples with states on the broken line through X parallel to the NCL will behave in a similar way. These states can be described by a stress state parameter S_σ given by

$$S_\sigma = \frac{\sigma'_x}{\sigma'_e} \qquad [15]$$

where σ'_e is the equivalent stress on the NCL at the same specific volume as that at X. The state parameter describes the distance of the state from the NCL. If the swelling index C_s is small, S_σ is approximately equal to the overconsolidation ratio, R_0. The concept of state is of fundamental importance for soils which are both frictional and which change volume during loading, as it combines both relative density and stress into a single parameter.

Dense and Loose States

After shearing, soils reach ultimate or critical states in which they continue to distort at constant state (i.e., at constant stress and volume). The relationship between specific volume and effective stress gives a critical state line (CSL) parallel to the normal compression line, as shown in Figure 6(B). The critical state line is given by

$$v = v_c - C_c \log \sigma' \qquad [16]$$

Soil states which are above the CSL are known as ‘loose of critical’ and the soil will compress on shearing. Soil states which are below the CSL are known as ‘dense of critical’ and the soil will dilate on shearing.

The CSL separates regions of fundamentally different behaviour of the same soil. A soil which has a relatively low specific volume and the grains are relatively closely packed will compress if the effective stress is very large. Similarly, a soil which has a relatively high specific volume and the grains are relatively loosely packed will dilate if the effective stresses are very small.

Sediments at great depth deform plastically. Near-surface soils often behave in a brittle manner and crack. Relative Density, or Liquidity Index, on its own is not sufficient to predict the behaviour on subsequent shearing; the effective stress must be taken into account as well.

Strength of Soil

Behaviour of Soil During Shearing

Figure 7A shows a block of soil with a constant normal effective stress σ' subjected to an increasing shear stress τ'.

The soil is drained and it distorts with a shear strain γ and a volumetric strain ε_v. If the soil is undrained, there are no volume changes but the pore pressures change. The block of soil represents conditions inside a slip zone in the slope illustrated in Figure 7B or in a foundation illustrated in Figure 7C. If the slope is created by excavation or erosion the normal stress decreases and, since soil strength is frictional, it will weaken, whereas below the loaded foundation the normal stress increases and the soil becomes stronger.

The behaviour of soil initially loose and initially dense of critical is illustrated in Figure 8. The loose soil (marked L) compresses during shearing even

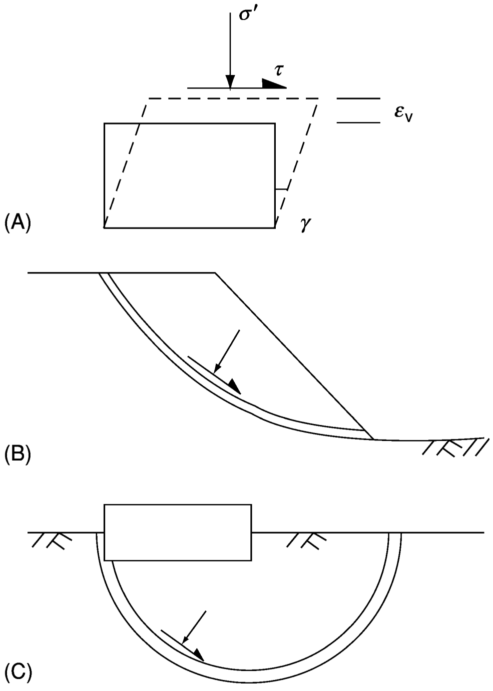


Figure 7 Shearing of soil.

though the normal stress remains constant and the dense soil (marked D) dilates. The rate of dilation is given by an angle of dilation ψ, given by

$$\tan\psi = -\frac{d\varepsilon_v}{d\gamma} \qquad [17]$$

(The negative sign is required as ψ is positive for negative (dilation) volumetric strains.)

Critical State Strength

The samples shown in Figure 8 have the same effective stress and they reach the same critical shear stress and the same critical specific volume after relatively large strains. Figure 9 shows critical states for a number of samples. There are unique relationships between the critical shear stress τ'_f, the critical normal stress σ'_f, and critical specific volume v_f, given by

$$\tau_f = \sigma'_f \tan\phi'_c \qquad [18]$$

$$\tau_f = v_c - C_c \log\sigma'_f \qquad [19]$$

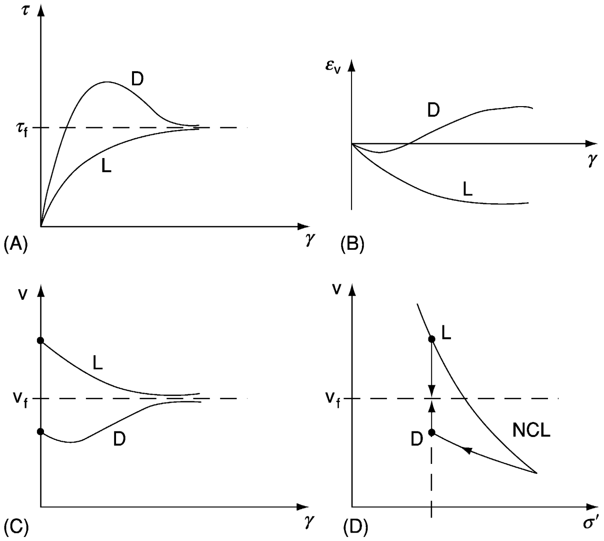


Figure 8 Stress and volume change in shearing soil.

These equations define a critical state line and the parameters ϕ'_c, C_c, and v_c are material parameters. (Critical state lines are usually shown as double lines, as in Figure 9). Typical values are given in Table 1. During shearing distortions, all soils will ultimately reach a critical state; if they did not they would continue to change state indefinitely, which is impossible. In simple soil mechanics theories, the critical states reached by a particular soil, given by eqns [18] and [19], are independent of the starting state and whether the soil is drained or undrained.

Undrained Strength

Figure 9B shows that the shear stress at failure, which is the shear strength, decreases as the specific volume at failure increases. If soil is undrained the water content and the undrained strength remain unchanged for any changes in total normal stress. The undrained strength.

$$\tau_f = s_u \qquad [20]$$

depends on the water content. In practice, samples are taken from the ground and tested without change of water content. The undrained strengths measured can be used for design so long as the water content in the ground does not change.

It is common knowledge that soils become weaker as their water content increases. This is shown in Figure 10 in which the undrained strength, with a logarithmic scale, decreases linearly with water content. The strength of soil at its Liquid Limit is approximately 1.5 kPa and the undrained strength

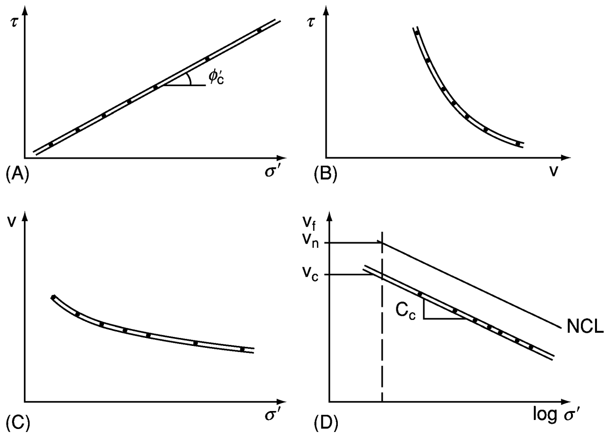


Figure 9 Critical states.

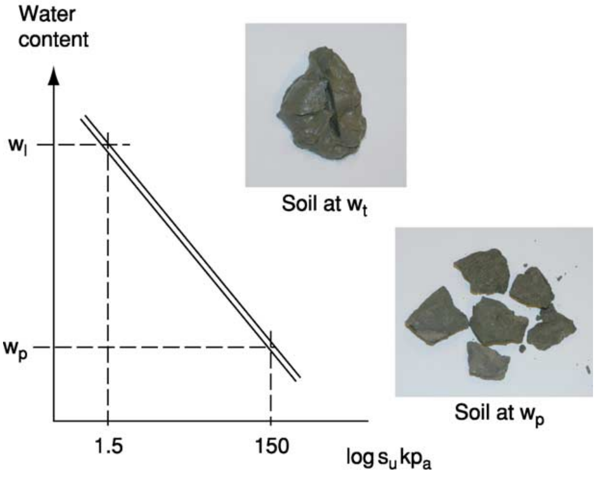


Figure 10 Undrained strength and water content.

of soil at its Plastic Limit is approximately 150 kPa (i.e., the strength of soil changes by about 100 times as the water content changes from the Liquid Limit to the Plastic Limit).

Peak Strength

Soils whose initial states are dense of critical have a peak strength before they reach a critical state, and they dilate during drained shear, as shown earlier in Figure 8. The peak strengths vary with effective normal stress and specific volume, as shown in Figure 11.

Samples which reach their peak states at the same specific volume have peak strengths on an envelope shown in Figure 11A. The envelope is often approximated by a straight line, shown in Figure 11A given by

$$\tau_p = c'_p + \sigma'_p \tan \phi'_p \qquad [21]$$

The peak friction angle, ϕ'_p, is a material parameter and, from Figure 11A $\phi'_p < \phi'_c$. The cohesion intercept, c'_p, is not a material parameter and its value depends on the specific volume. Moreover c'_p is not the strength at zero effective normal stress, as this must be zero for an uncemented granular material.

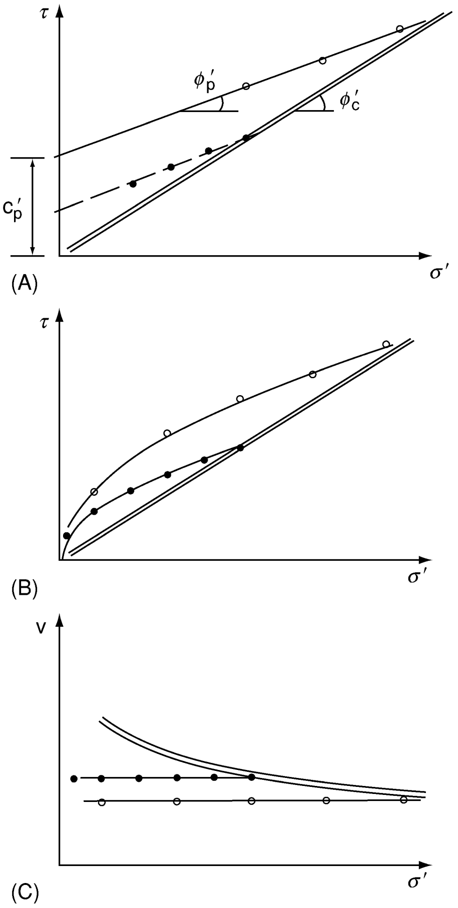


Figure 11 Peak strength.

The linear approximation for peak strength given by eqn [21] is applicable only within the range for which data are available. Figure 11B shows additional data at smaller normal effective stresses; there the envelope is now distinctly curved and passes through the origin. The curved peak failure envelope, shown in Figure 11C, can be represented by a power law of the form.

$$\tau_p = A\sigma'^b \qquad [22]$$

where b is a material parameter and A depends on the specific volume.

From analyses of the stresses and strains in the soil block, shown in Figure 7A, peak shear strength is given by

$$\tau_p = \sigma' \tan(\phi'_c + \psi) \qquad [23]$$

At the critical state, $\psi = 0$ and τ'_c is given by eqn [18]. At the peak state, the angle of dilation is at a maximum. The maximum rate of dilation is governed by the state parameter so the peak strength increases as the initial state moves away from the critical state line.

Equations [21, 22 and 23] are alternative theories for the peak strength of soils. They all contain a combination of material parameters and state dependent parameters. Equations [22 and 23] correctly give zero strength at zero effective stress. Equation [21] is most commonly applied in practice.

Stiffness of Soil

Figure 5A shows non-linear isotropic unloading and reloading behaviour. Similar non-linear behaviour occurs during shearing, as shown in Figure 12A. The tangent shear modulus G′ is the gradient of the stress-strain curve given by

$$G' = \frac{d\tau}{d\gamma} \qquad [24]$$

At the start of shearing near the origin the shear modulus is G'_o and at failure the shear modulus is zero.

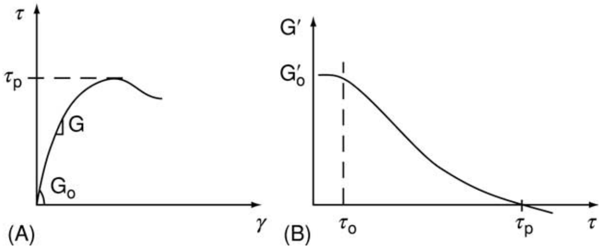


Figure 12 Stiffness and shear modulus.

Figure 12B shows the variation of shear modulus G′ with the progress of loading. There is a very small range up to a shear stress τ_o, in which G'_o is constant and the soil is linear, but over the remainder of loading the shear modulus decays with loading. For a particular soil the value of G'_o and the shear modulus at a particular strain, vary with the effective stress and with the state parameter.

For modest compression the bulk modulus, K′, and the one-dimensional compression modulus, M, both decay with normal stress in a manner similar to the decay of shear modulus with shear stress, shown in Figure 12B. At large compressive stresses the stiffness is the modulus corresponding to states on the NCL. At very large compressive stresses, the stiffness becomes very large as the specific volume approaches 1.0.

Consolidation

As soil is loaded or unloaded undrained, there are no volume changes but there are changes of pore pressure. These create excess pore pressures which are not in equilibrium with the surrounding pore pressures and so they dissipate with time. As they dissipate, under constant total stress, there are changes of effective stress which cause volume changes accompanied by drainage of water.

The basic theories for consolidation are for one-dimensional loading and drainage, illustrated in Figure 13A, in which all movements of soil and water are vertical. In practice this corresponds to conditions below a wide foundation or embankment.

Solutions for the rate of consolidation are given in terms of the degree of consolidation, U_t, and the time factor, T_v, given by

$$U_t = \frac{\rho_t}{\rho_\infty} \qquad [25]$$

$$T_v = \frac{c_v t}{H^2} \qquad [26]$$

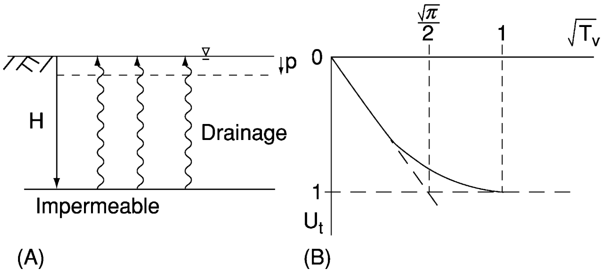


Figure 13 Consolidation.

where ρ_t is the settlement at time t, ρ_∞ is the settlement after a very long time, H is the length of the drainage path, and c_v is the coefficient of consolidation given by

$$c_v = \frac{Mk}{\gamma_w} \qquad [27]$$

where M is the one-dimensional modulus, k is the coefficient of permeability, and γ_w is the unit weight of water. The relationship between degree of consolidation and time factor is shown in Figure 13B.

The rate of consolidation depends on soil characteristics of stiffness and permeability and also on the geometry of the consolidating layer. This is given by the drainage path length H which is the greatest distance water must move to reach a drainage layer. Consolidation times can be significantly reduced by installing drains into the ground to reduce H.

Consolidation is the principal cause of the settlement of foundations and embankments on clays long after construction is complete.

Normalization and a State Boundary Surface

Figure 14A and B shows some different soil states. There are peak states corresponding to two different specific volumes; these are the same as those shown in Figure 11. There are paths for shearing of normally consolidated loose samples: path LD is for drained shearing and path LU is for undrained shearing.

These soil states involve three parameters, shear stress τ, normal stress σ', and specific volume v. Soil states can be represented by a three-dimensional surface using these axes. They may be represented on a two-dimensional graph using an appropriate normalizing procedure. There are several possibilities and one is to divide the shear and normal stresses by the equivalent stress σ'_e, shown in Figure 6A.

Figure 13C shows the states normalized by the equivalent stress. The NCL and the CSL reduce to single points. The peak states fall on a unique curve. The state paths for drained and undrained shearing of normally consolidated samples fall on a unique curve. The full curve represents a boundary to all possible states, known as a state boundary surface.

The concept of a state boundary surface is employed in advanced soil mechanics theories to develop complete constitutive relationships for soils. The surface is taken to be a yield surface and as a plastic potential surface from which plastic strains are determined. For states inside the boundary surface, the behaviour is taken to be elastic. One such theory is known as Cam Clay, for which the state boundary surface is represented by a logarithmic spiral curve.

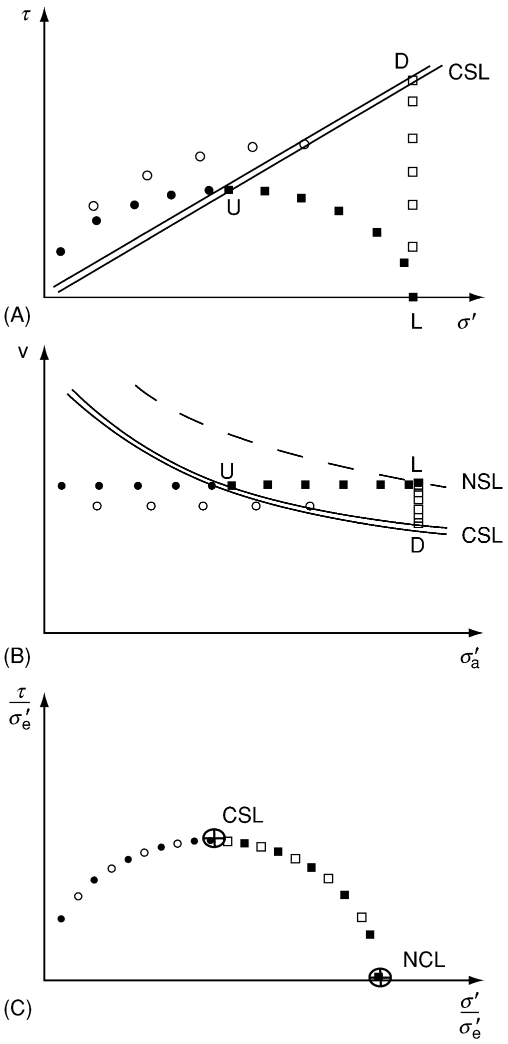


Figure 14 A state boundary surface.

Applications

The simple theories presented above for granular materials form the basis for analysis and design of engineering works which interact with the ground such as foundations, slopes, tunnels and retaining walls.

The basic theories for the mechanical behaviour granular materials are applicable equally to coarse-grained soils (sands and gravels) and fine-grained soils (clays). The principle factor to consider is the relative rate of loading and drainage. For routine analysis a particular case must be taken to be either fully drained or fully undrained.

If the soil is assumed to be fully drained, pore pressures can be determined and effective stresses calculated. Analyses are then carried out using *effective stresses* with effective stress strength and stiffness parameters. If the soil is assumed to be undrained, there are no changes in volume but there are changes in pore pressure which cannot be easily determined. In this case analyses have to be carried out using *total stresses* with undrained strength and stiffness parameters.

The critical state strength should be used to investigate ultimate failures. The peak strengths, with appropriate factors, should be used to investigate designs which are required to limit movements.

Simple analyses of foundation settlement are often carried out assuming one-dimensional conditions using the one-dimensional modulus, M, or using simple elastic theories using a shear modulus, G, and a bulk modulus, K. In all cases, it is necessary to take account of non-linear stress-strain behaviour and the appropriate drainage conditions. Simple analyses of rate of settlement due to consolidation can only be carried out assuming one-dimensional conditions.

The advanced soil mechanics theories, such as Cam Clay, are not used in simple analysis and design except for extremely simplified cases. Instead they form the basis for analyses using finite element or other comparable numerical methods.

See Also

Engineering Geology: Liquefaction; Made Ground; Problematic Soils; Subsidence. **Soils:** Modern; Palaeosols.

Further Reading

Atkinson JH (1993) *The Mechanics of Soils and Foundations.* London: McGraw-Hill.

Goodman RE (1999) *Karl Terzaghi: the Engineer as Artist.* American Society of Engineering Press, Reston, Virginia.

Heyman J (1972) *Coulomb's Memoir on Statics.* Cambridge: Cambridge University Press.

Lancellotta R (1995) *Geotechnical Engineering.* Balkema, Rotterdam.

Muir Wood DM (1990) *Soil Behaviour and Critical State Soil mechanics.* Cambridge: Cambridge University Press.

Powrie W (2004). *Soil Mechanics*, 2nd edn. Spon Press: London.

Schofield AN and Wroth CP (1968) *Critical State Soil Mechanics.* McGraw-Hill.

SOILS

Contents

Modern

G J Retallack, University of Oregon, Eugene, OR, USA

Introduction

There are many soil-forming processes, which in varying combinations create the large array of soils forming at the surface of the Earth. The study of soils is aided by the observation that soil-forming processes are slow and seldom go to completion. The parent materials of soils are modified over thousands of years by physical, chemical, and biological influences. However, few of these processes can be observed directly. Podzolization is one of the few soil-forming processes that is rapid enough to be recreated in the laboratory. Soil-forming processes that operate over thousands of years are studied using a space-for-time strategy (that is, studying soils of differing ages that are subject to the same soil-forming regime). A set of soils of different ages with comparable climates, vegetation, topographical positions, and parent materials is called a chronosequence (Figure 1). Mathematical relationships between the development of particular soil features and time are called chronofunctions, and include the increased clayeyness produced by the soil-forming process of lessivage (Figure 2). While specifying the rate and progress of soil formation, chronofunctions can also be used to infer the ages of landscapes from undated soils by comparison with dated soils. Such estimates of soil age can be important in the study of the neotectonic deformation of landscapes and their suitability for long-term installations such as dams and nuclear power plants. Soil fertility also varies with soil age, and chronofunctions can guide agricultural use and rehabilitation of soils.

Soil-forming processes vary not only with time but also with parent materials, topographical relief, vegetation, and climate. For example, the fragments of volcanic glass in certain kinds of air-fall tuff are distinct from the minerals of most soils, and they bestow high fertility and low bulk density on some volcanic soils (the process of andisolization). Waterlogging in low-lying parts of the landscape prevents the rusting of iron minerals and imparts a grey-green colour to the soil (the process of gleization). Leachates from highly acidic vegetation, such as pine forest, create soils in which clays are destroyed but quartz and haematite accumulate (the process of podzolization). Finally, climate is also an important factor in

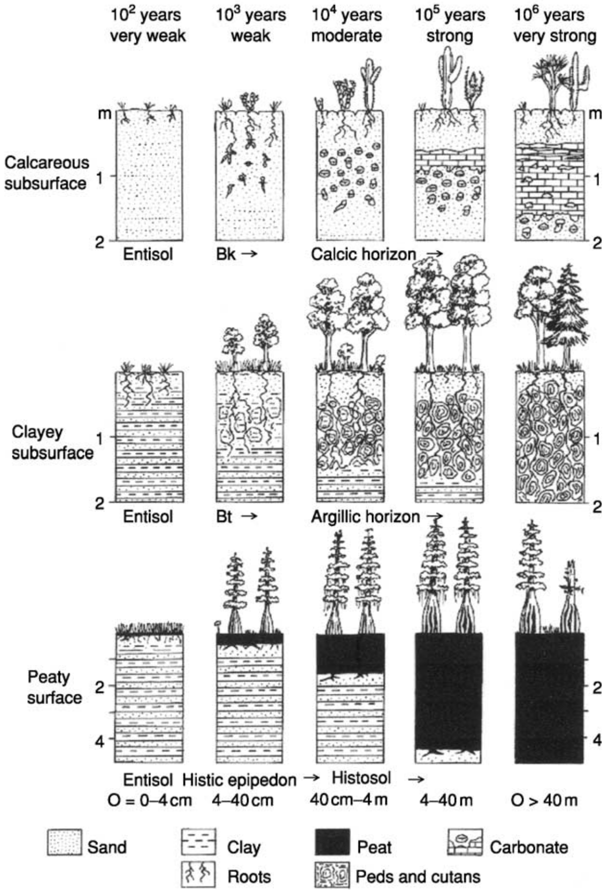


Figure 1 Soil development stages involving progressive calcification (top), lessivage (middle), and paludization (bottom). Reproduced with permission from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

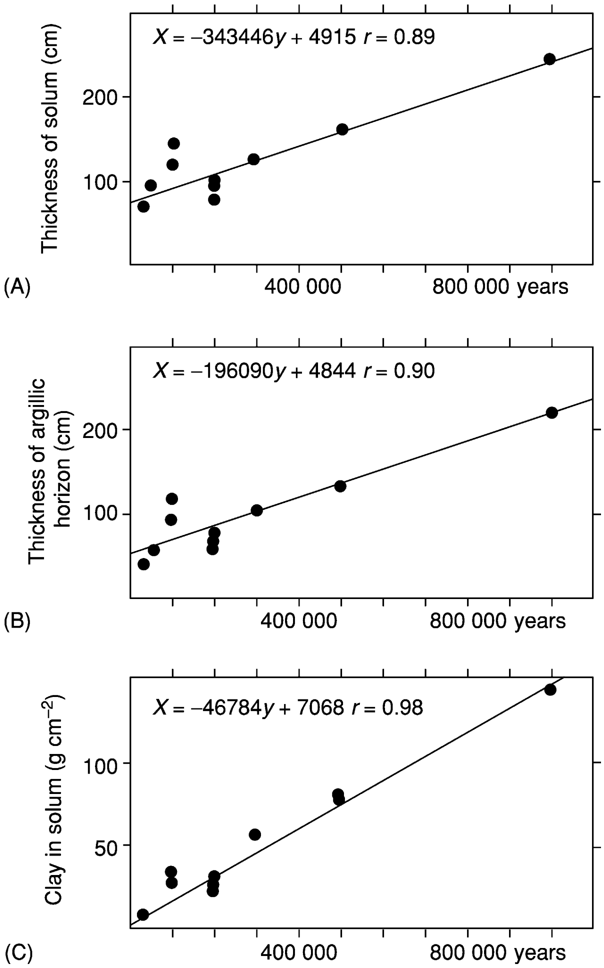


Figure 2 Chronofunctions for the progress of lessivage in soils of the Coastal Plain and Piedmont of south-eastern USA over time: (A) solum thickness; (B) thickness of the argillic horizon; and (C) the amount of clay in the solum. The solum is the A and B horizons; the argillic horizon is the Bt horizon; and the total profile is the A, B, and C horizons as defined in **Table 2**. Reproduced with permission from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell, using data from Markewich HW, Pavich MJ, and Buell GR (1990) Contrasting soils and landscapes of the Piedmont and Coastal Plain, eastern United States. *Geomorphology* 3: 417–447.

soil-forming processes, encouraging deeper and more thorough weathering in wetter and warmer climates (Figure 3).

The study of soil-forming processes has informed both soil taxonomy (Table 1) and soil-profile terminology (Table 2). The following outlines of soil-forming processes are presented in the order in which they would be encountered from warm wetlands to cold arid lands.

Gleization

Gleying or gleization is a process that produces and maintains unoxidized minerals in soils and is a term derived from a Russian term for the grey clay of swamps and bogs. Waterlogged peat-covered stagnant groundwaters allow the preservation of ferrous iron in clay minerals, such as grey smectite, carbonates, such as the siderite of freshwater bogs, and sulphides, such as the pyrite of mangrove swamps and salt marshes. In normally drained soils these minerals rust to produce red and brown clays, hydroxides such as goethite, and oxides such as haematite (Table 3). Goethite and haematite also form within gleyed soils when a short-term depression of the water table allows the atmospheric penetration of oxygen. Despite these red nodules and concretions, the dominant colour of gleyed soils is bluish or greenish grey (Figure 4).

Paludization

Paludization is literally ponding, but a pond would not be commonly understood as a soil. Paludization is soil flooding that is tolerated by swamp trees but not by most soil decomposers. Paludization is thus an accumulation of undecayed plant debris as peat in the waterlogged surface layer (O horizon of Table 2) of Histosols (Table 1). This process requires a balance between plant production and decomposition. If ponding is intermittent and the soil is moderately oxidized, usually because of a low subsidence rate, then fungal and other decay prevents the accumulation of plant debris. If, on the other hand, ponding is too deep or prolonged, because of high subsidence rates, then soil stagnation kills the roots of woody plants, thus cutting off the supply of vegetation for further peat accumulation. As swamp forests die from anoxia at the roots, peaty soils become overwhelmed by lakes, bayous, or lagoons. The rate of subsidence and accumulation of woody peats is generally between 0.5 mm and 1 mm per year, because of constraints on the growth rate of woody plants in low-fertility peaty substrates and the depth of penetration of air and decomposers within woody peats. Herbaceous plants and mosses are less constrained in their growth rates and form domed peats that rise well above the water table. Peat accumulation in both cases involves addition from the top, in the same way as sediment accumulation, and thus differs from soil-forming processes that modify pre-existing materials. The progress of paludization leads to progressively thicker peat (Figure 1).

Podzolization

Podzol in its original Russian means ‘under ash’ and refers to the light-coloured quartz-rich (E) horizon immediately beneath the humus. Many podzolic

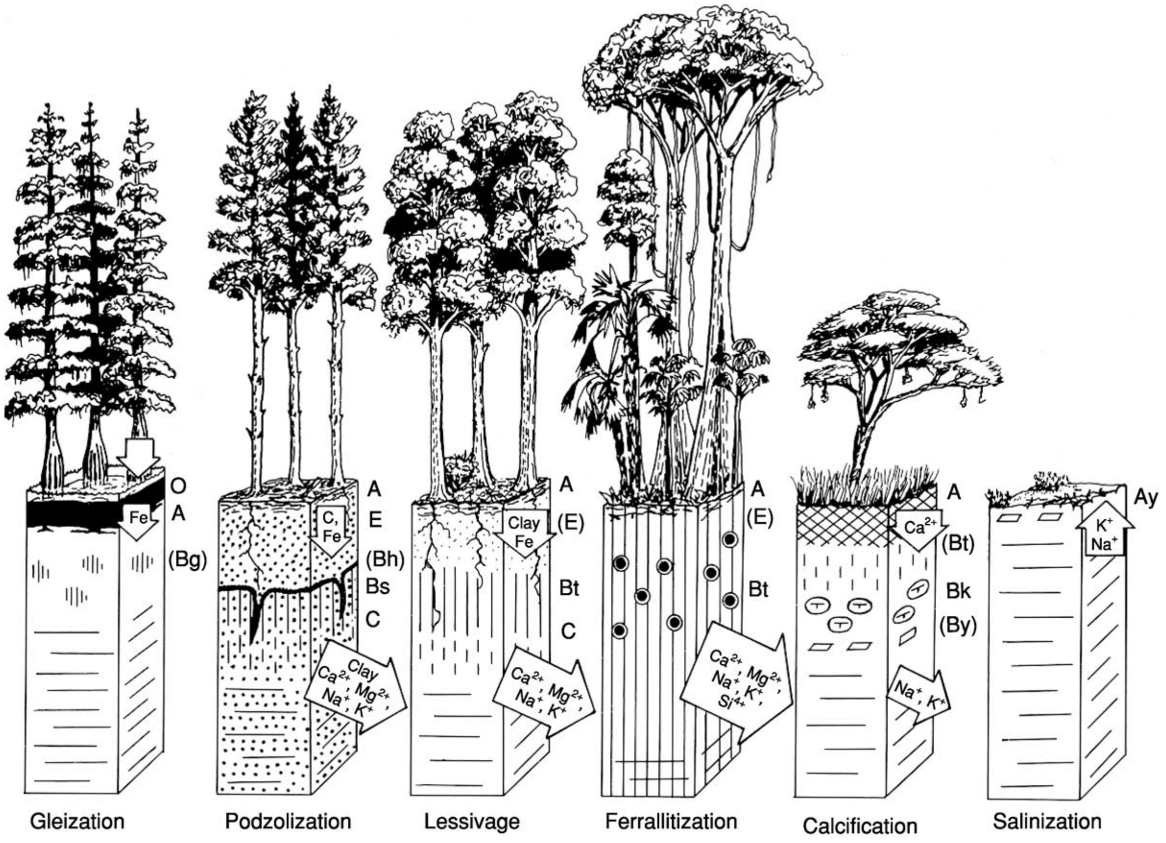


Figure 3 Selected common soil-forming processes arranged along a climatic gradient. The ecosystems depicted are (from left to right): bald cypress swamp, spruce forest, oak forest, tropical rain forest, *Acacia* savannah, and saltbush scrub. Horizon nomenclature is described in Table 1, and the large arrows indicate the movement of key soil components. Reproduced with permission from Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Table 1 Outline of soil taxonomy

Order	*Description*
Entisol	Very weakly developed soil with surface rooting and litter (A horizon) over weathered (C horizon) sediment with relict bedding or weathered igneous or metamorphic rock with relict crystals
Inceptisol	Weakly developed soil with surface rooting and litter (A horizon) over somewhat weathered (Bw horizon) clayey (Bt horizon) or calcareous (Bk horizon) subsurface
Andisol	Soil composed of volcanic ash with low bulk density and high fertility
Histosol	Peat (O horizon) over rooted grey clay (A horizon)
Spodosol	Quartz-rich clay-poor soil with bleached subsurface (E horizon) above a red-black iron–aluminia–organic cemented zone (Bs horizon)
Vertisol	Very clayey profile with common swelling clay (smectite), laterally variable thickness of surface (A horizon), and strongly slickensided subsurface (Bt horizon)
Mollisol	Grassland soil with thick crumb-textured carbon-rich surface (A horizon)
Gelisol	Permafrost soil with frost heave and other periglacial features
Aridisol	Desert soil with a shallow subsurface accumulation of pedogenic carbonate (Bk horizon) and soluble salts (By horizon)
Alfisol	Fertile forest soil with clay-enriched subsurface (Bt horizon) and high amounts of Mg, Ca, Na, and K
Ultisol	Infertile forest soil with clay-enriched subsurface (Bt horizon) and low amounts of Mg, Ca, Na, and K
Oxisol	Deeply weathered tropical soil, often highly ferruginous and aluminous, but with very low amounts of Mg, Ca, Na, and K

For technical limits of soil orders see Soil Survey Staff (2000) *Keys to Soil Taxonomy*. Blacksburg: Pocahontas Press.

soils are now included in the USDA (United States Department of Agriculture) soil order Spodosol (Table 1), which refers to the red, brown, or black (Bs) horizon below the light coloured near-surface layer. This striking differentiation between white near-surface and dark subsurface horizons is created by podzolization, which effectively leaches iron and organic matter from the upper horizons and reprecipitates them in a lower horizon. The resulting effect is as striking as the chromatographic separation of organic compounds, and podzolization is one of the few soil-forming processes that is rapid enough to have been recreated under controlled laboratory conditions. The process is particularly helped by highly acidic soil solutions (with a pH of less than 4) in well-drained soils of humid climates under acid-generating litter such as that of conifer forest (Figure 3). Under highly acidic conditions clay minerals are destroyed, so Podzols and Spodosols usually have a sandy texture.

Ferrallitization

The term ferrallitization is derived from iron (Fe) and aluminium (Al), which become enriched in minerals such as haematite, kaolinite, and gibbsite during intense weathering of well-drained tropical soils such as Oxisols (Figure 3). Much of the loss of major cations (Ca^{2+}, Mg^{2+}, Na^+, K^+) by hydrolysis requires carbonic acid derived from the carbon dioxide of soil respiration, yet the soil pH remains above 4, so that clays are not destroyed. Mitigation of acidity and deep oxidation of these soils may in part be due to the activity of termites and tropical trees, as ferrallitization is primarily found in soils under tropical rainforest. The broad-leaved trees of tropical rainforests produce less acidic litter than conifers and other plants, and litter decomposition rates are high on humid and warm forest floors. Furthermore, ferrallitic soils commonly contain abundant microscopic (125–750 μm) spherical to ovoid pellets of oxidized clay, like the faecal and oral pellets of termites. Some ferrallitic soils appear to have passed through the guts of termites many times. Termites are unique in having extremely alkaline midguts (with a pH of 11–12.5).

Table 2 Standard acronyms for soil-horizon description

Acronym	*Description*
O	Surface accumulation of peaty organic matter
A	Surface horizon of mixed organic and mineral material
E	Subsurface horizon rich in weather-resistant minerals, e.g. quartz
Bt	Subsurface horizon enriched in washed-in clay
Bs	Subsurface horizon enriched in organic matter, or iron or aluminium oxides
Bk	Subsurface horizon enriched in pedogenic carbonate
Bn	Subsurface horizon with domed columnar structure and sodium-clays
By	Subsurface horizon enriched in salts such as gypsum and halite
Bo	Subsurface horizon deeply depleted of Ca, Mg, Na, and K
Bw	Subsurface horizon mildly oxidized and little weathered
C	Mildly weathered transitional horizon between soil and substrate
R	Unweathered bedrock

For technical limits of soil orders see Soil Survey Staff (2000) *Keys to Soil Taxonomy*. Blacksburg: Pocahontas Press.

Biocycling

Biocycling includes a variety of processes in which nutrient elements are exchanged by soil biota without reincorporation into soil minerals. In tropical soils such as Oxisols (Table 1) this is a very efficient process in which the decay of leaves and wood is orchestrated by waves of bacteria, fungi, ants, and termites, which excrete and die to feed a copious network of epiphytes and tree roots. Effective biocycling explains the spectacular luxuriance of tropical-rainforest ecosystems despite their extremely nutrient-depleted and humus-poor mineral soils (Oxisols). Comparable mechanisms operate in swamp forests growing in peat (Histosols), which also experience severe mineral-nutrient limitations. These mineral nutrients include the major cations (Ca^{2+}, Mg^{2+}, Na^+, K^+), but these are seldom as limiting as nitrogen, which is derived largely from the microbial recombination of atmospheric nitrogen, or phosphorous, which is derived largely from the weathering of apatite. Biocycling of

Table 3 Common kinds of chemical reactions during weathering

Reaction	*Example*
Hydrolysis	$2NaAlSi_3O_8 + 2CO_2 + 11H_2O \rightarrow Al_2Si_2O_5(OH_4) + 2Na^+ + 2HCO_3^- + 4H_4SiO_4$ albite + carbon dioxide + water → kaolinite + sodium ions + bicarbonate ions + silicic acid
Oxidation	$2Fe^{3+} + 4HCO_3^- + 1/2O_2 + 4H_2O \rightarrow Fe_2O_3 + 4CO_2 + 6H_2O$ ferrous ions + bicarbonate ions + oxygen + water → haematite + carbon dioxide + water
Dehydration	$2FeOOH \rightarrow Fe_2O_3 + H_2O$ goethite → haematite + water
Dissolution	$CaCO_3 + CO_2 + H_2O \rightarrow Ca^{2+} + 2HCO_3^-$ calcite + carbon dioxide + water → calcium ions + bicarbonate ions

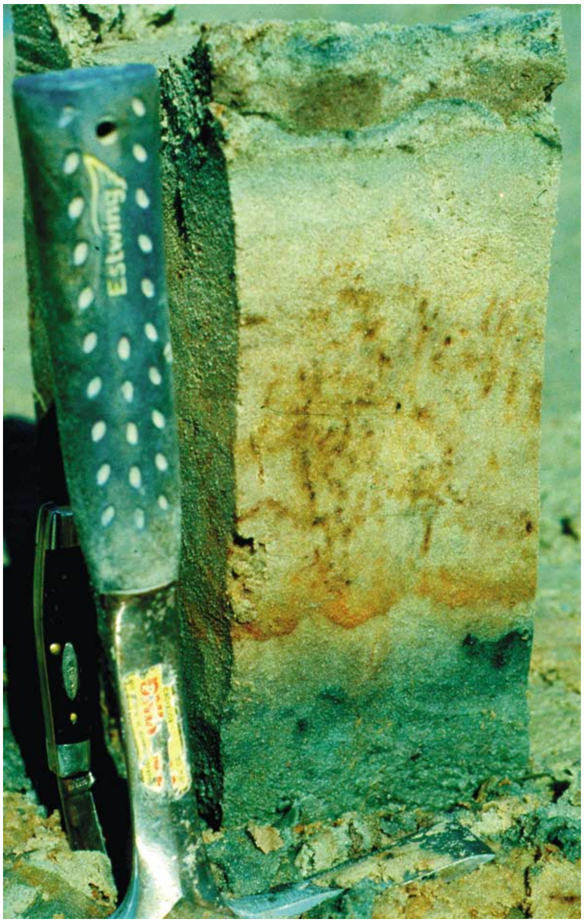


Figure 4 Red and brown mottles of goethite in the upper part of the profile and dark stains of pyrite formed by gleization in the lower part of the profile of a gleyed Inceptisol, excavated as a soil column from a salt marsh on Sapelo Island, Georgia, USA. Hammer handle is 25 cm long.

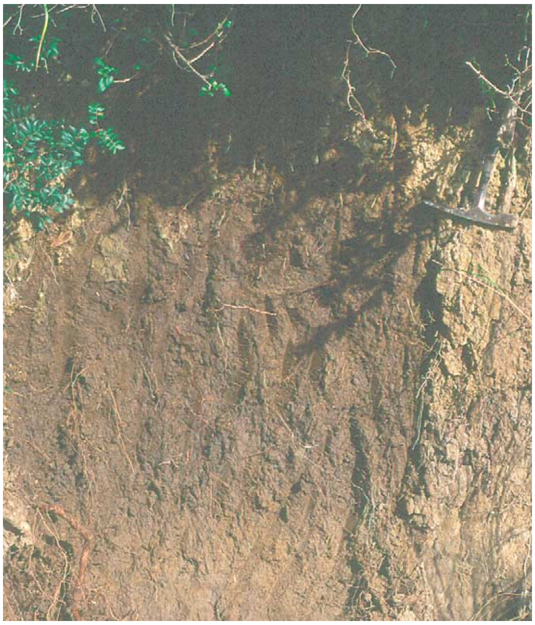


Figure 5 Light-brown near-surface (E) and dark-brown subsurface (Bt) horizons of an Alfisol produced by lessivage near Killini, Greece. Hammer handle upper right is 25 cm long.

nitrogen is especially important during the early development of soils such as Entisols and Inceptisols, which are developed over decades or centuries. Biocycling of phosphorous becomes increasingly important in very old soils such as oxisols and ultisols, which are depleted in apatite over thousands or millions of years.

Lessivage

Lessivage or argilluviation is the process of clay accumulation within a subsurface (Bt or argillic) soil horizon (Figures 1, 2 and 3). This is a common and widespread soil-forming process in the forested soils of humid climates, particularly Alfisols and Ultisols (Figure 5). The clay is primarily derived from a hydrolytic weathering reaction in which clays remain as a residuum and dissolved cations are removed in groundwater during the incongruent dissolution of feldspars and other minerals by carbonic acid (Table 3). Driving the reaction are abundant rainfall and high soil respiration rates fuelled by high primary productivity. Clay forms rinds around mineral grains of the sedimentary, igneous, or metamorphic parent material, but is also washed down cracks in the soil created by desiccation, roots, and burrows. This washed in or illuvial clay has a very distinctive banded appearance, which is obvious in petrographic thin sections. The clay is not washed any lower than the water table, where percolating rainwater ponds. Clay is less common near the surface of the soil, where unweathered grains are added by wind and water, and grains are leached of clay by plant acids. The net effect is a subsurface clayey horizon that becomes more clayey over time (Figures 1, 2 and 3).

Lixiviation

Lixiviation is a process of leaching of major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}) from soil minerals and their loss from the soil in groundwater. Lixiviation is a component of ferrallitization, podzolization, and lessivage, and represents the progress of the hydrolysis chemical reaction, in which hydronium ions (H^{+}) of a weak acid (usually carbonic acid) displace cations into solution and thus convert primary minerals such a feldspars into soil minerals such as clays (Table 3). The term lixiviation is primarily used to describe the beginnings of this process in soils such as Entisols and Inceptisols that have developed over only decades or centuries. Such young soils have not yet acquired the distinctive deeply weathered and oxidized horizons produced by ferrallitization in Oxisols, the distinctive leached (E) and enriched (Bs) horizons produced by podzolization in Spodosols, or the distinctive clay-enriched subsurface (Bt) horizons produced by lessivage in Alfisols and Ultisols.

Melanization

Melanization is a process of soil darkening due to the addition of soil organic matter. The process is best known in Mollisols, the fertile dark crumb-textured soils of grasslands (Figure 6). In these soils melanization is largely a product of the activities of grasses and earthworms. Earthworms produce faecal pellets rich in organic matter and nutrients such as carbonate. Earthworms also produce slime, which facilitates their passage through the soil. Root exudates from grasses are also added to soil crumbs. Many soils have dark humic near-surface horizons, but a peculiarity of grassland soils is that dark organic fertile crumb-textured soil extends to the base of the rooting zone, which can be more than a metre deep in soils under tall-grass prairie. Melanization also occurs in swamp and marsh soils (gleyed Inceptisols and Entisols), where the decay of humus is suppressed by poor oxidation and waterlogging. Unlike the alkaline crumb-textured melanized surface of grassland soils, the melanized surface of wetland soils is nutrient-poor, acidic, and has a massive to laminated fabric. Melanization is not usually applied to the precipitation of amorphous Fe–Mn oxides (birnessite) in gleyed soils, which can also produce dark soil. The creation of these Fe–Mn-stained (placic) horizons is a process of gleization rather than melanization.

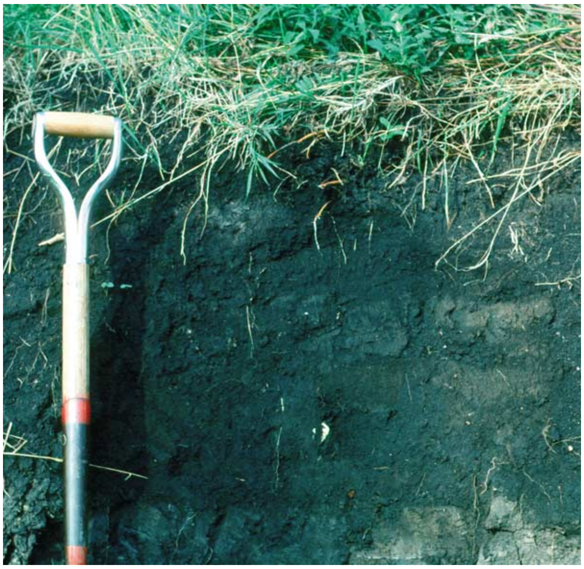


Figure 6 Dark organic-rich surface (mollic epipedon) of a Mollisol formed by melanization near Joliet, IL, USA. The shovel handle is 15 cm wide at the top.

Andisolization

Andisolization is the formation of fertile mineralogically amorphous low-density horizons within soils of volcanic ash (Andisols). Many volcanic ashes are composed largely of small angular fragments (shards) of volcanic glass. Unlike soil minerals such as feldspar, volcanic glass weathers, not to crystalline minerals such as clay, but to non-crystalline compounds such as imogolite. The loosely packed angular shards and colloidal weathering products create a soil of unusually low bulk density (1.0–1.5 $\mathrm{g\,cm^{-3}}$, compared with 2.5–3.0 $\mathrm{g\,cm^{-3}}$ for most common minerals and rocks). Furthermore, these colloidal compounds contain plant-nutrient cations, and particularly phosphorous, in a form that is more readily available to plants than those of other kinds of soils dominated by crystalline minerals such as apatite. Andisolization is not sustainable for more than a few thousand years unless there are renewed inputs of volcanic glass, because glass and other colloids (such as imogolite) weather eventually to oxides and clay minerals.

Vertization

Vertization is the physical soil overturning and mixing by means of the shrink–swell behaviour of clays. It occurs mostly in Vertisols but also in Entisols, Inceptisols, Mollisols, and Alfisols. It is especially characteristic of soils rich in swelling clays (smectites), which swell when wet and shrink when dry. Also characteristic is a climate with a pronounced seasonal contrast in precipitation. During the wet season the clays swell and buckle under the pressure of their inflation. During the dry season they open up in a system of cracks, which are then partly filled by wall collapse. This fill exacerbates the buckling in the next wet season so that the soil develops ridges or mounds with intervening furrows or pits, called gilgai microtopography. In a soil pit, the cracks of mounded areas divide areas of festooned slickensides under the furrows and pits in a distinctive arrangement called mukkara structure (Figure 7). Vertization is mainly a phenomenon of semiarid to subhumid regions. Soils of arid regions are generally not sufficiently clayey, whereas soils of humid regions are generally too deeply weathered to contain abundant smectite and are also stabilized by massive plant and animal communities.

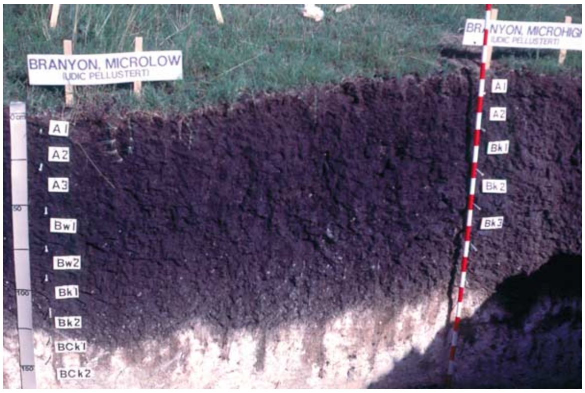


Figure 7 Gilgai microrelief (low to left, high to right) and its subsurface mukkara structure (festooned intersecting slickensided cracks) produced by vertization in the Branyon clay soil, a Vertisol, near New Braunfels, TX, USA. Scale to left shows 50 cm and 100 cm; red and white bands on pole to right are 10 cm wide.

Anthrosolization

Anthrosolization is the alteration of soil by human use, such as buildings, roads, cesspits, garbage dumps, terracing, and ploughing. Archaeological ruins and artefacts are important clues to prior occupation of a site, but many sites also contain impressive amounts of mollusc shells and mammal and fish bones. A distinctive soil structure of subsoil pockets of laminated clay between large soil clods is produced by moldboard ploughs. The primitive or ard plough also tends to disrupt the natural crumb structure to a fixed depth (plough line). Phosphorous content is an indicator of human use. Many soils have trace amounts of phosphorus (10–20 ppm by weight), but occupation floors and long-used garden soils and middens have large amounts of phosphorous (1000–2000 ppm). Anthrosolization is locally common worldwide in cities and fields, both ancient and new, but is scattered and local in deserts, polar regions, and high mountains.

Calcification

Calcification is the accumulation of calcium and magnesium carbonates in the subsurface (Bk) horizons of soils (Figures 1 and 3). The carbonate is usually obvious, appearing as soft white filaments, hard white nodules, and thick white benches within the soil. Calcification is largely a soil-forming process of dry climatic regions, where evaporation exceeds precipitation. It is characteristic of Aridisols but is also found in some Mollisols, Andisols, Vertisols, Inceptisols, and Alfisols. The source of the carbonate is the soil respiration of roots, soil animals, and microorganisms. Calcification requires soil respiration at levels greater than those in hyperarid soils, where halite and gypsum formed by salinization prevail, and less than those in humid soils, where lessivage prevails. The source of the cations of calcium and magnesium, which create the soil minerals calcite and dolomite, respectively, is the weathering of soil minerals by hydrolysis (Table 3). Some of these cations originate from feldspars and other minerals of the parent material, but dry regions of calcification have open vegetation and are often dusty, so that carbonate and feldspar dust is an important source of cations. Dissolved cations from hydrolytic weathering are commonly lost downstream in the groundwater in humid regions, but in arid lands the water table is commonly much deeper than the soil profiles, which are seldom wet much beyond the depth of rooting. The subsurface zone of groundwater evaporation and absorption is where the wisps of soil carbonate form, then coalesce into nodules and, eventually, thick layers.

Solonization

Solonization is a process by which clays rich in soda are formed within the soils of dry climates (Aridisols), where the hydrolytic mobilization of major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}) is weak. Hydrolysis removes cations from soils by lixiation in humid climates, but in dry climates the acidity created by soil respiration after rain storms is sufficient to remove cations from minerals such as feldspar without leaching them from the profile. Solonized soils commonly contain carbonate nodules of dolomite or low-magnesium calcite, formed by calcification, as well as salts of halite and gypsum, formed by salinization. Solonized soils have illitic clays rich in potassium and smectite clays rich in sodium, and the progress of solonization can be assessed by measuring the pH (which is usually around 9–10), by chemical analysis, or by X-ray diffraction to determine the mineral composition. A field indicator of solonization is the presence of domed columnar peds that run through most of the subsurface (natric or Bn) horizon of the soil (Figure 8). The sodium-smectite clays of solonized soils have some shrink–swell capacity, meaning that they form prismatic cracks as the soil dries out and swelling or domed tops to the prisms when the soil is wet. Solonization is common around desert playa lakes and salinas and in coastal soils affected by saltwater spray.

Solodization

Solodization is intermediate between solonization and lessivage, and creates profiles with acidic-to-neutral near-surface horizons but alkaline subsurface

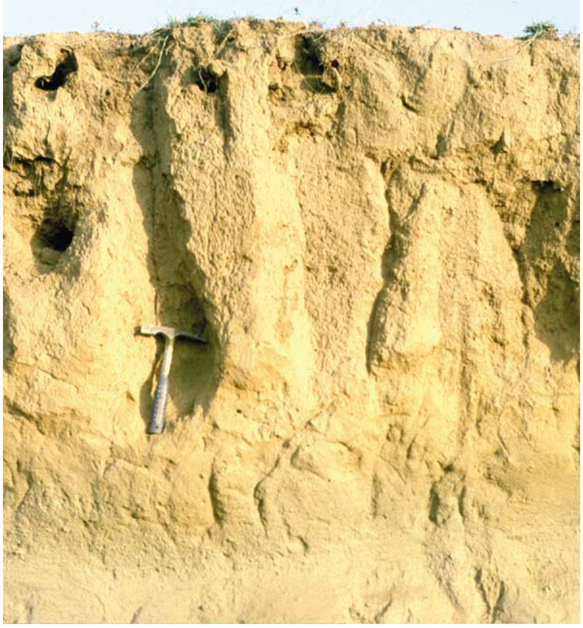


Figure 8 Domed columnar peds produced by solonization in an inceptisol near Narok, Kenya. Hammer handle is 25 cm.

horizons dominated by sodium-smectite. Solodized soils have domed columnar clayey peds in a subsurface (Bn) horizon, but these are sharply truncated by a granular leached (E) horizon. Solodization occurs in desert soils (Aridisols) with better vegetative cover and a more humid climate than solonized soils.

Salinization

Salinization is the precipitation of salts in soils (Figure 3) and is found mostly in desert soils (Aridisols). The most common salts are halite and gypsum, which can form either as clear crystals within soil cracks or as sand crystals that engulf the pre-existing soil matrix. Salts are easily dissolved by rain and so accumulate in regions where there is a marked excess of evaporation over precipitation, which is generally less than 300 mm per year. There is a strong relationship between mean annual precipitation and the depth of leaching of salts in soils. Salinized soils are sparsely vegetated or lack vegetation, and occur in playa lakes, sabkhas, and salinas. Although these are commonly regarded as depositional environments, they are significant soil environments as well.

Cryoturbation

Cryoturbation is the mixing of soils by the freezing and thawing of ground ice. The ice can form disseminated crystals, hair-like threads, thin bands, thick benches, or vertical cracks depending on the local climatic conditions. Soil mixing results from the expansion of water to ice during winter freezing and the relaxation of the deformation on summer melting. Ice-wedge polygons, for example, are wide polygonal cracks that are filled with ice in winter but can be filled with layered sediments in water during the summer in climates where the mean annual temperature is less than −4°C. Sand-wedge polygons form in colder climates where the mean annual temperature is less than −12°C; here, summer melting of ice is limited and sediment fills cracks between the ice and soil in a series of near vertical layers.

Conclusion

Soil-forming processes are varied and complex, and our understanding of them guides the classification, description, and management of soils. The processes are also of interest in simplifying the vast array of chemical reactions, biological processes, and physical effects that create soil. Some processes are more common and widespread than others. Lixiviation and its underlying hydrolysis chemical reaction is perhaps the most important weathering process on Earth, affecting geomorphology, sedimentation, ocean chemistry, and climate. Other processes are restricted to more specific climatic, biotic, geomorphological, geological, and temporal environments, but are no less important in their local environments.

Glossary

Alfisol A fertile forested soil with subsurface enrichment of clay.

Andisol A volcanic-ash soil.

Andisolization A soil-forming process that creates low-density non-crystalline fertile soil from volcanic ash.

Anthropic epipedon A soil surface modified by human use.

Anthrosolization A soil-forming process involving modification by human activities.

Argillic horizon A subsurface horizon of soil enriched in clay.

Argilluviation A soil-forming process that involves creating clay and washing it into a subsurface clayey horizon.

Aridisol A soil of arid regions, usually containing carbonate nodules.

Biocycling The recycling of nutrient elements by biota.

Birnessite A non-crystalline mixture of iron and manganese oxides.

Entisol A very weakly developed soil.

Ferrallitization A soil-forming process involving intense weathering that removes most elements other than iron and aluminium.

Gelisol A soil of permafrost regions, usually containing ground ice.

Gibbsite An aluminium hydroxide mineral ($Al(OH)_3$).

Gilgai A soil microrelief consisting of ridges or mounds alternating with furrows or pits.

Gleization A soil-forming process involving chemical reduction of the soil due to waterlogging.

Halite A salt mineral (NaCl).

Haematite An iron oxide mineral (Fe_2O_3).

Imogolite A colloidal weathering product of volcanic-ash soils.

Inceptisol A weakly developed soil.

Lessivage A soil-forming process that creates clay and washes it into a subsurface clayey horizon.

Lixiviation A soil-forming process that involves leaching nutrient cations from the soil.

Melanization A soil-forming process that involves darkening the soil with organic matter.

Mollic epipedon A humic fertile crumb-textured soil surface typical of grassland soils.

Mollisol A grassland soil with a humic fertile crumb-textured surface.

Mukkara A soil structure consisting of festooned and slickensided cracks between uplifted parts of the soil; the subsurface structures below gilgai microrelief.

Natric horizon A subsurface horizon of soil enriched in sodium-clay.

Oxisol A deeply weathered soil of tropical humid regions.

Paludization A soil-forming process involving peat accumulation in waterlogged soils.

Ped A clod, a unit of soil structure.

Placic horizon Iron- and manganese-stained bands and nodules in soils.

Plaggen epipedon A ploughed surface horizon of soils.

Podzol A sandy soil with a bleached near-surface horizon.

Podzolization A soil-forming process in which acid leaching creates a bleached sandy upper horizon and an iron- or organic-rich subsurface horizon.

Siderite An iron carbonate mineral ($FeCO_3$).

Solonization A soil-forming process that creates soda-rich clays and domed columnar peds in arid regions.

Spodosol A sandy clay-poor soil with an iron- or organic-rich subsurface horizon.

Ultisol A deeply weathered forest soil with subsurface enrichment in clay.

Umbric epipedon A humic acidic clayey massive-to-laminar soil surface found in wetland soils.

Vertisol Swelling clay soil.

Vertization A soil-forming process involving deformation and mixing due to the shrink–swell behaviour of clay during drying and wetting cycles.

See Also

Carbon Cycle. **Clay Minerals**. **Engineering Geology:** Ground Water Monitoring at Solid Waste Landfills. **Sedimentary Environments:** Deltas; Deserts. **Sedimentary Processes:** Glaciers. **Soils:** Palaeosols. **Weathering**.

Further Reading

Bockheim JG and Gennadiyev AN (2000) The role of soil forming processes in the definition of taxa in soil taxonomy. *Geoderma* 95: 53–72.

Bohn H, McNeal B, and O'Connor G (1985) *Soil Chemistry.* New York: Wiley.

Eisenbeis G and Wichard H (1987) *Atlas on the Biology of Soil Arthropods.* Berlin: Springer.

Jenny H (1941) *Factors of Soil Formation.* New York: McGraw-Hill.

Lündstrom US, Van Breeman N, and Bain D (2000) The podzolization process: a review. *Geoderma* 94: 91–107.

McFadden LD, Amundson RG, and Chadwick OA (1991) Numerical modelling, chemical and isotopic studies of carbonate accumulation in arid soils. In: Nettleton WD (ed.) *Occurrence, Characteristics and Genesis of Carbonate Gypsum and Silica Accumulations in Soils,* pp. 17–35. Special Publication 26. Madison: Soil Science Society of America.

Markewich HW, Pavich MJ, and Buell GR (1990) Contrasting soils and landscapes of the Piedmont and Coastal Plain, eastern United States. *Geomorphology* 3: 417–447.

Marshall TJ, Holmes JW, and Rose CW (1996) *Soil Physics.* Cambridge: Cambridge University Press.

Paton TR, Humphreys GS, and Mitchell PB (1995) *Soils: A New Global View.* London: UCL Press.

Retallack GJ (1997) *A Colour Guide to Paleosols.* Chichester: Wiley.

Retallack GJ (2001) *Soils of the Past.* Oxford: Blackwell.

Richter DD and Markewitz D (2001) *Understanding Soil Change.* Cambridge: Cambridge University Press.

Sanford RI (1987) Apogeotropic roots in an Amazon rain forest. *Science* 235: 1062–1064.

Soil Survey Staff (2000) *Keys to Soil Taxonomy.* Blacksburg: Pocahontas Press.

Washburn AL (1980) *Geocryology.* New York: Wiley.

Palaeosols

G J Retallack, University of Oregon, Eugene, OR, USA

Introduction

Palaeosols are ancient soils, formed on landscapes of the past. Most palaeosols have been buried in the sedimentary record, covered by flood debris, landslides, volcanic ash, or lava (Figure 1). Some palaeosols, however, are still at the land surface but are no longer forming in the same way that they did under different climates and vegetation in the past. Climate and vegetation change on a variety of time-scales, and the term relict palaeosol for profiles still at the surface should be used only for such distinct soil materials as laterites among non-lateritic suites of soils (Figure 2). Thus, not all palaeosols are fossil soils or buried soils.

An alternative spelling of paleosol has been adopted by the International Quaternary Association. Other terms such as pedoderm and geosol refer to whole landscapes of buried soils. These soil stratigraphical units are named and mapped in order to establish stratigraphical levels. The terms pedotype and soil facies are more or less equivalent and are used to refer to individual palaeosol types preserved within ancient buried landscapes. These terms are used to distinguish one type of palaeosol from another in environmental interpretations of palaeosols. Pedolith, or soil sediment, describes a sediment, as indicated by bedding and other sedimentary features, with distinctive soil clasts, such as ferruginous concretions. Pedoliths are uncommon in sedimentary sequences, because soils are readily eroded to their constituent mineral grains, which retain few distinctive soil microfabrics.

Recognition of Palaeosols

Palaeosols buried in sedimentary and volcaniclastic sequences can be difficult to distinguish from enclosing sediments, tuffs, or lavas and were not widely recognized before about 20 years ago. Three features of palaeosols in particular aid their identification: root traces, soil horizons, and soil structure.

Soil is often defined as the medium of plant growth. Geological and engineering definitions of soil are broader, but fossilized roots and traces of their former paths through the soil are universally accepted as diagnostic of palaeosols. Not all palaeosol root traces are permineralized or compressed original organic matter: some are tortuous infillings of clay with discoloured haloes or mineralized alteration (Figure 3). Both fossilized roots and root traces show the downward tapering and branching of roots. Soils also contain fossil burrows, but these are usually more sparsely branched and parallel-sided than root traces. The distinction between burrows and roots can be blurred in cases where soil animals feed on roots and where roots find an easier passage through the soft fill of burrows. For very old rocks, predating the Early Devonian evolution of roots, the criterion of root traces is of no use in identifying palaeosols.

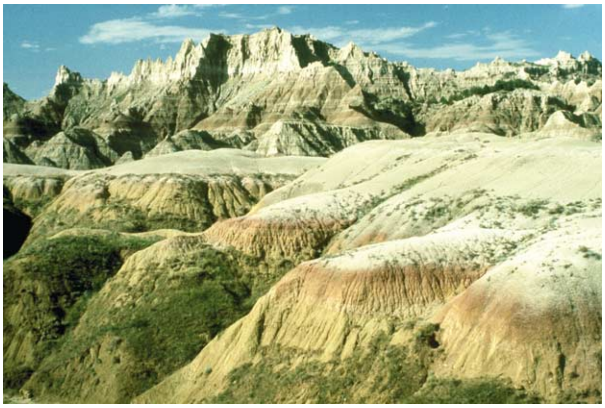


Figure 1 The subtle colour banding in these cliffs is the result of a sequence of 87 Eocene and Oligocene palaeosols in 143 m of nonmarine silty claystones exposed in the Pinnacles area of Badlands National Park, South Dakota, USA.

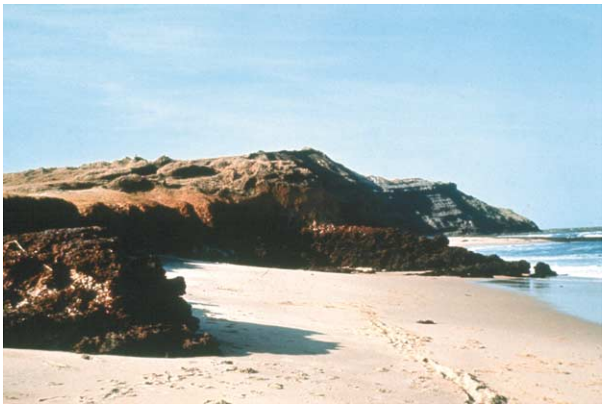


Figure 2 The red rock exposures to the left on the beach are a lateritic palaeosol of Middle Miocene age. Even though these horizons are at the surface, they are considered to be palaeosols because soil horizons of this type are not currently forming in this area. The red rock in the background is a sequence of Early Triassic palaeosols in Bald Hill Claystone, near Long Reef, New South Wales, Australia.

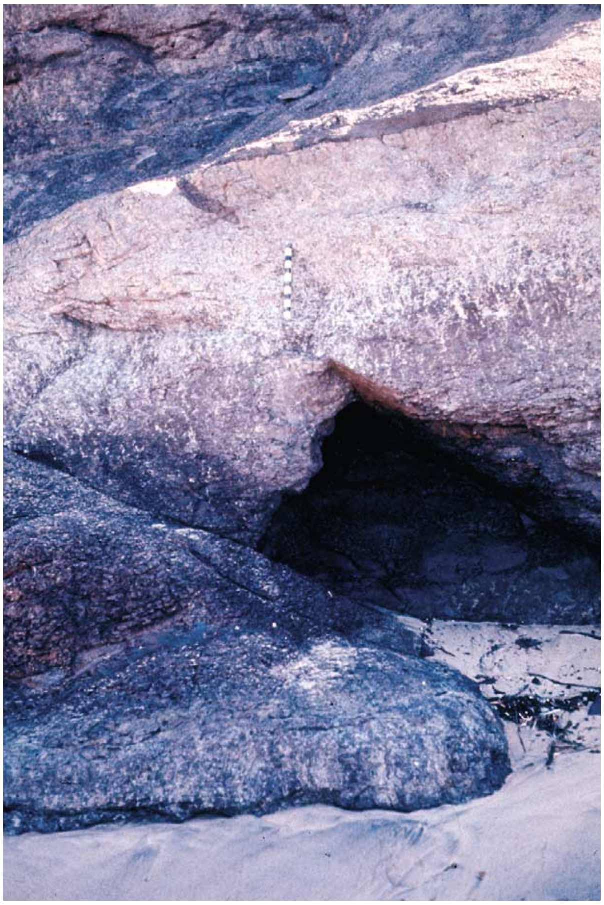


Figure 3 The sharply truncated top and abundant drab-haloed root traces (A horizon) petering out downwards into red claystone (Bt horizon) are soil horizons of a palaeosol (Long Reef clay palaeosol, Early Triassic, Bald Hill Claystone, near Long Reef, New South Wales, Australia).

Palaeosols also have recognizable soil horizons, which differ from most kinds of sedimentary bedding in their diffuse contacts downwards from the sharp upper truncation of the palaeosol at the former land surface. Palaeosol horizons, like soil horizons, are seldom more than a metre thick and tend to follow one of a few set patterns. Subsurface layers enriched in clay are called Bt horizons in the shorthand of soil science (Figure 3). Unlike a soil, in which clayeyness can be gauged by resistance to the shovel or plasticity between the fingers, clayeyness in palaeosols that have been turned to rock by burial compaction must be evaluated by petrographic, X-ray, or geochemical techniques. Subsurface layers enriched in pedogenic micrite are called Bk horizons in the shorthand of soil science and are generally composed of hard calcareous nodules or benches in both soils and lithified palaeosols (Figure 4).

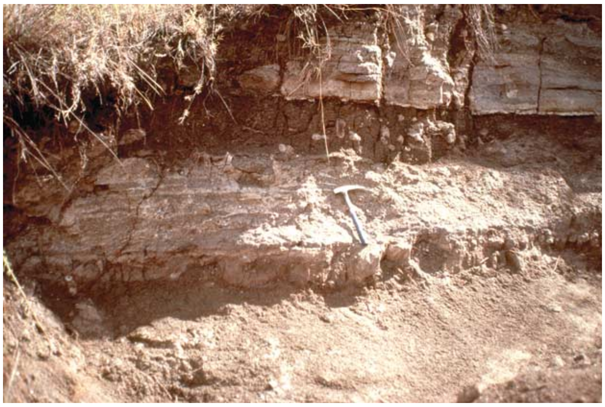


Figure 4 Two successive palaeosols overlain sharply by volcanic grits show crumb-structured organic surfaces (A horizon) over calcareous-nodule-studded subsurfaces (Bk horizon). In the upper right corner is a comparable modern soil (Middle Miocene fossil quarry near Fort Ternan, Kenya).

A final distinctive feature of palaeosols is soil structure, which varies in its degree of expression and replaces sedimentary structures such as bedding planes and ripple marks, metamorphic structures such as schistosity and porphyoblasts, and igneous structures such as crystal outlines and columnar jointing. Because they lack such familiar geological structures, palaeosols are commonly described as featureless, massive, hackly, or jointed. Palaeosols, like soils, have distinctive systems of cracks and clods. The technical term for a natural soil clod is a ped, which can be crumb, granular, blocky, or columnar, among other shapes. Peds are bounded by open cracks in a soil and by surfaces that are modified by plastering over with clay, by rusting, or by other alterations. These irregular altered surfaces are called cutans, and they are vital in recognizing soil peds in palaeosols that have been lithified so that the original cracks are crushed. The rounded 3–4 mm ellipsoidal crumb peds of grassland soils and palaeosols (Figure 4) are quite distinct from the angular blocky peds of forest palaeosols (Figure 3). Common cutans in soils and palaeosols include rusty alteration rinds (ferrans) and laminated coatings of washed-in clay (argillans). Cutans and other features of lithified palaeosols are best studied in petrographic thin sections and by electron microprobing and scanning electron microscopy. Some petrographic fabrics, such as the streaky birefringence of soil clays when viewed under crossed Nicols or sepic plasmic fabric, are diagnostic of soils and palaeosols.

Alteration of Soils after Burial

Palaeosols are seldom exactly like soils because of alteration after burial or exposure to additional weathering, and this can compromise their interpretation and identification with modern soils. Palaeosols, like sediments, can be altered by a wide array of burial processes: cementation with carbonate, haematite, or silica; compaction due to pressure or overburden; thermal maturation of organic matter; and metamorphic recrystallization and partial melting. These high-pressure and high-temperature alterations of palaeosols are not as difficult to disentangle from processes of original soil formation as are three common early modifications: burial decomposition, burial reddening, and burial gleization.

Some soils are buried rapidly by chemically reducing swamps or thick lava flows, preserving most of their organic matter. In contrast, many palaeosols are covered thinly by floodborne silt or colluvium, and their buried organic matter is then decomposed by aerobic bacteria and fungi deep within the newly forming soil of the palaeosol sequence. For this reason many palaeosols have much less organic carbon (fractions of a weight per cent) than comparable modern soils (usually 5–10% by weight of carbon at the surface). Thus palaeosol A horizons are seldom as dark as soil surface horizons, and must be inferred from the abundance of roots rather than from colour and carbon content.

Soils vary considerably in their degree of redness, but most palaeosols are red to reddish brown from haematite (iron oxide) or occasionally yellowish brown from goethite (iron hydroxide). Soils become redder from the poles to the tropics, from moderately drained to well drained sites, and with increasing time for development, as iron hydroxides are dehydrated to oxides. The dehydration of iron hydroxides continues with the burial of soils, so that red palaeosols are not necessarily tropical, unusually well drained, or especially well developed.

In river-valley and coastal sedimentary sequences with abundant palaeosols, formerly well-drained soils can find themselves subsiding below the water table with root traces and humus largely intact. Burial gleization is a process in which organic matter is used by microbes as a fuel for the chemical reduction of yellow and red iron oxides and hydroxides. Comparable processes of biologically induced chemical reduction are common in swamp soils, but superimposition of this process on the organic parts of formerly well-drained soils produces striking effects in some palaeosols. The whole A horizon is turned grey, with grey haloes extending outwards from individual roots, which diminish in abundance down the profile (Figure 3). Burial gleization is especially suspected when the lower parts of the profile are highly oxidized and have deeply penetrating roots, as in well-drained soils, and when there is no pronounced clayey layer that would perch a water table within the soil.

The combined effect of burial decomposition, dehydration, and gleization can completely change the appearance of a soil. The gaudy grey-green Triassic palaeosol shown in Figure 3, for example, was probably modified by all three processes from an originally dark brown over reddish brown forest soil.

Palaeosols and Palaeoclimate

Many palaeosols and soils bear clear marks of the climatic regime in which they formed. The Berkeley soil scientist Hans Jenny quantified the role of climate in soil formation by proposing a space-for-climate strategy. What was needed was a carefully selected group of soils, or climosequence, that varied in climate of formation but were comparable in vegetation, parent material, topographical setting and time for formation. He noted that mean annual rainfall and the depth in the profile to calcareous nodules decline from St Louis west to Colorado Springs, in the mid-western USA, but that temperatures and seasonality at these locations are comparable. Also common to all these soils is grassy vegetation on postglacial loess that is about 14 000–12 000 years old. From these soils he derived a climofunction or mathematical relationship between climate and soil features. A 1994 compilation of comparable data showed a clear relationship between the depth from the surface of the soil of carbonate nodules (D in cm) and the mean annual precipitation (P in mm) according to the formula:

$$P = 139.6 + 6.388D - 1.01303D^2$$

Such climofunctions can be used to interpret palaeoclimate from the depth within palaeosols of calcareous nodules (Figure 4), once allowance is made for reduction in depth due to burial compaction.

Climatic inferences also can be made from ice deformation features, concretions, clay mineral compositions, bioturbation, and chemical analyses of palaeosols. The thick clayey palaeosol shown in Figure 5 is riddled with large root traces of the kind found under forests and is very severely depleted in elemental plant nutrients such as calcium, magnesium, sodium, and potassium. Comparable modern soils are found at mid-latitudes, yet this palaeosol formed during the Triassic at a palaeolatitude of about 70° S. This palaeoclimatic anomaly indicates pronounced global warming, in this case a postapocalyptic greenhouse effect following the largest mass extinction in the history of life at the Permian-Triassic boundary.

Palaeosols and Ancient Ecosystems

Just as soils bear the imprint of the vegetation and other organisms they support, so many aspects of ancient ecosystems can be interpreted from palaeosols. The palaeosols shown in Figure 4, for example, have a dark crumb-textured surface horizon with abundant fine (1–2 mm) roots, comparable to the modern grassland soil seen forming on the outcrop to the upper left. Forest soils, in contrast (Figure 3), have large woody root traces, a blocky structure, and thick subsurface clayey horizons (Bt).

In some cases root traces in palaeosols are identifiable, although the species *Stigmaria ficoides* (Figure 6) is a form genus for roots of a variety of extinct tree lycopsids and not a precisely identified ancient plant. The tabular form of the roots of *Stigmaria* indicates a poorly drained soil, because roots do not photosynthesize, but rather respire using oxygen from soil air. Tabular, rather than deeply reaching, root traces (Figure 3) are characteristic of swamp palaeosols.

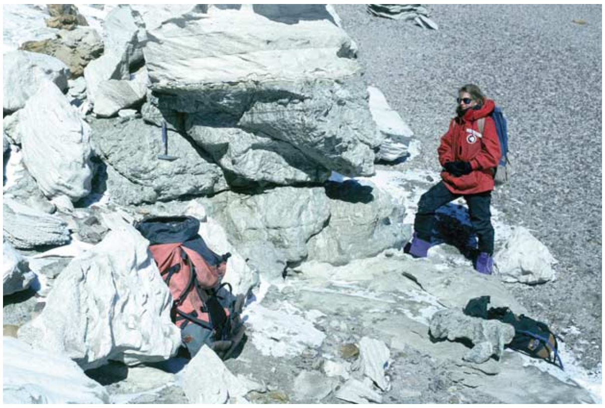


Figure 5 An unusually warm palaeoclimate is indicated by this palaeosol, which is unusually thick, clayey, and deeply weathered for its palaeolatitude of 70°S and is comparable to soils now forming no further south than 48°S (Early Triassic Feather Conglomerate, Allan Hills, Victoria Land, Antarctica).

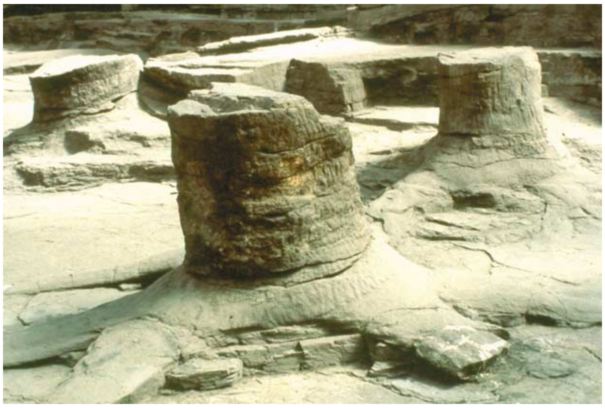


Figure 6 Swamp forests of tree lycopsids (*Stigmaria ficoides*) grew in waterlogged soils, in which lack of oxygen forced the roots to form planar mats rather than reaching deeply into the soil (Carboniferous Lower Limestone Coal Group, Victoria Park, Glasgow, Scotland).

Some palaeosols also contain fossil leaves, fruits, wood, stones, bones, and teeth. These are direct evidence of soil ecosystems. Unlike fossils in deposits of lakes and shallow seas, fossil assemblages in palaeosols have the advantage of being near the place where the organisms lived. However, the preservation of fossils in palaeosols is seldom as ideal as complete skeletons in river-channel deposits or compressed leaves in carbonaceous shales. The carbon and carbonate contents of palaeosols can be used to evaluate the Eh and pH, respectively, of the palaeosol preservational environments of the fossils.

Palaeosols and Palaeogeography

Just as soils vary from mountain tops to coastal swamps, so do palaeosols give clues to their ancient topographical setting. Many palaeosols within sedimentary sequences show clear relationships with deposits of palaeochannels and levees, so that their depositional subenvironment can be inferred from context. Water tables are close to the ground surface in many sedimentary environments, and palaeosols yield important information on their position relative to ancient water tables. Palaeosols formed below the water table include peats and are grey with chemically reduced minerals such as pyrite and siderite. Burrows of crayfish and other aquatic organisms are locally common in waterlogged soils, but burrows of most rodents and beetles are not. Root traces also do not penetrate deeply into waterlogged soils or palaeosols (Figure 6). Deeply penetrating roots and burrows and red oxidized minerals of iron or aluminium are common in formerly well-drained palaeosols (Figure 3). Palaeosols may also reveal upland sedimentary environments such as alluvial and colluvial fans, glacial moraines, river terraces, and erosional gullies (Figure 7).

Major geological unconformities often mark erosional landscapes of the past. Rocky cliffs and bedrock platforms are found along geological unconformities, but so are upland palaeosols. For example, the hilly erosional landscape of Lewisian Gneiss in northern Scotland had 1 km of relief (Figure 8).

Palaeosols and their Parent Materials

The parent material of a soil or palaeosol is the substance from which it formed and can usually be inferred from the less-weathered lower parts of the profile. The parent material may be precisely known if the palaeosol is on metamorphic or igneous rocks (Figure 8), because pedogenic minerals are easily

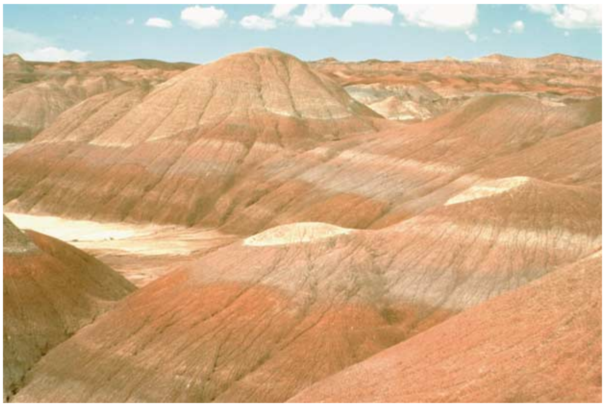


Figure 7 A palaeogully in a strongly developed sequence of palaeosols (dark coloured) is filled with alluvium including weakly developed palaeosols (Late Triassic Chinle Formation, Petrified Forest National Park, Arizona, USA). The hill in the foreground is 11 m high. Photograph courtesy of Mary Kraus.

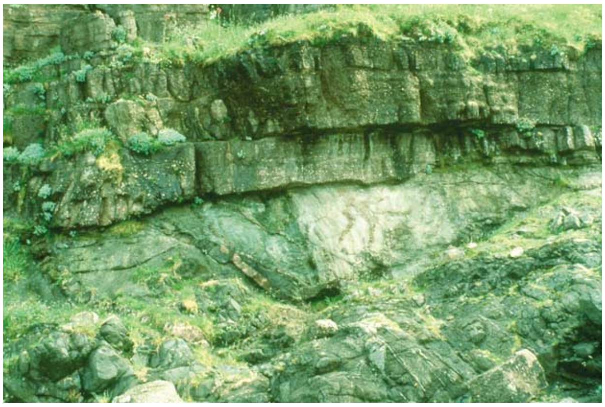


Figure 8 The bleached pink palaeosol formed on gneiss to the right (Sheigra palaeosol) is thicker and more deeply weathered than the light green palaeosol formed on amphibolite to the left (Staca palaeosol). Both palaeosols are overlain by red quartz sandstones of the Torridonian Group (Late Precambrian, near Sheigra, Scotland).

distinquished from igneous and metamorphic minerals. Parent material is more difficult to find in palaeosols that are developed from sedimentary parent materials, especially if sedimentary facies reveal erosional relief (Figure 7). In such settings, the sediment is derived from pre-existing soils, whose degree of weathering can be quite varied. The kinds of soils of sediment and rock also can be very different. If soil were a commercial product, economy would dictate manufacturing it from materials that are already similar in chemical composition and physical characteristics. Soils form more readily from sediments than from rocks. Perhaps the most distinctive of parent materials is volcanic ash, because it may consist of more volcanic glass than minerals. Volcanic glass weathers to noncrystalline amorphous substances such as imogolite, which confer high fertility from loosely bound phosphorous, potassium, and other plant nutrients. Such soils also have low bulk density and good moisture-retaining properties. Such soils around tropical volcanoes support intensive agriculture, despite the hazards of the nearby active volcano, because they are so much more fertile than surrounding soils. Comparable palaeosols are commonly associated with volcanic arcs of the past (Figure 1).

Palaeosols and their Times for Formation

Soils develop their profiles over time, although some soils, such as peats, also accumulate layer-by-layer in the manner of sediments. Each palaeosol within a sedimentary or volcanic sequence represents a short break in sedimentary accumulation, or diastem, whose duration can be calculated from key features of the soil. The peats that become coal seams in the geological record, for example, cannot accumulate at rates of more than 1 mm year^{-1} because the roots will be suffocated by stagnant water. Nor can they accumulate at rates of less than 0.5 mm year^{-1} because aerobic decay will destroy the organic debris as fast as it accumulates. Thus, the durations of coal-bearing palaeosols can be calculated from coal thickness, once compaction is taken into account. Calcareous soils and palaeosols accumulate carbonate at first in wisps and filaments, and later in nodules, which become larger and larger (Figure 4). The size of the nodules thus gives us an idea of the time over which they formed. The development of clayey subsurface horizons is comparable (Figure 3) in that clay becomes more and more abundant over time. The amount of washed-in clay can thus be a guide to the time over which palaeosols formed.

From the times for palaeosol formation and the thickness of rock for successive palaeosols it is possible to calculate rates of sediment accumulation. In the badlands of South Dakota, for example, the clayey lower part of the section accumulated at a slower rate than the ashy and silty upper part of the section (Figure 1). Variations in the rate of sediment accumulation can be used to address a variety of tectonic, volcanic, and sequence stratigraphical problems using palaeosols.

Glossary

Argillan Clay skin, a kind of planar feature in a soil or cutan formed of clay.

Burial decomposition An early diagenetic modification of a palaeosol in which buried organic matter is decayed microbially.

Burial gleization An early diagenetic modification of a palaeosol in which buried organic matter fuels microbial chemical reduction of iron oxides and oxyhydraes to ferrous clays, siderite or pyrite.

Climofunction A mathematical relationship between a soil feature and a measure of climate.

Climosequence A set of soils formed under similar vegetation, topographic setting, parent material and time, but varied climate.

Concretion A seggregation of materials in a soil, harder or more cemented than the matrix, with prominent internal concentric banding, for example iron-manganese concretion.

Cutan A planar feature within a soil formed by enrichment, bleaching, coating or other alteration, for example a clay skin (argillan).

Ferran Ferruginized surface, a kind of planar feature in a soil (cutan) formed by chemical oxidation.

Geosol A mappable land surface of palaeosols, a soil stratigraphic unit in the North American Code of Stratigraphic Nomenclature.

Nodule A segregation of materials in a soil, harder or more cemented than the matrix, with massive internal fabric, for example caliche nodule.

Palaeosol A soil of a landscape of the past: a past surficial region of a planet or similar body altered in place by biological, chemical or physical processes, or a combination of these.

Ped A natural aggregate of soil: stable lumps or clods of soil between roots, burrows, cracks and other planes of weakness.

Pedoderm A mappable land surface of palaeosols, a soil stratigraphic unit in the Australian Code of Stratigraphic Nomenclature.

Pedolith Soil sediment: a seadimentary rock dominated by clasts with the internal microfabrics of soils.

Pedotype A kind of palaeosol: an ancient equivalent of soil series of the United States Soil Conservation Service.

Perched water table Level of water ponded in a soil by an impermeable subsurface layer.

Sepic plasmic fabric Birefringence microfabric: appearance of the fine grained part of a soil or palaeosol in petrographic thin sections viewed under crossed Nicols of wisps or streaks of highly oriented and highly birefringent clay in a less organized dark matrix.

See Also

Carbon Cycle. **Clay Minerals**. **Palaeoclimates**. **Sedimentary Environments:** Depositional Systems and Facies; Alluvial Fans, Alluvial Sediments and Settings. **Sedimentary Processes:** Karst and Palaeokarst. **Sedimentary Rocks:** Evaporites. **Soils:** Modern. **Weathering**.

Further Reading

Delvigne JE (1998) *Atlas of Micromorphology of Mineral Alteration and Weathering*. Canadian Mineralogist Special Publication 3. Ottawa: Mineralogical Association of Canada.

Follmer LR, Johnson GD, and Catt JA (eds.) (1998) Revisitation of concepts in paleopedology. *Quaternary International* 51/52: 1–221.

International Subcommission on Stratigraphic Classification (1994) *International Stratigraphic Guide*. Boulder: Geological Society of America.

Jenny HJ (1941) *Factors in Soil Formation*. New York: Wiley.

Martini IP and Chesworth W (eds.) (1992) *Weathering, Soils and Paleosols*. Amsterdam: Elsevier.

Ollier C (1991) *Ancient Landforms*. London: Belhaven.

Ollier C and Pain C (1996) *Regolith, Soils and Landforms*. Chichester: Wiley.

Reinhardt J and Sigleo WR (1988) *Paleosols and Weathering through Geologic Time: Principles and Applications*. Special Paper 216. Boulder: Geological Society of America.

Retallack GJ (1983) *Late Eocene and Oligocene Paleosols from Badlands National Park, South Dakota*. Special Paper 193. Boulder: Geological Society of America.

Retallack GJ (ed.) (1986) Precambrian paleopedology. *Precambrian Research* 32: 93–259.

Retallack GJ (1991) *Miocene Paleosols and Ape Habitats of Pakistan and Kenya*. New York: Oxford University Press.

Retallack GJ (1997) *A Colour Guide to Paleosols*. Chichester: Wiley.

Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Retallack GJ, Bestland EA, and Fremd TJ (2000) *Eocene and Oligocene Paleosols of Central Oregon*. Special Paper 344. Boulder: Geological Society of America.

Thiry M and Simon-Coinçon R (eds.) (1999) *Palaeoweathering, Palaeosurfaces, and Related Continental Deposits*. Oxford: Blackwell.

Wright VP (ed.) (1986) *Paleosols: their Recognition and Interpretation*. Oxford: Blackwell.

SOLAR SYSTEM

Contents

The Sun

K R Lang, Tufts University, Medford, MA, USA

Physical Characteristics of the Sun

Distance to the Sun

The mean distance of the Sun from the Earth sets the scale of our Solar System and enables us to infer, from other observations, the luminosity, radius, effective temperature, and mass of the Sun. This distance is called the astronomical unit, or AU for short, with a value of $1\,\mathrm{AU} = 1.49597870 \times 10^{11}\,\mathrm{m}$. At that distance, light from the Sun takes 499.004 782 s to travel to the Earth. By way of comparison, light from the Sun's nearest stellar neighbour, Proxima Centauri (part of the triple star system Alpha Centauri), takes 4.29 years to reach us.

Absolute Solar Luminosity

The Sun's absolute, or intrinsic, luminosity is designated by the symbol $L_\odot$, where the subscript $\odot$ denotes the Sun. We can infer the Sun's luminosity from satellite measurements of the total amount of solar energy reaching every square centimetre of the Earth every second, obtaining $L_\odot = 3.854 \times 10^{26}$ W, where a power of $1\,\mathrm{W} = 1\,\mathrm{J\,s^{-1}}$.

Radius of the Sun

The Sun's radius, which can be inferred from its distance and angular extent, has a value of $R_\odot = 6.955 \times 10^{8}\,\mathrm{m}$. That is about 109 times the radius of the Earth.

The Sun's Effective Temperature

We can use the Stefan–Boltzmann law, together with the Sun's size and luminous output, to determine an effective temperature of 5780 K. The temperature of the Sun increases below and above the visible disk (Table 1).

Mass of the Sun

The Sun's gravitational pull holds the solar system together. That is why we call it a solar system: governed by the central Sun with its huge mass. This gravitational attraction keeps the planets in orbit around the Sun, with longer orbital periods at increasing distances from the Sun. And since we know the Earth's orbital period and mean distance from the Sun, we can weigh the Sun from a distance, obtaining its mass $M_\odot = 5.9165 \times 10^{11}\ (\mathrm{AU})^3/P^2 = 1.989 \times 10^{30}\,\mathrm{kg}$, where the constant is equal to $4\pi^2/G$, the universal constant of gravitation is G, the semi-major axis of the Earth's orbit is $1\,\mathrm{AU} = 1.4959787 \times 10^{11}\,\mathrm{m}$, and the orbital period of the Earth is $\mathrm{P} = 1$ year $= 3.1557 \times 10^{7}$ s.

The Sun does not just lie at the heart of our solar system; it dominates it. Some 99.8% of all the matter between the Sun and halfway to the nearest star is contained in the Sun. It is 332 946 times the mass of the Earth. All the objects that orbit the Sun—the planets and their moons, the comets, and the asteroids—add up to just 0.2% of the mass in our solar system.

Composition of the Sun

When the intensity of sunlight is displayed as a function of wavelength, in a spectrum, it exhibits numerous fine dark gaps of missing colours called

Table 1 The Sun's physical properties[a]

Mean distance, AU	1.4959787×10^{11} m	
Light travel time from Sun to Earth	499.004782 s	
Radius, $R_{\odot}$	6.955×10^{8} m (109 Earth radii)	
Volume	1.412×10^{27} m^{3} (1.3 million Earths)	
Mass, $M_{\odot}$	1.989×10^{30} kg (332 946 Earth masses)	
Escape velocity at photosphere	617 km s^{-1}	
Mean density	1409 kg m^{-3}	
Solar constant, $f_{\odot}$	1366 J s^{-1} m^{-2} = 1366 W m^{-2}	
Luminosity, $L_{\odot}$	3.854×10^{26} J s^{-1} = 3.854×10^{26} W	
Principal chemical constituents	(By number of atoms)	(By mass fraction)
	Hydrogen 92.1%	$X = 70.68\%$
	Helium 7.8%	$Y = 27.43\%$
	All other 0.1%	$Z = 1.89\%$
Age	4.566 billion years	
Temperature (center)	15.6 million K	
Temperature (effective)	5780 K	
Temperature (photosphere)	6400 K	
Temperature (chromosphere)	6000 to 20 000 K	
Temperature (corona)	2 million to 3 million K	
Rotation period (equator)	26.8 days	
Rotation period (60° latitude)	30.8 days	
Magnetic field (sunspots)	0.1 to 0.4 T = 1000 to 4000 G	
Magnetic field (polar)	0.001 T = 10 G	

[a] Mass density is given in kilograms per cubic metre (kg m^{-3}); the density of water is 1000 kg m^{-3}. The unit of luminosity is joules per second, power is often expressed in watts, where 1.0 W = 1.0 J s^{-1}.

absorption lines. Each chemical element, and only that element, produces a unique set, or pattern, of wavelengths at which the dark absorption lines fall. So these lines can be used to determine the chemical ingredients of the Sun. They indicate that hydrogen is the most abundant element in the visible solar gases. Since the Sun is chemically homogenous, except for its core, a high hydrogen abundance is implied for the entire star, and this was confirmed by subsequent calculations of its luminosity. Hydrogen accounts for 92.1% of the number of atoms in the Sun, and it amounts to 70.68% by mass.

Helium, the second-most abundant element in the Sun, accounts for 7.8% of the number of atoms in the Sun, and it amounts to 27.43% by mass. Helium is so rare on Earth that it was first discovered on the Sun. All of the heavier elements in the Sun amount to only 0.1% of the number of atoms, and just 1.89% by mass.

Rotation of the Sun

The Sun rotates, or spins, around a rotational axis whose top and bottom mark the Sun's north and south poles. Like Earth, the Sun rotates from west to east when viewed from above the north pole, but unlike Earth, different parts of the Sun rotate at different rates. We know from watching sunspots that the visible disk of the Sun rotates faster at the equator than it does at higher latitudes, decreasing in speed evenly towards each pole. Also, because the Earth orbits the Sun, we observe a rotation period that is about a day longer than the true value. The synodic rotation period of the visible solar equator, as observed from Earth, is 26.75 days, while the equatorial region of the visible solar disk is intrinsically spinning about the Sun's axis once every 25.67 days.

Scientists have used sound waves, generated inside the Sun, to show that the differential rotation of the Sun persists to about one-third of the way down inside the Sun, or 220 000 km from the visible disk. Lower down the rotation speed becomes uniform from pole to pole and the rotation rate remains independent of latitude. The Sun's magnetism is probably generated at the interface between the deep interior, which rotates with one speed, and the overlying gas that spins faster in the equatorial middle.

Solar Magnetic Fields

Detailed scrutiny indicates that the visible solar disk often contains dark, ephemeral spots, called sunspots, which can be as large as the Earth. The sunspots appear and disappear, rising out from inside the Sun and moving back into it. Most sunspots remain visible for only a few days; others persist for weeks and even months.

Sunspots contain magnetic fields as strong as 0.3 T, or 3000 G, thousands of times stronger than the Earth's magnetic field. The intense sunspot magnetism chokes off the upward flow of heat and energy from the solar interior, keeping a sunspot thousands of degrees colder than the surrounding gas.

The total number of sunspots visible on the Sun varies over an 11-year cycle. At the maximum in the cycle we may find 100 or more spots on the visible disk of the Sun at one time; at sunspot minimum very few of them are seen, and for periods as long as a month none can be found. Since most forms of solar activity are magnetic in origin, they also follow an 11-year cycle. Thus, the sunspot cycle is also known as the solar cycle of magnetic activity.

Sunspots are usually found in pairs or groups of opposite magnetic polarity. The magnetic field lines emerge from a sunspot of one polarity, loop through the solar atmosphere above it, and enter a neighbouring sunspot of opposite polarity. The highly magnetized realm in, around, and above bipolar sunspot pairs or groups is a disturbed area called an active region; it consists of sunspots and the magnetic loops that connect them.

Sunspots are usually oriented roughly parallel to the Sun's equator, in the east–west direction of the Sun's rotation. Moreover, sunspot pairs in either the northern or southern hemisphere have the same orientation and polarity alignment, with an exact opposite arrangement in the two hemispheres.

The Outer Solar Atmosphere

The visible photosphere, or sphere of light, is the level of the solar atmosphere from which we get our light and heat, and it is the part that we can see with our eyes. The thin chromosphere and extensive corona lie above the visible sharp edge of the photosphere. They can both be seen during a total solar eclipse, when the Moon blocks the intense light of the photosphere.

Telescopes called coronagraphs allow us to see the corona by using occulting disks to mask the Sun's face and block out the photosphere's glare. Modern solar satellites, such as the *Solar and Heliospheric Observatory* (*SOHO*), use coronagraphs to get clear, edge-on views of the corona.

The solar corona has a temperature of millions of degrees kelvin, hundreds of times hotter than the underlying visible solar disk whose effective temperature is 5780 K. Very hot material—such as that within the corona—emits most of its energy at X-ray wavelengths. Also, the photosphere is too cool to emit intense radiation at these wavelengths, so it appears dark under the hot gas. As a result, the million-degree corona can be seen all across the Sun's face, with high spatial and temporal resolution, in X-rays.

Since X-rays are totally absorbed by the Earth's atmosphere, they must be observed through telescopes in space. This has been done using a soft X-ray telescope on the *Yohkoh* spacecraft (Figure 1). *Yohkoh's* soft X-ray images have demonstrated that the corona contains thin, bright, magnetized loops that shape, mold, and constrain the million-degree gas. Wherever the magnetism in these coronal loops is strongest, the coronal gas in them shines brightly at soft X-ray wavelengths.

Not all magnetic fields on the Sun are closed loops. Some of the magnetic fields extend outward, within regions called coronal holes. These extended regions

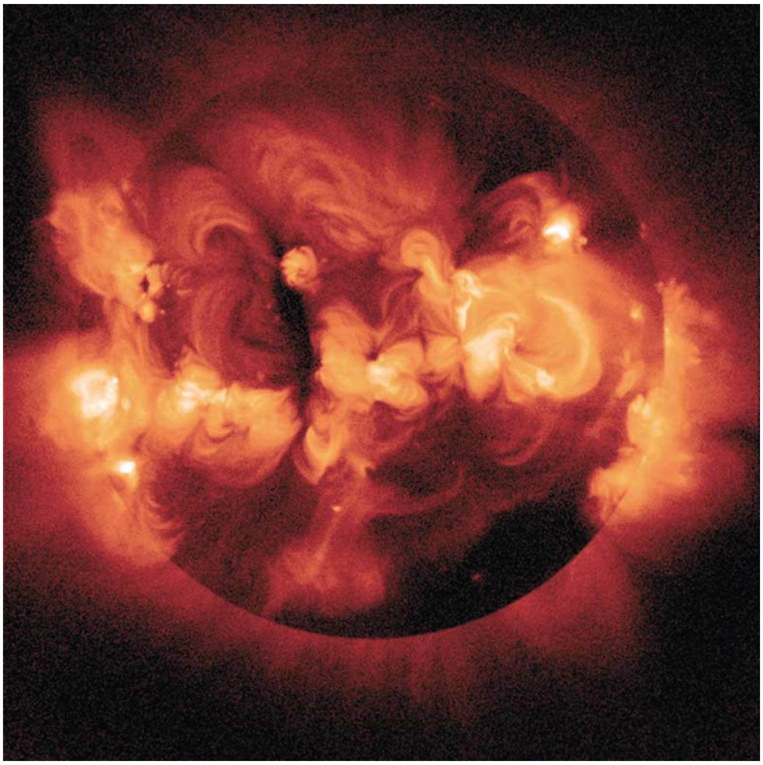


Figure 1 The Sun in X-rays. The bright glow seen in this X-ray image of the Sun is produced by ionized gases at a temperature of a few million degrees kelvin. It shows magnetic coronal loops which thread the corona and hold the hot gases in place. The brightest features are called active regions and correspond to the sites of the most intense magnetic field strength. This image of the Sun's corona was recorded by the Soft X-ray Telescope (SXT) aboard the Japanese *Yohkoh* satellite on 1 February 1992, near the maximum of the 11-year cycle of solar magnetic activity. Courtesy of Gregory L Slater, Gary A Linford, and Lawrence Shing, NASA, ISAS, Lockheed-Martin Solar and Astrophysics Laboratory, National Astronomical Observatory of Japan, and University of Tokyo.

have so little hot material in them that they appear as large dark areas seemingly devoid of radiation at X-ray wavelengths. Coronal holes are nearly always present at the Sun's poles, and are sometimes found at lower solar latitudes. The open magnetic fields in coronal holes do not return directly to another place on the Sun, allowing charged particles to escape the Sun's magnetic grasp and flow outwards into surrounding space.

Explosions on the Sun

Solar Flares

Sudden and brief explosions, called solar flares, rip through the atmosphere above sunspots, releasing an incredible amount of energy, amounting to as much as a million, billion, billion (10^{24}) joules in just a few minutes. All of this power is created in a relatively compact explosion, comparable in total area to an Earth-sized sunspot.

For a short time, usually about 10 minutes, a flare is heated to tens of millions of degrees kelvin. The explosion floods the solar system with intense radiation across the full electromagnetic spectrum, from the shortest X-rays to the longest radio waves, and hurls high-energy electrons and protons out into interplanetary space.

Despite the powerful cataclysm, most solar flares are only minor perturbations in the total amount of emitted sunlight. Routine visual observations of solar explosions are only made possible by tuning into the red emission of hydrogen alpha, designated Hα, at a wavelength of 656.3 nm, and rejecting all the other colours of sunlight.

Since solar flares are very hot, they emit the bulk of their energy at X-ray wavelengths, and for a short while, a large flare can outshine the entire Sun in X-rays. The energetic electrons that produce the impulsive, flaring X-ray emission also emit radio waves known as a radio burst to emphasize its brief, energetic, and explosive characteristics. A solar flare can also outshine the entire Sun at radio wavelengths.

There are more flares near the peak of the 11-year cycle of magnetic activity, but this does not mean that sunspots cause solar flares. They are instead energized by the powerful magnetism associated with sunspots. When these magnetic fields become contorted, they can suddenly and explosively release pent-up magnetic energy as a solar flare, with a main energy release in the corona just above sunspots. The energy is apparently released when magnetized coronal loops, driven by motions beneath them, meet to touch each other and connect. If magnetic fields of opposite polarity are pressed together, an instability takes place and the fields partially annihilate each other, releasing energy to power the explosion.

Coronal Mass Ejections

A coronal mass ejection (CME) is a giant magnetic bubble that rapidly expands to rival the Sun in size. Each time a mass ejection rises out of the corona, it carries away up to 50 billion tons (5×10^{13} kg) of coronal material. Its associated shocks also accelerate and propel vast quantities of high-speed particles ahead of them.

CMEs release about as much energy as a solar flare. However, most of the energy of a mass ejection goes into the kinetic energy of the expelled material, whereas a flare's energy is mainly transferred into accelerated particles that emit intense X-ray and radio radiation.

Coronal mass ejections are detected during routine visible-light observations of the corona from spacecraft such the *SOHO*. With a disk in the centre to block out the Sun's glare, the coronagraph is able to show huge pieces of the corona blasted out from the edge of the occulted photosphere (Figure 2).

Like sunspots, solar flares, and other forms of solar activity, coronal mass ejections occur with a frequency that varies in step with the 11-year cycle. A few coronal mass ejections balloon out of the corona per day, on average, during activity maximum, and the rate decreases by about an order of magnitude by sunspot minimum.

The triggering mechanism for CMEs seems to be related to large-scale interactions of the magnetic field in the low solar corona. This magnetism is continuously emerging from inside the Sun, and disappearing back into it, driven by the Sun's 11-year cycle of magnetic activity. The release of a coronal mass ejection appears to be one way that the solar atmosphere reconfigures itself in response to these slow magnetic changes.

The Sun's Winds

Basic Properties of the Solar Wind

The tenuous solar atmosphere expands out in all directions, filling interplanetary space with a ceaseless wind that is forever blowing from the Sun. This solar wind is mainly composed of electrons and protons, set free from the Sun's abundant hydrogen atoms, but it also contains heavier ions and magnetic fields. This perpetual solar gale brushes past the planets and engulfs them, carrying the Sun's atmosphere out into interstellar space at the rate of a million tons (10^6 tons $= 10^9$ kg) every second.

The relentless wind has never stopped blowing during the more than three decades that it has been

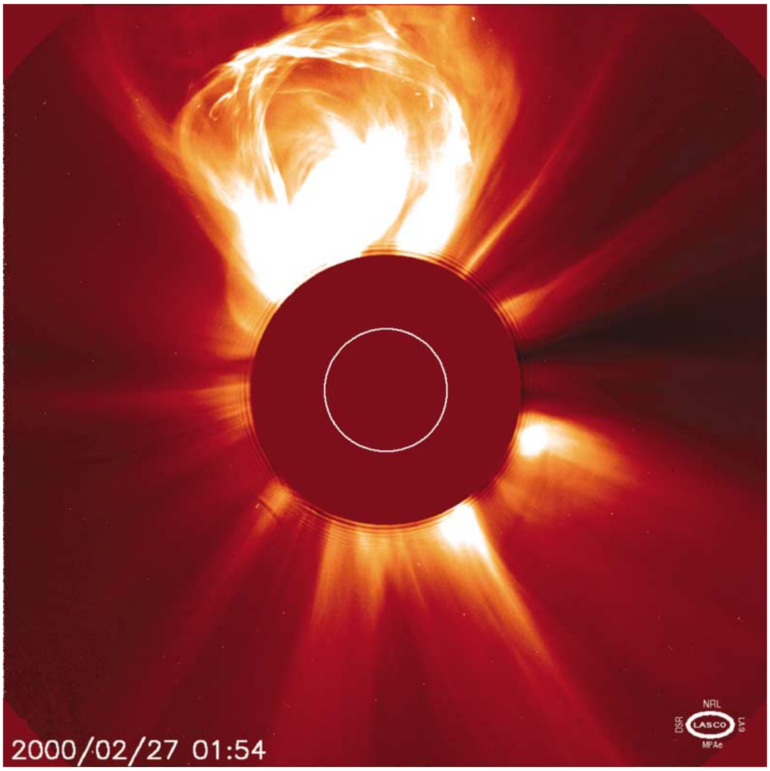


Figure 2 Coronal mass ejection. A huge coronal mass ejection is seen in this coronagraph image, taken on 27 February 2000 with the Large Angle Spectrometric Coronagraph (LASCO) on *the Solar and Heliospheric Observatory* (*SOHO*). The white circle denotes the edge of the photosphere, so this mass ejection is about twice as large as the visible Sun. The black area corresponds to the occulting disk of the coronagraph that blocks intense sunlight and permits the corona to be seen. About one hour before this image was taken, another *SOHO* instrument, the Extreme Ultraviolet Imaging Telescope (EIT), detected a filament eruption lower down near the solar chromosphere. Courtesy of the *SOHO* LASCO consortium. *SOHO* is a project of international cooperation between ESA and NASA.

Table 2 Mean values of solar-wind parameters at the Earth's orbit[a]

Parameter	*Mean value*
Particle density, *N*	$N \approx 10$ million particles m^{-3} (5 million electrons and 5 million protons)
Velocity, *V*	$V \approx 375\,000\ m\ s^{-1}$ and $V \approx 750\,000\ m\ s^{-1}$
Flux, *F*	$F \approx 10^{12}$ to 10^{13} particles $m^{-2}\ s^{-1}$
Temperature, *T*	$T \approx 120\,000$ K (protons) to 140 000 K (electrons)
Magnetic field strength, *H*	$H \approx 6 \times 10^{-9}\ T = 6\ nT = 6 \times 10^{-5}$ G

[a] These solar-wind parameters are at the mean distance of the Earth from the Sun, or at one astronomical unit, 1 AU, where $1\ AU = 1.496 \times 10^{11}$ m.

observed with spacecraft. Two winds are always detected—a fast, uniform wind blowing at about 750 km s^{-1} and a variable, gusty slow wind moving at about half that speed.

By the time it reaches the Earth's orbit, the solar wind has been diluted by its expansion into the increasing volume of space to about 5 million electrons and 5 million protons per cubic meter (Table 2).

The charged particles in the solar wind drag the Sun's magnetic fields with them. While one end of the interplanetary magnetic field remains firmly rooted in the photosphere and below, the other end is extended and stretched out by the radial expansion of the solar wind. The Sun's rotation bends this radial pattern into an interplanetary spiral shape within the plane of the Sun's equator. This spiral pattern has been confirmed by tracking the radio emission of high-energy electrons emitted during solar flares (Figure 3), as well as by spacecraft that have directly measured the interplanetary magnetism.

Origin of the Sun's Winds

The million-degree coronal gas creates an outward pressure that tends to oppose the inward pull of the Sun's gravity. At great distances, where the solar gravity weakens, the hot protons and electrons in the corona overcome the Sun's gravity and accelerate away to supersonic speed.

Instruments aboard the *Ulysses* spacecraft conclusively proved that a relatively uniform, fast wind pours out at high latitudes near the solar poles, and

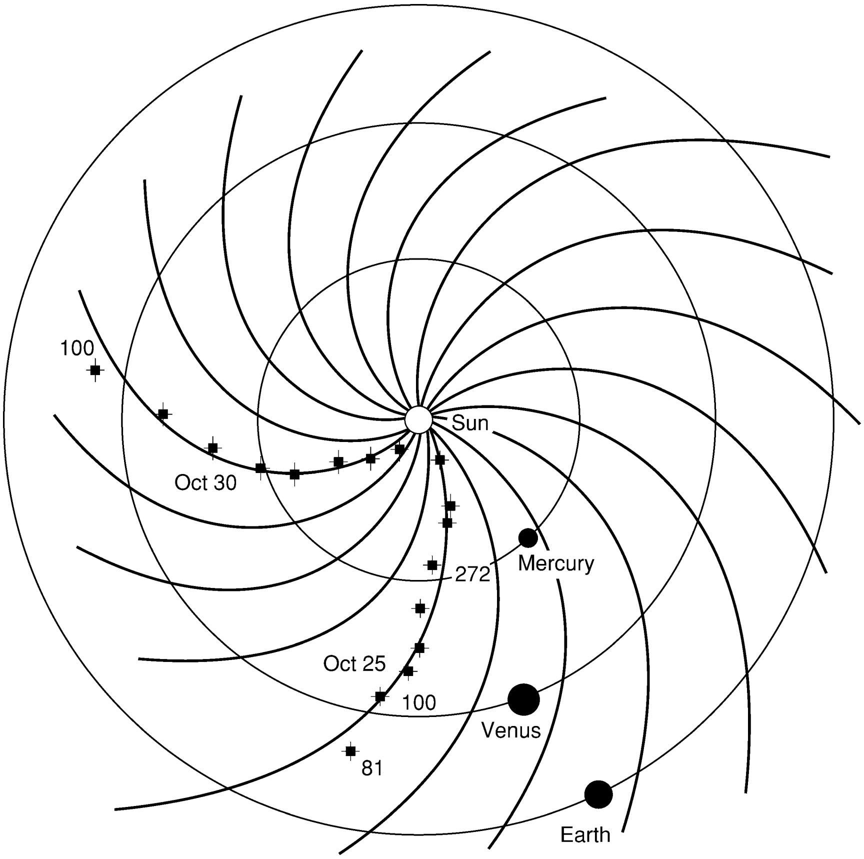


Figure 3 Spiral path of interplanetary electrons. The trajectory of flare electrons in interplanetary space as viewed from above the polar regions using the *Ulysses* spacecraft. As the high-speed electrons move out from the Sun, they excite radiation at successively lower plasma frequencies; the numbers denote the observed frequency in kilohertz (kHz). Since the flaring electrons are forced to follow the interplanetary magnetic field, they do not move in a straight line from the Sun to the Earth, but instead move along the spiral pattern of the interplanetary magnetic field, shown by the solid curved lines. The squares and crosses show *Ulysses* radio measurements of type III radio bursts on 25 October 1994 and 30 October 1994. The approximate locations of the orbits of Mercury, Venus, and the Earth are shown as circles. Courtesy of Michael J Reiner. *Ulysses* is a project of international collaboration between ESA and NASA.

that a capricious, gusty, slow wind emanates from the Sun's equatorial regions at activity minimum.

Comparisons with *Yohkoh* soft X-ray images showed that much, if not all, of the high-speed solar wind flows out of the open magnetic fields in polar coronal holes, at least during the minimum in the 11-year cycle of magnetic activity. In addition, instruments aboard *SOHO* have shown that the strongest high-speed flows gush out of a magnetic network at the bottom of coronal holes near the Sun's poles. Comparisons of *Ulysses* data with coronagraph images pinpointed the equatorial coronal streamers as the birthplace of the slow and sporadic wind during the minimum in the 11-year cycle.

The Heliosphere

A solar gale carries the Sun's rarefied atmosphere past the planets and out to the space between the stars, creating a large cavity in interstellar space called the heliosphere—from the Greek word 'helios' for 'Sun'. Within the heliosphere, physical conditions are dominated, established, maintained, modified, and governed by the magnetic fields and charged particles in the solar wind.

The solar wind's domain extends out to about 150 times the distance between the Earth and Sun, marking the outer boundary of the heliosphere and the edge of our solar system. Out there, the solar wind has become so weakened by expansion that it is no longer dense or powerful enough to repel the ionized matter and magnetic fields coursing between the stars.

The Sun–Earth Connection

Radiation from the Sun

The Sun emits radiation that carries energy through space as waves. Different types of solar radiation differ in their wavelength, although they propagate at the same speed—299 792 458 m s^{-1}, the velocity of light.

Our eyes detect a narrow range of these wavelengths that is collectively called visible radiation. From long to short waves, the colours of visible sunlight correspond to red, orange, yellow, green, blue, and violet. Red light has a wavelength of about 7×10^{-7} m, or 700 nm, and violet waves are about 400 nm long.

Although the most intense radiation from the Sun is emitted at visible wavelengths, it emits less luminous radiation at invisible wavelengths that include the infrared and radio waves, with wavelengths longer than that of red light, and ultraviolet (UV), X-rays, and gamma (γ) rays, which have wavelengths shorter than that of violet light.

Radio waves have wavelengths between 0.001 and 10 m, and they pass easily through the atmosphere, even on cloudy days or in stormy weather. The infrared part of the Sun's spectrum is located between the radio-wave region and the red part of the visible region. Atmospheric molecules, such as carbon dioxide and water vapor, absorb infrared radiation from the Sun.

The short-wavelength, ultraviolet radiation from the Sun is sufficiently energetic to tear electrons or atoms off many of the molecular constituents of the Earth's atmosphere, particularly in the stratosphere where ozone is formed.

The X-ray region of the Sun's spectrum extends from a wavelength of one-hundred billionth (10^{-11}) of a meter, which is about the size of an atom, to the short-wavelength side of the ultraviolet. X-ray radiation is so energetic that it is usually described in terms of the energy it carries. The X-ray region lies between 1 and 100 keV (kiloelectron volts) of energy, where 1 keV = 1.6×10^{-16} J. The atmosphere effectively absorbs most of the Sun's ultraviolet radiation and all of its X-rays.

Varying Solar Irradiance of Earth

The total amount of the Sun's life-sustaining energy is called the 'solar constant', and it is defined as the total amount of radiant solar energy per unit time per unit area reaching the top of the Earth's atmosphere at the Earth's mean distance from the Sun. Satellites have been used to accurately measure the solar constant, obtaining a value of $f_{\odot} = 1366.2\ \mathrm{W\,m^{-2}}$, where the power of one watt is equivalent to one joule per second and the uncertainty in this measurement is $\pm 1.0\ \mathrm{W\,m^{-2}}$.

The total power received at any square metre of the Earth's surface, known as the solar insolation, is much less than the solar constant. This is due to the absorption of sunlight in the terrestrial atmosphere, as well as the time of day.

The solar constant is almost always changing, in amounts of up to a few tenths of a per cent and on time-scales from 1 s to 20 years. This inconstant behaviour can be traced to changing magnetic fields in the solar atmosphere, and its variation tracks the 11-year cycle of magnetic activity (Figure 4).

There are enormous changes in the Sun's radiation at the short ultraviolet and X-ray wavelengths that contribute only a tiny fraction of the Sun's total luminosity. The ultraviolet emission doubles from the minimum to maximum of the 11-year cycle, while the X-ray brightness of the corona increases by a factor of 100.

Global Warming and Cooling by the Sun

The brightening and dimming of the Sun dominated our climate for two centuries, from 1600 to 1800. Cooling by hazy emission from volcanoes next played an important role, but the Sun noticeably warmed the climate for another century, from 1870 to 1970. After that, heat-trapping gases apparently took control of our climate. Global warming by the greenhouse effect is probably responsible for this recent, unprecedented rise in temperature. If current emissions of carbon dioxide and other greenhouse gases go unchecked, the average surface temperature of the globe will rise by about 2°C, making the Earth hotter than it has been in millions of years.

The varying Sun may offset some of this warming. Observations of other stars indicate that the Sun luminosity could vary by much more than that observed by satellites so far, producing dramatic changes in the Earth's climate on time-scales of hundreds and thousands of years. Radioactive isotopes found in both tree rings and ice cores indicate that the Sun's activity has fallen to unusually low levels at least three times during the past one thousand years, each drop apparently corresponding to a long, cold spell on Earth of roughly a century in duration.

Further back in time, during the past one million years, our climate has been dominated by the recurrent ice ages, each lasting about 100 000 years. These glaciations begin and end in a relatively short interval of unusual warmth, called an interglacial, lasting 10 000 or 20 000 years, when the glaciers retreat. We now live in such a warm interglacial interval. The rhythmic alteration of glacial and interglacial intervals is related to periodic alterations in the amount and distribution of sunlight received by Earth over tens of thousands of years.

Our Sun-Layered Atmosphere

Our thin atmosphere is pulled close to the Earth by its gravity, and suspended above the ground by

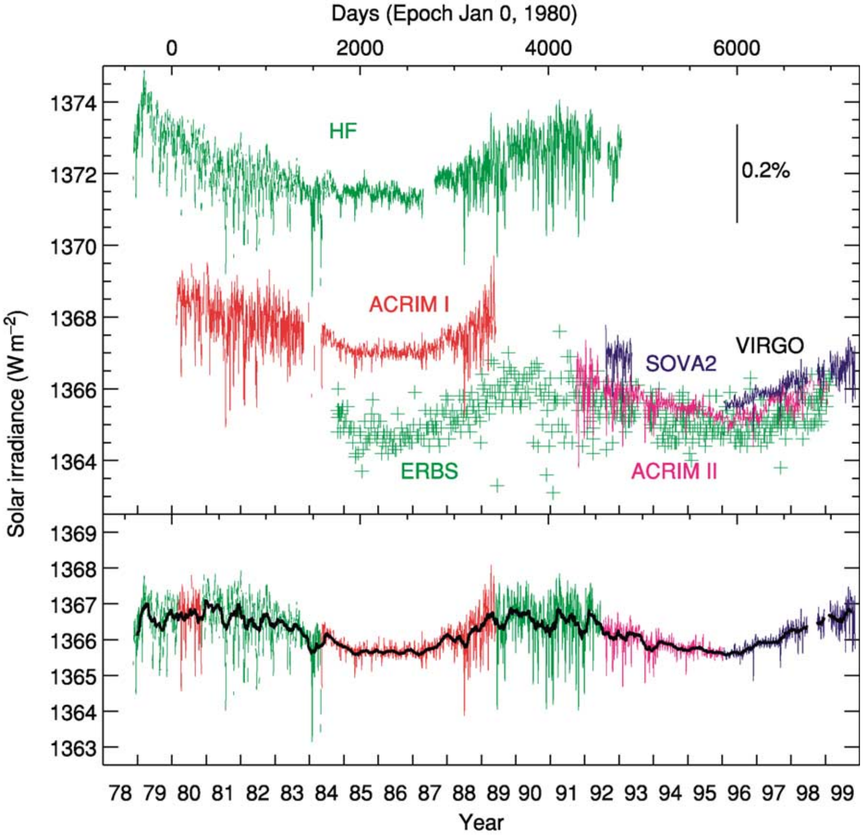


Figure 4 Variations in the solar constant. Observations with very stable and precise detectors on several Earth-orbiting satellites show that the Sun's total radiative input to the Earth, termed the solar irradiance, is not a constant, but instead varies over time-scales of days and years. Measurements from five independent space-based radiometers since 1978 (top) have been combined to produce the composite solar irradiance (bottom) over two decades. They show that the Sun's output fluctuates during each 11-year sunspot cycle, changing by about 0.1% between maximums (1980 and 1990) and minimums (1987 and 1997) in magnetic activity. Temporary dips of up to 0.3% and a few days' duration are due to the presence of large sunspots on the visible hemisphere. The larger number of sunspots near the peak in the 11-year cycle is accompanied by a rise in magnetic activity that creates an increase in luminous output that exceeds the cooling effects of sunspots. Here the total irradiance just outside our atmosphere, called the solar constant, is given in units of watts per square metre, where $1\,\mathrm{W} = 1\,\mathrm{J\,s^{-1}}$. The capital letters are acronyms for the different radiometres, and offsets among the various datasets are the direct result of uncertainties in their scales. Despite these offsets, each dataset clearly shows varying radiation levels that track the overall 11-year solar activity cycle. Courtesy of Claus Fröhlich.

molecular motion. The atmosphere near the ground is compacted to its greatest density and pressure by the weight of the overlying air. At greater heights there is less air pushing down from above, so the compression is less and the density and pressure of the air falls off into the near vacuum of space. The temperature of the air falls and rises in two full cycles at increasing altitudes, and the temperature increases are caused by the Sun's energetic radiation (Figure 5).

The temperature above the ground tends to fall at higher altitudes where the air expands in the lower pressure and becomes cooler. The average air temperature drops below the freezing point of water (273 K) about 1 km above the Earth's surface, and bottoms out at roughly 10 times this height.

The temperature increases at greater heights, within the stratosphere, where the Sun's invisible ultraviolet radiation warms the gas and helps make ozone. This ozone layer protects us by absorbing most of the ultraviolet and keeping its destructive rays from reaching the ground, where it can cause eye cataracts and skin cancer.

Due to the Sun's variable ultraviolet radiation, the total global amount of ozone becomes enhanced, depleted, and enhanced again from 1 to 2% every 11 years, modulating the protective ozone layer at a level comparable to human-induced ozone depletion by chemicals wafting up from the ground. Monitoring of the expected recovery of the ozone layer from outlawed, man-made chemicals will therefore require careful watch over how the Sun is changing the layer from above.

The temperature declines rapidly with increasing height just above the stratosphere, reaching the lowest levels in the entire atmosphere, but the temperature

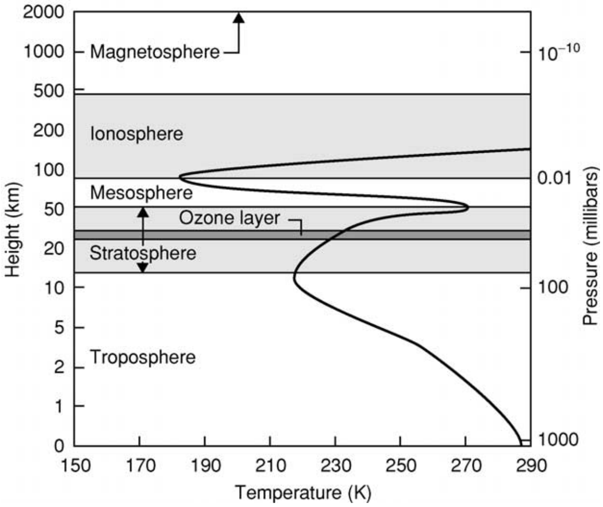


Figure 5 Sun-layered atmosphere. The pressure of our atmosphere (right scale) decreases with altitude (left scale). This is because fewer particles are able to overcome the Earth's gravitational pull and reach higher altitudes. The temperature (bottom scale) also decreases steadily with height in the ground-hugging troposphere, but the temperature increases in two higher regions heated by the Sun. They are the stratosphere, with its critical ozone layer, and the ionosphere. The stratosphere is mainly heated by ultraviolet radiation from the Sun, and the ionosphere is created and modulated by the Sun's X-ray and extreme ultraviolet radiation.

rises again within the ionosphere, reaching temperatures that are hotter than the ground. The ionosphere is created and heated by absorbing the extreme ultraviolet and X-ray portions of the Sun's energy. This radiation tears electrons off the atoms and molecules in the upper atmosphere, thereby creating ions and free electrons not attached to atoms.

At a given height in the ionosphere, the temperature, the density of free electrons, and the density of neutral, unionized atoms all increase and decrease in synchronism with solar activity over its 11-year cycle.

The Earth's Magnetosphere

Invisible magnetic fields, produced by currents in the Earth's molten core, emerge out of the Earth's south geographic polar regions, loop through nearby space, and re-enter at the north polar regions. The surface equatorial field strength is 0.000031 T, or 31 000 nT, and the field strength decreases at greater distances from the Earth.

Yet, the Earth's magnetism is strong enough to deflect the Sun's wind away from the Earth, forming the magnetosphere (Figure 6). The magnetosphere of the Earth, or any other planet, is that region surrounding the planet in which its magnetic field dominates the motions of energetic charged particles such as electrons, protons, and other ions. It is also the volume of space from which the main thrust of the solar wind is excluded.

The solar wind pushes the terrestrial magnetic field towards the Earth on the dayside that faces the Sun, compressing the outer magnetic boundary and forming a bow shock at about 10 times the Earth's radius. Also the Sun's wind drags and stretches the Earth's magnetic field out into a long magnetotail on the night side of our planet. The magnetic field points roughly towards the Earth in the northern half of the tail and away in the southern. The field strength drops to nearly zero at the centre of the tail where the opposite magnetic orientations lie next to each other and currents can flow.

Some of the energetic particles outside the magnetosphere do manage to penetrate it, especially in the magnetotail. When the solar and terrestrial magnetic fields touch each other in the magnetotail, it can catapult the outer part of the tail downstream and propel the inner part back towards Earth.

The inner magnetosphere is always filled with electrons and protons, trapped within two torus-shaped belts that encircle the Earth's equator but do not touch it. These regions are often called the inner and outer Van Allen radiation belts, named after James A Van Allen (1914–) who discovered them in 1958. The inner belt is about 1.5 Earth radii from planet centre, and the outer belt is located at about 4.5 Earth radii, where the Earth's radius is 6378 km.

Intense Geomagnetic Storms

Significant variations in the Earth's magnetic field, lasting seconds to days, are known as geomagnetic storms. The great, sporadic geomagnetic storms, which shake the Earth's magnetic field to its very foundations, can produce magnetic fluctuations as large as 1.6% at mid-terrestrial latitudes, or 500 nT, compared with the Earth's equatorial field strength of 31 000 nT.

Solar wind disturbances driven by exceptionally fast coronal mass ejections produce the most intense geomagnetic storms. The Earth intercepts about 70 coronal mass ejections per year when solar activity is at its peak, and less than 10 will have the punch needed to produce large, geomagnetic storms. These mass ejections plow through the solar wind, driving a huge shock wave far ahead of them. When directed at the Earth, these shocks ram into the terrestrial magnetic field and trigger the initial phase, or sudden commencement, of an intense geomagnetic storm a few days after the mass ejection leaves the Sun.

Strong interplanetary magnetic fields are also generated by fast coronal mass ejections (*see* **Magnetostratigraphy**). It is their intense magnetism and high speed that account for the main phase of a powerful

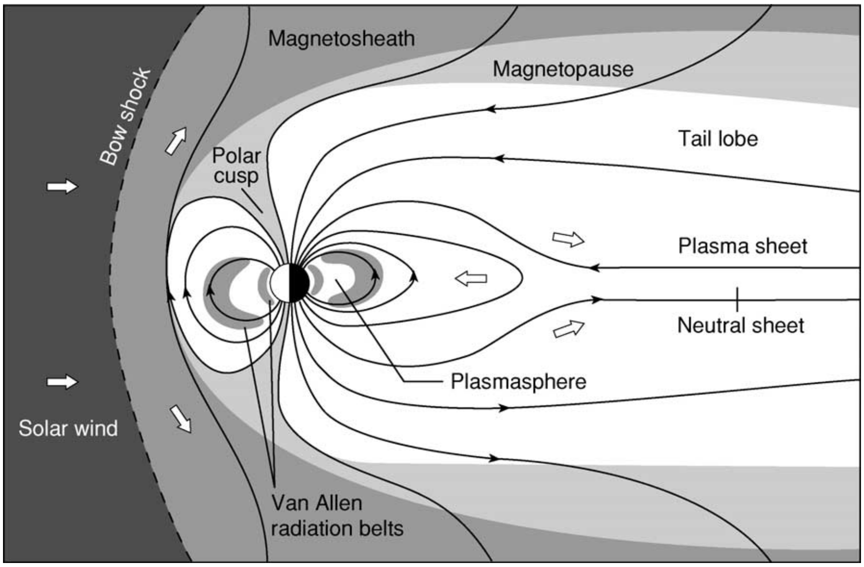


Figure 6 Magnetosphere. The Earth's magnetic field carves out a hollow in the solar wind, creating a protective cavity, called the magnetosphere. A bow shock forms at about 10 Earth radii on the sunlit side of our planet. Its location is highly variable since it is pushed in and out by the gusty solar wind. The magnetopause marks the outer boundary of the magnetosphere, at the place where the solar wind takes control of the motions of charged particles. The solar wind is deflected around the Earth, pulling the terrestrial magnetic field into a long magnetotail on the night side. Plasma in the solar wind is deflected at the bow shock (left), flows along the magnetopause into the magnetic tail (right), and is then injected back towards the Earth and Sun within the plasma sheet (centre). The Earth, its auroras, atmosphere, and ionosphere and the two Van Allen radiation belts all lie within this magnetic cocoon.

geomagnetic storm, provided that the magnetic alignment is right. The Earth's field is generally directed northwards in the outer dayside magnetosphere, so a fast coronal mass ejection is more likely to merge and connect with the terrestrial field if it points in the opposite southward direction.

Moderate Geomagnetic Activity

Moderate mid-latitude magnetic fluctuations of about 0.1%, or tens of nanoTesla, last a few hours, and they are most noticeable near the minimum of the 11-year solar activity cycle. They have a 27-day repetition period, corresponding to the rotation period of the Sun at low solar latitudes when viewed from the moving Earth.

The recurrent activity is linked to long-lived, high-speed streams in the solar wind that emanate from coronal holes. When this fast wind overtakes the slow-speed, equatorial one, the two wind components interact, producing shock waves and intense magnetic fields that rotate with the Sun, and periodically sweep past the Earth, producing moderate geomagnetic activity every 27 days.

The Auroras

The northern or southern lights, named the 'aurora borealis' and 'aurora australis' in Latin, are one of the most magnificent and earliest-known examples of solar–terrestrial interaction. They illuminate the cold, dark Arctic and Antarctic skies with curtains of green and red light that flicker across the night sky far above the highest clouds.

Spacecraft look down on the auroras from high above, showing an oval centred at each magnetic pole (Figure 7). An observer on the ground sees only a small, changing piece of the aurora oval.

The reason that auroras are usually located near the polar regions is that the Earth's magnetic fields guide energetic electrons there. The high-speed electrons move down along the Earth's magnetic field lines into the upper polar atmosphere, colliding with oxygen and nitrogen. The pumped-up atoms or molecules fluoresce, giving up the energy acquired from the electrons and emitting a burst of light.

The electrons that cause the auroras come from the Earth's magnetic tail and are also energized locally within the magnetosphere. The rare, bright, auroras seen at low terrestrial latitudes only become visible during very intense geomagnetic storms.

Space Weather

Down here on the ground, we are shielded from the direct onslaught of solar explosions and the solar wind by the Earth's atmosphere and magnetic fields, but out in deep space there is no protection. Energetic charged particles hurled out from intense solar flares

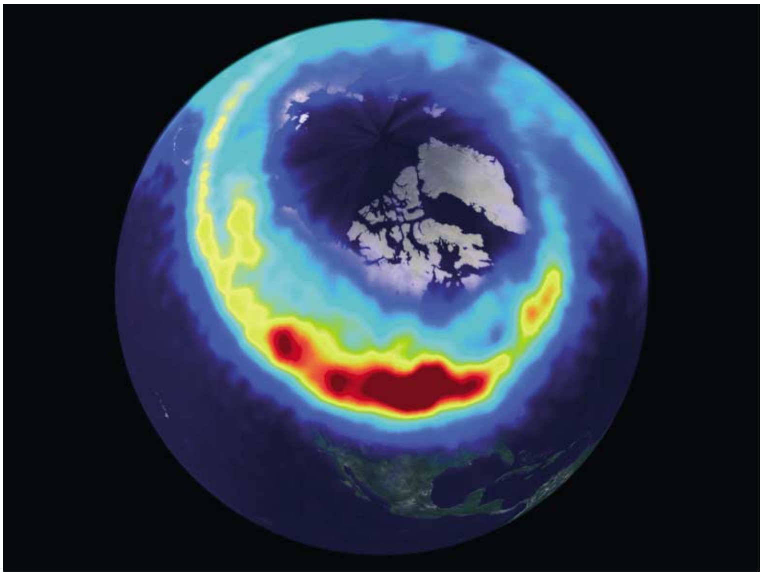


Figure 7 The auroral oval. The *POLAR* spacecraft looks down on an aurora from high above the Earth's north polar region on 22 October 1999, showing the northern lights in their entirety. The glowing oval, imaged in ultraviolet light, is 4500 km across. The most intense aurora activity appears in bright red or yellow. It is typically produced by magnetic reconnection events in the Earth's magnetotail, on the night side of the Earth. Courtesy of the Visible Imaging System, University of Iowa and NASA.

can seriously damage satellites, including their solar cells and electronic components, and even kill an unprotected astronaut.

The high-speed protons and electrons follow a narrow, curved path once they leave the Sun, guided by the spiral structure of the interplanetary magnetic field, so they must be emitted from active regions near the west limb and the solar equator to be magnetically connected with the Earth. Solar flares emitted from other places on the Sun are not likely to hit Earth, but they could be headed towards interplanetary spacecraft, the Moon, Mars, or other planets. The most energetic flare particles can travel from the Sun to the Earth in just 8 minutes, moving at nearly the velocity of light.

Coronal mass ejections move straight out of the Sun, energizing particles over large regions in interplanetary space. Mass ejections are most likely to hit the Earth if they originate near the centre of the solar disk, as viewed from the Earth, and are sent directly towards the planet. They take about 4 days to travel from the Sun to the Earth, moving at a typical speed of about 400 $\mathrm{km\,s^{-1}}$.

The strong blast of X-rays and ultraviolet radiation from a solar flare alters the Earth's atmosphere, transforming the ionosphere, which reflects radio waves to distant locations on Earth. During moderately intense flares, radio communications can be silenced over the Earth's entire sunlit hemisphere, disrupting contact with airplanes flying over oceans or remote countries.

The enhanced ultraviolet and X-ray radiation from solar flares also heats the atmosphere and causes it to expand, and similar or greater effects are caused by coronal mass ejections that produce major geomagnetic storms. The expansion of the terrestrial atmosphere brings higher densities to a given altitude, increasing the drag exerted on a satellite, pulling it to a lower altitude, and causing a premature and fatal spiral towards the Earth.

When a coronal mass ejection slams into the Earth, the force of impact can push the bow shock, at the dayside of the magnetosphere, down to half its usual distance of about 10 times the Earth's radius, compressing the magnetosphere below the orbits of geosynchronous satellites and exposing them to the damaging effects of the full brunt of the gusty solar wind.

During an intense geomagnetic storm, associated with a colliding coronal mass ejection, strong electric currents flow in the ionosphere. They induce potential differences in the ground below them, and produce strong currents in any long conductor such as a power line. These currents can blow circuit breakers, overheat and melt the windings of transformers, and cause massive failures of electrical distribution systems. A coronal mass ejection can thereby plunge major urban centres, like New York City or Montreal, into complete darkness, causing social chaos and threatening safety.

Our technological society has become so vulnerable to the potential devastation of these storms in space that national centres employ space weather forecasters to continuously monitor the Sun from ground and space to warn of threatening solar activity.

See Also

Earth: Orbital Variation (Including Milankovitch Cycles). **Gaia**. **Magnetostratigraphy**. **Palaeoclimates**. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Bone N (1996) *The Aurora: Sun-Earth Interactions*. New York: Wiley.

Calowicz MJ and Lopez RE (2002) *Storms from the Sun: The Emerging Science of Space Weather*. Washington, DC: Joseph Henry Press.

Golub L and Pasachoff JM (2001) *Nearest Star: The Surprising Science of Our Sun*. Cambridge, MA: Harvard University Press.

Kaler JB (1992) *Stars*. Scientific American Library. New York: WH Freeman.

Lang KR (1995) *Sun, Earth and Sky*. New York: Springer-Verlag.

Lang KR (2000) *The Sun from Space*. New York: Springer-Verlag.

Lang KR (2001) *The Cambridge Encyclopedia of the Sun*. New York: Cambridge University Press.

Lang KR (2003) *The Cambridge Guide to the Solar System*. New York: Cambridge University Press.

Odenwald S (2001) *The 23rd Cycle: Learning to Live with a Stormy Sun*. New York: Columbia University Press.

Phillips KJH (1992) *Guide to the Sun*. New York: Cambridge University Press.

Asteroids, Comets and Space Dust

P Moore, Selsey, UK

Introduction

Asteroids and comets must be regarded as minor members of the Solar System. Compared with planets they are of very low mass, and they have even been referred to as cosmic debris. The asteroids, dwarf worlds most of which are well below 1000 km in diameter, are found mainly between the orbits of Mars and Jupiter, though some stray from this 'main belt'; comets have been described as 'dirty snowballs', and though they may become very conspicuous in the sky they are very insubstantial. This article reviews the asteroids and comets, together with the large amount of thinly-spread material lying in the Solar System.

Distribution of the Asteroids

The Solar System is divided into two well-defined parts. There are four relatively small, rocky planets: Mercury, Venus, the Earth, and Mars. Then come the four giants: Jupiter, Saturn, Uranus, and Neptune. Between the orbits of Mars and Jupiter thousands of asteroids, otherwise known as minor planets, make up what is known as the main belt (Figure 1). Of the main belt asteroids, only one (Ceres) is as much as 900 km in diameter, and only one (Vesta) is ever visible with the naked eye. Some of the larger main belt asteroids are listed in Table 1.

Some small asteroids can leave the main belt, and swing closer to the Sun; they may even approach the Earth, and collision cannot be ruled out (it may even be that the impact of an asteroid, some 65 million years ago in Mexico, caused a climatic change and mass extinction, which included the dinosaurs). All of these Near Earth Approach (NEA) asteroids are very small indeed. There are asteroids known as Trojans which share the orbits of major planets; others have very eccentric orbits which take them into the far reaches of the Solar System, and recently it has been found that there are asteroid-sized bodies near and

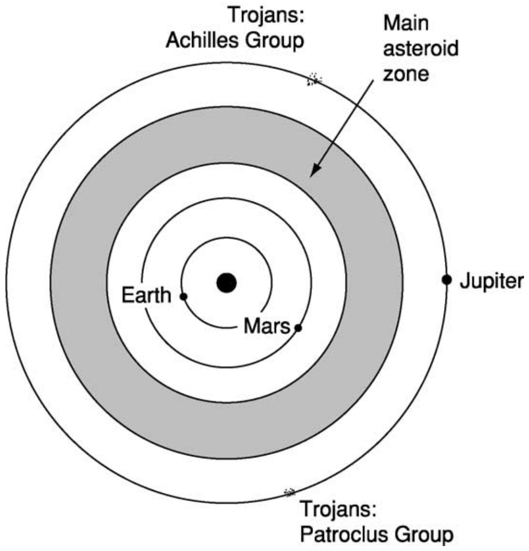


Figure 1 Distribution of asteroids.

Table 1 Some of the larger Main-Belt ateroids

Asteroid		*q*	*Q*	*Period, years*	*Orbital eccentricity*	*Orbital inclination*	*T*	*Diameter, km (max)*	*M*
1	Ceres	2.55	2.77	4.60	0.078	10.60	C	960	7.4
2	Pallas	2.12	2.77	4.62	0.234	34.80	CU	571	8.0
3	Juno	1.98	2.67	4.36	0.258	13.00	S	288	8.7
4	Vesta	2.15	2.37	3.63	0.090	7.14	V	525	6.5
5	Astrea	2.08	2.57	4.13	0.190	5.36	S	120	9.8
6	Hebe	1.94	2.43	3.77	0.202	14.79	S	204	8.3
7	Iris	1.84	2.39	5.51	0.229	5.51	S	208	7.8
8	Flora	1.86	2.20	3.27	0.156	5.89	S	162	8.7
9	Metis	2.10	2.39	3.69	0.121	5.59	S	158	9.1
10	Hygeia	2.76	3.13	5.54	0.120	3.84	C	430	10.2
72	Feronia	1.99	2.67	3.41	0.120	5.42	U	96	12.0
87	Sylvia	3.19	3.48	6.50	0.083	10.87	P	282	12.6
253	Mathilde	1.94	3.35	5.61	0.262	6.70	C	66	10.0
153	Hilda	3.10	3.97	7.91	0.142	7.83	P	222	13.3
279	Thule	4.22	4.27	8.23	0.011	8.23	D	130	15.4
511	Davida	2.61	3.18	5.66	0.177	15.93	C	324	10.5
704	Interamnia	2.61	3.06	5.36	0.148	17.30	D	338	11.0

q = perihelion distance, in astronomical units.
Q = aphelion distance, in astronomical units.
M = mean magnitude at opposition.
T = type (see Table 2).

well beyond the orbits of Neptune and Pluto. These make up what is known as the Kuiper Belt.

Discovery

A mathematical relationship, known as Bode's Law, led astronomers to believe that there should be another planet moving between the orbits of Mars and Jupiter. From 1800, a systematic search was carried out by a group of observers who called themselves the 'Celestial Police', and on 1 January 1801, the first asteroid, Ceres, was discovered by G Piazzi (who was not then a member of the group, though he joined later). Three more small bodies were found during the next few years: Pallas, Juno, and Vesta. It was not until 1845 that the next asteroid, Astræa, was found; others followed, and by now many thousands are known. When a new asteroid is discovered, it is given a temporary designation and then, when its orbit has been reliably worked out, a number. At first mythological names were used, but the supply of these names soon ran out; today the discoverer is entitled to suggest a name, which is then ratified by the International Astronomical Union.

Origin and Orbits

It is no longer thought that the asteroids are fragments of a large planet which broke up. It seems that no planet of appreciable size could form in this part of the Solar System, because of the disruptive influence of Jupiter. The asteroids in the main belt are not evenly distributed; Jupiter's gravitational pull tends to produce groups or 'families', made up of numbers of asteroids moving at around the same distance from the Sun (Figure 2). A family is named after one of its most prominent members, and does seem to be due to the disruption of a larger body, The Flora family has at least 400 members. There are also gaps in the main belt (the Kirkwood Gaps) which are almost empty, because of regular gravitational interactions with Jupiter. For instance, there is a gap at a distance of 375 million km from the Sun, where an asteroid would complete three orbits for every one orbit of Jupiter (Figure 3).

Types of Asteroids

Asteroids are divided into various types, according to their physical and surface characteristics. The main types are listed in Table 2 (Figure 4). There is certainly a link between comets and small asteroids; thus a tailed comet discovered in 1951 (Wilson-Harrington) was lost for years before being recovered in 1979 in the guise of an asteroid. It was given a number (4015) and now shows no sign of cometary activity.

Asteroid Surfaces and Composition

Details on some asteroids have been recorded. 3 Vesta has been imaged by the Hubble Space Telescope, and

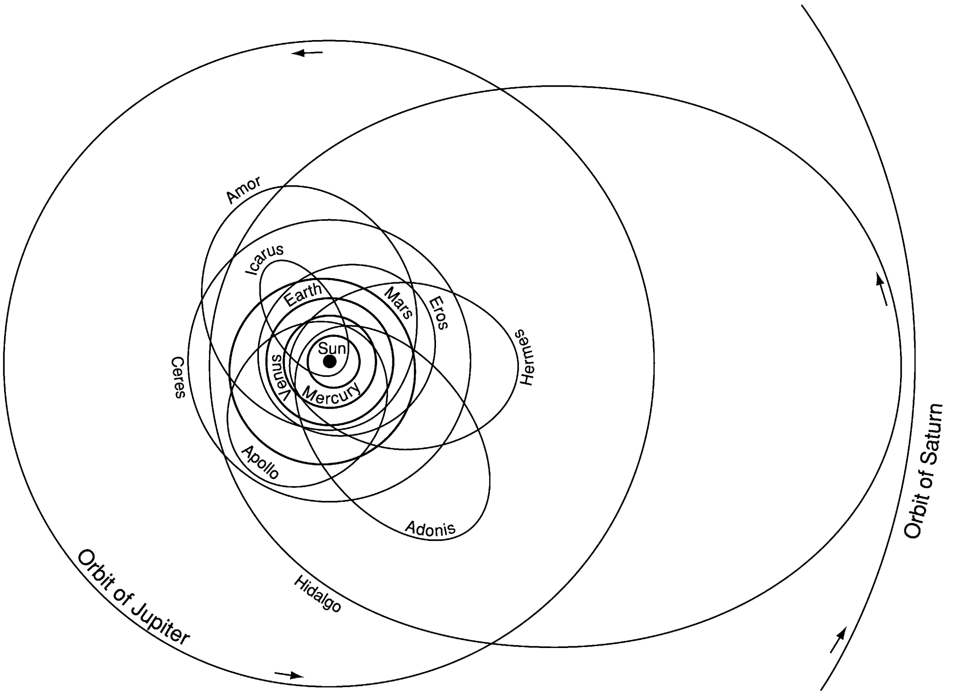


Figure 2 Orbits of some asteroids.

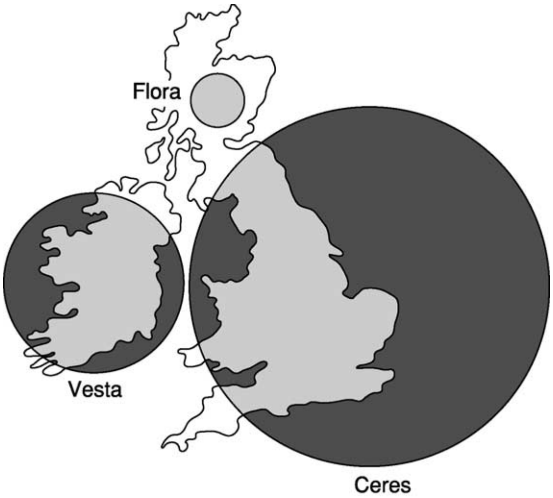


Figure 3 Sizes of same asteroids compared with British Isles Diameters: Ceres 970 km Vesta 288 × 230, Flora 204.

Table 2 Types of asteroids

C (Carbonaceous)	Most numerous, increasing in number from 10% at 2.2 a.u. up to 80% at 3 a.u. Low albedo; spectra resembles carbonaceous chondrites
S (Silicaceous)	Most numerous in the inner part of the main zone. Generally reddish, spectra resemble those of chondrites
M (Metallic)	Moderate albedoes; may be the metal-rich cores of larger parent bodies
E (Enstatite)	Relatively rare, high albedos; enstatite ($MgSiO_3$), is a major constituent
D	Low albedo; reddish; surfaces are 90% clays, with magnetite and carbon-rich substances
A	Almost pure olivine
P	Dark and reddish; not unlike Type B.
V	Igneous rock surfaces, very rare; 4 Vesta is the only large example
U	Asteroids which are regarded as unclassifiable

is geologically of great interest; there are two distinct hemispheres, covered with different types of solidified lava, and there is one huge impact crater. Some asteroids have been imaged from passing space-craft; 253 Mathilde (Figure 5) is very dark and irregular, and has been described as ‘a heap of rubble’, while 243 Ida is cratered and is accompanied by a tiny satellite, Dactyl. 216 Kleopatra has two lobes of similar size, and looks remarkably like a dog’s bone!

Asteroids closer-in than the Main Belt

These are of various types. Details are given in Table 3. All are small, usually only a few kilometres across, and are irregular in shape. The first to be discovered (in 1898) was 433 Eros; it is an Amor asteroid, so that its orbit crosses that of Mars but not that of the Earth. It can approach Earth at a distance of 23 million km. On 12 February 2001,

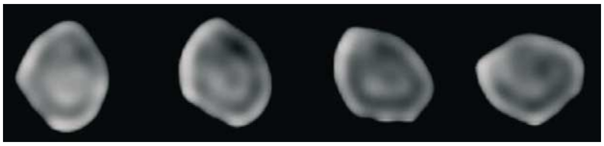


Figure 4 511 Davida, a Main Belt asteroid 320 km in diameter. This sequence of images was taken at the WM Keck Observatory on 28 December 2002, The rotation period is just over one hour; here Davida is seen from above its north pole as it spins counter-clockwise.

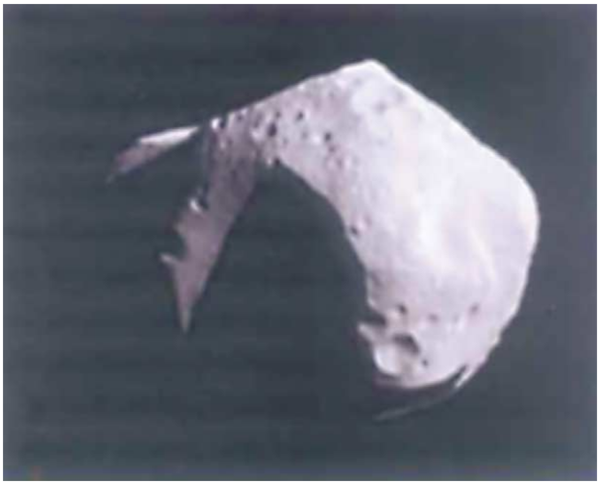


Figure 5 Asteroid 253 Mathilde, imaged by NEAR space-craft on 27 June 1997, from a range of 2400 km. There are large craters. The asteroid is very dark, with an average albedo of 4%. Mathilde's diameter is 50 × 50 × 70 km, rotation period 418 hours.

the space-craft NEAR-Shoemaker made a controlled landing on it; Eros proved to be a very primitive body, (Figure 2) and very ancient. Craters were plentiful, as well as rocks and boulders of all kinds, and superficial 'landslides' in the surface material were recognized.

Some small asteroids pass between the Earth and the Moon, and collision cannot be ruled out, and there are more potentially hazardous asteroids (PHAs) than used to be thought.

Asteroids Beyond the Main Belt

The Trojan asteroids move in the same orbit as Jupiter, though they keep either well ahead of or well behind the Giant Planet and are in no danger of being engulfed. Mars has several Trojans, and Neptune one. No true Earth Trojans are known, though 3753 Cruithne has almost the same orbital period and describes a curious sort of 'horseshoe' path with respect to the Earth. There are also asteroids, such as 944 Hidalgo and 5335 Damocles, with very eccentric orbits, very like those of comets. For example, Damocles has a period of 40.9 years; its orbit crosses those of Mars, Jupiter, Saturn, and Uranus, but is in no danger of collision as its orbital inclination is 61°. It is no more than 15 km in diameter.

Table 3 Asteroids closer-in than the Main Belt

Apohele type	Orbit entirely within that of the Earth, only one example is known, the tiny 2003 CR20
Aten type	Average distance from the Sun less than 1 a.u., though they may cross the Earth's orbit. All very small
Apollo type	Orbits cross that of the Earth; average distance from the Sun over 1 a.u.
Amor class	Orbits cross that of Mars, but not that of the Earth

The 'Centaur' asteroids remain well beyond the Main Belt; the first to be found (in 1977) was 2060 Chiron, which moves mainly between the orbits of Saturn and Uranus, in a period of 50 years. It shows traces of a coma at times, but seems much too large to be classed as a comet, even though it has been given a cometary number.

The Kuiper Belt

Many asteroidal bodies have been found near and beyond the orbits of Neptune and Pluto; the existence of such a belt was suggested by GP Kuiper (and earlier, less positively, by K. Edgeworth). Some are larger than any Main Belt asteroids; 50 000 Quaoar has a diameter of about 250 km, more than half that of Pluto. Other large Kuiper Belt objects are 28 978 Ixion (1200 km), 20 000 Varuna (900 km), and 38 093 Rhadamanthus (320 km). There are also asteroid-sized bodies which recede to immense distances from the Sun, and have orbital periods of hundreds of years. There are excellent reasons for suggesting that Pluto should be regarded as merely an exceptionally large Kuiper Belt object rather than as a bonafide planet. The Kuiper belt also includes some comets.

Comets

Comets are the most erratic members of the Solar System. They were once regarded as unlucky, and descriptions of them go back for thousands of years. Certainly a brilliant comet may look really spectacular, but by planetary standards all comets are of very low mass. They are true members of the Solar System, but in general their orbits are very eccentric, and their movements are quite unlike those of the planets.

Nature of Comets

The only fairly substantial part of a comet is the nucleus, made up of rocky fragments held together by frozen ices such as H_2O methane, carbon dioxide, and ammonia. When a comet is warmed as it approaches perihelion the rise in temperature leads to evaporation, so that the comet develops a head or coma, often with a tail or tails. Cometary tails always point away from the Sun, and are of two main types ion and dust tails. A gas or ion tail is due to particles being repelled by the solar wind, while with a dust tail the particles are driven out by the pressure of sunlight; this means that when a comet is moving outward, after perihelion, it travels tail-first. However, not all comets develop tails of any kind, and even a large comet will lose its tail when it has receded into the far part of the Solar System.

Nomenclature

Traditionally, a comet is named after its discoverer or discoverers; thus the brilliant comet seen in 1995 and 1996 was known as Hale-Bopp, since it was found independently by two American observers, Alan Hale and Tom Bopp. Occasionally the comet is known by the name of the mathematician who first computed its orbit, as with Comets Halley and Encke. There is also an official numbering system which relates to the date of discovery.

Orbits

Many comets have short periods – for example 3.3 years for Comet Encke. These short-period comets can be predicted, and some can be followed all round their orbits. Many have their aphelia near the distance of the orbit of Jupiter, making up what is termed Jupiter's comet family. Most of them are faint, and few attain naked-eye visibility. The only reasonably bright comet with a period of less than 100 years is Halley's (76 years), which last returned to perihelion in 1986–1987.

Long-period comets recede to great distances, and since their periods amount to many centuries they cannot be predicted, Hale-Bopp (Figure 6) will be back in about 2350 years, but for the next return of Comet Hyakutake, which was bright for a few weeks in 1996, we must wait for around 14 000 years. These orbits are almost parabolic, and indeed some comets are thrown into parabolic orbits after passing perihelion, so that they will never return. Arend-Rolànd, the bright comet of April 1957, is one example of this.

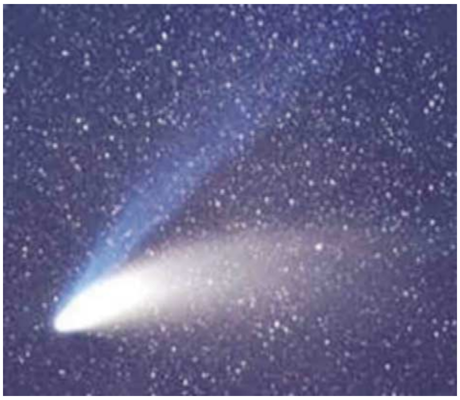


Figure 6 Comet Hale-Bopp, 1997. Note the straight ion tail, and the curved dust-tail. This was the most spectacular comet for many years.

Origin of Comets

It seems that short-period comets come from the Kuiper Belt. In general, their orbits are not highly inclined to the ecliptic, though some, notably Halley's Comet, have retrograde motion. Comets of much longer period are thought to come from the Oort Cloud, a huge spherical cloud of debris surrounding the Sun at a distance of over one light-year; its existence was suggested in 1950 by the Dutch astronomer JH Oort. It is, of course, unobservable from Earth. If an Oort Cloud comet is perturbed for any reason, it may swing in towards the Sun; it may then be perturbed into a short-period orbit, it may fall into the Sun and be destroyed, or it may simply return to the Oort Cloud. The orbital inclinations may be very high, and many long-period comets have retrograde motion.

It may be that the Oort Cloud comets were formed closer to the Sun than the Kuiper Belt objects. Low-mass objects formed near the giant planets would have been ejected by gravitational encounters. While Kuiper Belt objects, formed further out, were not so affected. Details of some notable comets are given in Table 4 (*see* **Solar System:** Meteorites).

Comets and Meteors

As a comet moves, it leaves a 'dusty trail', and if the Earth passes through such a trail we see a meteor shower. In many cases the parent comets are identifiable, for example the Orionid meteors, seen every October, come from Halley's Comet, while the August Perseids come from Comet Swift–Tuttle.

Some comets have been seen to disintegrate; thus Biela's Comet, which had a period of 6.6 years, broke in two during its return in 1846, and has not been seen since 1852, though for many years meteors appeared from the position where the comet ought to have been. Other periodical comets have been lost because their orbits have been so violently perturbed by planetary encounters. One comet, Shoemaker-Levy 9, in captured orbit around Jupiter, was destroyed in 1994 when it impacted Jupiter.

Halley's Comet

Named for Edmond Halley, who observed it in 1682 and was the first to realize that it was periodical

Table 4 Some Notable Comets

Periodical Comets

Comet		*P*	*q*	*Q*	*E*	*I*	*M*
2	Encke	3.28	0.33	2.21	0.850	11.9	11
26	Grigg-Skjellerup	5.10	0.99	2.96	0.664	6.6	12
10	Tempel 2	5.47	1.48	3.10	0.552	12.0	10
46	Wirtanen	5.46	1.07	3.10	0.657	11.7	16
9	Tempel 1	5.51	1.50	3.12	0.502	10.5	9
7	Pons-Winnecke	6.37	1.26	3.44	0.634	22.3	14
6	D'Arrest	6.51	1.35	3.49	0.614	19.5	6
21	Giacobini-Zinner	6.61	1.03	3.52	0.706	31.9	10
19	Borrelly	6.80	1.37	3.59	0.623	30.2	13
15	Finlay	6.95	1.09	3.64	0.699	3.7	13
4	Faye	7.34	1.59	3.78	0.578	9.1	8
36	Whipple	8.53	3.09	4.17	0.239	9.9	9
8	Tuttle	13.51	0.997	5.67	0.824	54.7	8
27	Crommelin	27.4	0.74	17.4	0.919	19.1	11
13	Olbers	69.6	1.18	32.6	0.930	44.6	5
1	Halley	76.0	0.59	35.3	0.967	162.2	4
109	Swift-Tuttle	135.0	0.96	51.7	0.964	113.4	4
153	Ikeya-Zhang	341	0.51	60	0.99	28.1	5

q – perihelion distance, astronomical units.
Q – aphelion distance, astronomical units.
E – orbital eccentricity.
I – orbital inclination, degrees.
M – absolute magnitude (the magnitude which the comet would have if seen from a distance of 1 a.u. from the Sun and 1 a.u. from the Earth.)
P – period, years.

(Figure 7). It was probably record by the Chinese as early as 1059 BC, and since 240 BC it has been seen at every return; it came to perihelion in 1066, and is shown in the famous Bayeux Tapestry. During the 1986 return several space–craft were sent to it, and one of these, Giotto, passed within 605 km of the nucleus. The nucleus was shaped rather like a peanut, and measured 15 × 8 × 8 km. Over 60 000 million comets of this mass would be needed to equal the Earth. The nucleus was dark-coated, and made up largely of water ice; dust-jets were active, though only from a small area on the sunward wide (Figure 8). The comet is now too faint to be detected, though it should be recovered before the next perihelion passage, due in 2061.

Great Comets

Really brilliant comets were seen fairly frequently during the nineteenth century, but were less common in the twentieth century (Figure 9). The brightest comet of near-modern times was probably that of 1843, which cast shadows and was visible in broad daylight. The last two really spectacular comets were those of 1910 – the Daylight Comet, seen shortly before Halley's – and 1965 (Ikeya–Seki), which faded quickly. Its period has been given as 880 years. Some Great comets are listed in Table 5.

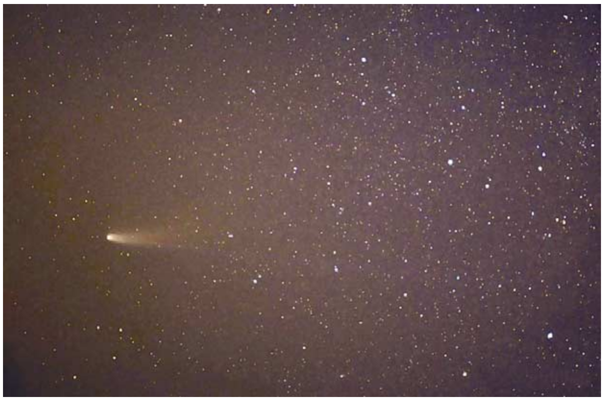


Figure 7 Halley's Comet, March 1986, (Photo by Tom Polaks with a 100 mm lens at f/2.8.) The faint globular cluster M75 is also shown. From the most left of the three conspicuous stars left and above Halley's head, go to the fainter star above and left. This star forms a fainter, nearly rectangular triangle with the other stars above and left of it. On the line connecting with the far left edge a star like spot anneals; this is M75.

Comet Hale-Bopp was not so brilliant as these, but was exceptionally beautiful, and was visible with the naked eye for over a year, from July 1996 to October 1997. It was enormous by cometary standards, with a 40 km nucleus, but unfortunately it did not come close to the Earth. There were both ion and dust tails, plus a third inconspicuous tail made up of sodium. It was last at perihelion about 4200 years ago, but planetary perturbations mean that it should return in about 2350 years, though of course all periods of this kind of length cannot be given accurately. Its orbital inclination is 89°, so that its path lies at almost a right angle to that of the Earth. During its period of visibility there were marked changes in the tails, and a spiral structure developed in the coma. Comet Ikeya-Zhang of 2002 was much less striking – it became no brighter than the fourth magnitude – but is notable because it was found to be a return of the comet of 1661, and is therefore the longest-period comet to have been seen at more than one apparition. It will be back once more in 2343.

Life in Comets?

The ‘panspermia’ theory was due to the Swedish scientist Svants Arrhenius, whose work won him the Nobel Prize for Chemistry in 1903. Arrhenius believed that life on Earth was brought here in a meteorite, but the theory never became popular, because it seemed to raise more problems that it solved. The same sort of theme has been followed up much more recently by Sir Fred Hoyle and C Wickramasinghe,

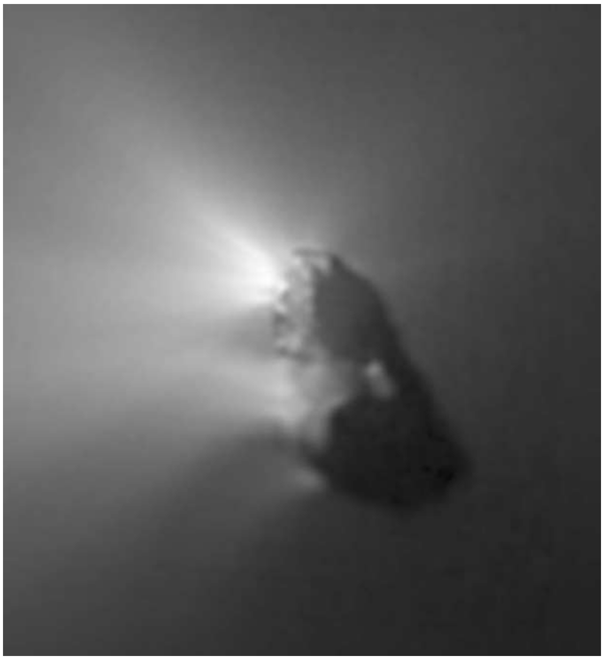


Figure 8 Head of Halley’s Comet, imaged from the Giotto space-craft. The dark coating and the active dust-jets are well seen. (Photograph from the HMC [Halley Multi-colour Camera]), Giotto passed 605 km from the nucleus on the night of, 13–14 March 1986.

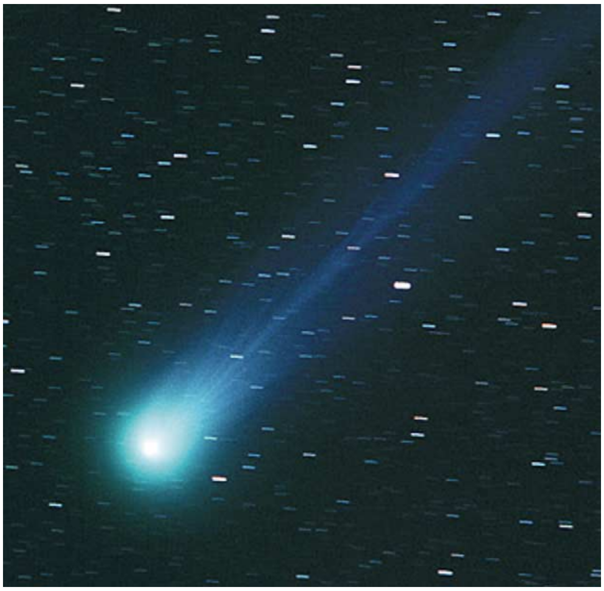


Figure 9 Comet Hyakitake, C/1996 B2. This beautiful comet was conspicuous object briefly in April–May 1996; it was obviously greenish, and had a long tail. It was in fact a small comet, but made a fairly close approach to the Earth. It will next come to perihelion in 14 000 years! time, look out for it then.

Table 5 Some Great Comets

Year	*Name*	
1744	de Chéseaux	Multi-tailed comet; max. magnitude − 7
1811	Flaugergues	Mag. 0; 24-degree ion tail. Period 3096 years
1843	Great Comet	Mag. −6. Sun-grazer. Period 517 years
1858	Donati	Mag, −1. Most beautiful of all comets, with ion and dust tails. Period 1951 years
1882	Great Comet	Reached mag. −4. Period 760 years
1910	Daylight Comet	Magnitude −4. Immensely long period
1927	Skjellerup–Maristany	Magnitude −6; 35-degree tail
1947	Southern Comet	Magnitude −5, 25-degree tail
1965	Ikeya–Seki	Magnitude −10; seen very near the Sun
1976	West	Magnitude −2. Multiple dust tail
1996	Hyakutake	Briefly reaches magnitude Q. Very long tail. The green comet
1997	Hale–Bopp	Magnitude −0.5; naked-eye object for over a year

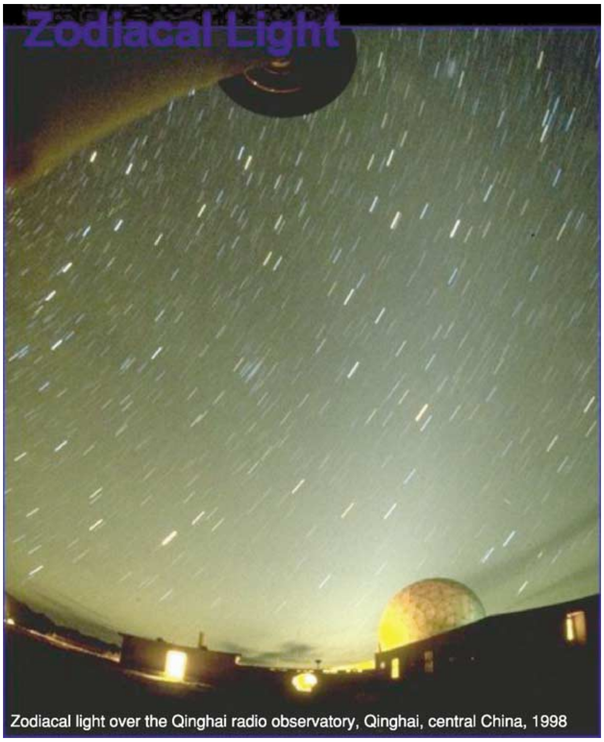

Zodiacal Light
Zodiacal light over the Qinghai radio observatory, Qinghai, central China, 1998

Figure 10 The Zodiacal Light. A typical display, photographed on 19 November 1998 over the Qinghai Radio Observatory near Delinghom Qinghai, Central China. (M Langbroek).

who claimed that comets can actually deposit harmful bacteria in the Earth's upper atmosphere, causing epidemics. Again there has been little support.

Space Dust

There is a great quantity of 'dust' in the Solar System, particularly near the main plane. It is the cause of the Zodiacal Light, (Figure 10) which may be seen as a cone of light extending upwards from the horizon for a fairly brief period either after sunset or before sunrise. Since it extends along the ecliptic, it is best seen when the ecliptic is nearly vertical to the horizon, in February to March and again in September–October. Cometary debris is a major contributary factor. It was first correctly explained by the Italian astronomer, GS Cassini, in 1683.

Another glow due to cosmic dust is the Gegenschein, seen as a faint patch exactly opposite to the Sun in the sky. It is extremely elusive, and is visible only under near-ideal conditions. The best opportunities occur when the anti-Sun position is well away from the Milky Way, from February to April and from September to November. Generally it is oval in shape, measuring about 10° by 22°, so that its maximum diameter is 40 times that of the full moon.

The Zodiacal Band is a very dim, parallel-sided band of radiance which may extend to either side of the Gegenschein, or prolonged from the apex of the Zodiacal Light Cone to join the Zodiacal Light with the Gegenschein. It also is due to sunlight being reflected from interplanetary particles near the main plane of the Solar System.

See Also

Solar System: The Sun; Meteorites; Mars; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus.

Further Reading

Bone N (1986) *Meteors*. London: Philip.

Bhandt G and Chapman D (1982) *Introduction to Comets*. Cambridge: Cambridge University Press.

Burnfam R (2000) *Great Comets*. Cambridge: Cambridge University Press.

Krishna S (1997) *Physics of Comets*. Singapore: World Scientific.

Kronk G (1988) *Comet Catalogue*. Enslow: Hillside NJ and Aldershot.

Kronk G (1988) *Meteor Showers*. Enslow: Hillside NJ and Aldershot.

Moore P (2001) *Astronomy Data Book*. London: Institute of Physics, Publishing.

Moore P (2003) *Atlas of the Universe*. London: Philip.

Norton CR (1992) *Rooks from Space*. Montana: USA Mountain Press Publishing.

Schmadel L (2002) *Dictionary of Minor Planet Names*. Berlin, Heidelberg, New York: Springer-verlag.

Kowal CT (1996) *Asteroids*. Wiley.

Whipple FL (1985) *The Mystery of Comets*. Cambridge: Cambridge University Press.

Yeomans K (1991) *Comets*. New York: Wiley Science Editions.

Meteorites

G J H McCall, Cirencester, Gloucester, UK

Introduction

Meteorites are bodies of metal or stony material mixed with metal which fall to Earth in sporadic and random arrival events, characterized by entry of a fireball or bolide streaking, often with punctuated explosive bursts, through the sky on their frictional passage through the Earth's atmosphere (Figure 1). The history of the gradual scientific acceptance of the reality of such events is followed by a brief description of the classification of various types of meteorite; the four age and time interval measurements significant for any meteorite; and the known or likely provenance in the bodies of the Solar System of the various types are then considered. After a brief mention of impact cratering and tektites, and also 'fossil' meteorites enclosed in ancient rocks, an account is given of the revolution in 'Meteoritics' (essentially an extension of geology, geochemistry, metallurgy, and physics into the realms of astronomy and planetology) during the latter half of the twentieth century. This is a result of space exploration and the recognition of hitherto unknown optimum collection regions (icebound Antarctica; the Nullarbor Plain, Australia; and other desert regions). Examples of some extensions of research into meteoritics in modern state-of-the-art science are listed.

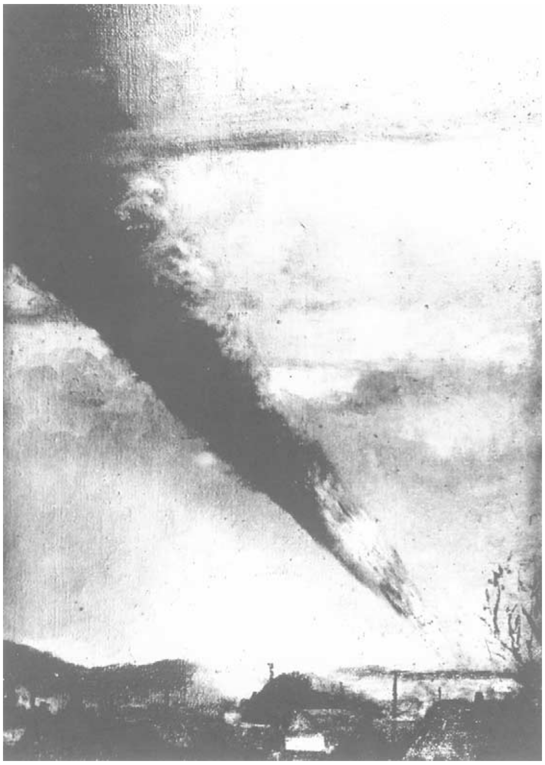


Figure 1 A painting by P.V. Medvedev of the fireball accompanying the Sikhot-Alin fall of 1949 (reproduced from McCall 1973).

Historical: the Fall of Stones and Metal from the Sky

Records of shooting stars, bright objects seen to dart across the night sky, go back to Egyptian papyrus writings of *ca*. 2000 BC and records of actual meteorites falling to Earth out of the sky go back almost as far – the fall of a black stone in the form of a cone, circular below and ending in an apex above, was reported in Phrygia about 652 BC, the familiar image of a stony meteorite such as the Middlesborough Meteorite (Figure 2) coming to us from the distant past. The Parian chronicle records falls of stones in Crete in 1478 BC and in 1168 BC of iron. In 618 BC, a fall of stones is reported to have broken

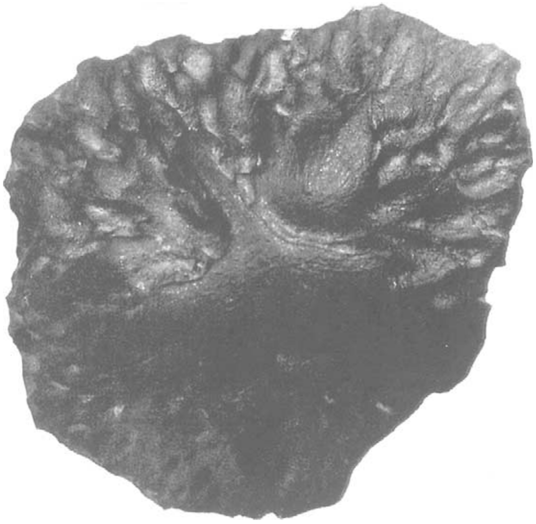


Figure 2 The Middlesborough, England, stone (fell 1881) showing the dark fusion crust and anterior surface in flight, the apex of the cone being in the direction of flight and the radiating flutings being produced by atmospheric ablation (from McCall 1973).

several chariots and killed ten men, a unique fatality. The sacred stone built into the Kaaba at Mecca is reported to have been long known prior to Islam and to have fallen from the sky. Such falls were given a religious significance, and officers of the Geological Survey in India had to go hot-foot to the site of a fall or the mass was either enshrined or broken into pieces to release evil spirits. American Indians confused later scientists by transporting masses long distances and burying them in cysts. Particularly pleasing is the custom in mediaeval France of chaining meteorites up to prevent them departing as swiftly as they arrived or from wandering at night. The earliest material from a fall preserved in western Europe is believed to be at Ensisheim, Alsace, stored in the local church since it fell in AD 1462.

Despite all these early records (and there are many more, in particular from Russia and China), scientists were slow to accept the process of rocky or metal material falling from the sky. Though there are records of the finds of irons and the falls of stones much earlier and the problem had been solved – Diogenes of Apollonia wrote “meteors are invisible stars that fall to Earth and die out, like the fiery, stony star that fall to Earth near the Egos Potamos River (in 465 BC): and natives in northern Argentina had led the conquistadors to buried masses of exotic iron, of supposed celestial origin in 1576 – scientific acceptance was widely achieved only in the last years of the eighteenth century Age of Enlightenment and the earliest years of the nineteenth century, with natives leading the explorer Pallas in Siberia to a buried stony-iron mass reputedly fallen from the sky; also falls were followed by material recovery at Wold Cottage, near Scarborough, Yorkshire and L/Aigle France.

The fall at Albareto, Italy, in 1766, had been well described by the Abbé Dominico Troili, but dismissed as the product of a subterranean explosion which hurled it high in the sky from a vent in the Earth. The stone which fell at Lucé, France in 1768, the first to be chemically analysed, was dismissed as neither from thunder, nor fallen from the sky, but as a piece of pyritiferous sandstone by a panel of august scientists! So it was the Pallas stony-iron meteorite (700 kg, ‘Krasnojarsk’), the subject, together with the Otumpa iron from South America, of a book published by E.F.F. Chladni in Riga in 1794, which really established the scientific reality of meteorite falls. Both were exotic, being found far from any known volcanic province, and by a process of elimination, he reached a single possible answer and further connected them with the phenomenon of fireball meteors. Russian scientific circles were distant from western Europe, and the English were really convinced only by the fall of a stone at Wold Cottage near Scarborough in 1795. This came into the possession of Joseph Banks, who recognized the similarity of the black fusion crust to the Siena fall material of 1794 in his possession. Edward Howard studied both and the presentation of his findings to the Royal Society in 1802–1803 convinced sceptics in England. Presentation to the Institut de France convinced several important scientists, but resistance to the idea was not finally overcome in that country until 3000 stones showered down on L’Aigle, Normandy and were described by Biot. Chladni’s work then received belated international acknowledgement, but decades would elapse before the connection with fireballs was completely established and a century before the origin of most of them through impacts between asteroids would be established.

Classification

The classification of meteorites has developed over the years and some new types and revisions of the system have inevitably arisen in the last half of the twentieth century with the prolific collection from optimum Antarctic and desert regions; despite this, the system remains workable though some revision might in time be necessary.

There are three principal classes:

Irons (Figures 3, 4)
Stony-irons (Figure 5)
Stones (Figure 2)

The latter are subdivided into (i) Chondrites, which display rounded bodies (chondrules) (Figures 6, 7),

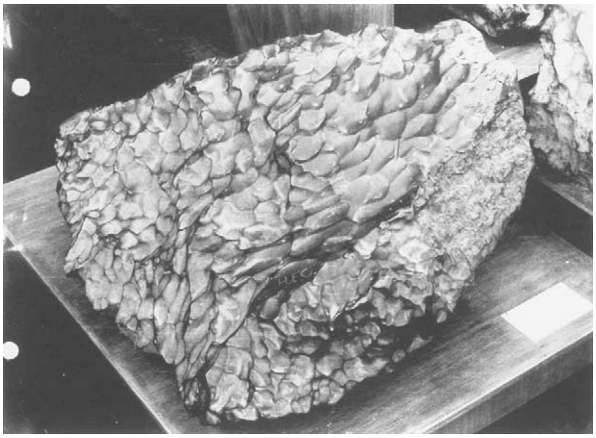


Figure 3 The Haig, Western Australia, iron (find 1951, 480 kg, III AB) with typical hackly markings on the surface.

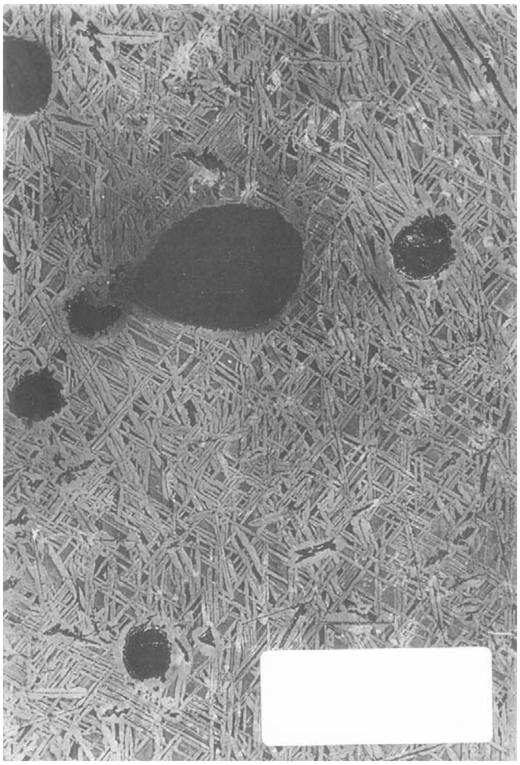


Figure 4 Cut and etched surface of the Mount Edith iron, Western Australia (find 1913, 160 kg, III AB) showing the Widmanstatten pattern and dark troilite (sulphide) nodules.

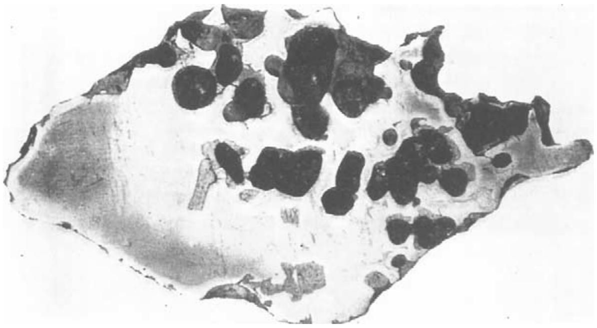


Figure 5 Cut surface of the Brenham, Kansas, pallasite stony-iron (find 1962, 22 and 9 kg), showing nickel-iron (light grey) and olivine (dark) (from McCall 1973).

believed to be relics of a very early accretionary stage in the formation of the asteroidal parent bodies (the chondrules may be wholly obliterated by recrystallization); and (ii) Achondrites, without chondrules, having textures resembling those of terrestrial igneous rocks (Figure 12). The classification used worldwide, as at 2003, is shown in Table 1a and 1b and the statistics of meteorite falls and finds in Table 2.

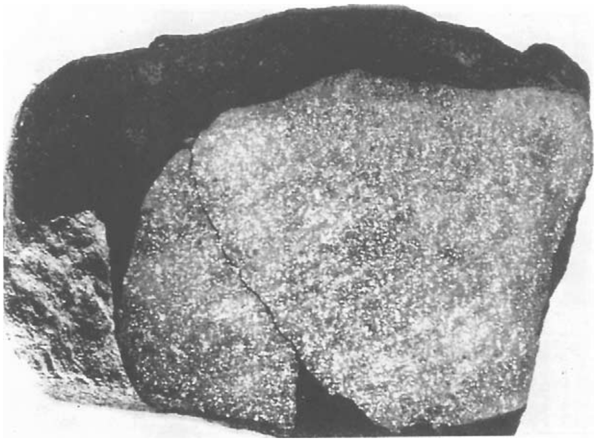


Figure 6 The Cocklebiddy, Western Australia, ordinary chondrite (fall 1949, 0.794 kg), cut face showing specks of light grey nickel-iron disseminated in a dark grey silicate matrix: the rounded chondrules are microscopic and thus not visible (from McCall 1973).

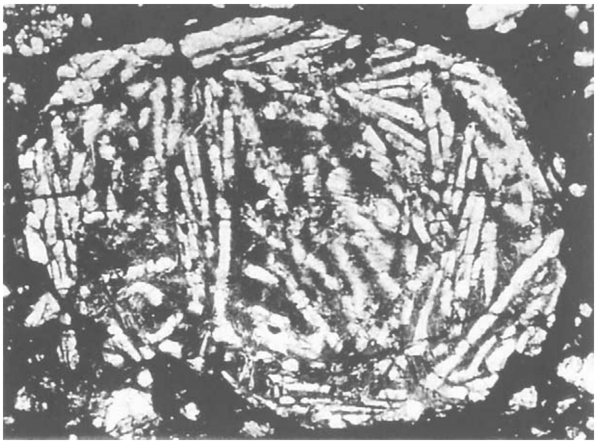


Figure 7 View in a microscope thin section across a chondrule (2 mm diameter) showing elongated olivine crystals and dark glass, within the rounded chondrule, which is set in an aggregate of olivine, pyroxene, and feldspar grains, opaque nickel-iron, sulphide and products of secondary weathering (Mulga South ordinary chondrite, Western Australia (from McCall 1973)).

Meteorites within Meteorites

Many meteorites are brecciated, probably mainly due to shock processes through collision with other meteorites in space, but some also carry other meteorite types as fragments within them. Chondrites may occur as fragments within dissimilar host chondrites. Even more spectacular are shocked eucrite achondrite bodies within the Mount Padbury stony iron (mesosiderite) and enstatite and carbonaceous and ordinary chondrite bodies within the Bencubbin stony iron meteorite, both found in Western Australia.

Table 1a Undifferentiated meteorites

Class	*Symbol*	*Example*
Carbonaceous chondrites	CI	Orgueil
	CM	Murchison
	CO	Ornans
	CV	Allende
	CK	Karounda
	CR	Renazzo
	CH	ALH 85085
Rumurutiites	R	Rumuruti
Kakangari-type chondrites	K	Kakangari
Ordinary chondrites	LL	Saint Mesmin
	L	L'Aigle
	H	Wiluna
Enstatite chondrites	EL	Eagle
	EH	Saint Sauveur

Carbonaceous chondrites: characterized by sparse to abundant chondrules set in a dark, friable matrix of carbon-rich compounds, phyllosilicates, mafic silicates, Ni-Fe metal, and glass. The letter symbols separate groups based on different mineralogy, relative abundance of different lithophile and siderophile elements, relative abundance and size of chondrules, and oxygen isotope signatures. Numerical suffixes 3, 2, and 1 denote progressive aqueous alteration and 3, 4, 5, and 6 progressive thermal alteration.

Rumurutiites: a new rare group of chondrites.

Kakangari-type chondrites: a small group of chondrites now separately defined.

Ordinary chondrites: chondrules are embedded in a finely crystalline matrix of mafic minerals, pyroxene, and olivine, together with NI-Fe metal and glass. Some are recystallised thermally and lose the definition of chondrules and the glass. The H, L, and LL groups differ in the magnesian/iron ratio in the ferromagnesian silicate minerals. The number suffixes 3–7 denote degree of thermal alteration (loss of original texture and recystallization).

Enstatite chondrites: these are chondrites with the Mg-rich pyroxene enstatite. The EL and EH groups have different relative abundances of silicates and metals. The numerical suffixes above (3–6) may be applied.

Table 1b Differentiated meteorites[*]

Class	*Symbol*	*Example*
Irons	I AB	Campo del Cielo
	I C	
	II AB	Sikhot-Alin
	II C	
	II D	
	II E	
	II F	
	III AB	Cape York
	III CD	
	III E	
	III F	
	IV A	Gibeon
	IV B	
Stony-irons	Mesosiderites	Mount Padbury
	Pallasites	Krasnojarsk
Stones (Achondrites)	Eucrites	Camel Donga
	Diogenites	Johnstown
	Howardites	Kapoeta
	Angrites	Angra dos Rios
	Ureilites	Novo Urei
	Aubrites	Aubres
	SNC Meteorites (Mars sourced?)	
	Shergottites	Shergotty
	Naklites	Nakhla
	Chassignite	Chassigny
	(Orthopyroxenite)	ALH 84001
	Basaltic and anorthositic achondrites (Lunar sourced)	ALH 85085
Primitive achondrites[*]	Brachinites	Brachina
	Winonaites	Winona

[*] The primitive achondrites have igneous textures with no chondrules, but their mineralogy and bulk chemistry shows little difference from ordinary chondrites. They are supposed to have undergone igneous processes but with no fractional crystallization, but partial melting and segregation of the phases to varying degrees.

The irons were formerly separated into *octahedrites* (kamacite plus taenite; on etching yield criss-crossing Windmanstatten patterns) (**Figure 4**): *hexahedrites* (mostly kamacite, yield only narrow thin Neumann lines on etching) and *ataxites* (no etch pattern). The Symbol classification above which replaced this metallurgical classification is still being modified and I AB and III CD have recently been grouped as I AB-III CD. These symbols reflect the differences in chemistry (nickel, gold, iridium content, etc.).

The eucrites have basaltic textures.

Many meteorites defy classification and are listed as unclassified. For example, the Bencubbin (find, Australia) meteorite appears to be a stony-iron but is in fact a mixture of four types, an iron, an enstatite achondrite, and two chondrites, one carbonaceous. It would seem to be the result of more than one collision, the first mixing occurring very early in its history (*ca.* 4500 Ma) and causing heating and melting.

Age

There are four periods of time that are significant in the history of any meteorite:

Terrestrial age: the time spent on Earth since fall. Obviously, the material from an observed fall has an immediately known terrestrial age. Cosmic-ray induced isotopes are used to obtain this age from such finds. We know from observed fall meteorites how much of these isotopes are in a meteorite when it arrives. A meteorite found later will have less isotopes because the Earth's atmosphere protected it after arrival, and unstable products of cosmic radiation, such as ^{14}C will decay, so that the difference between the normal content on arrival and that

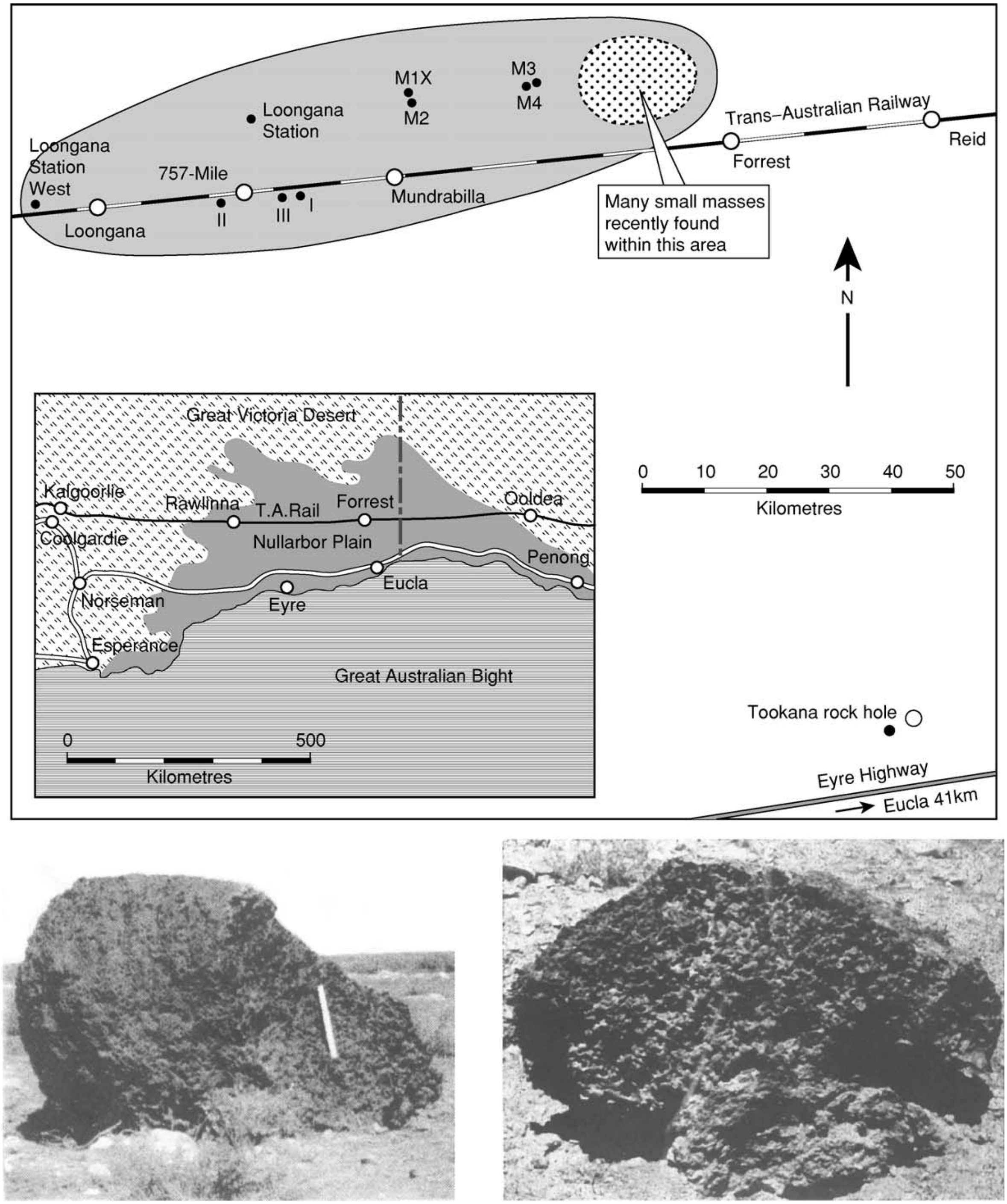


Figure 8 The distribution of the Mundrabilla irons on the Nullarbor Plain, Western Australia (rediscovered 1964 onwards) showing the typical dispersion ellipse. Below left: the M1 mass (est. 11 tonnes), as found, showing the space capsule shape with striations on anterior surface in atmosphere descent: also the curved face where the M2 mass separated. Below right, the M2 mass (est. 5 tonnes), showing the 10 cm pad of iron shale below, the product of a million years weathering by surface agents since fall (from McCall 1999, reproduced with permission of Palgrave Macmillan).

Table 2 The total of known meteorites up to the end of 1999

Class	*Fall*	*Find*	*Total*
Stones	940	20574	21514
Stony-irons	12	104	116
Irons	48	817	865
Unknown	5	7	12
Total	1005	21502	22507

(After MM Grady (2002)).

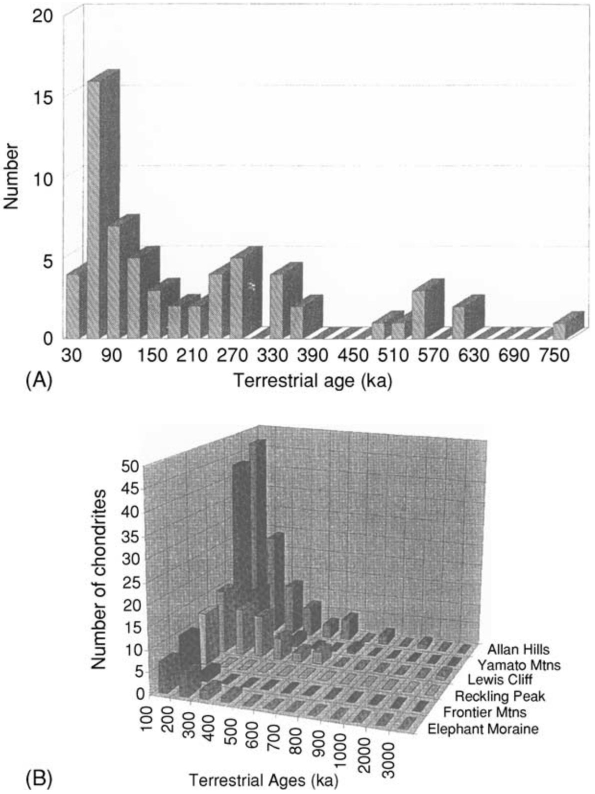


Figure 9 (A) Terrestrial age distribution for meteorites from the Allan Hills main icefield, Antarctica. (B) Terrestrial age distribution for 280 Antarctic meteorites sorted by stranding site. (A) from AJT Jull, S Cloudt, and E Cielaszyk; and (B) from ME Zolensky, in Grady *et al.* (1998). Published with the permission of the Geological Society Publishing House, Bath.

measured after the find can be used to determine the terrestrial age. As meteorites decay through natural weathering processes, these ages are usually values of tens of thousands of years, but in arid regions such as the Nullarbor Plain they are likely to be more, even a million years in the case of the large Mundrabilla iron (Figure 8); and in the Antarctic the ages taper off about 300 000 years though a very few have ages of one to three million years (Figure 9).

Cosmic ray exposure age: the time spent as a metre-scale meteoroid orbiting the Sun. Cosmic rays react with some atoms in iron or stony meteoroids and the quantity of gases formed depends on the chemical nature of the meteoroid and the duration of exposure to cosmic rays in space. The most usual measurements are of the quantity of neon gas resulting from this cosmic ray exposure. The evidence suggests that few stony meteorites survive in space without further collisional destruction and pulverisation for more than 40 million years, but iron meteorites are more robust, surviving up to 1000 million years.

Formation age: the age between the last high temperature episode in the parent body and the present. In the case of basaltic achondrites, this represents the time of crystallization from the liquid in a magma: chondrites, which have slightly greater formation ages, did not melt but were hot and recrystallised as solids soon after formation. The method involves the normal radioactive 'clocks' used by geologists, such as uranium-lead, the amount of lead produced by radioactive decay being an indicator of formation age. Values for chondrites are near to 4550 million years; some parent bodies were then heated and melted with fractional crystallization during the next 100 million years.

Formation interval: the time of the formation of the elements in stars (where almost all the elements except H and He were formed) and their incorporation in the parent body. This is done by measurement of the decay products of plutonium, an element which, because of its short half life, does not occur naturally. Plutonium was formed in a star about 150 million years before the formation of the asteroidal parent bodies of meteorites, but other elements were formed at different times.

Provenance

Asteroidal

Meteorites are nowadays accepted as fragments of strays from the asteroid belt between Mars and Jupiter. Prior to the mechanism being established of producing (due to collisions) eccentric elongated orbits for asteroids – replacing their quasi-circular orbits beyond Mars – the nucleii of comets, impoverished in volatiles by repeated passage round the Sun, were long favoured as their source, but petrological and mineralogical evidence is against this. The Farmington fall in Kansas in 1890 seems to have heralded the firm establishment of asteroidal source. Sixty reports of visual observations of this fireball, at 12.50 pm on a midsummer day and reportedly rivalling the Sun, were selected by scientists who deduced an orbit indicating that the parent asteroid was 1862 Apollo, Hermes, or 1865 Cerberus. Direct observation of fireballs by astronomers of the Sikhot-Alin, Siberia, 1949 and Pribram, Czechoslovakia, 1959 fireballs again strongly supported asteroidal sources and there have been many further supporting observations since (Figure 10). In recent years there have been numerous attempts to use optical and spectrographic methods to equate the reflectance and chemistry of asteroids with different classes of meteorites, but results seem to be inconclusive, possibly because of the operation of little understood space-weathering processes which affect the regolith surface of asteroids. Even a direct exploration mission to Eros in 2000–2001 (Figure 11) yielded no correlation and it must be borne in mind that there must be asteroids of classes never sampled by meteorites falling on the Earth. Several thousand asteroids are now known and it is estimated that there may be as many as 10 000 out there.

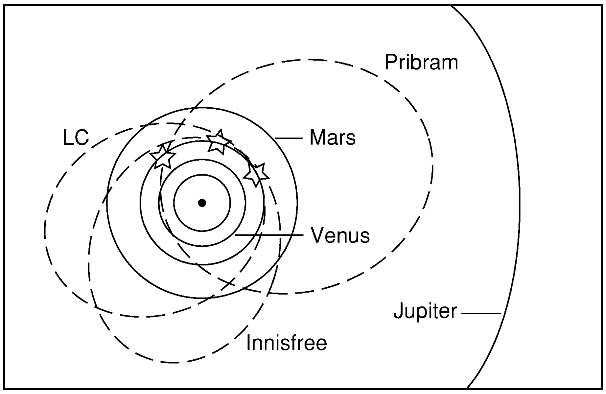


Figure 10 Orbits crossing that of the Earth derived photographically from the falls of the Pribram (Czechsoslovakia), Innisfree (Canada), and Lost City (USA) meteorites. (New figure, after Hutchison and Graham (1992).)

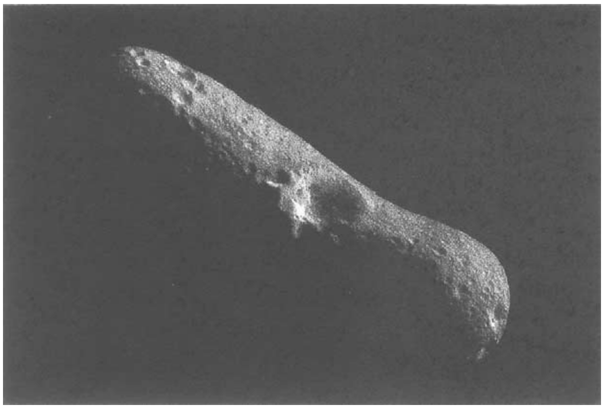


Figure 11 Asteroid 433 Eros (NEAR-Shoemaker multispectral NASA image). The large crater, Psyche, has a diameter of 5.3 km.

Even in these small parent bodies, though some did not reach 100°C, others heated to more than 1200°C, the temperature needed to form a basaltic-textured eucrite. The heat sources in these small bodies are not known for certain, but a source in extreme early heating of the Sun or internal short-lived radioactive isotopes such as ^{26}Al is favoured.

Martian Achondrites?

Some meteorites apparently do not originate from asteroids. The 'SNC' group of achondrites (Shergottites, Nakhlites, Chassignite) (Figures 12 and 13)

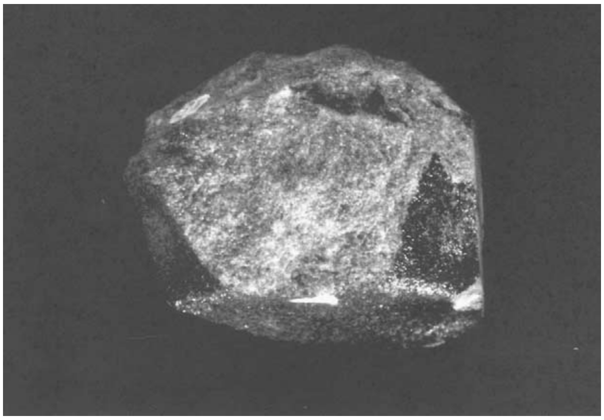


Figure 12 The Nakhla achondrite (fell 1911, Egypt, one of 40 stones, totalling 40 kg); one of the SNC (?Mars-sourced) meteorites (from McCall 1999, reproduced with permission of Palgrave Macmillan).

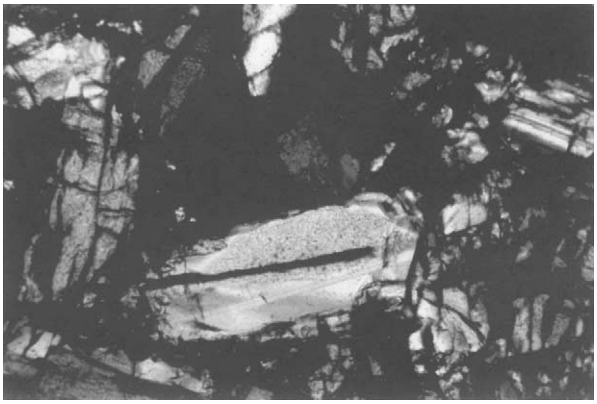


Figure 13 Thin section view of the microtexture of the Nakhla meteorite, a typical achondritic texture resembling that of terrestrial igneous rocks, formed by diopside pyroxene, olivine, and a few plagioclase crystals (×10) (from McCall 1999, reproduced with permission of Palgrave Macmillan).

were first thought to come from Mars because of the presence of oxidised iron and hydrated minerals. Later in the twentieth century, entrapped gases in these meteorites were found to be similar to the Martian atmosphere sampled by Viking missions. The ages of formation of these meteorites (see below) are not those of the asteroidal meteorites (*ca.* 4550 Ma), but fall into two groups – Nakhlites 180 and Shergottites 1300 Ma (equivalent to Earth's Jurassic and mid-Proterozoic). The widely accepted source of these meteorites is Mars – the source must surely be a planet, and the mechanism the spalling off the surface by large impacts (there are theoretical objections to volcanic ejection). However there are problems: the trapped atmosphere should be the planet's atmosphere 180 and 1300 Ma ago, not the present atmosphere, and atmosphere's change with time: also, why are the 26 SNC meteorites recovered to date all a limited range of familiar igneous rocks – Mars is a very diverse surfaced planet? A hypothetical geological history of Mars has been built up by scientists on the basis of these 26 meteorites, an edifice which direct exploration may surely demolish?

The joker in the pack is the famous ALH 84001 from Antarctica, a unique orthopyroxenite, which has a formation age similar to the asteroidal meteorites and contains the famous putative microfossils, the evidence about which seems now to favour inorganic rather than organic origin.

Lunar Achondrites

Lunar achondrite meteorites (Figure 14) so completely match lunar surface rock samples obtained by Apollo and Luna missions that there is no doubt as to their provenance. First found in Antarctica, they have been later recognized in an existing collection from Western Australia and also new finds in the Libyan desert. Volcanic ejection can be ruled out; isotopic evidence suggests that all were spalled off by geologically quite recent and relatively minor impacts on the surface of the Moon, but here there is a glaring unresolved problem. There is widespread scientific acceptance of a major impact bombardment of the Moon 3.9 Ma ago, forming innumerable and immense craters: this must have hurled vast volumes of rock out into space, sampling deep below the regolith and surficial breccia (which is all that has yet been directly sampled), there is no trace of this material in the varied log of meteorites. Where has it gone?

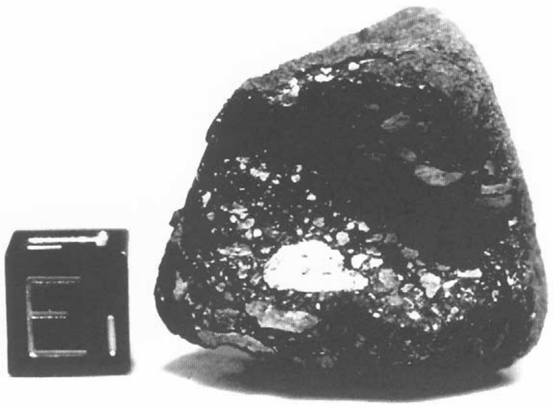


Figure 14 Lunar-sourced achondrite meteorite, ALH 81005 from Antarctica, discovered in 1981, after the first such discovery in 1979 by Japanese scientists in the Yamato Mountains. The structure is that of the lunar regolith breccias and a large white fragment of highlands anorthosite is visible. The cube has sides of 1 cm length (from McCall 1999, reproduced with permission of Palgrave Macmillan).

Cratering and Tektites

Meteorites normally land with little effect on the ground – even the 11 tonne Mundrabilla iron left no dent in the limestone surface – but multiple showers may produce small, simple craters (the 1947 Sikhot-Alin shower produced 106 associated with nickel-iron fragments). Larger masses have, in the quite recent geological past, produced kilometre-scale simple craters associated with nickel-iron (e.g., Canyon Diablo, Arizona; Wolfe Creek, Western Australia) and about 170 larger simple craters and more complex ring structures in the geological record are attributed to impact explosion processes involving larger masses, even asteroids. The largest, at 180 km diameter (Chicxulub, Yucatan, Mexico) has been associated with the Cretaceous-Tertiary boundary extinction of life (*see* **Impact Structures**). Only geochemical traces of the impactor have been discovered at such sites. Tektite showers were associated with a very small minority of such structures, but tektites are not meteorites, but are glassy objects melted from the impacted surface rocks, and spread over strewn fields at long distances from the impact sites (*see* **Tektites**).

Fossil Meteorites

The only recorded case of a meteorite being recorded in ancient rocks relates to limestone strata at a quarry near Goteborg, Sweden, where there are 12 horizons crowded with ordinary chondrite meteorites, which must have been derived from rains of stones 480 Ma ago, in the Ordovician, the stones falling onto the limey mud bottom of shallow sea. Meteorites do not fall repeatedly at the same place because of the Earth's rotation and this repetition is astonishing, as it implies repeated globally spread rains of meteorites over a period of about 1.75 million years.

The Twentieth-Century Revolution in Meteoritics

Up to the orbital flight of Sputnik heralding the space age in 1957, the study of meteorites was a quiet museum occupation. The scientific interest in meteorites then exploded because of what they might tell us about planets, satellites, and asteroids.

By coincidence, the year of Apollo XI, 1969, saw a Japanese party find nine meteorites on an area of bare ice in the Yamato Mountains. Antarctica. This 5 × 10 km area subsequently yielded 1000 meteorites. Blue ice areas and moraines in Antarctica have now yielded approximately 30 000 specimens representing some 20 000 falls. Two principle factors produce the optimum conditions for recovery: weathering is virtually nil in the arid climatic conditions and low temperatures prevailing, and the 'conveyor belt' situation on the ice sheet, snow falling and being buried and compacted to ice together with any meteorites on the surface, the snow moves coastwards and where nunataks (rocky peaks) obstruct its passage, the entrained and buried meteorites are excavated by wind action which removes the ice above (Figures 15 and 16).

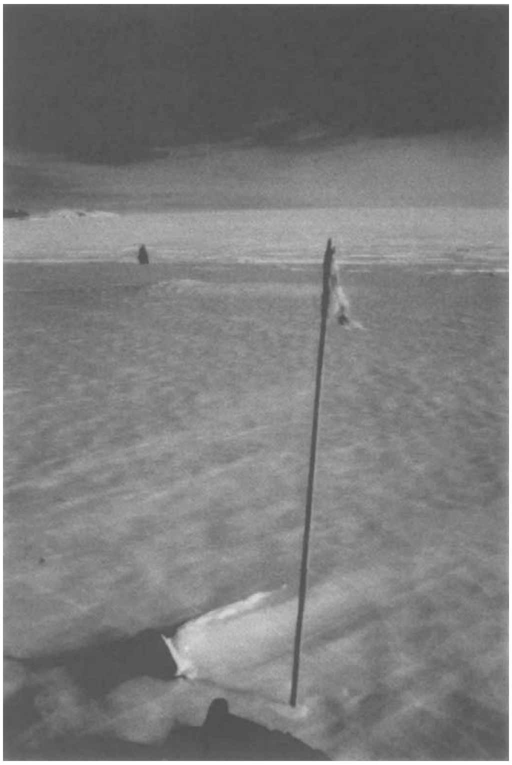


Figure 15 A meteorite as found on blue ice, its position flagged, Antarctica (from McCall 1999, reproduced with permission of Palgrave Macmillan).

By coincidence again, in the 1960s, rabbit trappers kept bringing in meteorite finds strewing the limestone surface of the arid Nullarbor Plain in Western Australia, and the writer of this entry, then working at the Western Australian Museum, wrote prophetically, "that the Nullarbor Plain must be littered with meteorites of all types". This was indeed so and systematic collection has so far yielded about 300 individual meteorites including two shower groups of more than 500 meteorite masses. Other desert areas of the world were then searched and Libya, Algeria, Morocco, and Oman have yielded several hundred finds, while desert areas in Roosevelt County, New Mexico have also proved productive. Neither Antarctica nor the desert areas are 'worked out' and many more finds will undoubtedly be made in the next years of this century. There are some desert areas in Asia, including the Gobi, that are not even searched so far, but a reconnaissance in the Gobi proved disappointing.

State-of-the-Art Research

Meteoritics is a major area of scientific research nowadays and as many as 500 scientists may attend the yearly meetings of the Meteoritical Society. Research topics are extremely varied and besides such related topics as impact processes; tektites; planetary, lunar, satellite, cometary, and asteroid exploration, topics bearing directly on meteorites may include:

- Ca-Al rich inclusions (CAIs) in meteorites, believed to be survivals from the accretion of the Solar System
- Isotope fractionation in pre-solar graphite in carbonaceous chondrites
- Isotope studies of chondrules and CAIs
- Modelling conditions for the launch-window of ?Martian meteorites
- Aqueous alteration of carbonaceous chondrites
- Presolar nano-diamonds in meteorites
- Xenon isotopes in nano-diamonds
- Trapped gases in ordinary chondrites
- Trace elements trapped in lunar meteorites

This random sample illustrates the diversity of research: the revolution in meteoritics described above has produced enough subject material to keep science busy for many decades, if not centuries, and more keeps coming in. The important point to remember that meteorites come in free of charge – they have been called the poor man's 'space probe'. Even the cost of searching after major bolide events, searching Antarctica and systematic searching of the

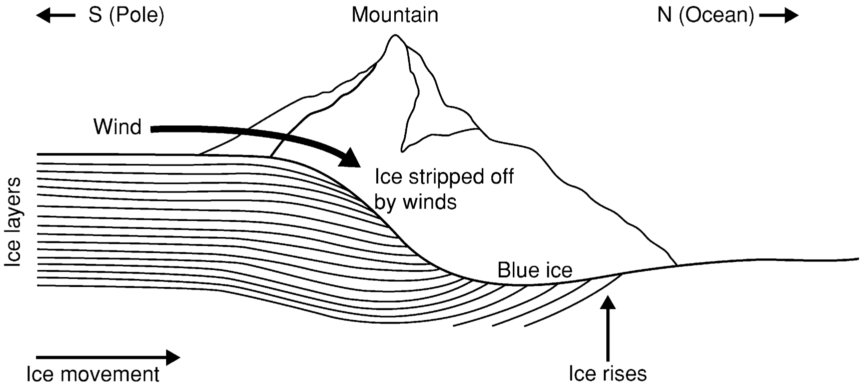


Figure 16 Diagram showing how ice, moving very slowly towards the coastal ice-front, is arrested by a rock nunatak, is stripped by wind action while stationary, to reveal entrained meteorites (new figure, after Hutchison and Graham 1992).

Nullarbor are infinitesimal when compared with the costs involved in direct space exploration.

See Also

Impact Structures. **Solar System:** Asteroids, Comets and Space Dust. **Tektites**.

Further Reading

Bevan AWR and Dehàeter JR (2002) *Meteorites: A Journey Through Space and Time*. Washington DC: Smithschian Institution Press.

Grady MM (2002) *Catalogue of Meteorites,* 5th edn. London: Natural History Museum.

Grady MM, Hutchison R, McCall GJH, and Rothery DA (1998) *Meteorites: Flux With Time and Impact Effects*, Special Publication No. 140. Bath: Geological Society Publishing House.

Hey MH (1966) *Catalogue of Meteorites,* 3rd edn. London: British Museum (Natural History).

Hutchison R and Graham A (1992) *Meteorites*. London: Natural History Museum.

Krinov EL (1960) *Principles of Meteoritics*. Oxford, London, New York, Paris: Pergamon Press.

Mason B (1962) *Meteorites*. New York, London: John Wiley & Sons.

McCall GJH (1973) *Meteorites and their Origin*. Newton-Abbot: David and Charles.

McCall GJH (1999) The Mundrabilla iron meteorite from the Nullarbor Plain, Western Australia: an update. In: Moore P (ed.) *2000 Yearbook of Astronomy,* pp. 156–166. London: Macmillan.

McCall GJH (1999) Meteoritics at the millennium. In: Moore P (ed.) *2000 Yearbook of Astronomy*, pp. 153–177. London: Macmillan.

McCall GJH (2001) *Tektites in the Geological Record*. Bath: Geological Society Publishing House.

McCall GJH and de Laeter JR (1965) *Catalogue of Western Australian Meteorite Collections*, Special Publication No. 3. Perth: Western Australian Museum.

Norton OR (2002) *The Cambridge Encyclopedia of Meteorites*. Cambridge: Cambridge University Press.

Olson RJ and Pasachoff JM (1998) *Fire in the Sky*. Cambridge: Cambridge University Press.

Zanda B and Rotaru M (2001) *Meteorites*. Cambridge: Cambridge University Press.

Mercury

G J H McCall, Cirencester, Gloucester, UK

Historical

Mercury, the closest planet to the Sun, was something of a mystery to ancient watchers of the sky, being visible to the naked eye only low down on the horizon close to sunset or sunrise – it is never seen more than 28° of arc from the Sun and is never seen against a dark sky. It was also some time before 'morning Mercury' and 'evening Mercury' where identified as the same planet. Nothing was known of its physical appearance until the advent of telescopes. Its phases and the blunting of its 'horns' (an optical effect) were then recognized (Figure 1). Johann Schroter (1745–1815) and W F Denning (1848–1931) claimed to have detected light and dark configurations, but their sketches bear no resemblance to the real surface as revealed by Mariner 10 in 1974. Denning also claimed to have detected a 25 h rotation period, now known to be erroneous. In 1953 A Dollfus recorded the presence of a tenuous atmosphere, which was later confirmed by Mariner 10, although it is even more tenuous than he supposed. The largest telescope cannot show Mercury as well as the Moon can be seen with the naked eye. Thus, accurate representation of a large part of its surface had to await Mariner 10, which reached a distance of 470 miles from the planet and transmitted images with a resolution of approximately 100 m showing a surface remarkably like that of the Moon, predominantly cratered with scarps, ridges, and plains.

'Vulcan': An Inner Neighbour Planet?

In 1958, Le Verrier received a report that a French amateur astronomer had discovered an innermost planet, closer to the Sun than Mercury. He had earlier found the movements of Mercury to suggest that such a planet existed, but in fact the anomalous movements have since been explained, and it is certain that 'Vulcan', the putative inner planet, does not exist, although some asteroids may pass closer to the Sun than Mercury on their orbits oblique to the ecliptic.

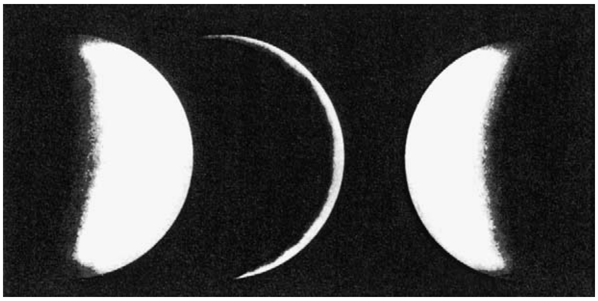


Figure 1 Phases of Mercury showing the optical effect of blunted 'horns'. Reproduced from Cross CA and Moore P (1977) *The Atlas of Mercury.* London: Mitchell Beazley Publications.

Statistics

Mercury is situated within the Solar System 57 850 000 km from the Sun. Its orbital eccentricity is 0.206, as determined by Antoniadi (1870–1943), the largest eccentricity of any planet except Pluto. It is, unlike Venus, brightest when gibbous. The equatorial diameter is 4880 km, intermediate between those of the Moon and Mars, more or less equal to that of Jupiter's satellite Callisto, and less than those of Ganymede (Jupiter) and Titan (Saturn). The escape velocity is 4.3 km s^{-1}, meaning that very little atmosphere is likely to be retained. Its density is surprising, at 5.4 g cm^{-3}; this high value compared with the Moon requires that a heavy iron-rich core takes up a higher relative proportion of the globe than in the case of the Earth. The mass of Mercury is 0.055 times that of the Earth, and its volume is 0.056 times that of the Earth. Its orbital period of 87.969 Earth days is not, as in the case of the Moon, synchronous with its rotation around its axis, which takes 58.65 Earth days. On Mercury there is no area of permanent daylight or night and no twilight zone. It has no satellite. There is a suggestion in the literature that it may have once been a satellite of Venus – the diameter ratio is not unlike that of the Earth and Moon.

Mariner 10 Mission

Technical Summary

All we know in any detail of Mercury is derived from the remarkable Mariner 10 mission, which visited both Venus and Mercury in 1974 on a gravity-assist trajectory. The mission lasted 17 months, and the same instruments were used throughout to obtain information about Earth, Moon, Venus, and Mercury – an advantage in making comparisons. There were two daylight-side encounters with Mercury as well as a night-side encounter, for orbital-change reasons, which allowed measurement of the night-surface temperature, the atmosphere, and the magnetic field. During 17 days of encounter only 17 h were spent close enough to obtain high-resolution images: 647 pictures were taken during the first daytime encounter and 300 during the second. The peculiar relation between the rotation period and the orbital period of Mercury meant that the same hemisphere was studied during both encounters. The sun rises and sets once during a Mercury day, which is two Mercury years long. During a Mercury day the planet rotates three times with respect to the stars.

Results

Mariner 10 imaged 40% of the lunar-like landscape, covering virtually a complete hemisphere. Despite the startling similarity, the high density of Mercury means that similarity to the Moon is only skin deep. In the 1970s the Moon was considered to supply a 'paradigm' for the understanding of other planets, but the high density and geochemical properties (volatiles, refractory minerals, FeO content in the crust) of Mercury revealed by Mariner 10 suggested that Mercury is the end member of an inner–outer progression of planets, whereas Mercury is anomalous. Of course, if the suggestion that Mercury is a displaced satellite of Venus is correct, then both the Moon and Mercury are anomalous. Only further missions to Mercury will answer this question.

The surface revealed by Mariner 10 has all the features of the Moon, except that it lacks extensive dark smooth plains (e.g. Imbrium) but there are quite substantial areas of lighter smooth-plain terrain, and the circular Caloris Planita feature (the largest single feature so far recognized at 1300 km in diameter) is of comparable extent to some maria and does show resemblance to lunar maria (Figure 2). The smooth-plain material does appear to have lapped over, obscured, and infilled large as on the Moon craters (Figure 3), which were formed in an older, rougher surface formation, analogous to the lunar highlands, although probably not of the same composition.

Craters dominate the entire mapped surface, and, as on the Moon, when one crater interferes with another the smaller crater is usually the intruder. Beethoven, the largest crater on Mercury, has a diameter of 625 km. Tolstoj (Figure 4), at 400 km, is about the same size as Mare Crisium on the Moon and is larger than any lunar crater. Some craters are double or have distorted circular outlines – Bach (225 km in diameter) shows both these features (Figure 5). Other craters have double walls. There are crater-sized rings outlined by annular grooves in otherwise flat plains. There are even circlets of small craters. There are prominent rayed craters like those of the Moon, which are apparently late-introduced features (e.g. Copley, Kuiper, Snori, Mena): Copley (Figure 6) is clearly later than the smooth plains; Mena (Figure 7) displays two anomalous features – one sector totally lacks rays and the rays are both curved in places and do not all emanate from a shared point focus, a characteristic seen at other rayed craters. These features are more consistent with rays being due to deposits along fracture lines than being ejection rays. Central peaks are common and may be single central peaks or off-centre single and clustered peaks.

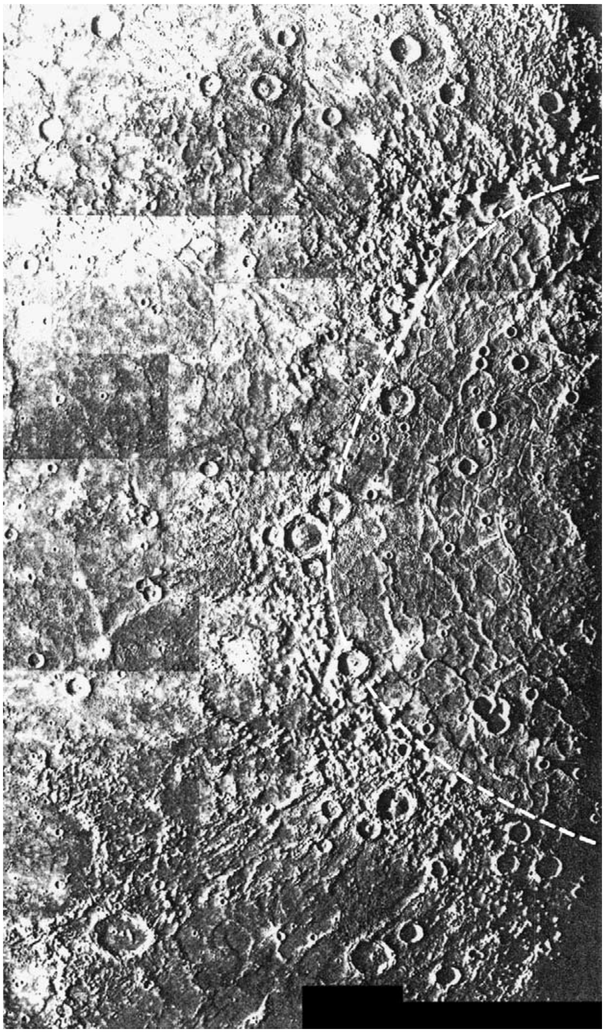


Figure 2 Caloris Planitia (dashed line) showing the concentric ridge pattern. The basin has a diameter of 1300 km. Mountain blocks at the margin rise to 1–2 km above the surrounding terrain and the peripheral linear ridge terrain extends to 100 km from the outer edge. Photograph from NASA image bank.

There are few 'Montes' on Mercury, the only such feature recorded so far being the edge of Caloris Planita. Linear scarps called 'Rupes' are, however, widespread. The albedo is different from that of the Moon – on the Moon iron-rich plain basalts and light feldspathic highland anorthosites make for a dark–light contrast, whereas the surface rocks of Mercury are all relatively light coloured because of their

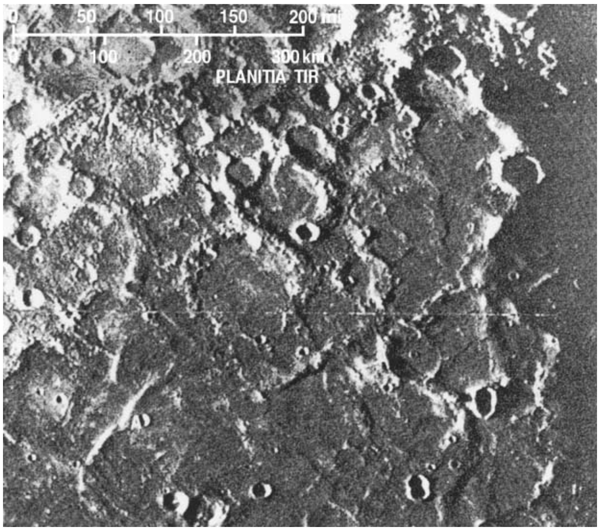


Figure 3 Part of Tir Planitia, on Mercury, showing the flooding of older large craters by smooth-plain material.

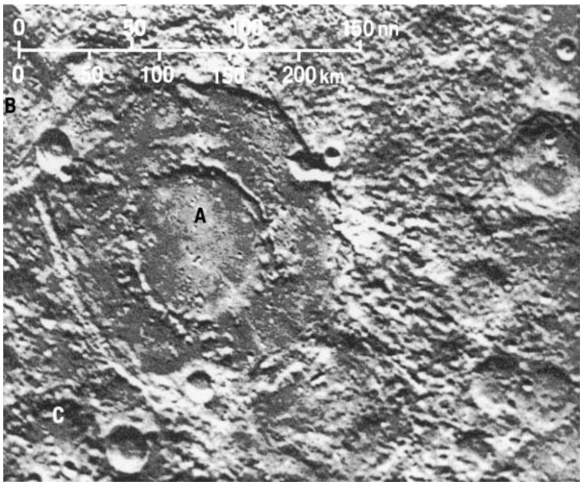


Figure 5 Bach, a double-ring crater with plain material in its floor; the shape of the outer ring is subpolygonal and one side has a wall formed by an almost straight groove.

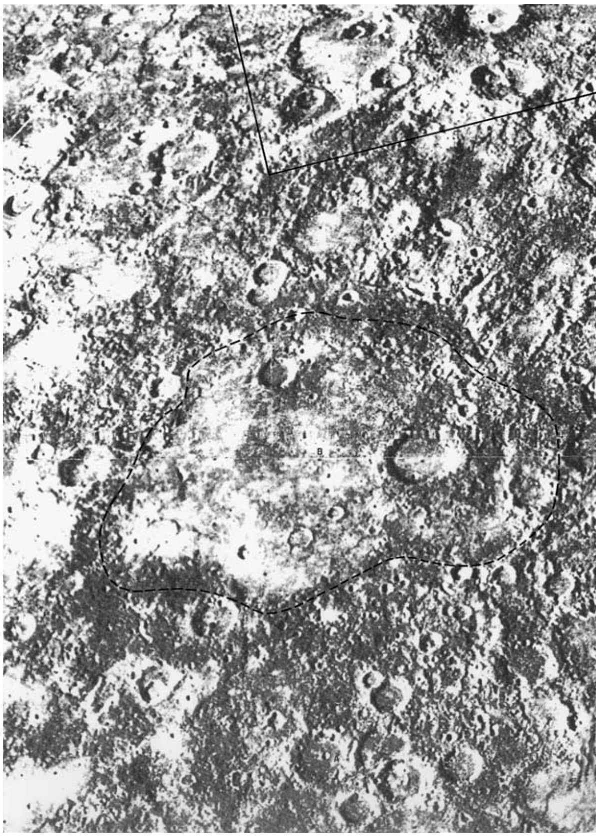


Figure 4 The large crater Tolstoj (outlined by the dashed line), which is comparable in size to the lunar Mare Crisium.

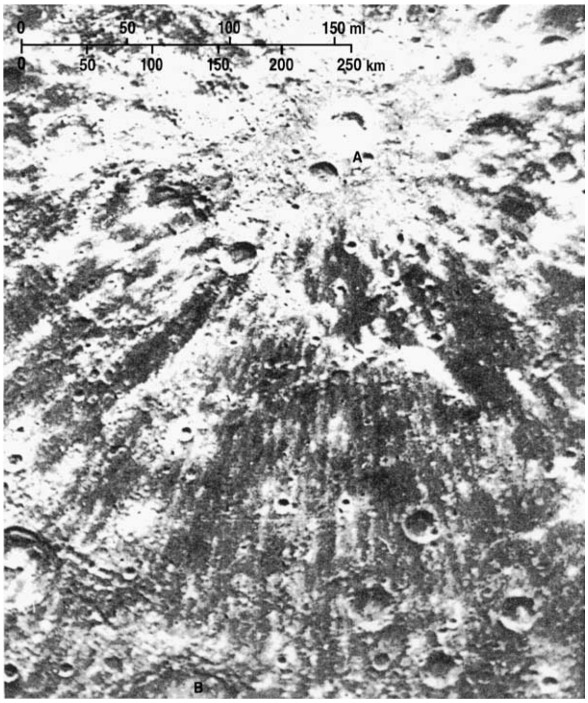


Figure 6 The rayed crater Copley, which is clearly younger than the smooth-plain material. The rays extend out into the south-east sector for 400 km. Note the irregularity and curved trace of the rays and the fact that rays overprint smooth-plain terrain.

iron-poor nature (Figure 8). At the Mercury conference in Chicago in 2001 there was a consensus that the FeO percentage in the rocks of Mercury averages around 3%. This is consistent with models in which the planet was assembled from planetesmals that were formed near the planet's current position. The magmas of Mercury may be similar in composition to the aubrites (enstatite achondrite meteorites), though these are demonstrably from their isotopic character asteroidal, not Mercurian, in provenance.

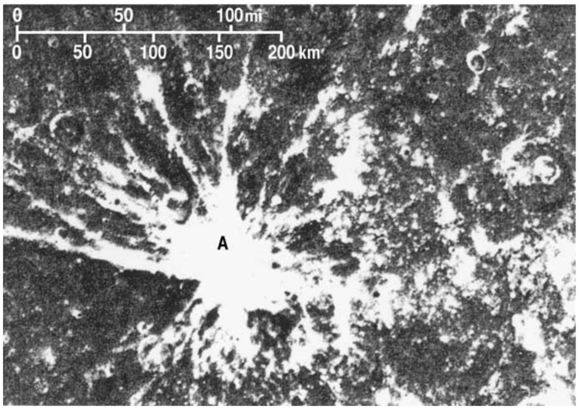


Figure 7 The rayed crater Mena, the rays of which neither emanate from a single focal point nor are straight; they extend outwards for more than 250 km. Photograph from NASA image bank.

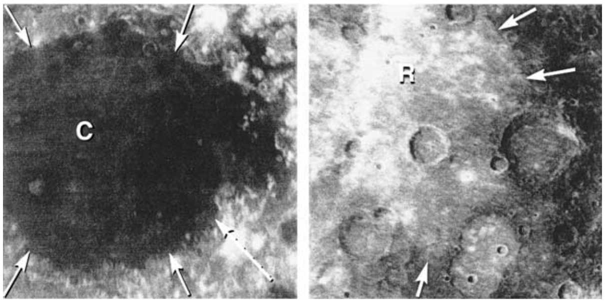


Figure 8 The contrast between the dark-floored Mare Crisium on the Moon and the similar smooth Rudaki plains of Mercury; both have embaying boundaries (arrowed). Both images were taken by Mariner 10 and are reproduced from Robinson MS and Taylor GJ (2001) Ferrous oxide in Mercury's crust and mantle. *Meteoritics and Planetary Science* 36: 842–847. © 2001 by the Meteoritical Society.

There are lobate scarps on the surface that may be due to shrinkage (thermal models predict 4–6% shrinkage), and these have been suggested to be the result of thrust faulting.

One entire side of Mercury remains to be seen, and this may be either similar to or very different from the cratered known surface.

The magnetic field was the biggest surprise revealed by Mariner 10. Though only amounting to 1% of the strength of our own planet's field, it is enough to indicate the existence of a core dynamo. Only the strength of the dipole component is at present known, but a solid inner core and liquid outer core are required by the present evidence. Convection in the outer core becomes more complex as the inner core grows. Thermal models suggest that the inner core of Mercury, if it exists, cannot be pure metal, and a non-metal such as sulphur is required to lower the crystallization temperature and density.

Volcanism on Mercury?

The impact-cratering paradigm is part and parcel of NASA's exploration and interpretation philosophy. For example, a presentation by Potts and others in 2002 made the assumption that the overall cratering results from 'bombardment time'. It remains possible, however, that many supposed impact craters, especially simple craters and some very complex structures, on surfaces of space bodies may have been too summarily dismissed as due to impact. Past volcanism is manifest on Mars, the Moon, and Venus, and there is active volcanism on Io. The widespread plains material on the surface of Mercury, as revealed by Mariner 10, though not as extensive as the larger lunar Maria, could be the result of primary volcanic flows or lobate crater ejecta. Study of the theoretical possibilities by Milkovich and others, published in 2002, indicates that widespread volcanism or no volcanism whatever or something in between are all possible. No volcanic features can be identified in the Mariner 10 images, although at the same image resolution few, if any, volcanic features would be identified on the Moon without the prior knowledge obtained by Apollo on-the-ground examination. High-resolution low-sun-angle images from Mariner 10 do show what appear to be flow fronts on Mercury; these could be volcanic lava flow fronts or ejecta flows. It seems likely that Mercury may show the same 'freezing' of the surfaces of lava flows as seen on the Moon, with lava of a basaltic (*sensu lato*) or a peculiar Mercurian petrological character preserving a plains surface among the craters and flooding some craters that were formed in the early history of the planet, perhaps around 4 Ga ago, with, as in the case of the Moon, very little having happened since then except for minor-scale impacts. This is, however, no more than informed speculation – Milkovich and others are correct in concluding that a clear assessment of the role of volcanism and whether it is primary or secondary (impact generated) must wait for better data.

Nevertheless, there are many features in the Mariner 10 database that appear to be incompatible with impact origin – for instance, the sprinkling of an area about 400 km across centred on the crater Zeami with a myriad of small craters, all virtually the same size. These have been dismissed as 'secondaries', but this explanation appears facile. The scalloped walls of the largest crater, Beethoven (Figure 9), are anomalous in an impact crater, and it lies in the centre of a cluster of similarly scalloped-walled craters that include double and triple rings. Unfortunately,

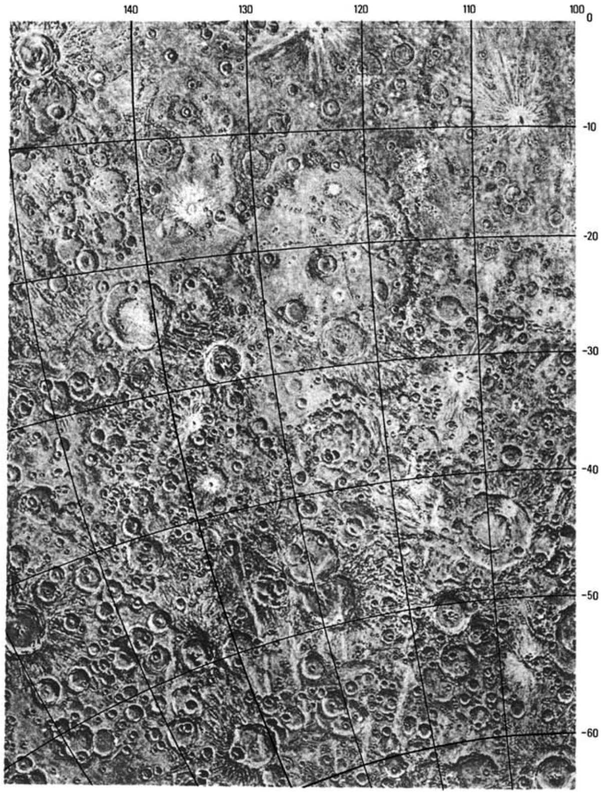


Figure 9 The area around Beethoven, the largest Mercurian crater, showing its scalloped walls and its position within a cluster of scalloped-walled craters including doubles and triples. Photograph from NASA image bank.

rigorous analysis of cratering features appears to have lapsed amongst planetologists, with the convenient assumption that impact craters can be secondarily modified in every conceivable way – and so all can be dismissed as products of the 'Great Bombardment'.

It must be concluded at the post-Mariner 10 state of knowledge that Mercury probably has had no volcanic activity, like the Moon, for nearly 4 Ga, but that the heavily cratered 'lunar-like' terrain cannot, as yet, be entirely dismissed as impact generated, and volcanism (either primary and endogenous or secondary, exogenous, and impact-generated) may have contributed significantly to the early formation of the crater dominated surface of Mercury. Future Mercury-directed missions will hopefully resolve this problem.

The success of Mariner 10

Mariner 10 told us just enough for us to realize how interesting it would be to plan return missions, more technically equipped and specifically designed, and building on Mariner 10, to answer outstanding questions, but Mercury remained the elusive planet until 2001 when the two new 'Messenger' and 'Bepi Colombo' missions were proposed.

The Future: 'Messenger' and 'Bepi Colombo'

It requires considerable energy to put a spacecraft into orbit around the innermost planet. NASA's 'Messenger' will use multiple gravity-assist encounters when it is launched in 2004 to reach Mercury in 2009. The European Space Agency and Japan Institute of Space and Astronomical Science's 'Bepi Colombo', to be launched in 2009 and arrive in 2012, will use propulsion technology, which is costlier and riskier but reduces the transit time from 5 years to 2.5 years. Messenger will study the nature of the surface, geochemistry, the Space environment, and ranging. Bepi Combo's remit is not fully worked out but will include geochemistry. It will deploy two orbiters and a lander. Thus, the nature of the unexplored side of the planet at least will be revealed when the orbiters send back the data.

The thermal environment provides the biggest challenge, with the Sun being more than 11 times as intense as on Earth and temperatures reaching up to 400°C on the sunlit side. The high temperature causes stress in equipment and may inhibit full uptake of geochemical data. The collection of mineralogical data is similarly inhibited by the blocking out of some infrared bands. Bepi Colombo will carry an actively cooled infrared spectrometer to counter this. Use of conventional solid-state detectors is impossible without power-hungry active cooling. Solar panels decay under such high temperatures. Despite these constraints, both missions will use photon (gamma, X-ray, optical) and neutron spectrometers to provide impressive geochemical information.

These two missions will research the magnetic field and its implications for the core configuration. Experiments will determine whether there is an inner core, decoupled from the rest of the planet. These experiments are important in understanding the geophysical properties of the planet and its volatile inventory, sulphur being a volatile element.

Space weathering by micro-impacts is likely to be greatly enhanced on Mercury compared with the Moon because of the flux of incoming particles close to the Sun (*ca.* X 20% has been suggested) but, because of the magnetic field, this effect may be limited to an equatorial belt. Mercury has a long comet-like sodium tail, which is probably caused by particle sputtering.

Ground-based radar observations suggest that there may be water-ice in the polar regions – high radar reflectivity suggests ice, possibly mantled by dust. There is similar evidence from craters where the surface is in 'shadow' from direct solar heating. Thermal models, however, predict that ice would not be stable there. The two missions will investigate this problem.

The tenuous atmosphere will also be the subject of investigation. Ground-based spectroscopy has detected a high sodium to potassium ratio. This is the case in the Moon's exosphere, and there it is related to the ratio found in the returned lunar-surface rocks. However, the Mercurian ratio is very large and shows diurnal variations; it appears to be related to the magnetic field rather than to the surface rocks.

Data return to Earth will also be constrained, because Mercury and the Sun interfere with radio transmission during part of the duty cycle. Antenna design is critical and constrained by weight limits and the fact that pointing antennae will tend to fail owing to the thermal cycling. Two fixed-phase array antennae on Messenger will limit the data return rate, and the same constraints will affect Bepi Colombo's planetary orbiters.

The two missions will overlap in their remits. Unfortunately, there seems to be little scope for short-term adjustment of the remit of the later Bepi Colombo mission on the basis of Messenger findings as the launch of the former in 2009 will coincide with the duty of the latter.

The orbits of both missions will, of necessity, be highly eccentric, and periapsis for Messenger will be over the north pole, so the southern hemisphere will be less well mapped. Thus, it would be ideal if Bepi Colombo had its periapsis at the south pole.

Conclusion

The idea that Mercury is a displaced satellite of Venus, though perhaps unlikely for astronomical reasons, must, unlike the 'Vulcan' concept, be taken seriously, in view of the surficial similarity to our Moon. If true, it would relegate the popular but criticized 'big planet collision' model for the origin of our Moon to the outer limits of credibility, for two such collisions of like dimensions are beyond the realms of probability.

The surface rocks of Mercury have been likened to the aubrite meteorites (enstatite achondrites) in their low FeO content (though we know from isotopic evidence that these achondrites do not come from Mercury). If the surface of Mercury was moulded by a giant bombardment as is widely supposed for the surface of the Moon, then the vast amount of internal rock material ejected into Space must be somewhere. Like the ejecta from the Moon's supposed Great Bombardment, this is as yet unrecognized among Space bodies. Because the Mercury ejecta was sent out closer to the Sun, total spiralling into the Sun is more conceivable as a reason for this absence, but it is still to be expected that some of it would have formed breccias by collision with asteroids. Once we know more about the nature of Mercurian rocks it will be possible to search for such foreign material in asteroidally sourced brecciated meteorites.

The renewed interest in Mercury is welcome, for the astonishing resemblence of its surface to that of the Moon revealed by Mariner 10 does suggest that when we know more about Mercury we will be able to extend this knowledge to the prime conundrum of the Earth–Moon system, and we may have to reject out of hand models for the Earth–Moon system that are at present widely supported (as we had to throw out the concept of lunar tektites after the Apollo and Lunar missions). Geologists would value more than anything a Luna-style retrieval and return of a rock sample from Mercury, but, alas, it appears that the technical obstacles are overwhelming. Yet there seem to be no limits to the ingenuity of Man. One thing is certain, there will never be a 'one step for mankind' on Mercury, such are the physical constraints.

See Also

Earth Structure and Origins. **Impact Structures**. **Solar System:** Meteorites; Venus; Moon; Mars; Jupiter, Saturn and Their Moons. **Volcanoes**.

Further Reading

Cross CA and Moore P (1977) *The Atlas of Mercury.* London: Mitchell Beazley Publications.

Hunten DM and Sprague AL (2002) Diurnal variation of sodium and potassium at Mercury. *Meteoritics and Planetary Science* 37: 1191–1195.

Koehn PL, Zurbuchen TL, Gloeckler G, Lundgren RA, and Fisk LA (2002) Measuring plasma environment at Mercury: the fast plasma spectrometer. *Meteoritics and Planetary Science* 37: 1173–1189.

Kracher A (2002) Mercury 2001 conference Field Museum, Chicago, 2001, October 4–5. Illinois. *Meteoritics and Planetary Science* 37: 307–309.

McCall GJH (2000) The Moon's origin: constraints on the giant impact theory. In: Moore P (ed.) *2001 Yearbook of Astronony,* pp. 212–217. London: Macmillan.

McCall GJH (2002) Back to the elusive planet. *Geoscientist* 12: 19.

Milkovich SM, Head JW, and Wilson J (2002) Identification of Mercurian volcanism. *Meteoritics and Planetary Science* 37: 1209–1222.

Moore P (1961) *Astronomy*. London: Oldbourne.

Peale SJ, Phillips RJ, Solomon SC, Smith DE, and Zuber MT (2002) A procedure for determining the nature of Mercury's core. *Meteoritics and Planetary Science* 37: 1269–1283.

Potter AE, Killen RM, and Morgan TH (2002) The sodium tail of Mercury. *Meteoritics and Planetary Science* 37: 1165–1172.

Potts LV, von Freese RRB, and Shum CK (2002) Crustal properties of Mercury by morphometric analysis of multi-ring basins on the Moon and Mars. *Meteoritics and Planetary Science* 37: 1197–1207.

Robinson MS and Taylor GJ (2001) Ferrous oxide in Mercury's crust and mantle. *Meteoritics and Planetary Science* 36: 842–847.

Sprague AL, Emery JP, Donaldson KL, *et al.* (2002) Mercury: mid-infra-red (3–13.5 μm) observations show heterogeneous composition, presence of intermediate and basic soil types, and pyroxene. *Meteoritics and Planetary Science* 37: 1255–1268.

Venus

M A Ivanov, Russian Academy of Sciences, Moscow, Russia
J W Head, Brown University, Providence, RI, USA

Introduction

Venus, similar to Earth in many ways, also shows many differences and provides insight into different paths of evolution that can be taken by Earth-like planets. The atmosphere of Venus is predominantly carbon dioxide; surface temperatures exceed the melting point of lead and surface pressures are almost 100 times that of Earth's atmosphere. The crater retention age of the surface of Venus is very young geologically, similar to that of the Earth; however, plate tectonics does not seem to be recycling the crust and lithosphere at present. The surface is dominated by regional volcanic activity and vertical crustal accretion, and regional tectonism appears to have been much more pervasive in the earliest part of the preserved stratigraphic record, dating from less than a billion years ago. The characteristics and distribution of superposed impact craters suggest that a major resurfacing event, perhaps catastrophic in nature, occurred on Venus in its relatively recent geological history. Venus may thus be characterized by relatively recent episodic heat loss, rather than the more monotonic loss thought to be typical of the other Earth-like planets. Despite the fact that the majority of the preserved geological record on Venus dates from the last ~20% of its history, Venus may provide insight into processes, such as the formation of continents, that operated in the first half of Earth history.

Venus is the second largest terrestrial planet by size after Earth, and in major characteristics is close to our planet: The radius of Venus is 6051.8 km (0.95 of Earth's radius), its mass is 4.87×10^{27} g (0.81 of Earth's mass), bulk density is $5.24\,\mathrm{g\,cm^{-3}}$ (0.95 of Earth's density), and surface gravity is $8.87\,\mathrm{m\,s^{-2}}$ (0.91 of Earth's gravity). For decades, Venus was considered as a 'twin' planet to Earth. Current knowledge of Venus geology is derived from several interplanetary missions, including landers and orbiters, as well as Earth-based observations. In the mid-1970s, the former Soviet Union conducted a series of successful landings; the Soviet landers transmitted panoramas of the surface of Venus, in addition to data on the near-surface environment, on surface rocks, and the chemistry of the atmosphere. The Pioneer Venus was the first American orbiter of Venus; launched by the United States in 1978, the Pioneer Venus collected data on global topography and gravity. The fundamental findings of this mission were that the global Venus hypsogram, in contrast to that of Earth, is characterized by one peak corresponding to the mean planetary radius (MPR), about 6051 km (Figure 1A), and that the gravity and topography of Venus are highly correlated. Three major topographic provinces characterize the surface of Venus (Figure 1B): lowlands (below MPR, ~11% of the surface), midlands (0–2 km above MPR, ~80% of the surface), and highlands (>2 km above MPR, ~9% of the surface). The spatial resolution of the Pioneer Venus imaging radar was too low to describe morphology of the surface in detail.

The systematic photogeological study of Venus began when high-resolution radar images were collected by the Soviet Venera-15/16 orbiters and by Earth-based radar observations from Arecibo Observatory. At a resolution of ~1–2 km, Venera-15/16 mapped the surface in the northern hemisphere above ~30° N; images from the Arecibo telescope covered a large area between 65° S–65° N and 270° E–30° E. In the early 1990s, the United States Magellan orbiter provided almost complete coverage (~97% of the surface) of Venus, providing high-resolution images (100–200 m) and medium-resolution altimetry

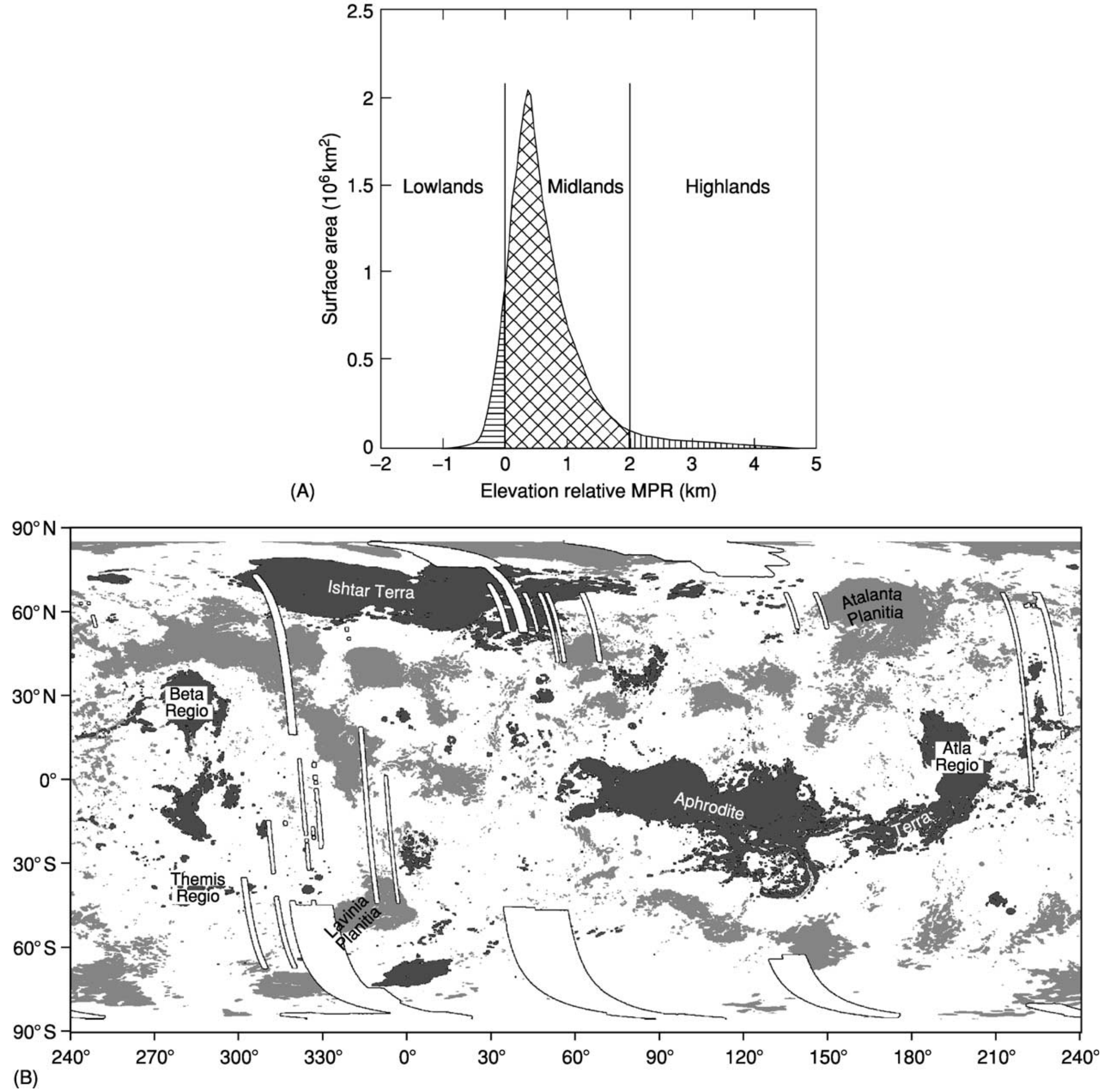


Figure 1 Characteristics of the global altimetry of Venus. (A). The Venus hypsogram (the fraction of surface area vs. elevation) has one peak, implying the absence of the surface elevation dichotomy that characterizes the distribution of the surface elevation on Earth (high-standing continents and low-lying ocean floor). The mean planetary radius (MPR) of Venus is 6051.84 km. (B) The map in simple cylindrical projection, showing the areal distribution of the three major topographic provinces of Venus. Lowlands (light grey, below 0 km), midlands (white, 0–2 km), and highlands (dark grey, above 2 km). The majority of the surface is within the midlands.

(~20-km footprint). Magellan also collected data on the Venus gravity field.

Surface Conditions and Rock Composition

In contrast to Earth, a very dense atmosphere (the pressure at the surface is ~95 bar) consisting primarily of CO_2 (Table 1) blankets Venus. The relative role of three major contributors to the atmosphere, primordial nebular material, volcanic outgassing, and influx of volatiles by comets, in the formation of the present atmosphere is an open question. Although the current atmosphere is very dry, a minute quantity of water is still detectable. An important feature of water in the Venusian atmosphere is that the deuterium/hydrogen (D/H) ratio is 150 ± 30 times higher than is found in terrestrial water. If water on Earth represents a sample of primitive water on Venus, the Venusian D/H ratio suggests that, depending on the escape flux of hydrogen and deuterium, originally Venus had 260 to 7700 times the current amount of water. Such an amount is equivalent to a global layer of water 4 to 115 m deep.

The dense and dry atmosphere on Venus results in a strong greenhouse effect that governs the conditions on the surface, leading to very high near-surface temperatures (~740 K) and equalizing the temperatures over the planet. Important consequences are the absence of both free water and water erosion, along with significantly reduced wind activity. Thus, the principal factors resurfacing Venus are volcanism and tectonics. Volcanism is the main mechanism driving the growth of the Venusian crust. The chemical compositions of the surface rocks have been measured at seven landing sites (Tables 2 and 3). The rock chemistry correlates with the compositions of terrestrial basalts, suggesting that volcanism on Venus is mostly basaltic.

Surface Population of Impact Craters

A study of the spatial density and distribution of impact craters is the principal means of understanding the age of the surface and the history of resurfacing of planetary bodies. There are 968 impact craters on the Venusian surface, making the mean crater density ~2 craters per 10^6 km^2. This value implies that the surface of Venus is relatively young; depending on the models of the flux of projectiles crossing the orbit of Venus, the surface is estimated to be from 300 to 750 My old.

A discovery of fundamental importance is that Venus lacks the densely cratered terrain characterizing significant portions of the surface of the Moon, Mars, and Mercury (Figure 2). Thus, Venus does not display the surface age dichotomy typical of the smaller terrestrial planets. Detailed study of the spatial distribution of surface craters, either by Monte Carlo simulations or by near-neighbor analyses, has shown that the distribution of impact craters is very close to a completely spatially random distribution (Figure 2). Another important characteristic of impact craters on Venus is their state of preservation. A global survey of craters has revealed that a majority appear to be morphologically pristine (Figure 3A), with only a small percentage either embayed by volcanic flows (about 2.5%; Figure 3B) or deformed tectonically (about 12%; Figure 3C). The small total number of impact craters on Venus, their predominantly pristine morphology, and their almost completely random spatial distribution may be explained by two alternative end-member models. In the first model, the catastrophic resurfacing model, the craters on Venus are considered to belong to a production population, with the planet now in the stage of accumulation of craters. This model proposes that the observable geological history of Venus (the last ~10–15% of the total history) began after a major planet-wide and relatively short ‘catastrophic’ episode of resurfacing that reset the crater record. The present crater population is thought to be accumulating since that time, in an environment of sluggish endogenous activity. Alternatively, the equilibrium-resurfacing model proposes that the actual crater population is in equilibrium with the current volcanic and tectonic activity that occurs within small areas ~400 km across. These two models imply fundamentally different histories and modes of geology on Venus. In the catastrophic model, the geological history is considered to be episodic, whereas the equilibrium-resurfacing model implies a steady-state (more Earth-like) history.

Table 1 Composition of the atmosphere of Venus

Atmosphere component	*Elevation above the surface (km)*		*Content*
	From	*To*	
CO_2	1.5	63	97 ± 4 vol. %
N_2	1.5	63	1.35–4.5 vol. %
H_2O	45	54	500–10000 ppm
H_2O	25	45	~500 ppm
H_2O	0	15	~20 ppm
O_2	0	42	<20 ppm
CO	0	42	28 ± 14 ppm
SO_2	0	42	130 ± 60 ppm
Ar	1.5	24	~100 ppm

Table 2 Contents of radiogenic elements in rocks on the surface of Venus[a]

Component	*Lander*				
	Venera 8	*Venera 9*	*Venera 10*	*Vega 1*	*Vega 2*
K_2O (wt.%)	4.8 ± 1.4	0.6 ± 0.1	0.4 ± 0.2	0.54 ± 0.26	0.48 ± 0.24
U (ppm)	2.2 ± 0.7	0.6 ± 0.2	0.5 ± 0.3	0.64 ± 0.47	0.68 ± 0.38
Th (ppm)	6.5 ± 0.2	3.7 ± 0.4	0.7 ± 0.3	1.5 ± 1.2	2.0 ± 1.0

[a] Determined by gamma spectroscopy during lander missions.

Major Surface Tectonic, Volcanic, and Volcano-Tectonic Features

Image and altimetry data reveal a rich array of volcanic and tectonic features on Venus, implying a complex geological history. In the practical absence of erosion on Venus, the features shaping its surface are well preserved and directly portray the volcanic and tectonic processes that have formed them. The most important terrains, tectonic structures, and volcanic material units making up the surface of Venus are listed in Table 4.

Table 3 Contents of major petrogenic elements in rocks on the surface of Venus[a]

	Lander		
Component (wt.%)	*Venera 13*	*Venera 14*	*Vega 2*
SiO_2	45.1 ± 3.0	48.7 ± 3.6	45.6 ± 3.2
TiO_2	1.59 ± 0.45	1.25 ± 0.41	0.2 ± 0.1
Al_2O_3	15.8 ± 3.0	17.9 ± 2.6	16.0 ± 1.8
FeO total	9.3 ± 2.2	8.8 ± 1.8	7.7 ± 1.1
MnO	0.2 ± 0.1	0.16 ± 0.08	0.14 ± 0.12
MgO	11.4 ± 6.2	8.1 ± 3.3	11.5 ± 3.7
CaO	7.1 ± 0.96	10.3 ± 1.2	7.5 ± 0.7
K_2O	4.0 ± 0.63	0.2 ± 0.07	0.1 ± 0.08
S	0.65 ± 0.4	0.35 ± 0.31	1.9 ± 0.6
Cl	0.3	0.4	0.3

[a] Determined by X-ray fluorescence during lander missions. Note: sodium is not analyzed by the X-ray fluorescence technique.

Major Tectonic Features

Tessera terrain The tessera terrain (Figure 4) is among the most deformed regions on Venus. Tesserae are defined as radar-bright, topographically elevated, equidimensional or elongated areas that are complexly deformed by at least two sets of coupled contractional (ridges) and extensional (grooves) tectonic structures. Tesserae cover ~8% of the surface of Venus; the size of individual tessera occurrences varies from a few hundred kilometres up to several thousand kilometres. The largest tesserae occur on the surface of major plateau-shaped highlands of Venus (Figure 1B). The smaller tesserae are within the midlands and only a few small fragments of a tessera occur on the floor of large lowland basins. Tessera terrain is concentrated in the equatorial zone and at high northern latitudes; there is a distinct paucity of tessera terrain south of about 30° S (Figure 5), where the vast low-lying plains dominate the surface of Venus. The elevated regions where the major rift zones are associated with coronae and large volcanoes represent another type of territory where

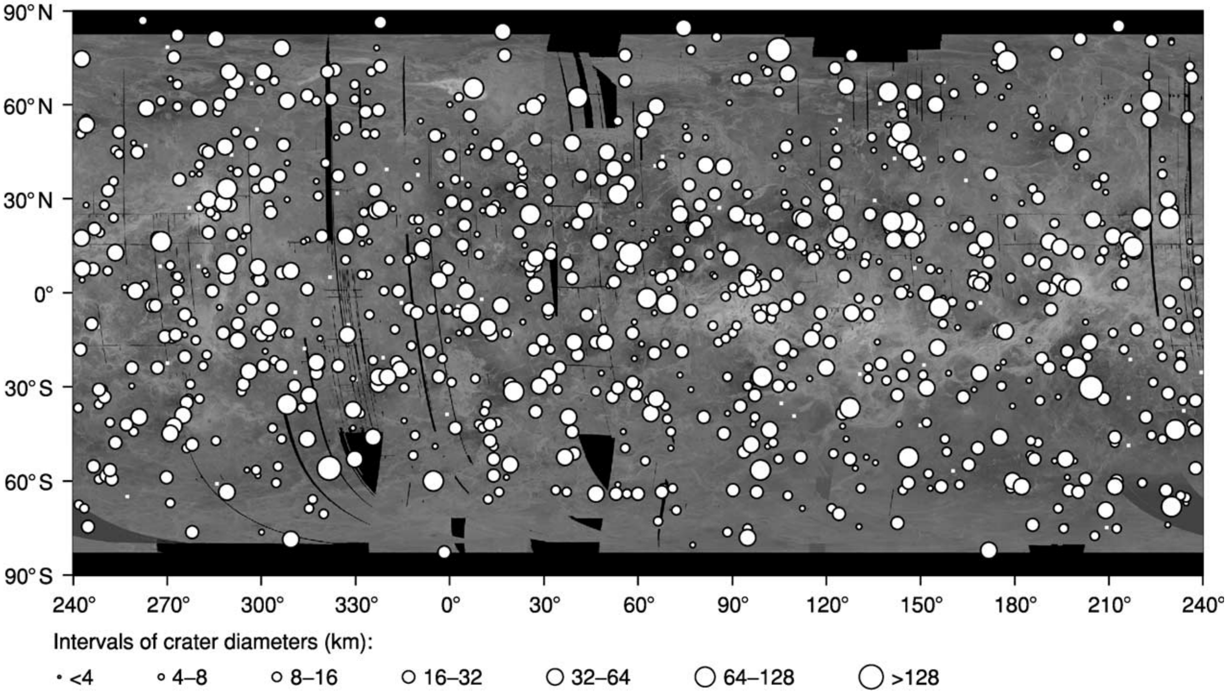


Figure 2 The spatial distribution of impact craters (white dots) on the surface of Venus is very close to a spatially random distribution. There is no evidence for the bimodal distribution of crater density characterizing the surface of Mercury, Moon, and Mars, where the heavily cratered terrains are in contact with lightly cratered plains. Crater size on Venus corresponds to the different intervals of diameters indicated at the bottom of the map. The map is in simple cylindrical projection.

tessera terrain is rare or absent. The surrounding plains embay practically all tessera massifs, meaning that the tessera is the oldest recognizable unit on Venus.

Densely fractured areas On the surface of Venus small areas are heavily dissected by numerous subparallel, densely packed, narrow and short lineaments (Figure 6). The lineaments that are wide enough to be imaged appear as fractures and imply a tectonic environment of extension during their formation. The dense fractures typically deform low-relief areas with an overall smooth surface, suggesting that these were initially lava plains. Densely lineated plains predominantly occur within midlands and there are almost no outcrops of the unit within vast areas of regional plains in lowland basins. The densely fractured areas and lineated plains occupy a small percentage of the Venusian surface and typically form small (tens of kilometres across) equidimensional, elongated, and arc-like occurrences (Figure 7) that sometimes form rims of coronae and corona-like features. Where the plains and tesserae are in contact, there is evidence for embayment of the tesserae by the material of densely lineated plains (Figure 8), suggesting that the plains are younger. Other plains units embay occurrences of densely lineated plains, which means that this unit represents one of the oldest terrains on Venus.

Ridge belts and ridged and grooved plains In some areas on Venus there are distinct belts consisting of swarms of contractional ridges (Figure 9). The linear and curvilinear ridges, 10–15 km wide and several tens of kilometres long, have smooth surfaces, rounded hinges, and appear to be symmetrical in cross-section. The ridge belts and individual ridges deform ridged and grooved plains and gradually merge with the surface of these plains, suggesting that they are tectonic facies of the same material unit. The ridged and grooved plains and ridge belts comprise less than 5% of the surface of Venus and

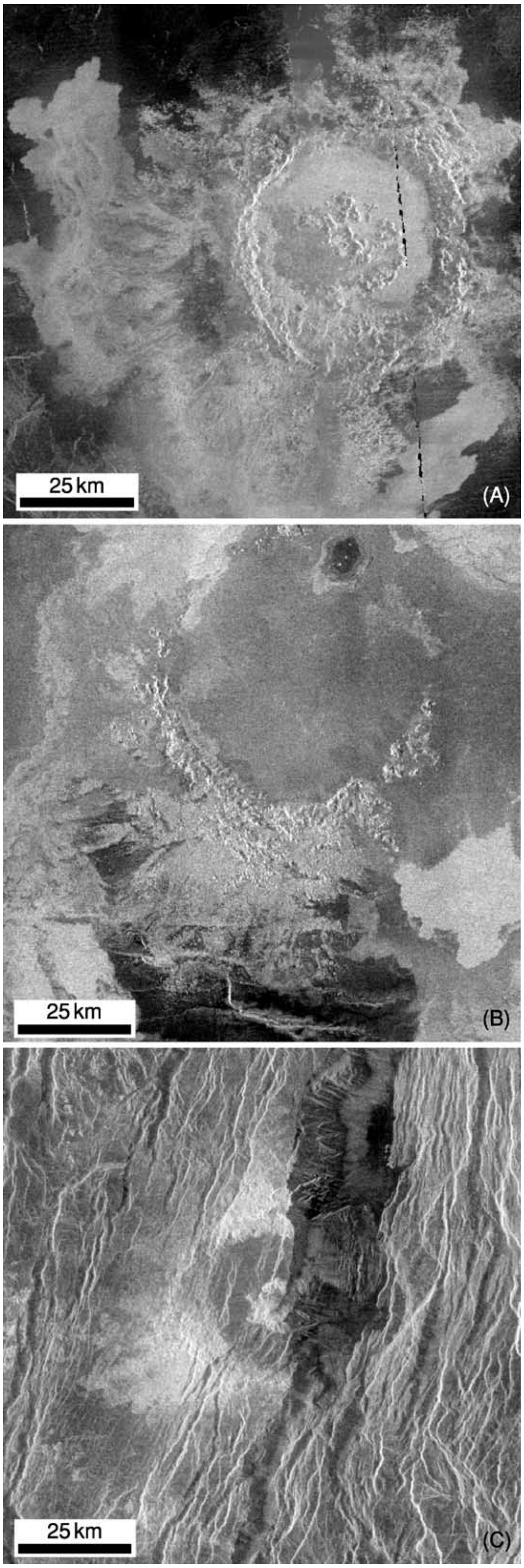


Figure 3 Morphology of impact craters on Venus. (A) Impact crater Caldwell (23.6° N, 112.4° E, 51 km) is characterized by a pristine morphology, a radar-bright floor, and extensive outflows (south and west of the crater) superposed on surrounding plains. Fragment of C1-MIDR.30N117; illumination is from the right. (B) Impact crater Gautier (26.3° N, 42.8° E, 59.3 km) is heavily embayed by volcanic flows; only the southern half of the crater rim is visible. Dark material at the bottom of the image is regional wrinkle-ridged plains on which ejecta from the crater are superposed. Younger lobate plains embay both the crater materials and the regional plains. Fragment of C1-MIDR.30N045; illumination is from the left. (C) Impact crater Balch (29.9° N, 282.9° E, 40 km) is severely deformed by tectonic structures of the Devana Chasma rift zone in Beta Regio. Only the western half of the crater is visible (center of the image); the eastern half is almost completely destroyed. A small fragment of the easternmost part of the rim is visible among individual graben of the rift zone. The position of the eastern rim suggests horizontal extension across the rift. Fragment of C1-MIDR.30N279; illumination is from the left.

Table 4 Classification of the major features and terrains on Venus

				Topographic characteristics			
Feature/terrain	*Shape of occurrences*	*Dimensions*	*Areal distribution*	*Background*	*Relief*	*Associated tectonics*	*Associated volcanism*
Regional tectonic features							
Tessera	Equidimensional, elongated, irregular shape	A few 100s up to ~1000s of km across	Megaclusters, arcs; mostly in northern hemisphere	Midlands and highlands; rare in lowlands	Major regions are a few km high	Crisscrossing ridges and grooves	No evidence
Areas of dense fractures	Equidimensional, elongated, arc-like	10s to a few 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Little relief	Dense narrow and short fractures	No evidence for contemporaneous volcanism
Ridge belts	Compact, elongated belts	100s to 1000s of km long, 10s of km wide	Major area is in a fan-shaped area in the northern hemisphere	Midlands and lowlands; rare in highlands	Rises, several 100s of m	Linear ridges a few km wide and tens of km long	Associated with specific volcanic plains
Fracture belts	Belts and arc- and star-like occurrences	100s to 1000s of km long, 10s of km wide	Belts, swarms in both hemispheres	Midlands to lowlands	Rises (100s of m high) in places; central valley	Dense fractures and graben up to 100s of km long	In places, evidence for contemporaneous volcanism
Rift zones	Elongated zones	1000s of km long, 100s of km wide	Equatorial zone, Beta–Atla–Themis regiones	Highlands to midlands	Deep (a few km) depressions	Fractures and graben	Volcanic plains in places
Regional volcanic plains							
Densely lineated plains	Equidimensional, elongated, arc-like	10s to a few 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Low relief	Dense narrow and short fractures	Plains volcanism, no discernible sources
Ridged and grooved plains	Elongated, belt-like, equidimensional	100s (up to 1000s) of km long, 10s of km wide	Major area is in a fan-shaped area in the northern hemisphere	Midlands to lowlands	Low relief	Deformed by ridges of ridge belts	Plains volcanism, no discernible sources
Shield plains	Equidimensional patches	10s to 100s of km across	Dispersed patches; clusters in places	Midlands; rare in lowlands	Low relief	Moderately deformed by wrinkle ridges	Plains volcanism with abundant small edifices
Wrinkle-ridged plains (regional plains)	Vast extensions	Up to 1000s of km across	Equidimensional and elongated regional lowlands	Lowlands and midlands	N/A	Moderately deformed by wrinkle ridges	Plains volcanism, no discernible sources
Lobate plains	Equidimensional and elongated	100s of km across	Patchy regional distribution	Highlands; rarely in midlands	N/A	In places, cut by rift structures	Plains volcanism through distinct sources
Regional volcano-tectonic features							
Lakshmi Planum	Circular highland plateau	A few 1000s of km across	Unique feature in the northern hemisphere	Midlands	Several km high	Highest mountain ranges outside	Abundant volcanism inside
Coronae and arachnoids	Circular, equidimensional	100s to 1000s of km across; average is ~300 km	Chains, clusters, isolated features	Midlands; some are in highlands	Relief varies from local lows to local highs	Annulus of fractures; ridges are subordinate	Lava flows and lava fields
Novae	Star-like, equidimensional	10s to 100s of km across; average is ~200 km	Isolated occurrences	Midlands; some are in highlands	Predominantly, topographic highs (100s of m)	Radial pattern of fractures and graben	Individual lava flows
Large shield volcanoes	Equidimensional	100s of km across	Isolated occurrences	Midlands to highlands	Several km high	In places, rift-related graben	Abundant lava flows on flanks

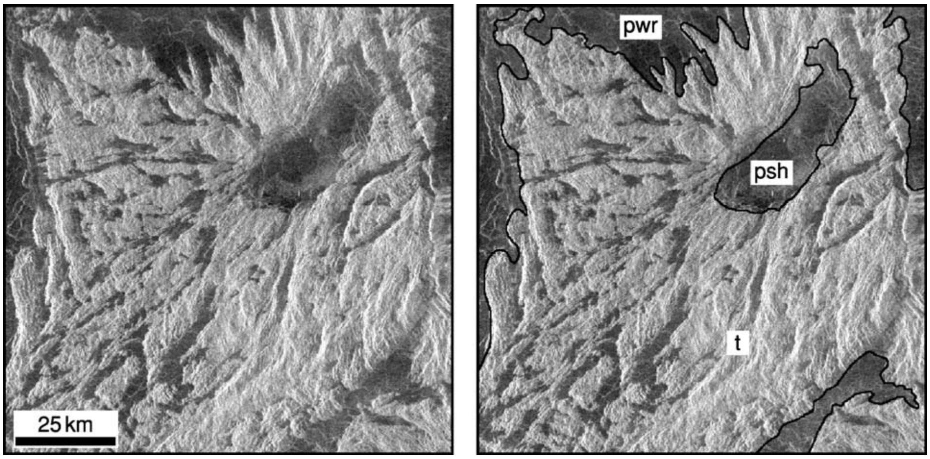


Figure 4 The surface of a tessera terrain (t) is deformed by several sets of tectonic features that completely obscure the morphological nature of the pre-existing terrain. The tessera is embayed by all plains units on Venus. The units in this example are shield plains (psh) and wrinkle-ridged plains (pwr). Fragment of C1-MIDR.30N009; illumination is from the left.

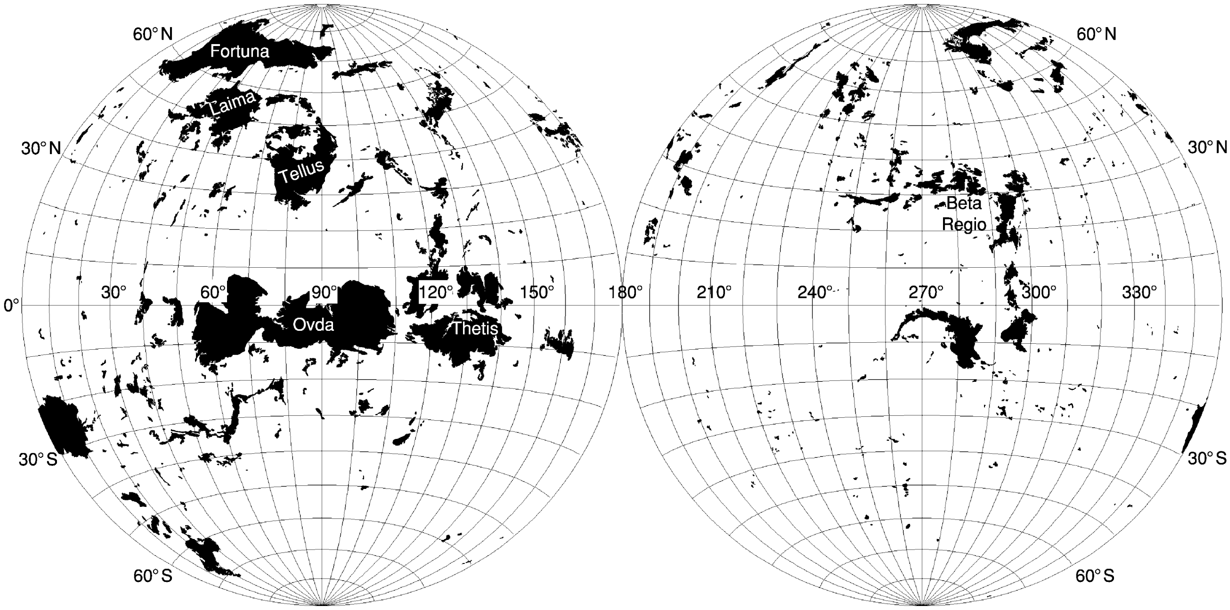


Figure 5 Map showing the global distribution of all occurrences of tessera terrains on Venus. The equatorial zone (Ovda and Thetis regiones) and high northern latitudes (Fortuna, Laima, and Tellus tesserae) have a higher density of tessera massifs. The hemisphere centred at about 230° E is dominated by major rift zones and large volcanoes and shows the clear scarcity of tesserae. The paucity of tesserae is also noticeable south of about 30° S. The map is in Lambert equal-area projection.

occur predominantly in midlands, but in some cases are found on the floor of lowlands (e.g., Lavinia and Atalanta planitiae) (Figure 10). The most prominent zone of ridge belts extends for thousands of kilometres in the northern hemisphere, where individual belts are several hundreds of kilometres long and tens of kilometres wide (Figure 11). The plains embay tesserae and densely lineated plains, suggesting the younger age of both emplacement and deformation (Figure 12A and B).

Groove belts Groove belts are swarms of curvilinear lineaments that are usually wide enough to be resolved as fractures and graben (Figure 13), manifesting the tectonic environment of extension across long (hundreds of kilometres) and broad (up to a few hundred kilometres) zones. Individual structures of the belts can reach several tens of kilometres in length and are up to 1–2 km wide. The belts occupy a small percentage of the surface and in places form prominent zones thousands of kilometres long (Figure 10).

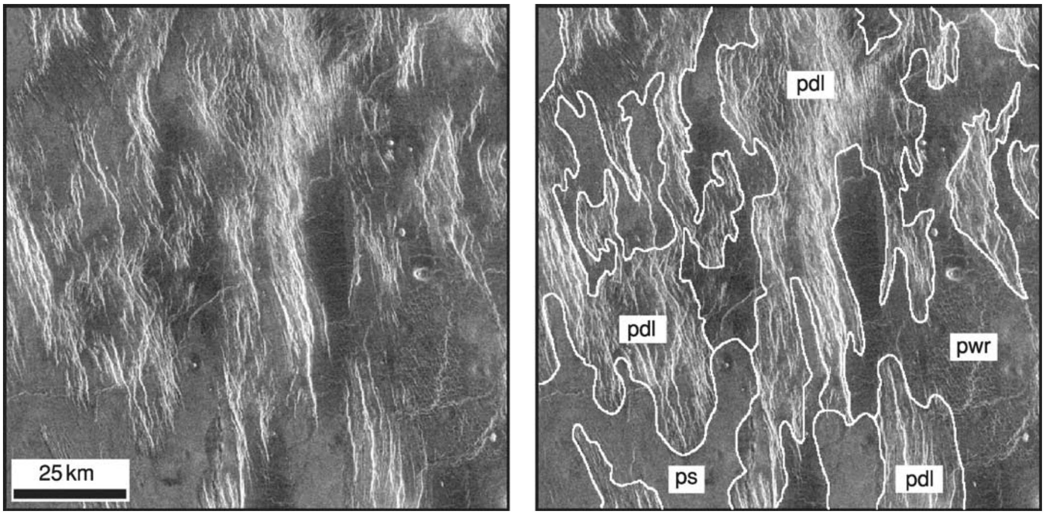


Figure 6 Areas of dense fracturing. Typically, small occurrences of these densely lineated plains (pdl) are heavily deformed by numerous narrow and densely packed lineaments (very narrow fractures). Fragments of densely lineated plains are embayed by varieties of lava plains, such as smooth plains (ps) and regional wrinkle-ridged plains (pwr). Fragment of C1-MIDR.30N027; illumination is from the left.

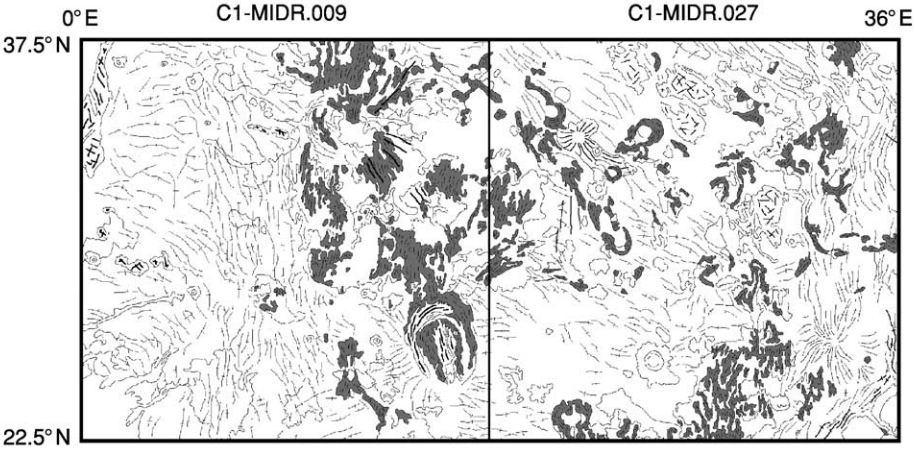


Figure 7 An example of the spatial distribution of densely lineated plains. Small fragments of densely lineated plains (dark grey) occur in clusters and as isolated patches within regional plains. In places, fragments of densely lineated plains occur within and near coronae and corona-like features (north and south-east of C1-MIDR.30N009).

Within the belts, the fractures are often anastomosing and sometimes form elliptical and circular coronae and corona-like features (Figure 14). Where groove belts are in contact with other units, they cut tesserae, densely lineated plains, and ridge belts, but are mostly embayed by younger plains such as shield plains and regional wrinkle-ridged plains (Figure 15).

Rift zones The most spectacular deformational belts on Venus are zones a few hundred kilometres wide and up to thousands of kilometres long, consisting of fractures and wide graben that can reach hundreds of kilometres in length and tens of kilometres in width (Figure 16). These features imply that the zones were formed by tensional stresses and in many aspects they resemble continental rifts on Earth. Topographically, the rift zones on Venus are troughs up to several kilometres deep that usually occur within regionally elevated areas. Preferentially, the rifts occur in a giant triangle-like area thousands of kilometres across, between Beta, Atla, and Themis regiones (the BAT province; Figure 17), where relatively young volcanic and tectonic activity on Venus is concentrated. The rift zones tend to occur with large dome-shaped rises; they are in close spatial association with lava plains and individual large lava flows, which appear to be

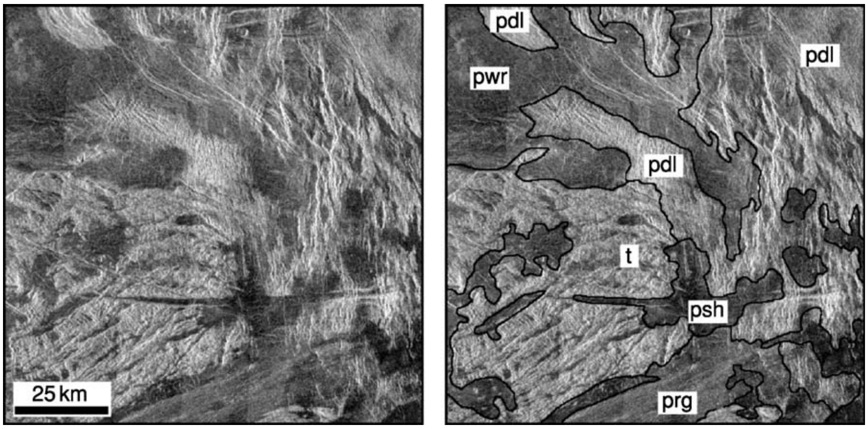


Figure 8 The relationships between tesserae (t) and densely lineated plains (pdl). A tessera has multiple sets of tectonic structures and a pdl is dissected by one set of narrow parallel lineaments. The complex pattern of deformation of tesserae is confined within its occurrences and does not penetrate into fragments of densely lineated plains. The pdl-type structures deform tessera massifs (centre of the image). The younger shield plains (psh), wrinkle-ridged plains (pwr), and ridged and grooved plains (prg) embay both tesserae and densely lineated plains. Fragment of C1-MIDR.30N125; illumination is from the left.

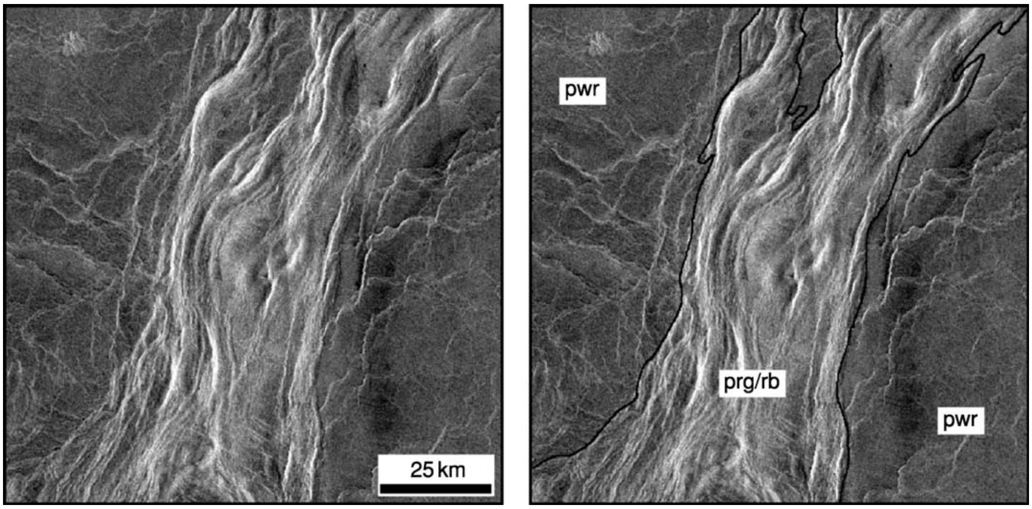


Figure 9 This typical ridge belt (rb) represents a small fragment of a larger elongated occurrence of ridged and grooved plains (prg). The curvilinear ridges typical of ridge belts are broader and less sinuous than wrinkle ridges within regional plains (pwr) are. Regional plains embay the ridge belt, implying that the belt was formed by contractional deformation of the material of ridged and grooved plains before emplacement of regional plains. Fragment of C1-MIDR.30N153; illumination is from the left.

the youngest volcanic features on Venus, and there is evidence for partly contemporaneous formation of the rifts and young volcanic plains.

Regional Plains on Venus

Due to conditions on Venus, volcanism is the prime factor contributing to growth of the crust on the planet, and extensive lava plains make up the vast majority of the surface. Several distinct units form extensive regional plains that are moderately deformed by tectonic structures.

Shield plains Shield plains are characterized by numerous small (up to 10 km across) shield- and cone-like features that are interpreted to be volcanic edifices (Figure 18). The surface of the plains is morphologically smooth but is sometimes deformed by wrinkle ridges. Shield plains cover about 10–15% of the surface of Venus and typically occur as small equidimensional areas several tens of kilometres across. Less frequently, occurrences of the plains are larger and can reach a few hundred kilometres (Figure 19A and B). The overall relief of the plains

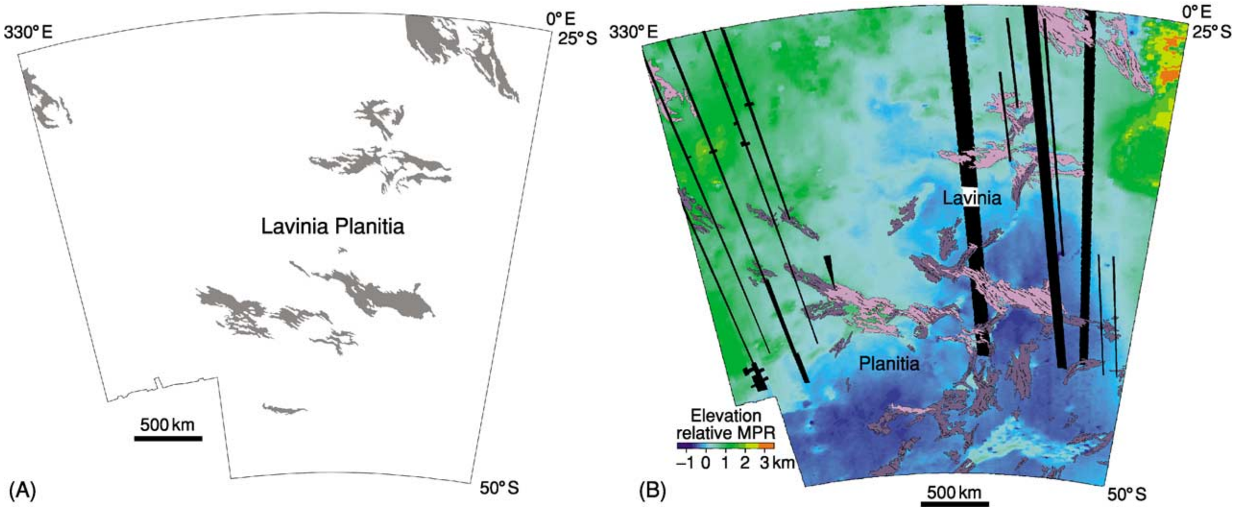


Figure 10 (A) An example of the spatial distribution of groove belts. Groove belts (dark grey) make major deformational belts on the floor of Lavinia Planitia. Individual occurrences of the belts can reach a 1000 km in length and several hundred kilometres in width. The map is in Lambert conformal projection. (B). An example of the spatial distribution of major deformational belts within the lowland of Lavinia Planitia. Occurrences of ridge belts (dark purple) are oriented in a north-eastern direction. Groove belts (pink) are orthogonal to the strike of ridge belts. Both types of belts are concentrated within the deepest portion of Lavinia Planitia and the ridge belts are parallel to the elongation of the floor of the lowland. Colours show distribution of elevation relative to mean planetary radius (MPR; 6051 km). Black strips are data gaps. The map is in Lambert conformal projection.

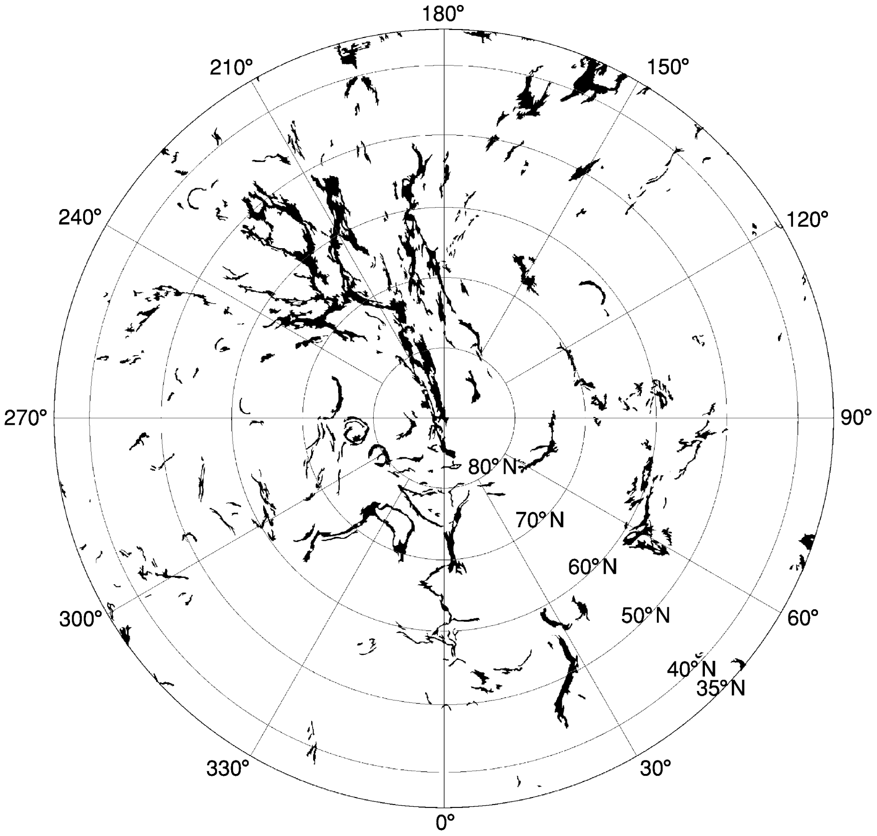


Figure 11 Distribution of ridged and grooved plains and ridge belts (black areas) in the northern hemisphere of Venus above 35° N. The most important occurrence of ridge belts is fan-shaped and centred at about 210° E. The map is in polar stereographic projection.

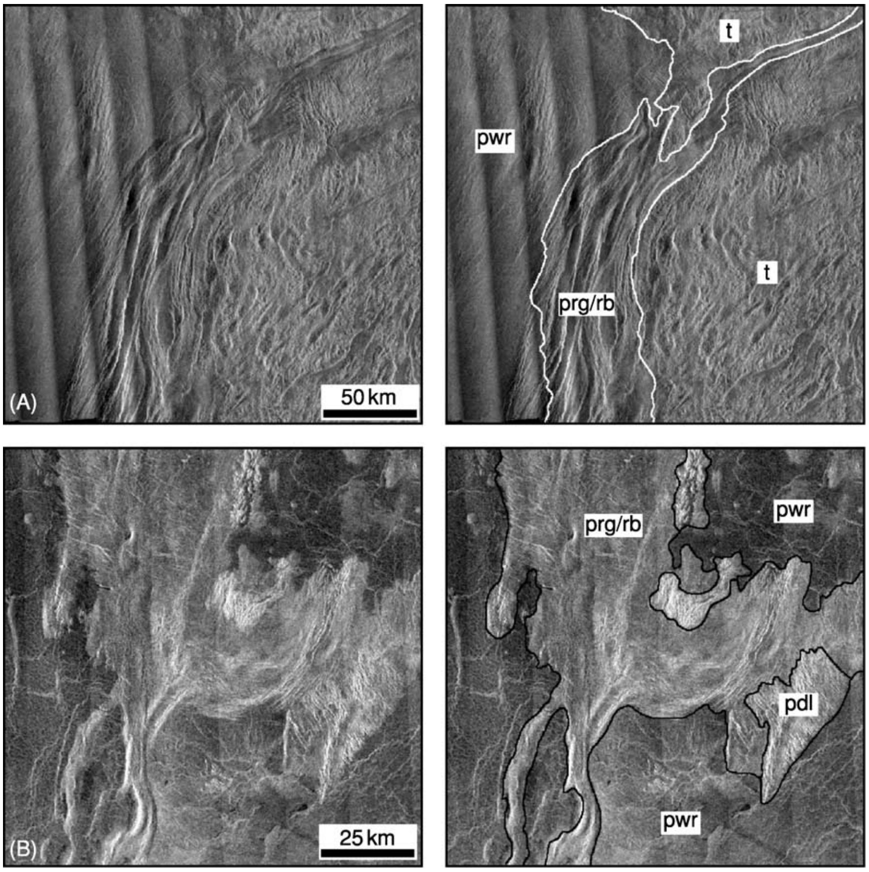


Figure 12 The relationships of ridge belts and ridged and grooved plains with other units. (A) Occurrences of ridged and grooved plains deformed into ridge belts (prg/rb) run along the edge of extensive tessera regions (t). The complex pattern of deformation within a tessera is confined within its massifs and appears to be cut by the system of ridges of the ridge belt. The surface of the regional plains (pwr) is moderately deformed, and plains material embays ridges of the belt. Fragment of C1-MIDR.00N112; illumination is from the right. (B) Material of the ridges and grooved plains, which is deformed in places into ridge belts (prg/rb), embays small fragments of densely lineated plains (pdl). Regional wrinkle-ridged plains (pwr) broadly embay both prg/rb and pdl. Fragment of C1-MIDR.30N153; illumination is from the left.

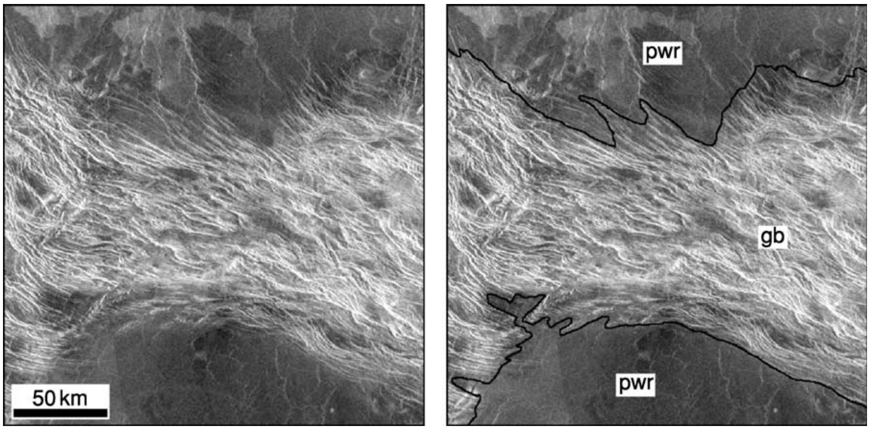


Figure 13 Typical groove belt consists of numerous linear and curvilinear fractures and graben that almost completely destroy pre-existing materials (pwr, wrinkle-ridged plains; gb, groove belt). Fragment of C1-MIDR.45S350; illumination is from the left.

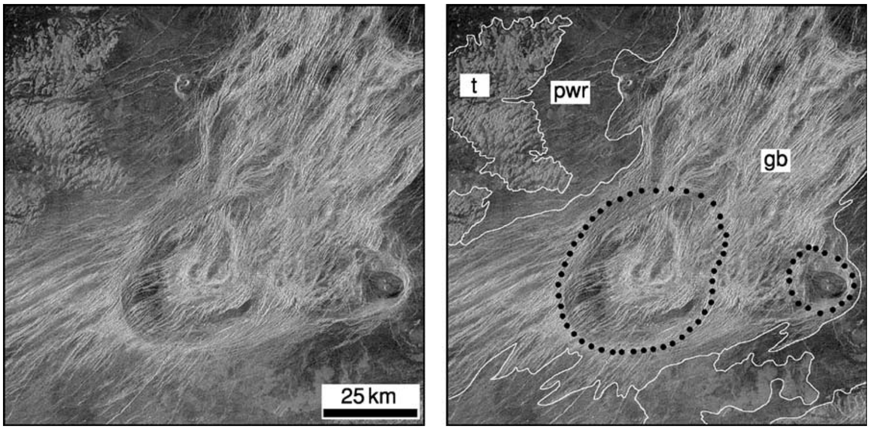


Figure 14 A fragment of a groove belt (gb), the structures of which outline coronae and corona-like features (dotted lines in the right image). t, tessera; pwr, wrinkle-ridged plains. Fragment of C1-MIDR.30N261; illumination is from the left.

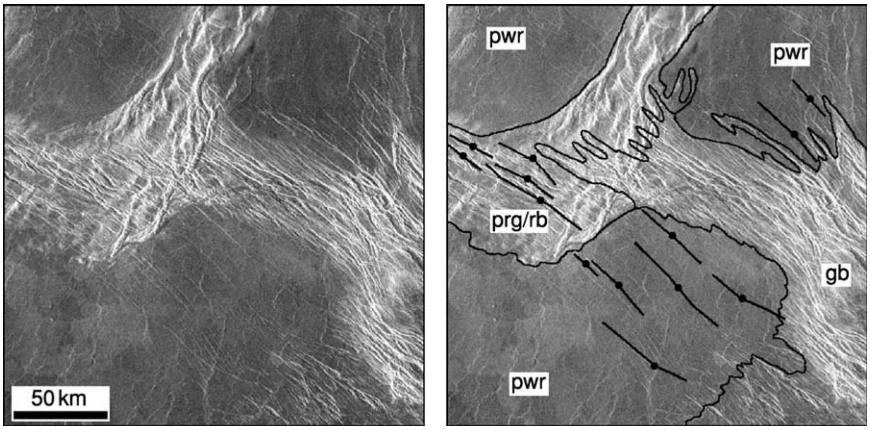


Figure 15 Relationships between groove belts (gb) and ridge belts (rb) typically show that fractures and graben of groove belts cut structures of ridge belts. Material of regional wrinkle ridged plains (pwr) embays both types of deformational belts. prg, ridged and grooved plains. Fragment of C1-MIDR.45S350; illumination is from the left.

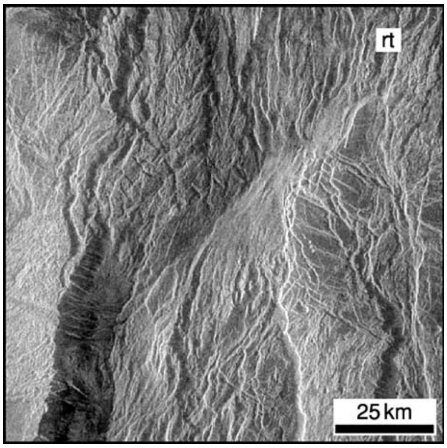


Figure 16 A portion of the Devana Chasma rift zone (rt) that cuts through the central portion of Beta Regio. The rift zone consists of a great number of fractures and graben, between which remnants of pre-existing terrains are visible. Fragment of C1-MIDR.30N279; illumination is from the left.

appears to be hilly due to abundant shield features; occurrences of shield plains tend to be slightly higher compared to the surrounding regional plains. Although shield plains occur at different elevations, they preferentially occupy regional slopes away from old elevated terrains such as tesserae or ridge belts. The plains embay older, heavily tectonized units (Figure 20A) but are predominantly embayed by regional wrinkle-ridged plains (Figure 20B).

Regional wrinkle-ridged plains Regional wrinkle-ridged plains have morphologically smooth surfaces that are moderately deformed by numerous low, narrow, and sinuous wrinkle ridges (Figure 21). Wrinkle-ridged plains make up ~50–55% of the surface, appear as a regional background (with other units and structures being either older or younger), and can be traced almost continuously around the planet. The surface of the plains usually has a

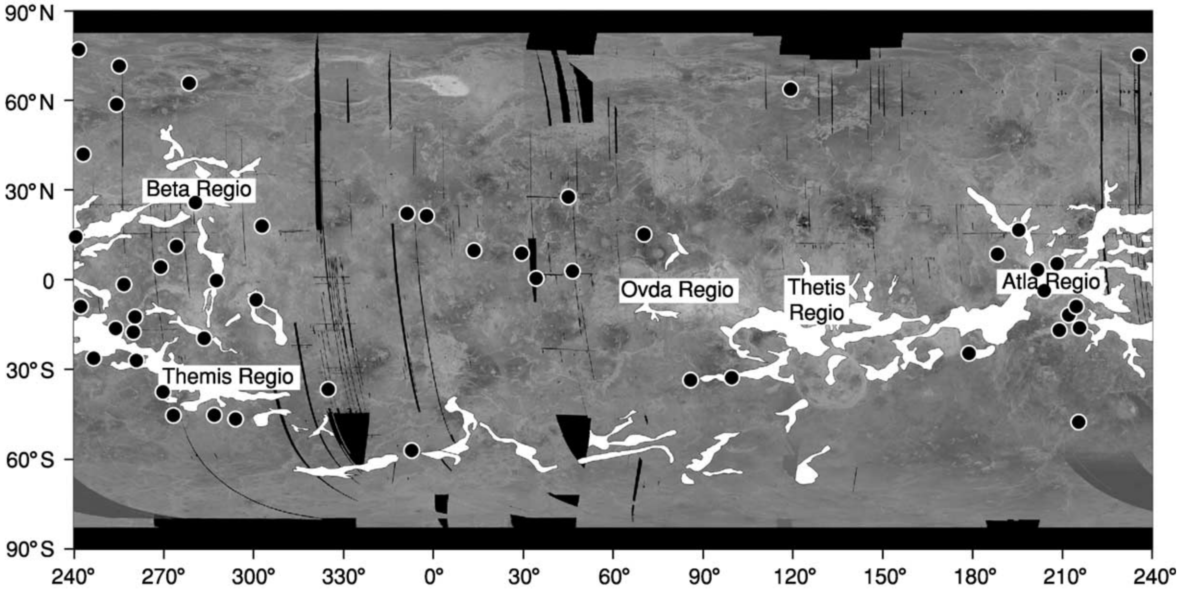


Figure 17 The global distribution of major rift zones (white) on Venus. The main concentration of rifts is within a giant triangle-shaped area between Beta, Atla, and Themis regiones. The photobase is a low-resolution synthetic aperture radar image of the Venus globe in simple cylindrical projection.

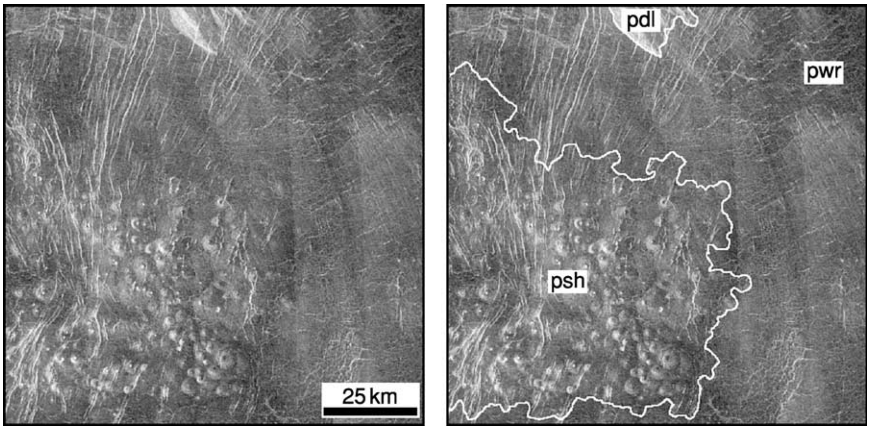


Figure 18 The surface of shield plains (psh) is characterized by a large number of small shield-like features interpreted as volcanic edifices. Many of the shields have a bright dot in the centre; this is interpreted as a central pit (crater). The occurrence of shield plains is visible in the central part of Boala Corona. pdl, Densely lineated plains; pwr, wrinkle-ridged plains. Fragment of C1-MIDR.30N135; illumination is from the left.

relatively low and uniform radar albedo with no visible flowlike features, which precludes identification of the sources of the plains material. A specific characteristic of the plains is the presence of long and narrow channels on their surface. The longest channel, Baltis Vallis, is about 7000 km long (Figures 21 and 22). In some places, the radar albedo of regional plains is distinctly higher and plains occur as relatively bright areas hundreds of kilometres across. These areas often surround distinct volcanic centers, such as large volcanoes and some coronae, and they form a distal apron of volcanic materials around them. Although in many cases there is evidence for embayment of the darker plains by the material of the brighter ones (Figure 23), the same family of wrinkle ridges appears to deform both varieties of regional plains (Figure 23). Regional plains cover the surface of large equidimensional basins (Figure 1) and make up the majority of midlands, but are noticeably less abundant within either plateau-shaped or dome-shaped highlands such as Ovda or Beta regiones.

Large volcanoes and lobate plains The large volcanoes are equidimensional mountains several hundred

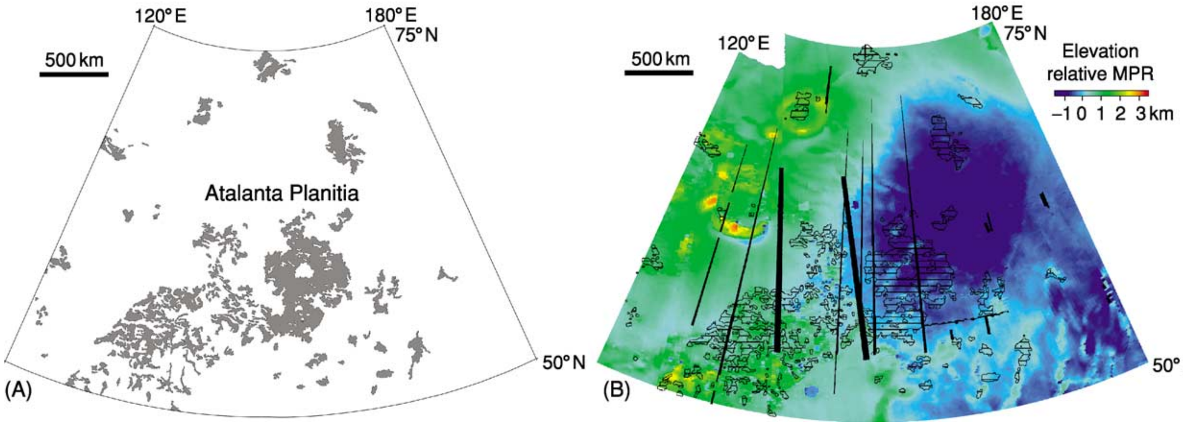


Figure 19 (A) The spatial distribution of shield plains (dark grey) within Atalanta Planitia. Fragments of the plains vary in size from several tens of kilometres up to a few hundred kilometres and occur in clusters and as individual patches. The map is in Lambert conformal projection. (B) The spatial distribution of shield plains (ruled pattern) within the Atalanta Planitia basin. Fragments of the plains vary in size from several tens of kilometres up to a few hundred kilometres and occur mostly on the regional slope of the Atalanta lowland, where groups of older units such as tesserae and densely lineated plains collectively form local highs. Colours show distribution of elevation relative to mean planetary radius (MPR; 6051 km). Black strips are data gaps. The map is in Lambert conformal projection.

kilometres across and a few kilometres high. Sometimes, a broad caldera-like feature is present at the summit of the volcano; in these cases, numerous radar-bright and dark flow-like features interpreted as lava flows always cover the slopes (Figure 24). The individual lava flows are superposed on each other and collectively form extensive lava plains (lobate plains), occurrences of which have lobate boundaries and can be several hundred kilometres across. There are 168 large volcanoes on Venus and their diameters vary from 100 to 1000 km. Typically, the large volcanoes associate with major rift zones and many of them occur within the Beta–Atla–Themis region (Figure 17).

Major Volcano-Tectonic Features

Lakshmi Planum Lakshmi Planum is a high-standing (2–4 km above MPR) plateau almost completely surrounded by mountain ranges (Figure 25). These ranges, the highest mountains on Venus, average 7–8 km in height (some reach 11 km). Lakshmi Planum, which is a few thousand kilometres wide, is so dissimilar to other types of highlands on Venus that it can be considered a specific class of topographic province. The interior of the Planum, flat and slightly tilted to south, is covered by smooth volcanic plains that are morphologically similar to vast regional wrinkle-ridged plains elsewhere on Venus. These plains embay both the tessera-like terrain within the plateau and the individual ridges at the base of the surrounding mountains. Two major volcanic structures, Colette and Sacajawea paterae, dominate the interior of Lakshmi and are the centers of the abundant lava flows that are superposed on the wrinkle-ridged plains inside the Planum; these plains are similar to the lobate plains on the slopes of large volcanoes.

Coronae, arachnoids, and novae Coronae, arachnoids, and novae are circular or quasi-circular features tens to hundreds of kilometres across. Coronae (Figure 26) and arachnoids (Figure 27) are characterized by concentric deformational annuli that predominantly consist of extensional structures (fractures and graben) and sometimes ridges (contractional features) and novae form starlike patterns of radial fractures; graben (Figure 28). In the catalogue of volcano-tectonic landforms compiled by Crumpler and Aubele in 2000, 209 coronae, 265 arachnoids, and 64 novae are listed. All of these features are thought to be the surface manifestations of mantle diapirs at different stages of evolution. Coronae and other circular volcano-tectonic features occur predominantly in the midlands and only a few of them are either within the high-standing plateaulike highlands and dome-shaped rises or in the lowlands (Figure 29). The topographic configuration of these structures varies from rimmed depressions to plateaus to dome-shaped features. Many coronae and novae are surrounded by prominent lava flows, suggesting that these features are distinct volcanic centers. However, there are neither medium-sized nor large volcanoes in association with these features. Lava flows are rarely associated with arachnoids and these structures appear mostly as tectonic structures. In many cases, regional wrinkle-ridged plains embay tectonic

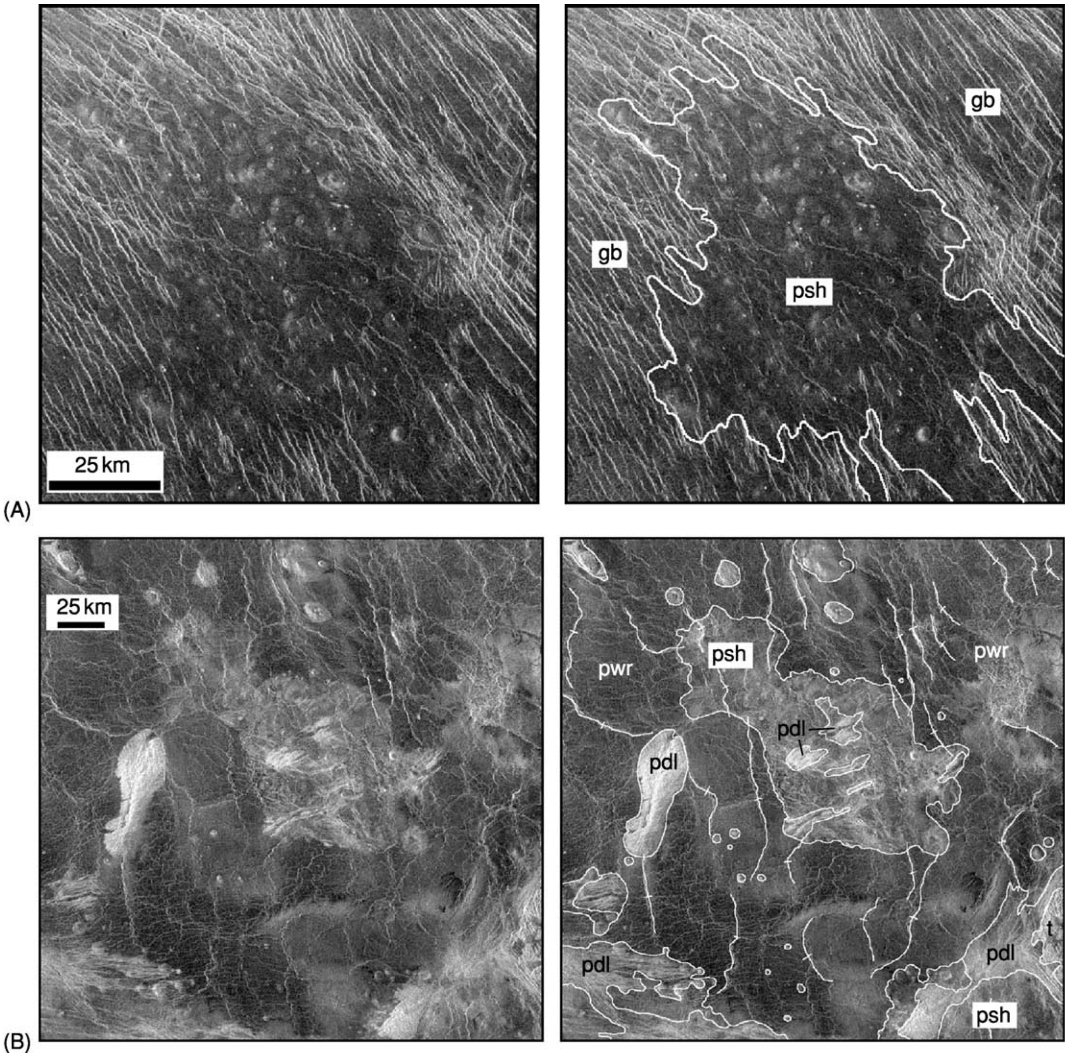


Figure 20 Relationships of shield plains (psh) with other units and structures. (A) Some portion of a groove belt (gb) is covered by deposits of shield plains that are obviously younger. A few shields, however, are cut by the fractures of the belt, suggesting that the formation of groove belt and shield plains partly overlapped in time. Fragment of C1-MIDR.30N333; illumination is from the left. (B) Occurrences of shield plains within regional plains are often characterized by a specific pattern of deformation confined within shield plains and a radar albedo that is different from the albedo of regional plains. Small individual shields that are seen within regional plains have morphological characteristics similar to these, typical of the main occurrences of shield plains. This suggests that the individual shields represent kipukas of more widespread shield plains covered by a mantle of wrinkle-ridged plains (pwr).

structures of coronae and arachnoids, and fractures and graben of novae commonly cut the plains.

Major Topographic Features

The fact that the gravity and topography fields of Venus are highly correlated suggests that Venus may not have a low-viscosity layer, as in the asthenosphere of Earth, and that the Venusian mantle is strongly coupled with the lithosphere. Thus, the mantle circulation on Venus could be directly responsible for the formation of large-scale tectonic and topographic features. In the almost complete absence of erosion, on the other hand, the large-scale topographic features on Venus should much better reflect the balance between lithospheric buoyancy and mantle dynamics, compared to Earth. Thus the distribution of the major topographic features on Venus combines the present pattern of mantle convection with contributions from extinct patterns. The global altimetry data collected by Pioneer Venus, Venera-15/16, and Magellan show that three principal topographic provinces characterize the surface of the planet (Figure 1B). Lowlands (<0 km) make up ~11% of the surface and consist of equidimensional basins and elongated depressions thousands of kilometres across. Their surface

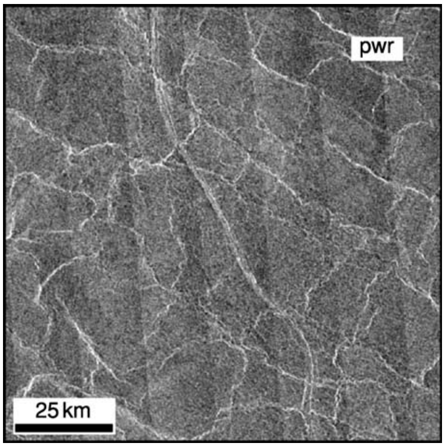


Figure 21 Regional plains with wrinkle ridges (pwr). The plains have a generally smooth surface with a relatively low and uniform radar albedo. Numerous narrow and sinuous ridges deform the surface of the plains. In the centre of the image, a narrow channel-like feature that cuts the surface of the plains is visible. These channels are typical features on the surface of regional plains. Fragment of C1-MIDR.30N153; illumination is from the left.

is predominantly covered with regional wrinkle-ridged plains. Midlands constitute the majority of the surface of Venus (~80%), occur at elevations between 0 and 2 km, and host the richest variety of terrains, units, and structures.

Highlands are above 2 km and comprise ~9% of the surface. The highlands include two distinct classes of first-order features that are thousands of kilometres across. The first class consists of relatively steep-sided plateaulike features, the surface of which is typically covered by tesserae (e.g., Ovda Regio and Fortuna Tessera in eastern Ishtar Terra). These features appear to be isostatically compensated at relatively shallow depth, several tens of kilometres, suggesting that they are areas of thickened crust and probably relate to ancient regimes of mantle convection. This is consistent with the stratigraphic position of tesserae, which are the oldest terrain on the surface of Venus. The second class of highlands includes dome-shaped rises that are typically rifted and topped by large volcanoes (e.g., Atla and Ulfrun regiones).

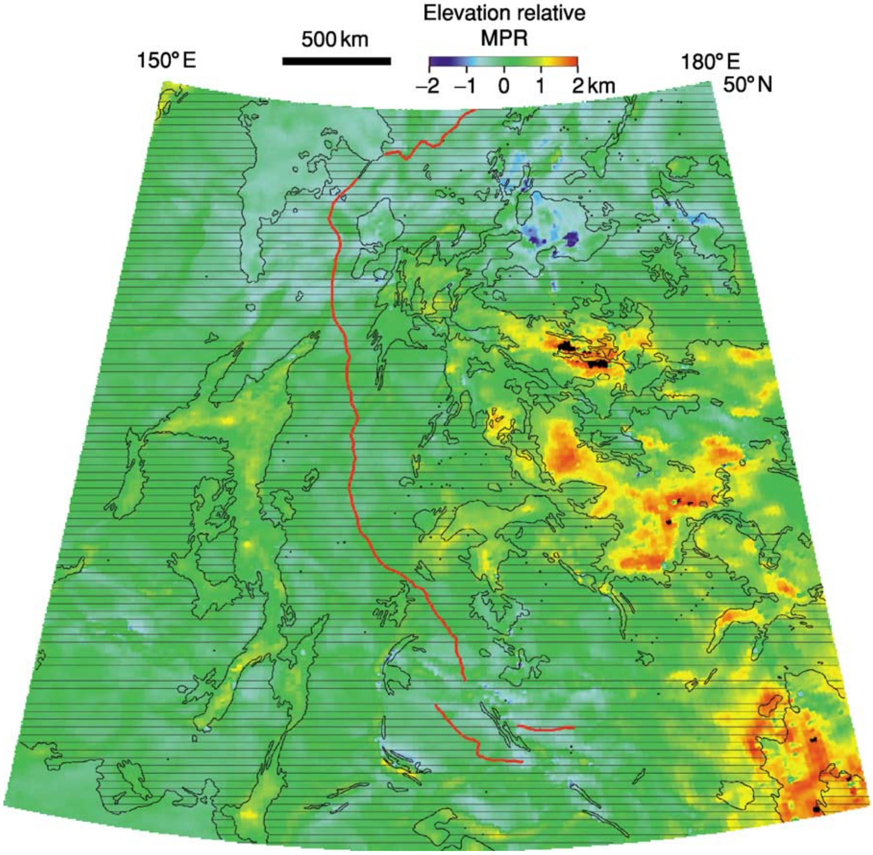


Figure 22 Spatial distribution of regional wrinkle-ridged plains (ruled pattern). The plains occupy the relatively low portion of the area (Ganiki Planitia) between elevated territories to the west and east, where the older units and structures such as tesserae and ridge belts are exposed. The thick red line in the centre of the map is Baltis Vallis, which runs along the major continuous lowland. Colours show the distribution of elevation relative to mean planetary radius (MPR; 6051 km). The map is in Lambert conformal projection.

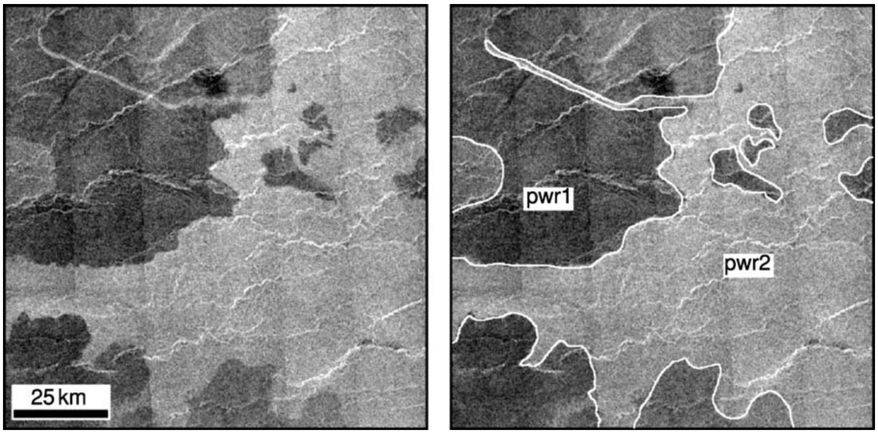


Figure 23 Varieties of regional wrinkle-ridged plains. The upper member of the plains (pwr2) has a uniform and relatively higher radar albedo compared to the lower member of the plains (pwr1). Material of the brighter plains fills a portion of the lava channel (upper left), implying that pwr2 plains are younger. The same pervasive network of wrinkle ridges, however, deforms both varieties of the plains. Fragment of C1-MIDR.30N153; illumination is from the left.

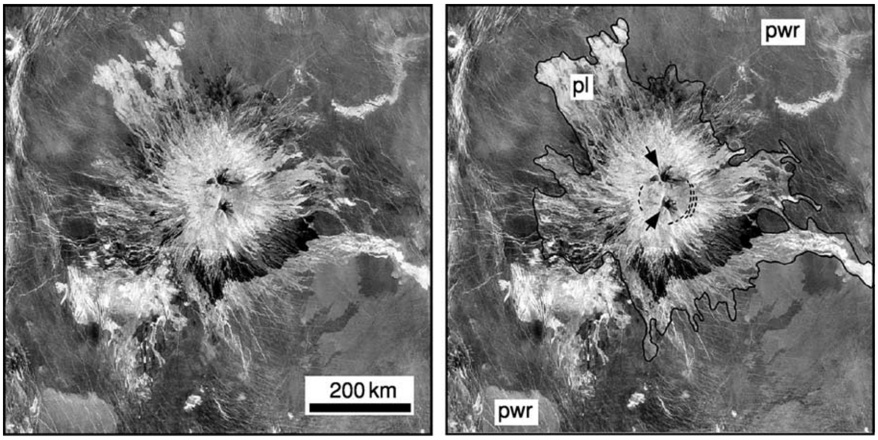


Figure 24 A large volcano (Sapas Mons) and lobate plains. Sapas Mons is a distinct volcanic centre from which issue a large number of radar-bright and radar-dark lava flows. In places, the flows merge with each other and form extensive lobate lava plains (pl) that are clearly superposed on the background of regional wrinkle-ridged plains (pwr). Dashed lines (right image) show a series of arcuate graben in the summit area of Sapas Mons and black arrows indicate two steep-sided domes. Mosaic of C1-MIDR.15N180, C1-MIDR.15N197, C1-MIDR.00N180, and C1-MIDR.00N197; illumination is from the left.

These rises appear to be compensated at much deeper levels, hundreds of kilometres, suggesting their dynamical support by active mantle upwelling. This is consistent with the geological characteristics of the rises, such as young rift structures and abundant young volcanism emerging through distinct sources.

Heat Loss Mechanisms

The style of volcanic and tectonic activity and the distribution of major topographic provinces are the specific manifestations of heat loss mechanisms operating on a planetary body. The global survey of the surface of Venus by Magellan has showed that except for a few possible sites, evidence for subduction is absent on the surface. Thus, the principal heat loss mechanism of Earth-like plate tectonics (lateral crust recycling, or "lateral" heat loss) apparently does not work on Venus. The alternative is a different orientation of the principal vector of heat loss mechanisms, vertical instead of horizontal. This means that Venus should be characterized by vertical crust accretion/recycling, or 'vertical' heat loss mechanisms. The manifestation of these is mantle upwelling and downwelling. The question of crucial importance in this context is the continuous or discontinuous nature of these mechanisms. Did they operate in a steady-state mode or did the temporal pattern of heat loss consist

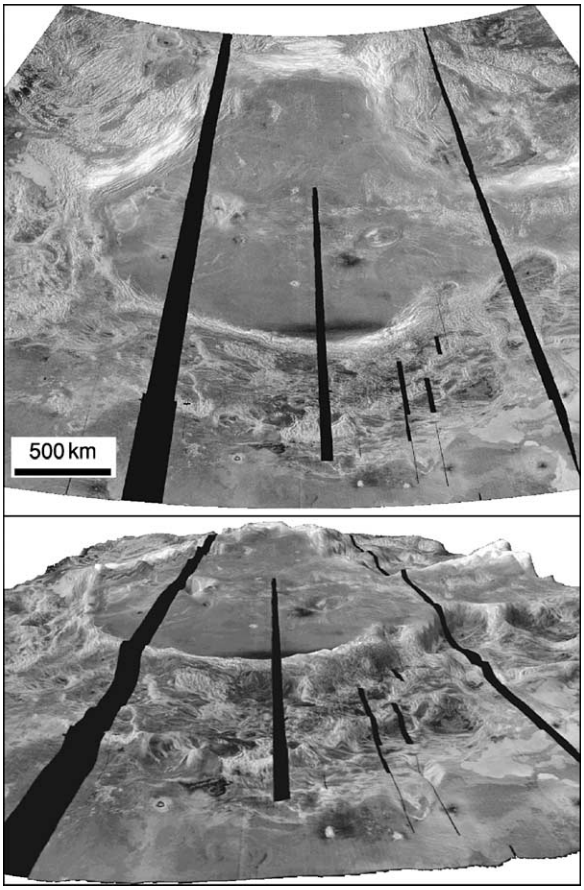


Figure 25 Lakshmi Planum, plan view (top) and perspective view (bottom). High mountain ranges (Danu Montes, south; Akna Montes, west and north-west; Freyja Montes, north) almost completely border the interior of Lakshmi. Two large volcanic centres, Colette and Sacajawea Paterae, are distinct sources of relatively young volcanic materials (lobate plains) within the Planum. The plan view is in Lambert conformal projection.

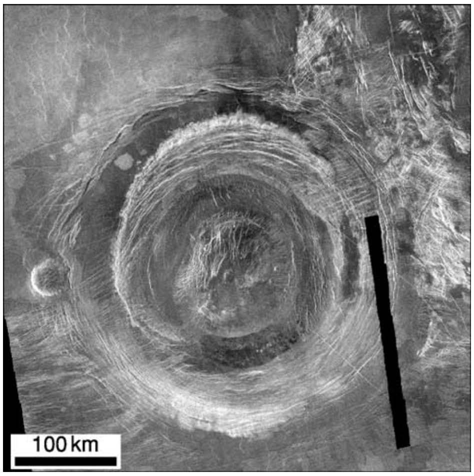


Figure 26 Corona Aramaiti is a typical example of this class of volcano-tectonic structures. The corona has an outer and inner rim and a relatively flat floor populated with small shields. Within the outer rim, contractional ridges (northern portion of the rim) and extensional fractures (southern portion of the rim) are seen. Extensional features dominate the inner rim. A swarm of narrow lineaments (fractures and graben) appears to cut the outer rim (lower left and upper right) but disappears within the inner rim and the floor of the corona. Fragment of C1-MIDR.30S082; illumination is from the left.

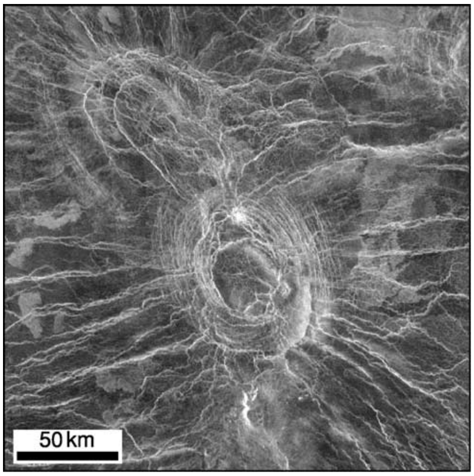


Figure 27 An example of an arachnoid. Swarms of concentric arcuate lineaments interpreted as fractures outline the doubled core of this feature (centre and upper left). Wrinkle ridges within regional plains appear to be focused at the arachnoid and form a radial pattern of structures around it. Fragment of C1-MIDR.45N011; illumination is from the left.

of a series of 'bursts' of endogenous activity intermittent with epochs of volcanic and tectonic quiescence, or was there one major change from a vigorous to a sluggish character of mantle convection?

The spatial distribution and morphology of impact craters place important constraints on the possible mode of heat loss/crust recycling on Venus. The catastrophic model of resurfacing is consistent with the characteristics of the crater population whereas the model of equilibrium resurfacing requires geological activity within small, ~400-km-diameter spots corresponding to the mean crater-to-crater distance. Thus, the characteristic horizontal scale of the mantle convection is also small, much smaller that the typical dimensions of many major features, both topographic and morphologic, on the surface. Thus, although the hypothesis of catastrophic resurfacing is an end-member model and almost certainly is incorrect in some details, it appears to describe better the geological situation on the surface of Venus.

There are two variants of the catastrophic resurfacing. In the first, the vertical accretion of crust and growth of the lithosphere lead finally to gravitational instability and large-scale delamination within the lithosphere. Depending on the rheological properties of the material, horizontal scales of the instabilities, and the time-scale at which the instabilities exist, this process may lead either to transient plate tectonics or to large-scale mantle overturn. Both scenarios imply a cyclic nature of the heat loss mechanisms and may lead to a planet-wide resurfacing event on the surface.

The second variant of the catastrophic resurfacing is based on the secular cooling of the interior of the planet during most of the geological history of Venus. According to this scenario, the observable geological history of Venus begins after the transition from vigorous mantle convection under thin lithosphere to stagnant lid convection under thick lithosphere. In this model, tessera terrain is the remnant of the previous, thin-lithosphere regime, and the rifted dome-shaped rises topped with large volcanoes are the manifestation of the current regime of mantle convection under thicker lithosphere.

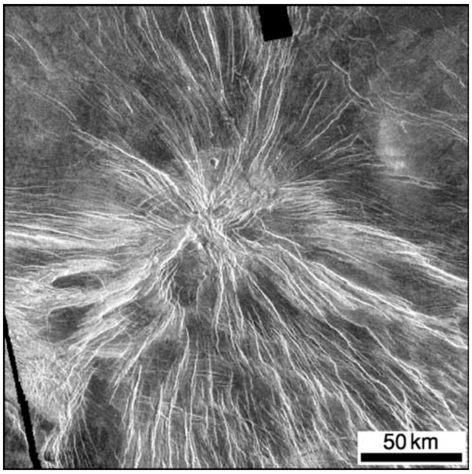


Figure 28 A typical nova is characterized by a star-like pattern of broad radial fractures and graben originating at its centre. Fragment of C1-MIDR.30S279; illumination is from the left.

Models of Geological History on Venus

The high quality and global coverage of the Magellan data provide the possibility of detailed geological mapping of the surface based on defining distinct units and structures and establishing their relative ages. The results of the mapping efforts have led to two proposed end-member models for the correlation of regionally observed sequences of units and structures. In the first model, sequences of distinctive units mapped in different regions appear to have similar repetitive sequences in different places. This model has been called a “directional” geological history, implying a specific set of global trends in the evolution of Venus. For example, the consistently oldest relative age of tesserae suggests that a tectonic style of tessera formation has changed, with subsequent tectonic styles that led to the formation of other types of terrains. The important attribute of this model is its ‘synchronism’, implying that the sequences of events observed in different regions are broadly synchronous globally. For instance, regional wrinkle-ridged plains

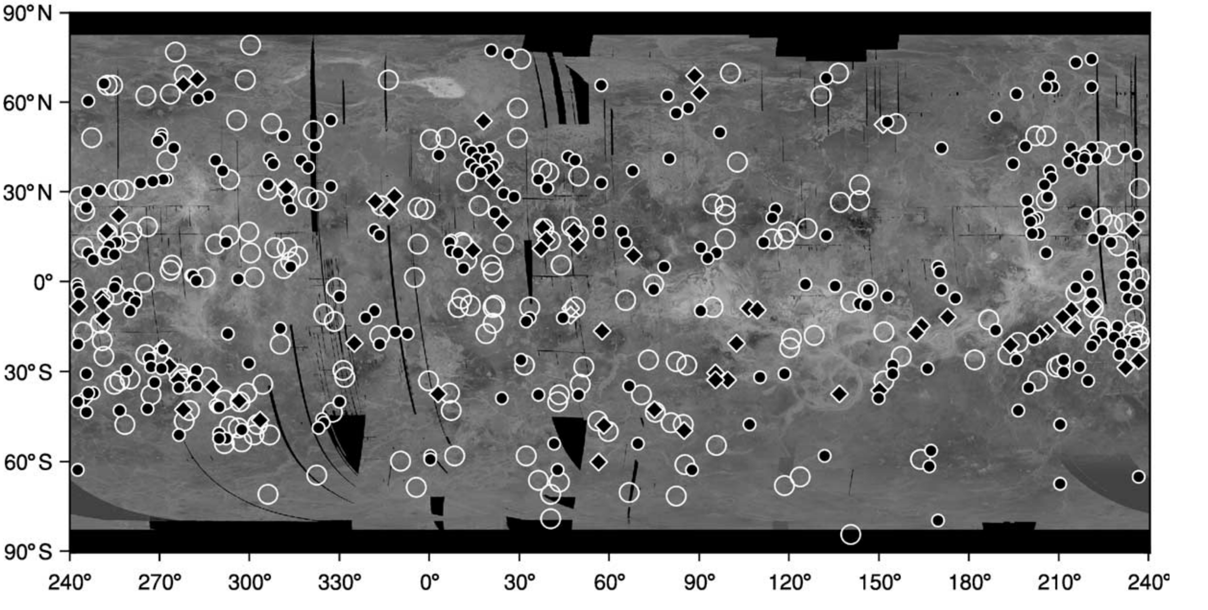


Figure 29 The spatial distribution of coronae (large circles), arachnoids (black dots), and novae (black diamonds) on the surface of Venus. The majority of these features, especially the novae, is concentrated in the Beta, Atla, and Themis regiones province and tends to be off of the large plateau-shaped tessera highlands, such as Ovda and Thetis regiones. The map is in simple cylindrical projection.

appear to represent a broadly similar unit with a distinct stratigraphic position (either postdating or predating groups of other units) that can be traced continuously around the globe of Venus.

In the alternative model of geological history, the observed sequence of units is interpreted to be due to specific volcano-tectonic regimes that occur at different times on different parts of Venus, similar to Wilson cycles on Earth (individual plate-tectonic cycles that are repeated at different times and in different places on Earth). In this model for Venus, the local sequence of units represents only local or regional time-dependent sequential styles of endogenous activity. This is a 'non-directional' model of geological history, implying that the individual sequences represent local conditions occurring at different times in different places. Because the sequences of units and structures are almost the same in different regions on Venus, this model indicates that similar sequences of events resulted in similar stratigraphic columns occurring in these areas throughout the visible part of geological history. Another aspect of this model is its 'non-synchronous' nature, implying that the sequences of units/events are non-synchronous globally and that similar stratigraphic columns in specific regions are shifted relative to each other in terms of their age.

Further Investigation

The data collected during the exploration of Venus reveal the uniqueness of this planet. Venus does not have the surface age dichotomy characterizing the Moon and smaller planets such as Mercury and Mars. This fundamental characteristic implies that Venus, like Earth, has a prolonged history of geological activity that did not significantly decrease in intensity early in the evolution of the planet. In contrast to Earth, where the global heat loss mechanism is governed by plate tectonics, vertical crust accretion/recycling appears to be the principal style of geological activity on Venus.

Although current knowledge of Venus is great, there are still several major issues about its geology that are open to debate and further investigation. What is the evolution of the heat loss mechanisms on Venus? What are the paths of the evolution of the large-scale topographic features on the planet? How did the properties of the Venus lithosphere change as a function of time? Why are the Earth and Venus, the 'twin' planets, so different? What is the role of water in the evolution of both planets? Why is there little evidence on Venus for the presence of non-basaltic continental crust, which constitutes the major part of crustal material on Earth? How and when did the present atmosphere form and how has it evolved with time? How has the atmosphere interacted with the surface in recent and more ancient history of Venus? Obtaining answers to these questions requires continued exploration and key datasets, including seismic data, global high-resolution topography, *in situ* analysis of ancient terrains such as tesserae, and samples returned to terrestrial laboratories.

See Also

Earth Structure and Origins. **Solar System:** Mercury; Moon; Mars.

Further Reading

Barsukov VL, Basilevsky AT, Burba GA, *et al.* (1986) The geology and geomorphology of the Venus surface as revealed by the radar images obtained by Venera 15 and 16. *Journal of Geophysical Research* 91: D399–D411.

Bougher SW, Hunten DM, and Philips RJ (eds.) (1997) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment.* Tucson: University of Arizona Press.

Crumpler LS and Aubele JA (2000) Volcanism on Venus. In: Sigurdsson H, Houghton BF, McNutt SR, Rymer H, and Stix J (eds.) *Encyclopedia of Volcanoes,* pp. 727–769. San Diego: Academic Press.

Ford PG and Pettengill GH (1992) Venus topography at kilometer-scale slopes. *Journal of Geophysical Research* 97: 13103–13114.

Hansen VL, Willis JJ, and Banerdt WB (1997) Tectonic overview and synthesis. In: Bougher SW, Hunten DM, and Phillips RJ (eds.) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment,* pp. 797–844. Tucson: University of Arizona Press.

Hauck SA, Phillips RJ, and Price MH (1998) Venus: Crater distribution and plains resurfacing models. *Journal of Geophysical Research* 103: 13635–13642.

Head JW, Crumpler LS, Aubele JC, Guest JE, and Saunders RS (1992) Venus volcanism: classification of volcanic features and structures, associations, and global distribution from Magellan data. *Journal of Geophysical Research* 97: 13153–13197.

Ivanov MA and Head JW (2001) Geology of Venus: mapping of a global geotraverse at 30° N latitude. *Journal of Geophysical Research* 106: 17515–17566.

Masursky H, Eliason E, Ford PG, *et al.* (1980) Pioneer-Venus radar results: geology from the images and altimetry. *Journal of Geophysical Research* 85: 8232–8260.

McKinnon WB, Zahnle KJ, Ivanov BA, and Melosh HJ (1997) Cratering on Venus: models and observations. In: Bougher SW, Hunten DM, and Phillips RJ (eds.) *Venus II Geology, Geophysics, Atmosphere, and Solar Wind Environment,* pp. 969–1014. Tucson: University of Arizona Press.

Parmentier EM and Hess PC (1992) Chemical differentiation of a convecting planetary interior: consequences for a one-plate planet such as Venus. *Geophysical Research Letters* 19: 2015–2018.

Phillips RJ, Raubertas RF, Arvidson RE, *et al.* (1992) Impact craters and Venus resurfacing history. *Journal of Geophysical Research* 97: 15 923–15 948.

Schaber GG, Strom RG, Moore HJ, *et al.* (1992) Geology and distribution of impact craters on Venus: what are they telling us? *Journal of Geophysical Research* 97: 13 257–13 301.

Schubert G and Sandwell TD (1995) A global survey of possible subduction sites on Venus. *Icarus* 117: 173–196.

Simons M, Solomon SC, and Hager BH (1997) Localization of gravity and topography: constraints on the tectonics and mantle dynamics of Venus. *Geophysical Journal International* 131: 24–44.

Solomon SC, Smrekar SE, Bindschadler DL, *et al.* (1992) Venus tectonics: an overview of Magellan observations. *Journal of Geophysical Research* 97: 13 199–13 255.

Stofan ER, Sharpton VL, Schubert G, *et al.* (1992) Global distribution and characteristics of coronae and related features on Venus: implications for origin and relation to mantle processes. *Journal of Geophysical Research* 97: 13 347–13 378.

Strom RG, Schaber GG, and Dawson DD (1994) The global resurfacing of Venus. *Journal of Geophysical Research* 99: 10 899–10 926.

Sukhanov AL (1992) Tesserae. In: Barsukov VL, Basilevsky AT, Volkov VP, and Zharkov VN (eds.) *Venus Geology, Geochemistry, and Geophysics (Research Results from the USSR)*, pp. 82–95. Tucson: University of Arizona Press.

Surkov YA, Moskalyova VP, Kharyukova AD, Dudin AD, Smirnov GG, and Zaitseva SE (1986) Venus rock composition at the Vega 2 landing site. *Proceedings of the Lunar and Planetary Science Conference, Part 1, Journal of Geophysical Research* 9(supplement): E215–E218.

Turcotte DL (1995) How does Venus lose heat? *Journal of Geophysical Research* 100: 16 931–16 940.

Moon

P Moore, Selsey, UK

Introduction

The Moon is our companion in space. It is usually regarded as the Earth's satellite, though since it has 1/81 the mass of the Earth it may be better to class the Earth–Moon system as a double planet. The Moon is a world of craters, mountains, and wide plains always referred to as seas (maria), though there has never been any water in them; the craters are generally accepted as having been produced by impacting meteorites, and some are well over 200 km in diameter. There have been six successful manned missions to the Moon, and many unmanned probes have been sent there. This article presents a general survey of the Moon, and summarises what has been learned from the lunar space-craft.

Origin

The Moon is so close to the Earth that even with the naked eye the surface markings are obvious. Physical and orbital data are given in Table 1.

For many years it was believed that the Earth and the Moon were one body, and that rapid rotation resulted in a portion being flung off to form the Moon – leaving the hollow now filled by the Pacific Ocean. This theory has long since been rejected, and only two serious theories remain. It is possible that the Earth and the Moon were formed at the same time and in the same region of the 'solar nebula', the cloud of material surrounding the young Sun; certainly the Earth and Moon are of the same age – about 4.6 thousand million years. However, most authorities now favour the 'giant impact' theory; the original

Table 1 Lunar data

Distance from Earth centre to centre:	max 406 697 km (apogee) mean 384 400 km min 356 410 km (perigee)
surface to surface:	max 398 581 km (apogee) mean 376 284 km min 356 410 km (perigee)
Orbital period:	27.321661 days
Axial rotation period:	synchronous
Synodic period:	29.53 days (29 d 12 h 44 m 2 s.9)
Mean orbital velocity:	1.023 km/s
Orbital eccentricity:	0.0549
Mean orbital inclination:	5° 9′
Diameter:	
equatorial	3746 km
polar	3470 km
Oblateness:	0.002
Mean apparent diameter from Earth:	31′5″
Reciprocal mass, Earth-1:	81.301
Density, water-1:	3.342
Volume, Earth-1:	0.0203
Escape velocity:	2.38 km/s
Surface gravity, Earth-1:	0.1653
Mean albedo:	0.067
Atmospheric density:	10^{-14} that of the Earth's air at sea-level.

body was struck by a 'planetary-sized body larger than Mars', so that the Moon condensed from the debris ejected during the collision.

Movements and Rotation

It is usually said that 'the Moon revolves round the Earth'. In fact the two bodies revolve together round the barycentre (the centre of gravity of the Earth–Moon system), but as the barycentre lies 1700 km below the Earth's surface the simple statement is good enough for most purposes.

The lunar orbit is not circular; the distance from Earth (centre to centre) ranges between over 400 000 km at furthest recession (apogee) and less than 360 000 km at closest approach (perigee). The orbital period is 27.32 days, but this is not the same as the synodic period, or interval between successive full moons or successive new moons, because the two bodies are moving together around the Sun; the synodic period is 29.53 days. It is often said that the Moon is 'new' when it appears as a slender crescent in the evening sky, but this is not strictly true; new moon occurs when the Moon lies between the Sun and the Earth, and its dark side is facing us, so that the actual new moon cannot be seen at all unless it passes directly in front of the Sun and produces a solar eclipse. A solar eclipse does not happen every month, because the lunar orbit is inclined at an angle of over 5°, and most new moons pass unseen either above or below the Sun.

During the crescent stage, the unlit hemisphere can usually be seen shining dimly. This is because of 'Earthshine' – light reflected on to the Moon from the Earth.

The axial rotation period of the Moon is exactly the same as its orbital period, so that the Moon always keeps the same hemisphere turned toward the Earth, and part of the surface is permanently turned away from us. This is not mere coincidence, it is the result of tidal effects over the ages. Originally the Moon spun much more quickly, but the tidal pull of the Earth slowed it down until the rotation had become 'captured' or synchronous. Note, however, that the Moon does not keep the same hemisphere turned sunward, so that day and night conditions there are the same all over the globe – apart from the fact that from the far side, the Earth can never be seen.

From Earth it is in fact possible to see 59% of the Moon's surface, though of course no more than 50% at any one time. This is because the Moon rotates at a constant speed, but the orbital velocity varies; the Moon moves quickest when near perigee. This means that the rotation and the position in orbit become 'out of step', and the Moon seems to rock very slowly to and fro. This libration in longitude means that narrow zones are brought alternately in and out of view. There are other minor librations, and only 41% of the surface is visually inaccessible, so that before the Space Age nothing definite was known about it.

Because of tidal effects, the Moon is receding from the Earth at the rate of 3.8 centimetres per year, and on average the Earth's rotation period is lengthening at a rate of 0.0000002 second per day.

Structure and Atmosphere

The outer surface layer of the Moon is termed the regolith; it is a loose layer or debris blanket probably, up to 10 metres deep in places, continually churned up by the impacts of micrometeorites (Figure 1). It is often referred to as 'soil', but this is misleading, as it contains no organic material. It overlies the rocky crust, which is on average just over 60 km in depth – thicker on the far side of the Moon than on the Earth-turned hemisphere. However, no unbrecciated rock outcrop has been encountered on Apollo missions and it remains uncertain how deep the zone of brecciation penetrates. At a fairly shallow level there are areas of denser material, which have been located because they affect the movements of orbiting spacecraft, they are called mascons (mass concentrations), and lie below some of the regular maria and very large basins. Below the crust lies the mantle, the structure of which may be fairly uniform; at a depth of around 1000–1200 km there seems to be a region where the rocks are hot enough to be molten. Finally, there may be a metallic core, no more than 1000 km in diameter; results from the Lunar Prospector spacecraft of 1998–1999 led to an estimate of an iron-rich core between 440 and 900 km in diameter. Most of what we know about the lunar interior comes from studies of moonquakes, which do occur and have been recorded by instruments left on the surface by the Apollo astronauts; some are shallow, but most occur in a zone from 800 to 1000 km below the surface. By terrestrial standards they are very mild, and never exceed a value of 3 on the Richter scale (*see* **Tectonics:** Earthquakes).

The Moon's low escape velocity means that there is only a trace of atmosphere, made up chiefly of helium and argon. If the entire atmosphere were compressed to a density equal to that of the Earth's air at sea-level, it would just about fill a cube with a side length of 65 metres. There is no detectable magnetic field, though the remnant magnetism of some rocks indicates that a definite field may have existed over 3.5 thousand million years ago.

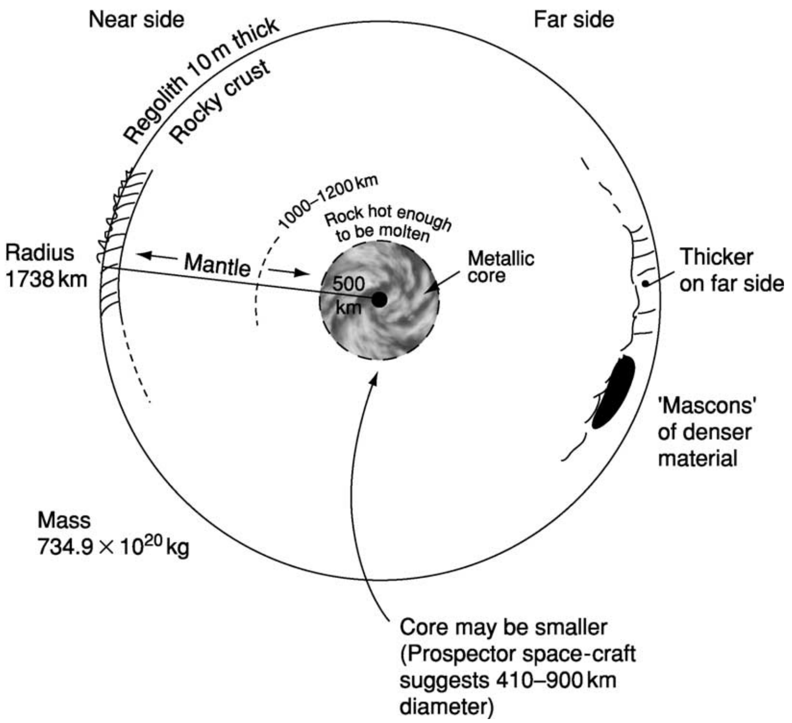


Figure 1 Structure of the Moon.

Table 2 Selected list of successful lunar missions

	Launch date	*Landing date*	
Name	*United States*		*Area*
Ranger 7	28 July 1964	31 July 1964	Mare Nubium hard lander, 4306 images returned.
Surveyor 1	30 May 1966	2 June 1966	Mare Nubium, near Flamsteed. Controlled landing. 11/150 images returned.
Surveyor 2	7 Jan 1968	9 Jan 1968	Nrim of Tycho. Controlled landing: 21 274 images
Orbiter 1	10 Aug 1966	–	207 images. (Uncontrolled impact, 29 Oct 1966.)
Orbiter 5	1 Aug 1967	–	212 images returned. (Impact, 31 Jan 1968.)
Clementine	25 Jan 1994	–	Mapping. Left orbit 3 May 1994.
Prospector	6 Jan 1998	–	Mapping, analysis. Uncontrolled impact, 11 July 1999.
Apollo manned missions			
Number	*Landing date*	*Area*	*Astronauts*
11	19 July 1969	M. Tranquillitatis	N. Armstrong, E. Aldrin, J. Collins
12	19 Nov 1969	Oc. Procellarum	C. Conrad, A. Bean, R. Gordon
14	2 Feb 1971	Fra Mauro	A. Shepard, E. Mitchell, S. Poosa
15	10 July 1971	Hadley-Apennines	D. Scott, J. Irwin, A. Worden
16	21 Aug 1972	Descartes formation	J. Young, C. Duke, T. Mattingly
17	11 Dec 1972	Taurus–Littrow	E. Cernan, H. Schmitt, R. Evans

Lunar Missions

Many space-craft have been sent to the Moon, quite apart from the manned Apollo missions. A list of the most important probes is given in Table 2.

The first successful missions were Russian, in 1959; in October of that year Luna 3 sent back the first images of the far side, always turned away from the Earth. On 3 February 1966, Luna 9 made the first controlled landing, showing that the surface was firm and disposing of an earlier theory that the maria at least were covered with deep dust. Controlled landings were made by the American Surveyors

(1966–1968) and the entire surface was mapped by the five Orbiters (1966–1967). The Apollo Programme extended from 1968 to 1972; of the seven planned landings, only Apollo 13 was unsuccessful. The Russians sent two automatic 'rovers', the Lunokhods (1970 and 1973). The latest lunar mapping probes have been the American Clementine (launched 1994) and Prospector (launched 1998, deliberately crashed on to the surface in 1999). Altogether 382 kg of samples have been returned to Earth, mainly by the Apollo astronauts but with small amounts from four Russian sample-and-return missions.

Surface Features

The most obvious features are, of course, the 'seas' (maria), which cover about 17% of the surface. A list of the major maria is given in Table 3. Most of these form a connected system, though the well-defined Mare Crisium is separate. Some of the maria are fairly regular in outline; others are very irregular. The largest of the regular maria is the Mare Imbrium, with a diameter of over 1000 km, bounded in part by the mountain ranges of the Apennines, Alps, and Carpathians, though the irregular Oceanus Procellarum has a longest diameter of over 2500 km and an area of well over 2 million square km. Of the major seas, only the Mare Orientale extends on to the far hemisphere of the Moon; there are no large maria wholly on the areas never visible from Earth. A small part of the Mare Orientale can be seen from Earth under favourable conditions of libration.

The whole lunar scene is dominated by craters, ranging from tiny pits up to huge enclosures well over 200 km in diameter (Figure 2). Basically they are circular, though often distorted by later formations, and near the limb they are foreshortened so much that it is sometimes hard to distinguish a crater wall from a ridge. A typical crater has a sunken floor, often with a central peak, and walls which rise to only a modest height above the outer surface. In profile, a lunar crater resembles a saucer rather than a steep-sided mine-shaft, and large formations would be better known as walled plains. Some have dark, relatively smooth floors, such as Plato in the region of the Alps (Figure 3), others have massive central structures which, however, never equal the height of the surrounding rampart. Some craters, notably Tycho in the southern uplands and Copernicus in the Oceanus Procellarum, are the centres of systems of bright rays which extend for hundreds of kilometres, and cross all formations in their path. A crater is at its most spectacular when seen at the terminator (the boundary between the sunlit and night hemispheres), as its floor will be wholly or partly filled with shadow. Near full moon there are almost no shadows, and the scene is dominated by bright rays, so that even a large crater will be difficult to identify unless it has a dark floor or very bright walls.

Large craters are often found in chains, such as the prominent. Ptolemæus, Alphonsus, and Arzachel, near the centre of the Earth-turned hemisphere, and Theophilus, Cyrillus, and Catharina, near the border of the Mare Nectaris. Chains of small craters are

Table 3 Selected list of Lunar maria

Name		*Diameter, km*	
Mare Australe	Southern Sea	600	Irregular, patchy area near SE limb
Mare Crisium	Sea of Crises	500	Well-defined, separate
Mare Fœcunditatis	Sea of Fertility	900	Irregular; confluent with M. Tranquillitatis
Mare Frigoris	Sea of Cold	1600	Elongated, irregular, in places narrow
Mare Humboldtianum	Humboldt's Sea	270	Limb sea beyond Endymion
Mare Humorum	Sea of Humours	390	Regular; leads off Mare Nubium
Mare Imbrium	Sea of Showers	1120	Regular: area 863 000 sq km.
Mare Nectaris	Sea of Nectar	330	Leads off Mare Tranquillitatis. Fairly regular
Mare Nubium	Sea of Clouds	750	Ill-defined border
Mare Orientale	Eastern Sea	340	Vast ringed structure; mainly on far side
Oceanus Procellarum	Ocean of Storms	2570	Irregular
Mare Serenitatis	Sea of Serenity	700	Regular. Few craters
Mare Smythii	Smyth's Sea	370	Well-defined limb sea
Mare Tranquillitatis	Sea of Tranquillity	870	Adjoins M. Serenitatis
Mare Vaporum	Sea of Vapours	250	SE of the Apennines
Lacus Mortis	Lake of Death	150	Adjoins Lacus Somniorum
Lacus Somniorum	Lake of the Dreamers	380	Irregular darkish area leading off Mare Serinitatis
Palus Putredinis	Marsh of Decay	160	Part of Mare Imbrium
Palus Somnii	Marsh of Sleep	290	Curiously-coloured area near Mare Crisium
Sinus Æstuum	Bay of Heats	290	Fairly dark area leading off Mare Nubium
Sinus Medii	Central Bay	260	Almost central on the disk
Sinus Roris	Bay of Dew	400	Joins Mare Frigoris to Oceanus Procellarum

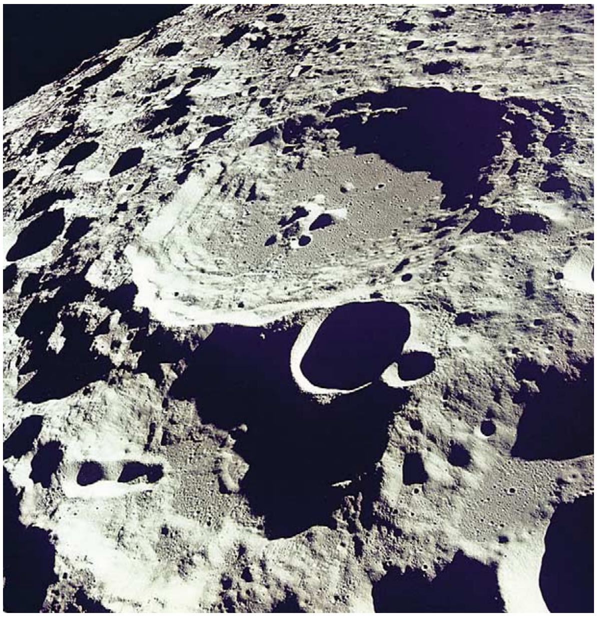


Figure 2 Craters on the 'far side' of the Moon. The hemisphere always turned away from Earth is as crater-scarred as the familiar hemisphere, though it lacks 'seas' similar in type to Mare Imbrium.

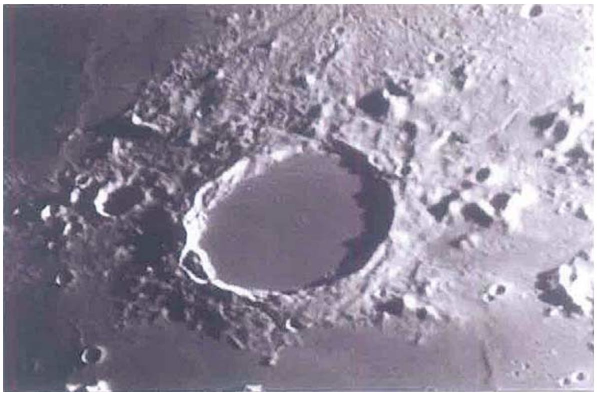


Figure 3 Plato, one of the most distinctive craters on the Moon. It is 109 km in diameter, and very regular. Its very dark grey floor makes it easy to locate whenever it is sunlit.

common, and so are crater-pairs, sometimes separated and sometimes joined. The brightest crater is the 40 km Aristarchus, which has even been mistaken for a volcano in eruption; the huge walled plain, Grimaldi, near the western limb, has the darkest floor.

The system of naming craters after people (usually, though not always, astronomers) was introduced by the Italian Jesuit Riccioli, who drew a lunar map in 1651. The system has been extended since, and is universally accepted. A selected list of some prominent craters is given in Table 4. One crater, the 84 km Wargentin, is filled with lava, so that it is in effect a large plateau.

Other features of the surface include wrinkle-ridges, crossing the maria; valleys, such as the magnificent valley which cuts through the Alps near Plato; low swellings or domes, with gentle slopes and often with summit pits, together with isolated peaks and the crack-like features known as rills, rilles, or clefts. Of special interest is the Straight Wall, in the Mare Nubium. In fact, it is not a wall, but a fault, appearing dark before full moon, because it casts a shadow, and bright after full moon, when its inclined face is illuminated.

Lunar Rocks

All the rocks brought back for analysis are breccias of igneous rocks; the Apollo astronauts brought back 2196 samples (Figure 4) (*see* **Analytical Methods:** Geochronological Techniques). Sedimentary and metamorphic rocks were absent. In the lavas, basalts are dominant; the youngest has been given a radiometric age of 3.08 thousand million years, while the oldest dates back 3.85 thousand million years. The basalts contain more titanium than terrestrial lavas – over 10% in the Apollo 11 samples – and there are small amounts of metallic iron. Many lunar rocks are also comparatively low in sodium and potassium, but one particular type of basalt is rich in potassium (chemical symbol K), the Rare Earth elements, and phosphorus (P), so that it is known as KREEP. Anorthosite – rock containing the minerals plagioclase, pyroxene, and/or olivine in various proportions – is plentiful; one specimen collected by the Apollo 15 astronauts is radiometrically dated as 4 thousand million years old, and is white. It is known as the Genesis Rock! A sample collected from Apollo 12 is about the size of a lemon; it consists largely of SiO_2, and is rich in uranium, potassium, and thorium, making it exceptionally radioactive. It is composed of a dark grey breccia, a light grey breccia, and a vein of solidified lava (Figure 5).

It is now known that some meteorites found on Earth have come from the Moon. Most are breccias, of the same type as mare basalts the anarthosites of the highlands or the regolith. Such lunar-sourced meteorites have been found in Antartica, Australia, North Africa and Oman (*see* **Solar System:** Meteorites). The subsurface rocks of the Moon have so far proved elusive to manned exploration, but lunar meteorite Dhofar 310 recovered from the Oman desert is reported by S.I. Demidova and others in 2003, to be a polymict breccia with deep-seated lunar crustal

Table 4 Some important craters

Name	*Latitude*	*Longitude*	*Diameter, km*	
Albategnius	11.7 S	4.3 E	114	Adjoins Hipparchus
Alphonsus	13.7 S	3.2 W	108	Ptolemæus chain
Anaxagoras	73.4 N	10.1 W	50	Ray-centre
Archimedes	29.7 N	4.0 W	82	On Imbrium; trio with Aristillus, Autolycus
Aristarohus	23.7 N	47.4 W	40	Brilliant walls and central peak
Aristillus	33.9 N	1.2 E	55	Archimedes group
Aristotle	50.2 N	17.4 E	87	Pair with Eudoxus
Arzachel	18.7 S	1.9 W	96	Archimedes group
Autolycus	30.7 N	1.5 E	39	Archimedes group
Bailly	66.5 S	69.1 W	287	Field of ruins, near S. limb
Bessel	21.8 N	17.9 E	15	Bright; on Serenitatis
Bullialdus	20.7 S	22.2 W	60	On Nubium; massive walls, central peak
Catharina	18.1 S	23.4 E	104	Theophius group
Clavius	58.8 S	14.1 W	245	Southern highlands
Copernicus	9.7 N	20.1 W	107	Great ray-crater
Cyrillus	13.2 S	24.0 E	98	Theophius group
Democritus	62.3 N	35.0 E	39	Highlands N of Mare Frigoris
Dionysius	2.8 N	17.3 E	18	Brilliant crater on edge of Tranquillitatis
Doppelmeyer	28.5 S	41.4 W	63	Bay leading off Humorum
Encke	4.7 N	36.6 W	28	On Procellarum; Kepler area
Endymion	53.9 N	57.0 E	123	Near Humboldtianum; darkish floor
Eratosthenes	14.5 N	11.3 W	58	End of Apennines
Fra Mauro	6.1 S	17.0 W	101	On Nubium, group with Bonpland, Parry
Fracastorius	21.5 S	33.2 E	112	Great bay off Nectaris
Gassendi	17.6 S	40.1 W	101	Edge of Humorum
Grimaldi	5.5 S	68.1 W	172	W of Procellarum, very dark floor
Hercules	46.7 N	39.1 E	69	Pair with Atlas
Hevel	2.2 N	67.6 W	115	Grimaldi chain
Hipparchus	5.1 S	5.2 E	138	Pair with Albategnius
Hyginus	7.8 N	6.3 E	9	Depression in Vaporum; great crater-rill
Janssen	45.4 S	40.3 E	199	Southern Uplands, rim broken by Fabricius
Kepler	8.1 N	38.0 W	31	In Procellarum; ray-centre
Langrenus	8.9 S	61.1 E	127	Patavius chain
Longomontanus	49.6 S	21.8 W	157	Clavius area
Macrobius	21.3 N	46.0 E	64	Crisium area
Maginus	50.5 S	6.3 W	194	Clavius area
Maurolycus	42.0 S	14.0 E	114	Stöfler group
Menelaus	16.3 N	16.0 E	26	Edge of Serenitatis; brilliant
Moretus	70.6 S	5.8 W	111	Southern uplands
Newton	76.7 S	16.9 W	78	Moretus area
Olbers	7.4 N	75.9 W	74	Grimaldi area, ray-centre
Petavius	25.1 S	60.4 E	188	Langrenus chain
Phocylides	52.7 S	57.0 W	121	Schickard area
Piccolomini	29.7 S	12.7 E	47	End of Altai Scarp
Pitatus	29.9 S	13.5 W	108	Sedge of Nubium; passes connected with Hesiodus
Plinius	15.4 N	23.7 E	43	Between Serenitatis and Tranquillitatia
Posidonius	31.8 N	29.9 E	95	Edge of Senenitatis
Ptolemæus	9.3 S	1.9 W	464	Trio with Alphonsus and Arzachel
Purbach	25.5 S	2.3 W	115	Walter group
Pythagoras	63.5 N	63.0 W	142	NW of Iridum
Riccioli	3.3 S	74.6 W	139	Adjoins Grimaldi; dark patches on floor
Rømer	25.4 N	36.4 E	39	Taurus area
Schickard	44.3 S	55.3 W	206	Great walled plain
Schiller	51.9 S	39.0 W	180 × 97	Schickard area; fusion of two rings
Stadius	10.5 N	13.7 W	60	‘Ghost ring’ near Copernicus
Stevinus	32.5 S	54.2 E	74	Petavius area; pair with Snellius
Taruntius	5.6 N	46.5 E	56	On Fœcunditatis
Thales	61.8 N	50.3 E	31	Near Strabo; ray centre
Theophilus	11.4 S	26.4 E	110	Trio with Gyrillus and Catharina
Triesnecker	4.2 N	3.6 E	26	Vaporum area; great rill system

Continued

Table 4 Continued

Name	*Latitude*	*Longitude*	*Diameter, km*	
Tycho	43.4 S	11.1 W	102	Southern highlands; brightest ray-centre
Vendelinus	16.4 S	61.6 E	131	Petavius chain
Walter	33.1 S	1.0 E	128	Trio with Regiomontanus and Purbach
Wargentin	49.6 S	60.2 W	84	Schickard area; the famous plateau
Zucchius	61.4 S	50.3 W	64	Schiller area; pair with Segner

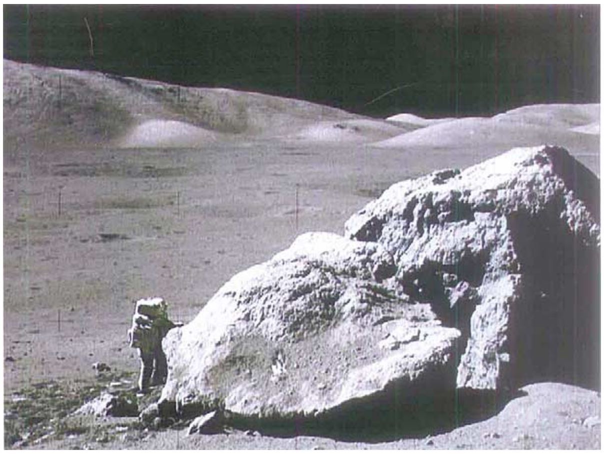


Figure 4 Astronaut Schmitt, of Apollo 17, standing by a huge boulder. Schmitt was a professional geologist who had been trained as an astronaut specially for the mission; December 1972, He and his companion, Eugene Cernan, are (so far) the last men to have been to the Moon. Courtesy of NASA.

material within it as clasts – granulites and igneous rocks of anorthosite, gabbronorite and troctolite composition, also minor dunite and pyroxenite. A unique Al-rich orthopyroxenite/Al-spinel clast is compatible with pressure at a depth of ~20 km within the lunar crust at its source.

Origin of the Craters

It was long believed that the craters were of volcanic origin, similar in type to terrestrial calderæ, but it is now widely accepted that the vast majority are of impact origin, and we have at least a reasonable idea of the sequence of events.

When the Moon came into existence as a separate body, the heat generated melted the outer layers, and for a time there must have been a magma ocean many kilometres deep. Eventually, a crust was formed, thicker on the far side of the Moon than on the Earth-facing side; by cosmic standards it did not take long for the axial rotation to become synchronous. At that stage there was a vast amount of debris moving round the Sun, and the newly-formed planets and satellites swept it up. Between 4400 and 3900 million years ago came the Great Bombardment, when meteorites rained down on the Moon to produce the first major basins such as the Mare Tranquillitatis. Then, between 3900 and 3800 million years ago, came the tremendous impact which resulted in the Imbrian basin and affected the whole of the Moon. As the Great Bombardment ceased there was widespread vulcanism, with magma pouring out from below the crust and flooding the basins to produce structures such as the Mare Orientale. The lava flows ended rather suddenly; as the outpouring slackened, many craters were left undamaged, so that the youngest, such as Tycho and Copernicus, are unflooded. On the far hemisphere, with its thicker crust, there was less flooding, which explains the absence of major Maria and the presence of large,

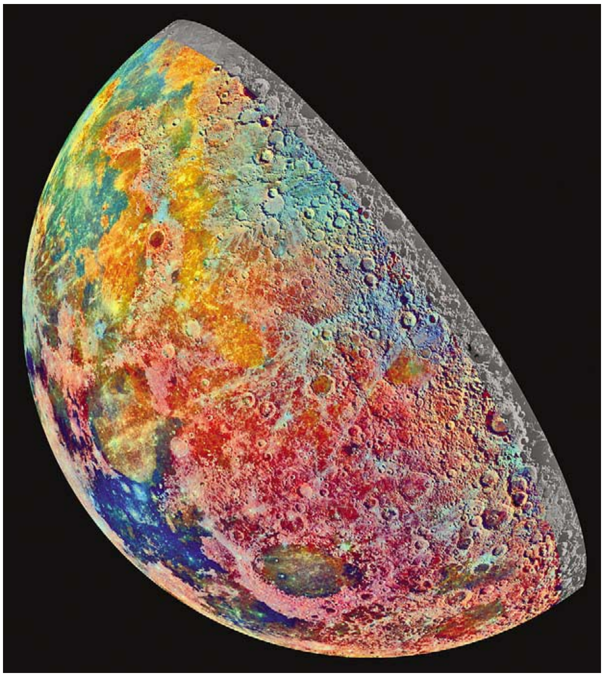


Figure 5 The minerals of the Moon. This is a mosaic of 53 images, obtained in 1992 by the Galileo space-craft. The exaggerated false colour shows the differences in surface structure. Blue to orange indicate volcanic lava showing the dark blue Mare Tranquillitatis (lower left) is rich in titanium. Near the bottom of the image, right at Mare Tranquillitatis, is Mare Crisium, pink colours indicating material of the lunar highlands.

light-floored walled plains known as palimpsests. Since then, the Moon has seen little activity, though it has been claimed that Copernicus is no more than a thousand million years old and Tycho even younger. One thing is certain: the dinosaurs would have seen the Moon looking very much as it does today!

The lunar cratered surface is remarkably like that of the planet Mercury (*see* **Solar System:** Mercury).

Transient Lunar Phenomena (TLP)

Many observations have been made of elusive glows and obscurations in parts of the Moon, notably in and near the brilliant crater Aristarchus. The reality of these events was demonstrated in 1992, when A. Dollfus, using the 83 cm refractor at the Meddon Observatory (Paris) saw and photographed moving glows inside the large walled plain Langrenus (Figure 6). TLP are almost certainly due to dust lifted above the surface under the effect of gas escaping from below the crust.

Ice on the Moon?

Some of the polar craters are deep by lunar standards, and their floors are always in shadow, so that they remain very cold indeed. The temperatures may be as low as −230°C. In 1966, NASA made the unexpected announcement that ice had been found in the bottoms of these deep polar formations.

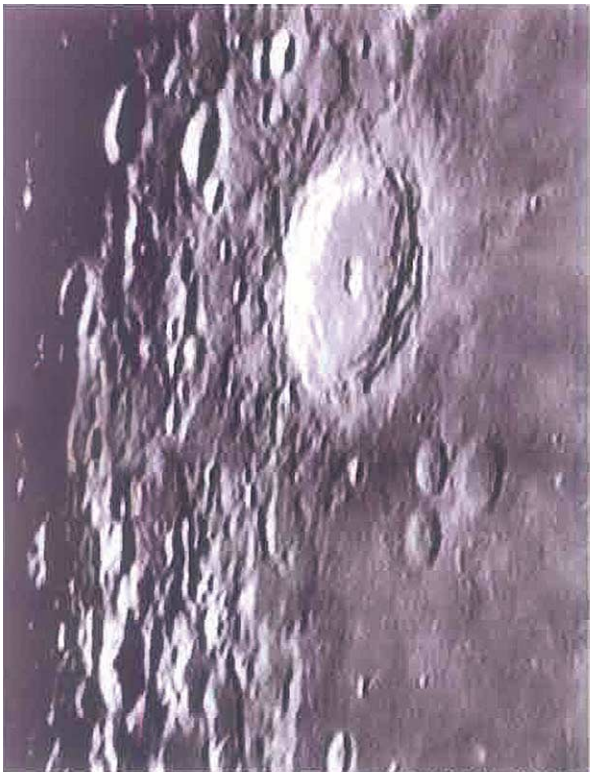


Figure 6 The lunar crater in Langrenus, diameter 127 km, with high walls and a central peak. It is one member of a large chain of craters, including Petavius and Vendelinus.

The results came from an unmanned probe, Clementine, which had been orbiting the Moon since 1994. It carried a neutron spectrometer, reported effects which indicated the presence of hydrogen, which could combine with oxygen to produce water. Extravagant claims were made, and one NASA scientist went so far as to comment that there are a bunch of craters filled with water ice could a significant resource that would allow a modest amount of colonization for many years. Water would now be mined directly on the Moon instead of having to be shipped from Earth. Yet how could the ice have got there? Rock samples had shown no sign of hydrated material, and the impact of an icy comet would have generated a great deal of heat. Similar results were announced from the next probe, Prospector, launched into lunar orbit in 1998, but many authorities were sceptical. Finally, on 31 July 1999, a test was carried out. Prospector had come to the end of its career, and was deliberately crashed into a polar crater to see whether any signs of water could be found. Predictably, the results were negative. Further tests of the same kind are being planned, but it must be said that the idea of lunar ice seems decidedly far-fetched.

Life on the Moon?

There seems no chance that life has ever appeared on the Moon; it has been sterile throughout its long history. The crews of the first two Apollos, 11 and 12, were quarantined on return from the Moon to make sure that they had brought back nothing harmful, but quarantining was then abandoned as being unnecessary.

Occultations

As the Moon moves, it may pass in front of a star and hide or occult it. The star shines steadily until it is covered. When it snaps out abruptly; this was one of the early proofs that the Moon has practically no atmosphere.

Eclipses

Eclipses of the Moon are caused by the Moon's entry into the cone of shadow cast by the Earth. The supply of direct sunlight is cut off, but in general the Moon does not disappear, because some sunlight is refracted on to it by the layer of atmosphere surrounding the Earth; the Moon becomes dim and often coppery in colour.

Table 5 Lunar eclipses, 2004–2020

Date	*Type*	*Time of mid-eclipse, GMT*	*Duration of totality minutes*	*Percentage of Moon eclipsed*
2004 May 4	Total	20.32	38	100
2004 Oct 28	Total	03.05	40	100
2005 Oct 17	Partial	12.04	–	6
2006 Sept 7	Partial	18.52	–	18
2007 Mar 3	Total	23.22	37	100
2008 Feb 21	Total	03.27	24	100
2008 Aug 16	Partial	21.11	–	81

Table 6 The Danjon scale

0	Very dark; Moon almost invisible at totality
1	Dark grey or brownish, details barely identifiable
2	Dark or rusty red, with a dark patch in the middle of the shadow: brighter edges
3	Brick red, sometimes a bright or yellowish border to the shadow
4	Coppery or orange-red; very bright, with a bluish cast and varied hues

Lunar eclipses may be either total or partial. Because the Sun is a disk, not a point source of light, there is an area to either side of the main shadow cone (the umbra), through which the Moon has to pass; this region is termed the penumbra, and produces only a slight dimming of the surface. Some eclipses are penumbral only. The last eclipses were those of 16 May and 9 November 2003; a list of umbral eclipses up to 2008 is given in Table 5.

No two eclipses are alike; everything depends upon conditions in the Earth's upper air through which the sunlight has to pass. If there is an unusual amount of dust, following an event such as a volcanic eruption, the eclipses is liable to be dark. Observers use a scale given by A. Danjon (Table 6).

Lunar eclipses do not happen at every new moon, because of the inclination of the lunar orbit. Astronomically, they are not important, but they are certainly beautiful to watch, some displaying a wonderful range of hues.

See Also

Analytical Methods: Geochronological Techniques. **Earth Structure and Origins**. **Solar System:** Meteorites; Mercury; Jupiter, Saturn and Their Moons; Neptune, Pluto and Uranus. **Tectonics:** Earthquakes. **Tektites**.

Further Reading

Harland B (1977) *Exploring the Moon*. Springer/Praxis.

Lindsay H (2001) *Tracking Apollo to the Moon*. New York, Berlin: Heidelberg, Springer-Verlag.

Massey H, *et al.* (eds.) (1997) *The Moon: A New Appraisal*. London: The Royal Society.

Moore P (2002) *Patrick Moore on the Moon*. London: Cassell.

Schultz P (1976) *Moon Morphology*. University of Texas Press.

Shepard A and Slayton D (1994) *Moon Shot*. Turner Publishing Inc.

Taylor S (1975) *Lunar Science*. Oxford: Pergaman Press.

Wilhelus E (1999) *To a Rocky Moon*. Arizona: University of Arizona Press.

Wood CA (2003) *The Modern Moon*. Sky Publishing Corp.

Westfall JE (2000) *Atlas of the Lunar Terminator*. Cambridge: Cambridge University Press.

Rukl A (2001) *Atlas of the Moon*. London: Hamlyn.

Mars

M R Walter, A J Brown, and S A Chamberlain,
Macquarie University, Sydney, NSW, Australia

Introduction

Mars, the fourth planet from the Sun, is Earth's second closest planetary neighbour. As a nearby terrestrial planet with an atmosphere, it shares some geological processes with Earth. However, our impressions of Mars are most clearly drawn into perspective by considering the differences between the geology of Earth and the geology of Mars.

Mars has an elliptical orbit, ranging from 207 to 249 million km from the Sun. Earth is much closer to the Sun (147 to 152 million km) and therefore is warmer. It takes 687 Earth days for Mars to orbit the Sun (670 Martian days, or 'sols'). A day on Mars lasts 24 h and 37 min. The diameter of the planet is 6780 km, about half that of Earth. Its surface area is about the same as that of the land area on Earth. At present, its axis is inclined at 25° to the ecliptic (the plane of rotation around the Sun), much like Earth's, and so it has similar seasons. Compared to the northern hemisphere, southern-hemisphere winters on Mars are more intense; springs and summers are shorter but, because they occur when the planet is closer to the Sun, peak temperatures are as much as 30°C higher than northern-hemisphere spring and summer temperatures are. This is demonstrated by the fact that the southern polar cap has been known to disappear entirely during southern Summer, though the northern polar cap has never done so. Southern hemisphere winters on Mars are longer and colder than those of the northern hemisphere since Mars is further from the Sun during this period.

Interior of Mars

Some of the physical characteristics of Mars are listed in Table 1. An interpretation of the interior of Mars is presented at Figure 1. The interior structure of Earth is well constrained due to our ability to measure seismic reflections throughout the mantle and into Earth's core. At present, without similar seismic information on Mars, it is possible only to postulate what the Martian interior may be like – for example, does it have a solid core, or a partly liquid core like Earth has? The current view is that Mars has a liquid outer and a solid inner core with a radius of approximately 1700 km. The resultant reduced mantle depth compared to Earth implies that the Martian interior is likely to have cooled substantially faster than Earth did, and convection of the mantle, if it ever occured, may now have slowed or ceased. This implies that volcanism on the Martian surface is less likely today, and may provide an explanation for the current lack of a strong magnetic field on Mars.

The Mars Global Surveyor orbiter did find enigmatic strong local magnetism in rocks of the southern highlands, but an extremely weak overall global magnetic field on Mars. The absence of crustal magnetism near large impact basins such as Hellas and Argyre implies cessation of internal dynamo action during the Early Noachian epoch (similar to 4 billion years ago). Although massive tectonic events most probably caused the formation of Valles Marineris, plate tectonics as is understood on Earth is unlikely to have played a major role in Martian geology. This implies that granites, which are dependent on recycling of the crust in the presence of water, are unlikely to form on Mars. To date, this has been borne out by spectroscopic investigations of the planet.

Table 1 Physical characteristics of Earth and Mars

Characteristic	*Mars*	*Earth*
Orbit (million km)	207–249	147–152
Year (in Earth days)	687	356.25
Day (hours)	24.6	23.9
Mean radius (km)	3390	6371
Core radius (km)	~1700	3485
Average mantle depth (km)	1690	2886
Present obliquity to orbit (degrees)	25.19	23.45
Surface temperature variations (degrees C)	−100 to +17	−82 to +54

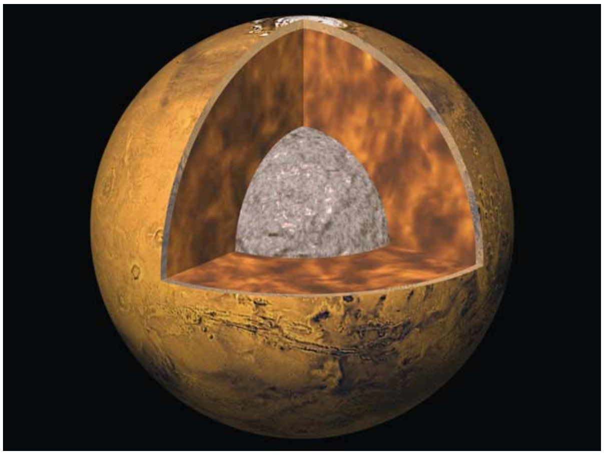


Figure 1 An interpretation of the interior of Mars, showing the core, mantle, and crust. Image courtesy of Calvin J. Hamilton.

Martian Atmosphere and Aeolian Processes

The atmosphere is 95% carbon dioxide, 2.7% nitrogen, 1.6% argon, and 0.13% oxygen, plus minute traces of other gases. In contrast, Earth's atmosphere is 78.1% nitrogen, 20.9% oxygen, 0.93% argon, and 0.03% carbon dioxide. Mars' atmosphere is thin, with the pressure at the surface of the planet (5.6 mbar) being less than one-hundredth of that on Earth. Despite the thin atmosphere of Mars, aeolian (wind-driven) processes have played a large role in shaping many surface features on Mars. These include dunes, yardangs, and etched and eroded terrains. Active dust devils have even been observed by the Mars Orbiter Camera on Mars Global Surveyor (Figure 2). Due to the low gravity of Mars, these features have greater height than their counterparts on Earth have. Dust, blown by the wind over the entire planet, tends to blanket rock features with a homogeneous layer that hampers identification of rock outcrops by orbital and telescopic spectroscopic methods. Dust storms, which periodically turn the Martian atmosphere into an opaque red layer, appear to be coupled with the heating of the atmosphere

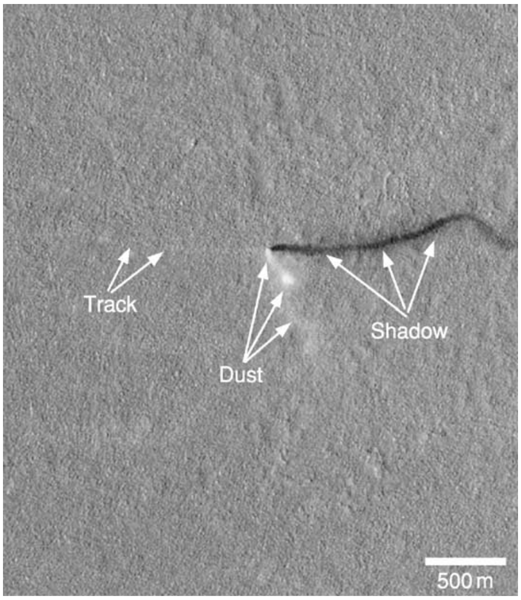


Figure 2 An image of a dust devil taken by the Mars Orbiter Camera. NASA/JPL/Malin Space Science Systems.

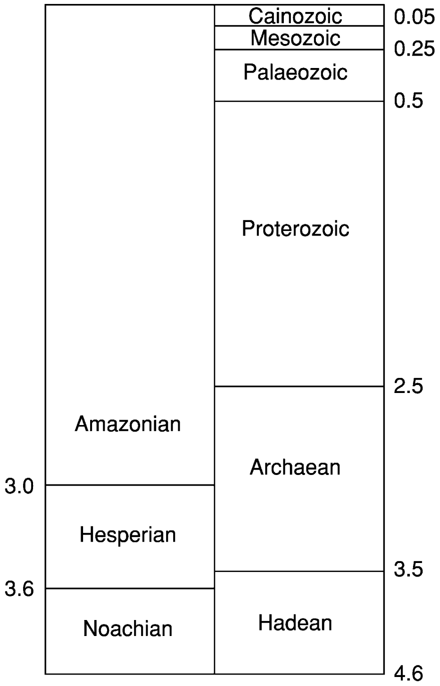


Figure 3 Geological time-scales on Mars (left) and Earth (in billions of years).

by the Sun as Mars approaches perihelion (i.e., the closest part of its elliptical orbit).

Obliquity and Climate Variations

It has been calculated that over the past 10 million years, the angle of the spin axis of Mars (the 'obliquity') to the ecliptic, the plane of rotation around the Sun, has ranged from 13° to 4°. The obliquity varies chaotically, on a time-scale of hundreds of thousands to millions of years. In contrast, the obliquity of Earth is stabilized by the presence of the Moon. When the obliquity of Mars is at a minimum, the poles would have permanent caps of frozen carbon dioxide, because, as on Earth, little of the Sun's warmth would reach the poles; when the obliquity of Mars is at a maximum, the polar caps would melt in summer. At times of high obliquity, the water and carbon dioxide stored at the poles would vaporize and be released into the atmosphere, possibly raising the pressure to high enough levels to make liquid water stable for short times.

Cratering Record

The lack of a thick atmosphere and active fluvial processes relative to Earth gives Martian impact craters great preservation potential. Assuming no preference for the location of impact sites, observations of the distributions of craters on the Martian surface can be used to date the surface units. The great Martian impact basins of Hellas and Argyre are thought to have formed early in the history of the Solar System. By correlating the crater distributions with those seen on the Moon (assuming Mars and the Moon have similar impact histories), it is possible to assign ages to regions of the Martian crust, providing a rough guide to when areas of Mars were last resurfaced. Figure 3 relates the Martian geological time-scale to the major geological time periods of Earth. All values are approximate and are given in billions of years. It should be noted that since the current dating of the Martian surface depends on interpretation of impact craters, and these may be incorrect for several reasons, chief amongst them the possibility of one impact causing multiple craters, and also that volcano calderas may be mistaken for craters.

Global Hemispheric Dichotomy and Crustal Thickness

The analysis of images returned by the Viking orbiters of the 1970s revealed a hemispheric dichotomy on Mars. When these results were coupled with those of the Mars Global Surveyor laser altimeter instrument in the late 1990s, a picture of two halves of Mars emerged. The young and smooth northern lowlands have developed separately from the old, cratered southern highlands. The origin of this

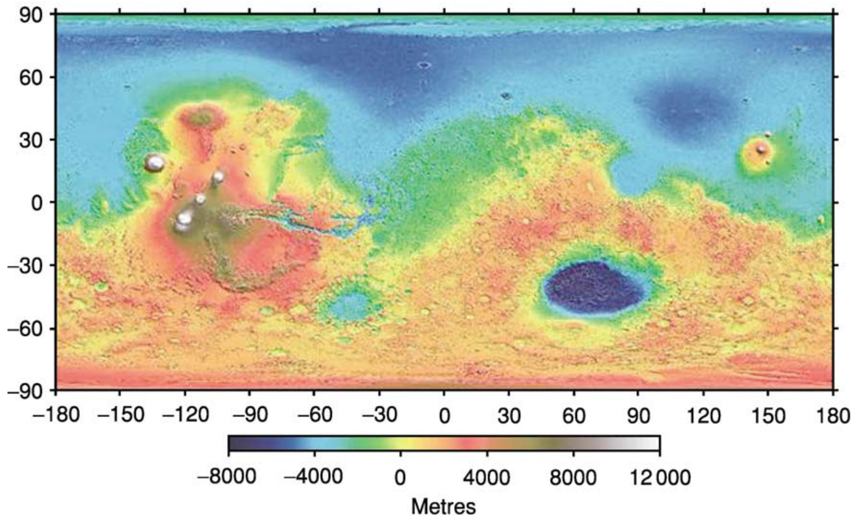


Figure 4 The elevation map of the topography of the Martian surface as determined by the Mars Orbital Laser Altimeter. Image courtesy of NASA.

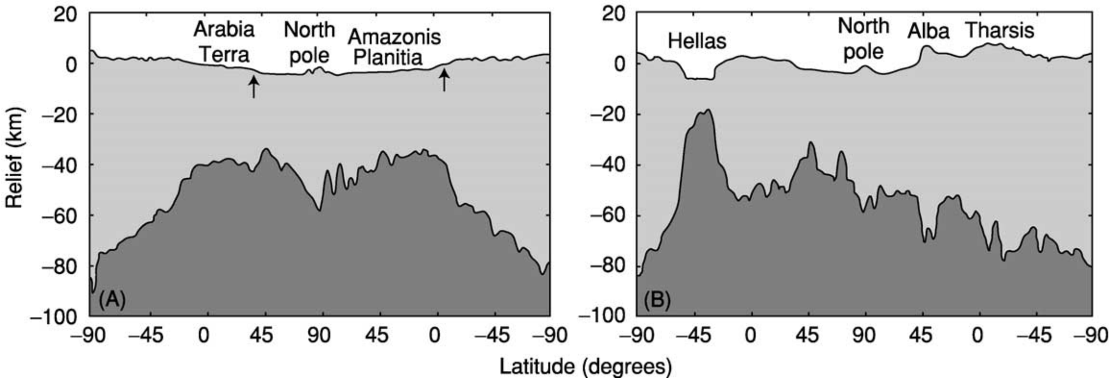


Figure 5 The estimated crustal depth on Mars. (A) Longitude line from 0° to 180°; areas indicate the regions of hemispheric dichotomy: note no major crustal depth variations that might indicate a collision origin of the dichotomy. (B) Longitude line from 70° to 250°. Reprinted with permission from Zuber MT, *et al.* (2000) *Science* 287: 1788–1793. Copyright 2004.

dichotomy is still the subject of debate. Suggestions have included the former presence of a shallow ocean in the northern hemisphere, or low-viscosity lava flows recently covering the northern plains.

An impact origin for the global dichotomy appears to have been ruled out by topography and gravity data obtained by Mars Global Surveyor. The data show that the crust thickness variations are fairly smooth across the dichotomy boundary, as shown in Figure 4. The crustal thickness shows a minimum crustal average depth of 50 km (Figure 5). This compares with approximate Earth values of 30 km for oceanic crust and 80 km for continental crust. The evidence from the Shergottite meteorites, as well as remote sensing from spacecraft and Earth, suggest that the upper crust of Mars is of volcanic origin and mainly of basaltic or andesitic origin, but there is very little in-depth information about the composition.

Mineralogy and Petrology

Thick, homogeneous layers of dust on the Martian surface have hampered efforts to map the surface mineralogy of Mars. The majority of orbital mapping has been conducted by the National Aeronautics and Space Administration (NASA) Viking mission, Mars Global Surveyor, and Mars Odyssey and European Space Agency (ESA) Mars Express missions. This has been complemented by results of landed missions, including Viking 1 and 2, Pathfinder (including the rover Sojourner), and the Mars Exploration rovers

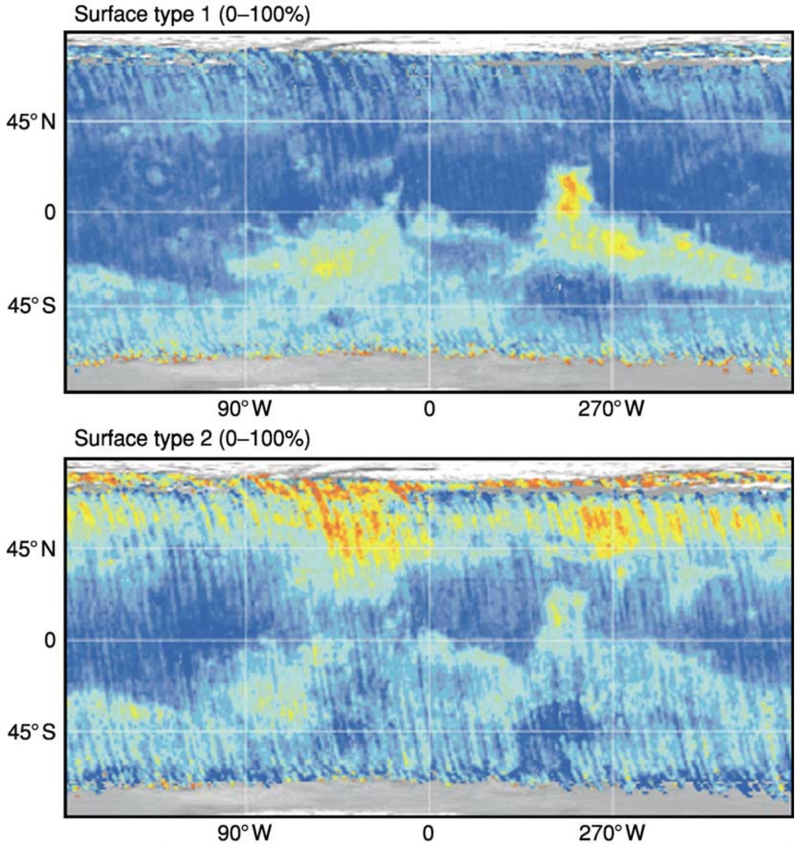


Figure 6 Map of surface types 1 and 2 as determined by thermal emission spectrometry. Red and yellow colours indicate high abundances; blue and light blue indicate low abundances. Type 1 is a basaltic-type rock. Type 2 is a more silicic rock that matches the composition of andesite on Earth, though it is probably formed by different processes. Image courtesy of NASA.

(Opportunity and Spirit). Orbital thermal infrared spectroscopy has revealed two different predominant rock types which have been interpreted as the volcanic rocks basalt and andesite. Spatially limited areas of hematite and olivine have been recognized by remote sensing from orbit (Figure 6).

The detection of olivine, which weathers easily in the presence of water, supports the generally held view that water has not existed in its liquid form on the Martian surface for large periods of time. Five of the six possible modes of formation for the detected hematite deposits, however, involve water. A possible resolution to this conundrum is the proposal that frozen glacial deposits or permafrost existed early in the history of Mars, then melted slowly over time to produce water that seeped into the ground and was later heated by a shallow magma, thereby reacting in the subsurface to form the oxidized iron mineral, hematite. Although these analyses effectively represent point measurements on widely dispersed locations on the Martian surface, they give a landed context for the orbital observations. Generally, the Mars Pathfinder analyses are thought to represent the more andesitic basalt member discovered by Mars Global Surveyor, whereas Viking analyses match the more mafic basalt end member.

The results of soil and dust geochemical analyses on Mars by orbital, landed, and telescopic missions have revealed soils relatively high in Fe and S and low in Si and Al. The mineralogy of the soil is not well defined, but is inferred to consist of poorly crystalline weathering products of basalts, analogous to palagonite, which is a hydrous weathering product of basaltic glass. The presence of large amounts of S and Cl in the dust suggests the presence of soluble salts such as sulphates, possibly deposited in the soil as a result of deposition of volcanic aerosols from the atmosphere. To the current time, only small amounts of carbonate have been detected in the homogeneous dust covering Mars. No outcrops of carbonate have been discovered. The lack of carbonates is a strong argument against a long-lived ocean or standing water, which might be expected to deposit carbonate in the presence of the CO_2-rich Martian atmosphere, just as limestone forms in the shallow coastlines on Earth.

Water on Mars

At low latitudes, daily temperatures on Mars range from about −100° to +17 °C, and the average is −60°C. Because of the low pressure exerted by the very thin atmosphere, at these temperatures liquid water is everywhere unstable. Consequently, the water ice at the poles sublimes (goes straight from ice to vapour) into the atmosphere. Down to latitudes of about 40°, ice can exist in the ground as ‘permafrost’ as shallow as 1 m. Water ice has been detected at the north pole when it is exposed as the overlying carbon dioxide ice (‘dry ice’) sublimes in summer.

Since the 1970s, when Viking orbiters sent back high-resolution images of the surface of Mars, scientists have puzzled over features that resemble the gullies and water-eroded valleys seen on Earth. The lack of liquid water on the Martian surface today begs the question as to how the gully and valley features formed. Some scientists have postulated that a thicker atmosphere in the past may have led to liquid water being stable on the Martian surface. Probable water-influenced features on Mars have been recognized on three different scales and are commonly grouped into ‘outflow channels’, ‘valley systems’, and ‘gullies’, in order of descending scale.

Giant outflow channels on Mars are several tens to hundreds of kilometres across and many thousands of kilometres in length. They are mostly Hesperian in age and commonly start in chaotic terrain, as seen in the image of Hydaspis Chaos (Figure 7). They are usually associated with streamlined islands and terraces, indicating massive fluid flow, with the probable source being water, although some researchers have suggested CO_2 as a possible fluid.

Valley networks, such as those seen in the Thaumasia region in Figure 8, superficially resemble branching valley fluvial systems here on Earth; however, the lack of fine-scale structure, low drainage densities, and differing morphology argues against a rainfall and subsequent surface runoff origin. Instead, groundwater flow from seeps or hydrothermal systems is a plausible mode of formation. Most of the valley networks are Noachian in age, but there are younger systems, including those developed on Amazonian-age terrain around volcanic centres in the Tharsis region.

Martian gully systems often develop in the rims of impact craters, and have fine-scale features that suggest they are very young. They do not occur near the equator, appearing only at latitudes poleward of 30° in both hemispheres. A possible mode of formation is by melting of water beneath snow or ice packs. The lack of gullies around the equator may indicate that

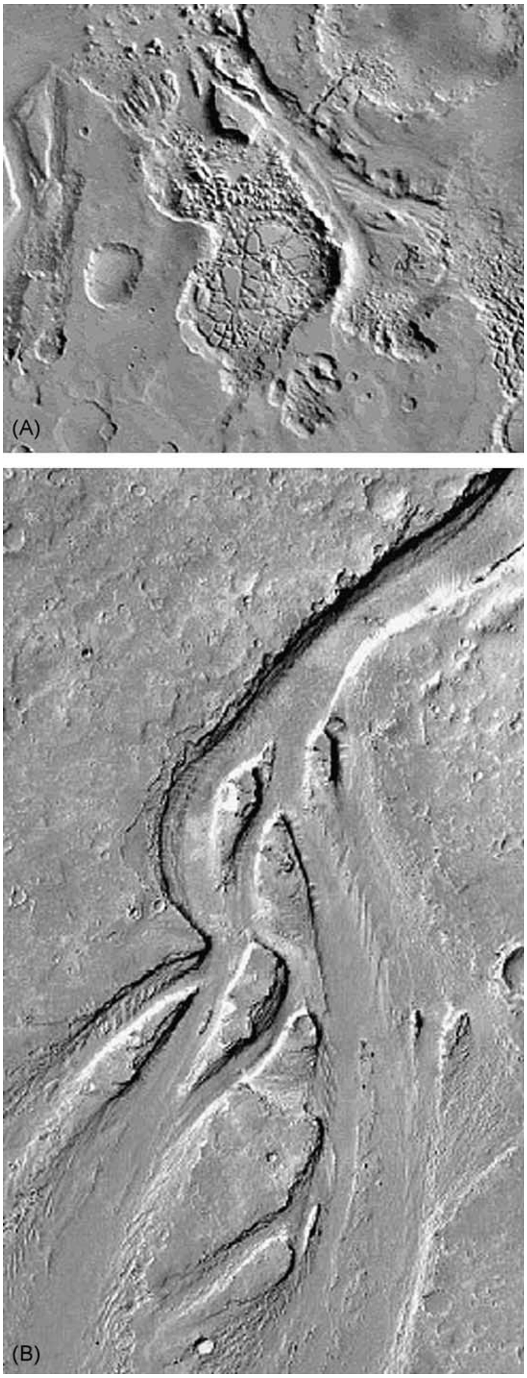


Figure 7 (A) Hydaspis Chaos, the source of Tius Valles, as imaged by Viking, showing a typical giant outflow channel. Note streamlined island features. The image is approximately 200 km across (north is up). (B) Evidence of erosion by a liquid (water?) in the Athabasca part of the Marte Vallis channel and a streamlined island system, as imaged by the Mars Global Surveyor M21-01914. (B) Reproduced with permission from Hartmann WK (2003) *A Travellers Guide to Mars*. New York: Workman Publishing. NASA/JPL/Malin Space Science Systems.

formation of ice in impact craters was insufficient to produce these deposits.

The gamma ray spectrometer (GRS) on the orbiting spacecraft Mars Odyssey has been able to map hydrogen abundance in the top metre of the surface of Mars. The most likely source of the hydrogen in the Martian crust is molecular water or hydroxyl (OH^-) ions in weathered rocks. The distribution of hydrogen in the Martian surface is shown in Figure 9.

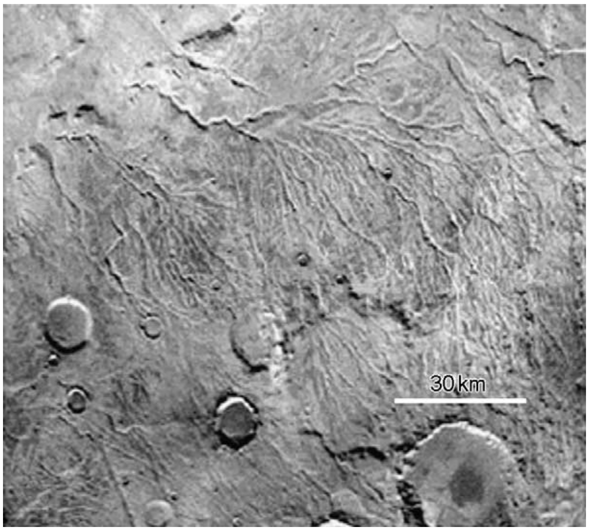


Figure 8 Martian valley networks in the Thaumasia region, as observed by Viking. Image is 100 km across. Image courtesy of NASA.

Large-Scale Features

Due to the smaller size of Mars in comparison to Earth, and the apparent lack of plate tectonics in the past 4 billion years, some topographic features have taken on exaggerated forms. The Tharsis bulge is a large region of volcanism and deformation crossing the equator. The formation process of the Tharsis uplift is still under discussion; it may be due to convection in the interior of the planet or to there being a region of intense volcanism. The Tharsis complex consists of two broad rises, the largest southern rise containing three of the largest volcanoes on Mars, Ascraeus Mons, Pavonis Mons, and Arsia Mons. The smaller northern rise is dominated by the volcanic construct Alba Patera.

Olympus Mons (Figure 10) is situated to the west of the Tharsis region and though not surficially connected to the region, it is thought to be connected in origin. Olympus Mons is the largest known volcano in the Solar System. Like some of Earth's volcanoes, it

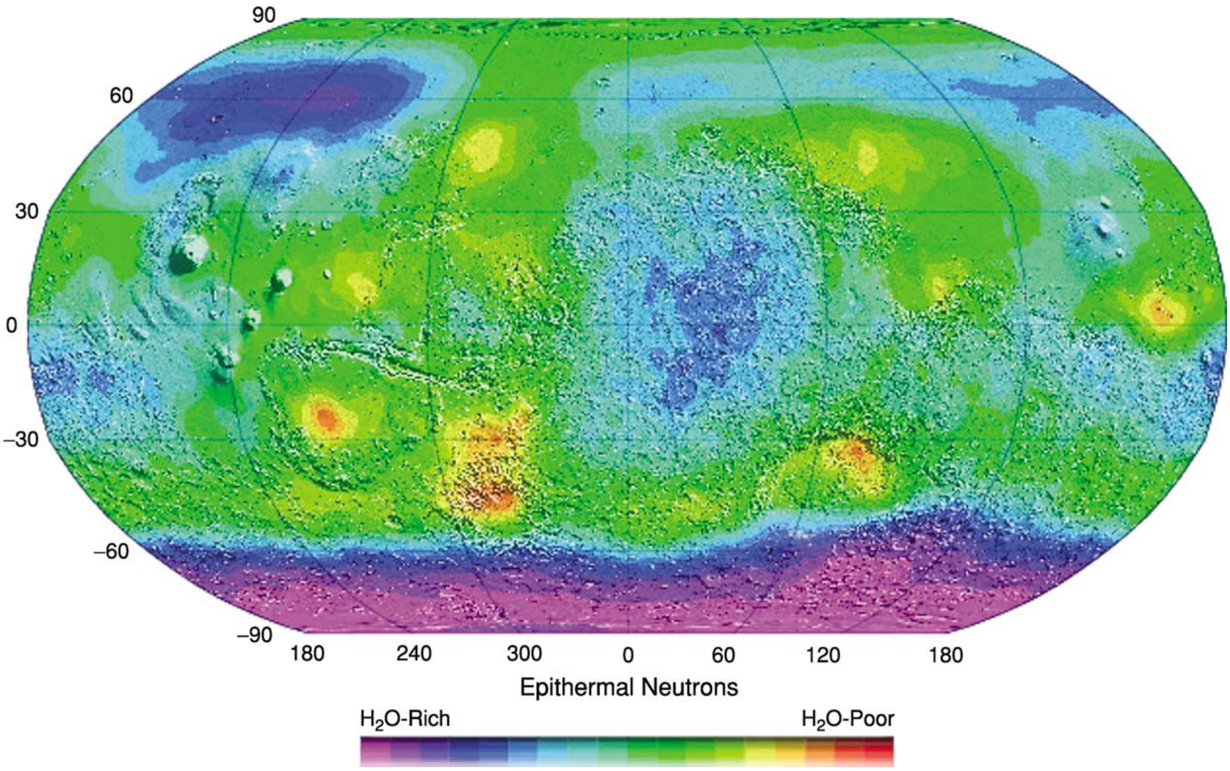


Figure 9 Map of hydrogen (or epithermal neutrons) in the shallow Martian crust (late southern summer), as detected by the gamma ray spectroscope on Mars Odyssey. Image courtesy of Dr William V. Boynton, University of Arizona.

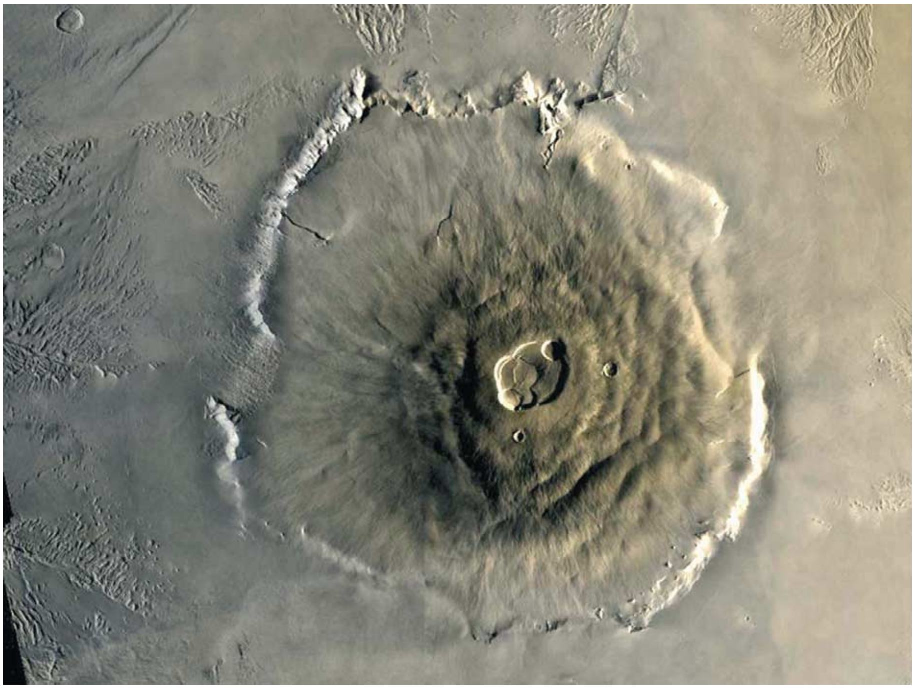


Figure 10 Image of Olympus Mons, as observed by the NASA Viking Orbiters. Image courtesy of NASA.

is thought to have been an area of the crust situated above a region of advecting hot mantle material, known as a hotspot. Unlike Earth analogues, Olympus Mons continued to build, because at the time of formation Mars lacked the plate tectonics to move the volcano away from the source. At 21 183 m above the global reference datum, Olympus Mons is 2.5 times higher than Earth's largest shield volcano, Mauna Loa, and 100 times its volume.

Valles Marineris, named for the spacecraft that first discovered it (Mariner 9), is a giant canyon system that extends eastward some 4500 km from the central regions of the Tharsis complex. The troughs are generally about 50 to 100 km wide and the depth of the floor drops down to 5310 m below the global reference datum (six to seven times as deep as the Grand Canyon). This feature is thought to have formed as the uplift of Tharsis created tension in the cooled crust.

Martian Polar Regions

The polar regions are broadly composed of four geological units: the basal plains, the polar layered deposits, the residual ice, and the seasonal frost. The basal plains differ north from south, In the south, they are believed to date from the Noachian Period. There is evidence of intense cratering, contractional deformation, resurfacing of low areas, and local dissection producing valley networks. There is no evidence for polar deposits, ice sheets, or glaciation at these times. During the Hesperian, these deposits were altered by waning impacts and volcanism. In the north, the basal plains are much younger, being buried by a water or debris ocean during the Hesperian Period.

The polar layered deposits began building up during the late Hesperian or later and appear to be similar at both poles. Polar layered deposits extend further than do the icecaps that overlie them (to a greater and more asymmetric extent in the south). The deposits are composed of dust, water ice, and other volatiles such as CO_2, in varying ratios for each layer. Each of these layers, shown in Figure 11, was deposited under differing environmental conditions.

Due to the small number of superimposed craters at each polar icecap, the caps have been interpreted to be of late Amazonian age, with the southern cap being slightly older than the northern. This means that the Martian environment has only recently allowed the current polar icecaps to form. This is likely to be due to the changing obliquity (0°–60°) of the planet, causing the icecaps to reform continuously. There is evidence of advance and retreat at both poles, but exact timing remains unknown. The residual ice in the north appears to be mainly H_2O, whereas the south polar surface residual ice appears to be CO_2. Little is known of the bulk composition of residual ices at either pole, because thin overlying layers of CO_2 ice can mask signatures of the composition beneath. The seasonal frost is composed of CO_2

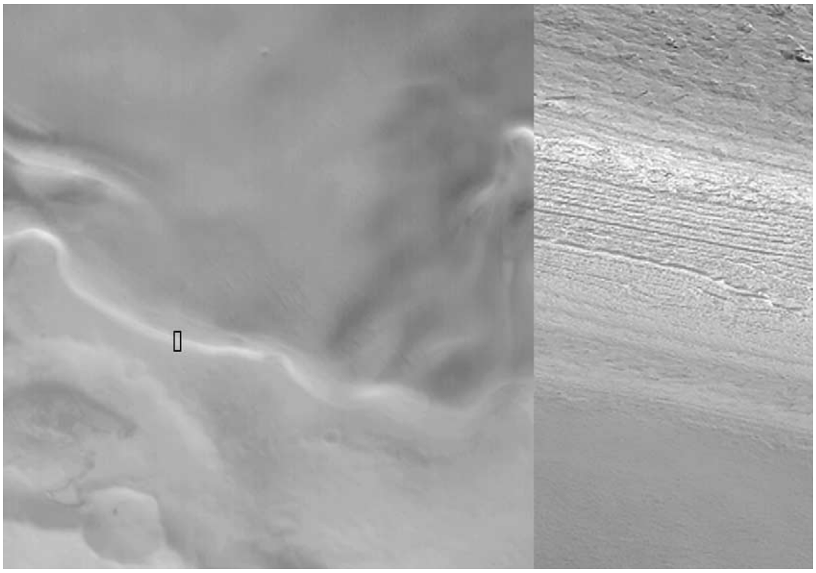


Figure 11 Martian layered polar terrain, as imaged by the Mars Orbiter Camera. Image on the right is 750 m across and sits within the small area outlined in the context image on the left. NASA/JPL/Malin Space Science Systems.

ice that condenses from the atmosphere during winter and sublimes again in the Martian summer.

Satellites of Mars

Mars has two small satellites, Phobos (27 km maximum diameter) and Deimos (25 km maximum diameter), both irregular potato-shaped bodies. Both have cratered surfaces and Deimos is also strongly grooved. Their origin is obscure – they may be captured asteroids. The orbit of Phobos is unstable and it will eventually crash into Mars.

Shergottite–Nakhlite–Chassigny Meteorites

Meteorites, widely accepted as being from Mars have been collected here on Earth. They are collectively called SNC meteorites after the discovery locations of the first three meteorites – Shergotty, Nakhla and Chassigny. They are believed to come from Mars because: 1.) their young age of formation relative to the Solar System, suggesting they were formed on a rocky planet, and 2.) trapped gases inside solidified molten rock produced by high shock pressures in some of the meteorites display chemical and isotopic characteristics matching those of the Martian *atmosphere* (as reported by the Viking landers). The trapped gases are interpreted to represent samples of the atmosphere when these rocks were ejected from Mars by impacts. The Martian meteorites have contributed greatly to knowledge about Mars, but naturally have also raised many scientific questions. The inability to pinpoint a source location on Mars is the greatest limitation on the utility of the SNC meteorites, but the ability to analyse the rocks with the most modern of terrestrial scientific equipment is a great advantage lacking as we do any rock samples directly recovered from the planet.

At the current time, there are 30 known SNC meteorites. The reported crystallization ages of SNC meteorites range from 154 Ma to 1.3 Ga, with the exception of ALH84001, which has a crystallization age of about 4 Ga. ALH84001 was discovered in Antarctica in 1984 and was held in storage by NASA for some years. In 1996, it was reported that a team of NASA scientists had found possible signs of life in small amounts of carbonate within ALH84001. Six lines of evidence were cited that would be explained best by invoking a biological origin. These lines of evidence have been debated within the scientific community in the intervening years and thus the question of life on Mars still remains open, lacking rock samples directly recovered from the planet.

Mars Exploration Rover Missions

The Mars Exploration Rovers, Spirit and Opportunity, have added to the data from the planet's surface. Designed as mobile remote geologists, they were equipped with various instruments to study the rocks they found, including panoramic cameras, infrared spectrometers, abrasion tools and a microscopic camera. '*Opportunity*' has been successful in

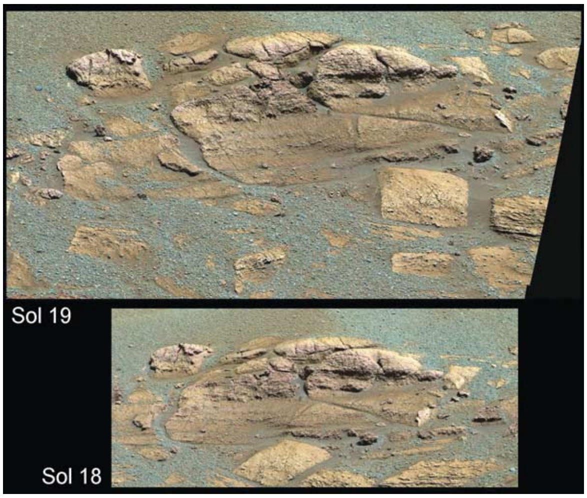


Figure 12 Images of the rock 'El Capitan', the first bedrock to be investigated by a Mars landed rover, on two days (sols). The outcrop is approximately 10 cm high. Image courtesy of NASA.

studying a number of fascinating outcrops which are only now starting to be interpreted. An image of the first competent rock outcrop found by a rover on Mars is shown in false colours in Figure 12. It reveals layered successions which could be volcanic or sedimentary in nature. Interplanetary investigations like these will continue to reveal the true extent of apparent differences in the geology of Earth and Mars.

See Also

Sedimentary Processes: Glaciers. **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Venus. **Volcanoes**.

Further Reading

Acuna MH, Connerney JEP, Ness NF, *et al.* (1999) Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* 284(5415): 790–793.

Barlow NG (1988) Crater size/frequency distributions and a revised Martian relative chronology. *Icarus* 75: 285–305.

Boyce JM (2002) *The Smithsonian Book of Mars*. Washington, DC and London: Smithsonian Institution Press.

Boynton WV, Feldman WC, Squyres SW, *et al.* (2002) Distribution of hydrogen in the near surface of Mars: Evidence for subsurface ice deposits. *Science* 297(5578): 81–85.

Christensen PR (2003) Formation of recent martian gullies through melting of extensive water-rich snow deposits. *Nature* 422(6927): 45–48.

Clifford SM and Parker TJ (2001) The evolution of the Martian hydrosphere: Implications for the fate of a primordial ocean and the current state of the northern plains. *Icarus* 154(1): 40–79.

Gibson EK, McKay DS, Thomas-Keprta KL, *et al.* (2001) Life on Mars: evaluation of the evidence within Martian meteorites ALH84001, Nakhla, and Shergotty. *Precambrian Research* 106(1–2): 15–34.

Gulick VC (1998) Magmatic intrusions and a hydrothermal origin for fluvial valleys on Mars. *Journal of Geophysical Research-Planets* 103(E8): 19 365–19 387.

Hartmann WK (2003) *A Travellers Guide to Mars*. New York: Workman Publishing.

Kieffer HH, Jakosky BM, Snyder CW, and Matthews MS (1992) *Mars*. Tucson: The University of Arizona Press.

Kolb EJ and Tanaka KL (2001) Geologic history of the polar regions of Mars based on Mars Global Surveyor data: II. Amazonian period. *Icarus* 154: 22–39.

McKay DS, Gibson EK, ThomasKeprta KL, *et al.* (1996) Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science* 273(5277): 924–930.

Nimmo F and Stevenson DJ (2000) Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *Journal of Geophysical Research-Planets* 105(E5): 11969–11979.

Sleep NH (1994) Martian plate-tectonics. *Journal of Geophysical Research-Planets* 99(E3): 5639–5655.

Squyres SW and Kasting JF (1994) Early Mars – how warm and how wet. *Science* 265(5173): 744–749.

Tanaka KL and Kolb EJ (2001) Geologic history of the polar regions of Mars based on Mars Global Surveyor Data: I. Noachian and Hesperian periods. *Icarus* 154: 3–21.

Treiman AH, Gleason JD, and Bogard DD (2000) The SNC meteorites are from Mars. *Planetary and Space Science* 48(12–14): 1213–1230.

Zuber MT, Solomon SC, Phillips RJ, *et al.* (2000) Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* 287(5459): 1788–1793.

Jupiter, Saturn and Their Moons

P Moore, Selsey, UK

Introduction

Jupiter and Saturn, the largest and most massive planets in the Solar System, have no visible solid surfaces and are therefore not of real concern to the geologist. However, their satellite systems are of immense interest. Jupiter has four major satellites (the Galileans) which are of planetary size, and which differ markedly from each other geologically; there are also over four dozen small satellites, almost certainly captured asteroids, which seem to be icy. Saturn has one very large satellite (Titan), which has a dense atmosphere obscuring the surface; there are also eight medium-sized icy satellites as well as numerous ex-asteroids. These various bodies are discussed in this article.

Jupiter

Jupiter is the giant of the Sun's system. It is more massive than all the other planets combined, and is generally the brightest object in the sky apart from the Sun, the Moon, and Venus. Physical and orbital data are given in Table 1.

Telescopically, Jupiter shows a yellowish, obviously flattened disk, crossed by dark belts and bright zones (Figure 1). It has always been assumed that the belts are regions of descending gases while the bright zones are regions where gas is rising from the interior, though some recent (2003) observations may indicate that the reverse is true. There are generally two very prominent belts, the North Equatorial and the South Equatorial, with others in higher latitudes. Jupiter does not rotate in the way that a solid body would do. The region between the north edge of the South Equatorial Belt and the south edge of the North Equatorial Belt (System I) has a mean period of 9 h 50 m 30 s, while the period of the rest of the planet (System II) is 9 h 55 m 41 s, but individual features have periods of their own. Radio methods indicate that the interior (System III) has a period of 9 h 55 m 29 s, though this is subject to some uncertainty.

Much of our knowledge of Jupiter has been derived from five space-craft: two Pioneers, two Voyagers, and Galileo. Details are given in Table 2. Useful data were also obtained from two fly-by probes, Ulysses in February 1992 (the solar polar probe) and Cassini in December 2000 (en route for an encounter with Saturn).

Jupiter is assumed to have a hot silicate core at a temperature of at least 20 000°C, probably rather more. The core is surrounded by a thick shell of liquid metallic hydrogen, which is itself surrounded by a shell of liquid molecular hydrogen; above lies the atmosphere, made up chiefly of hydrogen and helium, with hydrogen compounds such as methane and ammonia. Windspeeds in the visible clouds are high, and the surface details are always changing. Of special note is the Great Red Spot, which has been visible for most of the time since regular observations began in the seventeenth century. Once thought to be a glowing volcano, it is now known to be a phenomenon of Jovian meteorology – a high-level anticyclonic vortex, elevated by 8 km above the adjacent cloud deck.

Jupiter's magnetic field is much the strongest in the Solar System. The magnetic axis is inclined to the rotational axis at an angle of 9.6°; the polarity is opposite to that of the Earth's field. The planet is a powerful radio source, and is surrounded by zones of radiation which would quickly be fatal to an astronaut unfortunate enough to enter them. There is a system of dark, obscure rings, probably formed from material released from the small inner satellites by meteoritic impact. They are quite unlike the glorious icy rings of Saturn.

Table 1 Data for Jupiter

Distance from the Sun	Max 815 700 000 km (5.455 astronomical units) Mean 778 350 000 km (5.203 a.u.) Min 740 900 000 km (4.951 a.u.)
Orbital period	11.86 years (4332.59 days)
Rotation period	System I, 9 h 50 m 30 s System II, 9 h 55 m 41 s System III, 9 h 55 m 29 s
Axial inclination	3° · 4
Orbital inclination	1°18′15″8
Orbital eccentricity	0.048
Diameter	Equatorial 142 884 km Polar 133 708 km
Oblateness	0.065
Density, water-1	1.33
Mass, Earth-1	317.89
Volume, Earth-1	1318.7
Escape velocity	60.22 km/sec
Surface gravity Earth-1	2.64
Mean surface temperature	−150°C
Albedo	0.43

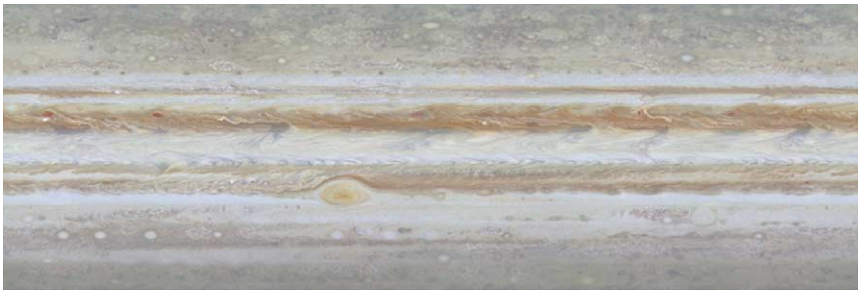


Figure 1 Linear zones of light and dark, rising and descending gases on Jupiter's surface. Reproduced from NASA.

Table 2 Space-craft to Jupiter

Name	*Launch date*	*Encounter date*	*Nearest approach, km*	
Pioneer 10	2 Mar. 1972	3 Dec. 1973	131 400	Fly-by
Pioneer 11	5 Apr. 1973	2 Dec. 1974	46 400	Fly-by. Went on to Saturn
Voyager 1	5 Sept. 1977	5 Mar. 1979	150 000	Images of Jupiter and the Galileans. Went on to Saturn
Voyager 2	20 Aug. 1977	9 July 1979	714 000	Complemented Voyager 1. Went on to Saturn, Uranus, Neptune
Galileo	18 Oct. 1989	7 Dec. 1995	Entry	Orbiter and entry probe

Data were also obtained from Ulysses (1992) and Cassini (2001).

Satellites

The four Galilean satellites – Io, Europa, Ganymede, and Callisto – were observed by Galileo as long ago as 1610 (hence the name of the space-craft). Four small inner satellites were discovered between 1892 and 1979, and there are many small bodies moving round Jupiter beyond the path of the outermost Galilean, Callisto. The total number of known satellites by April 2004 was 62. Data for all the satellites over 8 km in diameter are given in Table 3.

Io Io is violently volcanic. In March 1979, S Peale and his colleagues in America suggested that since Io's orbit is not perfectly circular, the interior might be 'flexed' by the gravitational pulls of Jupiter, and also the other Galileans, sufficiently to produce active surface volcanoes. Only a week later the first volcanic plume was detected on an image from the Voyager 1 space-craft, and many dozens have since been identified, both from the space probes and with the Hubble Space Telescope. Lava-flows and lava lakes are plentiful; the average temperature of the lavas is about 1600°C. Many explosive eruptions are driven by sulphur dioxide gas emission; the surface is remarkably colourful, with yellow, orange, red, and black areas (Figure 2). The surface is 'young' and virtually without surviving impact craters. One volcano, Loki, is the most powerful in the Solar System, emitting more heat than all the Earth's active volcanoes combined. Io is connected to Jupiter by a strong flux tube, and has a marked effect upon the Jovian radio emission; material sent out by the volcanoes is spread along the orbit, producing a torus.

Europa Europa is only slightly smaller than Io, and rather further from Jupiter, but the two satellites are very different. Europa has a smooth, icy surface with very limited vertical relief and few impact craters, though one of these, Pwyll, shows bright rays extending outward and crossing all other features. There are plains, chaotic areas, and low ridges, together with shallow pits. Detailed views from space-craft (particularly Galileo) show what look remarkably like icebergs, and it is widely believed that an ocean of salty water lies below the visible surface, with the icebergs floating around (Figure 3). Fragmented blocks of ice seem to look very like the blocks in the Earth's polar seas during a springtime thaw.

Europa does not have a strong internal magnetic field, but it orbits within Jupiter's magnetosphere, and the instruments on Galileo have detected an induced magnetic field which produces significant effects linked with the rotational period of the planet. Jupiter's magnetic field at Europa changes direction every $5\frac{1}{2}$ h, and this indicates the existence of a layer of electrically conducting material, such as salty water, not far below the icy surface of Europa. It seems that the ocean may lie at a depth of less than 100 km. If it really does exist (and as yet there is no

Table 3 Satellites of Jupiter

Satellite	*Mean distance from Jupiter, km*	*Orbital period, days*	*Diameter, km (equator)*	*Density water-1*	*Orbital eccentricity*	*Orbital inclination*	*Escape velocity km s^{-1}*
Small inner satellites							
Metis	128 100	0.294	60	2.8	0.001	0.021	0.025
Adrastea	128 900	0.300	26	2?	0.002	0.027	0.014
Amaithea	181 100	0.498	262	1.8	0.003	0.389	0.084
Thebe	221 900	0.674	110	1.5	0.018	0.070	0.043
Galileans							
Io	421 800	1.769	3660	3.6	0.004	0.036	2.56
Europa	671 100	3.551	3130	3.0	0.009	0.470	2.02
Ganymede	1 070 400	7.154	5268	1.9	0.002	0.195	2.74
Callisto	1 882 700	16.689	4821	1.1	0.007	0.281	2.45
Outer prograde satellites							
Themisto	7 507 000	130.0	9	2?	0.242	43.08	Low
Leda	11 165 000	240.0	16	2.7	0.164	27.46	0.01
Himalia	11 461 000	250.6	186	2.8	0.162	27.50	0.12
Lysithea	11 717 000	259.2	38	3.1	0.212	28.30	0.02
Elara	11 741 000	259.6	78	3.3	0.217	26.63	0.05
Outer retrograde satellites							
Ananke	21 276 000	610.5	28	2.7	0.244	148.9	0.02
Carme	23 404 000	702.3	48	2.8	0.253	164.9	0.03
Pasiphaë	23 624 000	708.0	58	2.9	0.109	151.4	0.03
Sinope	23 939 000	724.5	38	3.4	0.250	158.1	0.24

Ten of the other small retrograde satellites have been named; all are below 8 km in diameter. Harpalyke, Praxidike, Iocaste, Chaldene, Isonoe, Erinome, Tayrete, Kalyke, Megaclite, and Callirrhoë.
? = doubtful value.

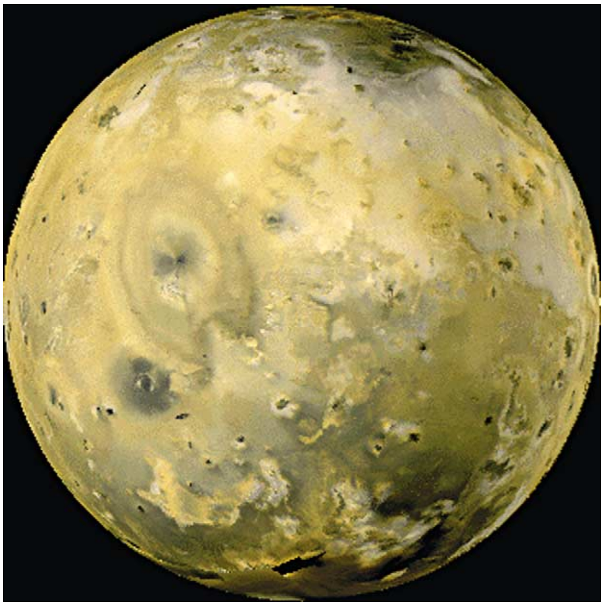


Figure 2 Io, imaged by the Galileo probe, the colour is very accurate. The volcanoes, such as Pele, are violently active. Reproduced from NASA.

final proof) there will be tidal effects. There have been the inevitable speculations about possible life-forms, but conditions in such a strange, sunless sea would not appear to be inviting!

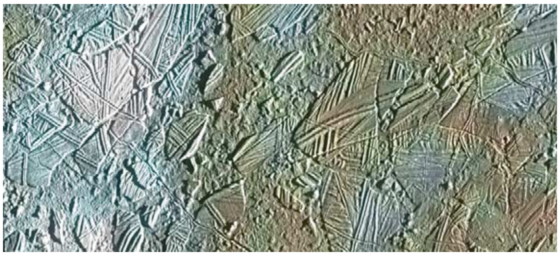


Figure 3 Europa, imaged from the Galileo space-craft in orbit round Jupiter. The icy surface, unlike any other in the Solar System, may cover a salty ocean. Hubble Space Telescope Image, NASA.

Ganymede Ganymede is the largest satellite in the Solar System, and is actually larger than the planet Mercury, though not so massive; the mean density of the globe is less than twice that of water. There is an excessively tenuous atmosphere, and a marked magnetic field. Presumably this indicates the presence of a metallic core surrounded by a mantle composed of a mixture of rock and ice; the surface is icy. There are two types of surface terrain, about equal in area, very ancient, thickly cratered dark regions and somewhat younger light regions, marked with an extensive array of ridges and grooves. Evidently there has been marked tectonic activity in the past. In general, the craters have little vertical relief. The largest indiviual feature is a dark plain which has been given the appropriate name of Galileo.

Callisto Callisto is rather smaller than Ganymede, and is less dense, so that ice is a major constituent of the globe. The surface is heavily cratered, and there are two huge ringed plains, Valhalla and Asgard. Unexpectedly, it has been found that the local magnetic field fluctuates in the same way as that of Europa, and there may be a similar salty ocean deep inside the globe. Space-craft data indicate that the interior of Callisto is made up of compressed rock and ice, with the percentage of rock increasing with depth.

Amalthea Amalthea the only other Jovian satellite over 200 km in diameter, was discovered in 1892 by EE Barnard, using the 91 cm refractor at the Lick Observatory (this was the last visual discovery of a planetary satellite). It is irregular in form; the surface is red, due possibly to contamination from the volcanoes of Io. In November 2002, the Galileo space-craft flew past Amalthea and found that the density is very low; indeed the satellite has been described as an 'ice rubble pile'. The rotation is synchronous, with the longest axis pointing to Jupiter. Images from Galileo show that there are two craters, Gaea and Pan, which are very large relative to the overall diameter of Amalthea; there are ridges, troughs, and two bright patches which seem to be hills. Amalthea is one of four satellites known to move within the orbit of Io; the others are Metis, Adrastea, and Thebe. They are icy in nature, and impact craters have been imaged.

Outer Icy Satellites By April 2004, the total number of known Jovian satellites had risen to 62, but few of these were more than a few kilometres in diameter. Several satellites, including Himalia (diameter 184 km) move at between 11 and 12 million kilometres from Jupiter; these have direct motion, and those that have been imaged show the usual impact craters. Much further out move small satellites, many of which have retrograde motion; the largest, known before the space-probe era, are Ananke, Carme, Pasiphaë, and Sinope. They are so far from Jupiter that their orbits are not even approximately circular, and no two cycles are alike. There seems no doubt that all the outer icy satellites are captured asteroids; their small size means that they are very difficult to record.

Saturn

Saturn the largest and most massive planet in the Solar System, apart from Jupiter, is distinguished by its magnificent icy ring system, making it probably the most beautiful object in the entire sky. Physical and orbital data are given in Table 4.

Saturn's globe shows belts not unlike those of Jupiter, but much less prominent, and sensibly curved. Predictably, the globe is flattened; this is because of the rapid rotation. It is thought that there is a silicate core at a temperature of perhaps 15 000°C, overlaid by layers of metallic hydrogen, molecular hydrogen, and then the atmosphere, which is not unlike that of Jupiter.

There is no surface feature comparable with the Great Red Spot on Jupiter, but prominent, temporary white spots are seen occasionally, as in 1933, 1960, and 1990 (Figure 4). The 1933 spot was discovered on 3 August by WT Hay, using a 15 cm refractor; it remained identifiable until 13 September, and was considered to be of an eruptive nature. The 1990 spot, discovered by S Wilber, was in the same latitude.

Table 4 Data for Saturn

Distance from the Sun, km	Max 1 506 400 (10.069 astronomical unite) Mean 1 426 800 (9.359 a.u.) Min 1 347 600 (9.008 a.u.)
Orbital period	29.4235 years (10 746.94 days)
Equatorial rotation, period	10 h 13 m 59 s
Axial inclination	26.73°
Orbital eccentricity	0.05555.
Orbital inclination	2°29′21″
Diameter; km	Equatorial 120, 536 polar 108, 728
Oblateness	0.098
Mass, Earth-1	95.17
Volume, Earth-1	752
Escape velocity	35.26 km/s
Surface gravity, Earth-1	1.19
Mean surface temperature	−180°C

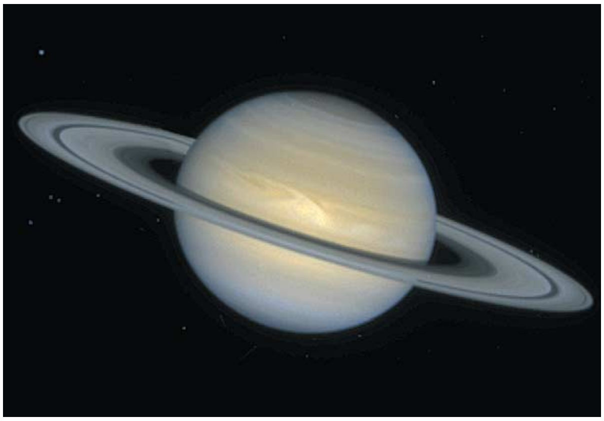


Figure 4 Saturn, image from the Hubble Space Telescope, 1 December 1994; showing the bright white spot, it consists of condensed ammonia ice crystals, and had changed little since its discovery in September 1994.

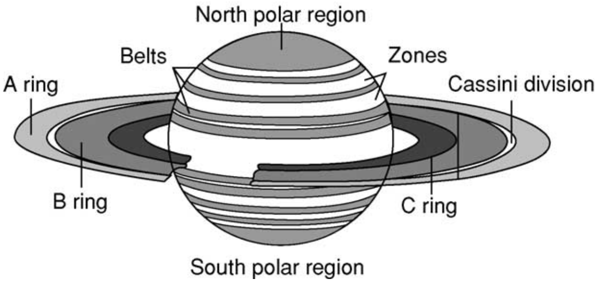


Figure 5 Diagram of Saturn and its ring-system. There are some less prominent rings behind the main system.

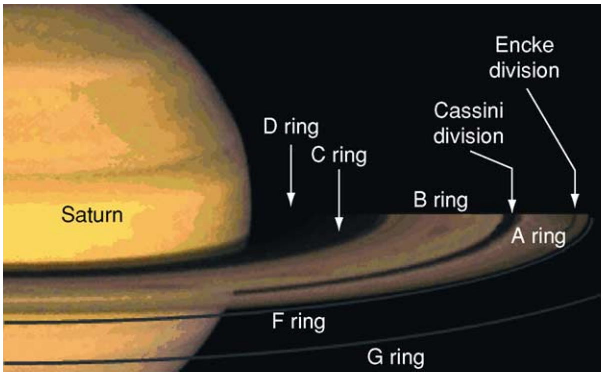


Figure 6 Diagram showing the main ring system. The bright rings are A and B, C is the Crepe or Dusky Ring. D is not well-defined. Reproduced from Kennod R (1990) *The Journeys of Voyager – NASA reaches for the planets, BDD*, New York.

Table 5 The ring system of Saturn

Feature	*Distance from centre of Saturn, km*
Inner edge of ring D	66 900
Outer edge of ring D	73 150
Inner edge of ring C	74 510
Outer edge of ring C	92 000
Inner edge of ring B	92 000
Outer edge of ring B	117 500
Centre of Cassini division	119 000
Inner edge of ring A	122 200
Centre of Encke division	135 700
Outer edge of ring A	136 800
Centre of ring F	140 210
Centre of ring G	168 000
Inner edge of ring E	180 000
Brightest part of ring E	230 000
Outer edge of ring E	480 000

Within a few days the spot had been spread out by Saturn's strong equatorial winds, and by October had been transformed into a bright zone all round the equator. Extra outbreaks were seen in it, clearly indicating an uprush of material from below. Saturn has a strong magnetic field, though less powerful than that of Jupiter; the rotational axis and the magnetic axis are almost coincident.

There are three main rings (Figure 5); A and B, separated by a gap known as the Cassini Division is honour of its discoverer, and an inner semi-transparent ring, C (the Crepe Ring) (Figure 6). Details of the ring system are given in Table 5. The rings are made up of ice particles, from grains up to several metres across; the brightest ring, B, shows curious 'spokes', presumably particles elevated away from the main ring plane by magnetic or electrostatic forces. The irregular F ring, outside the main system, is stabilized by two small 'shepherd' satellites, Prometheus and Pandora. Though the system is very extensive, with a total diameter of 270 000 km, it cannot be more than 200 m thick, and if all the particles could be combined they would make up an icy satellite less than 300 km across.

Three space-craft have now encountered Saturn; Pioneer 11, and the two Voyagers. Data are given in Table 6. The Cassini/Huygens probe, launched in 1997, was scheduled to reach its target in late 2004.

Satellites

The satellite system of Saturn is very different from that of Jupiter. There is one very large satellite, Titan, and seven icy satellites with diameters between 200 and 1600 km; the rest are much smaller. By May 2003, the total number of known satellites had risen to 31. Data are given in Table 7.

Titan Apart from Ganymede, Titan is the largest, satellite in the Solar System. Its surface is permanently hidden by its dense atmosphere, made up chiefly of nitrogen with appreciable amounts of methane and ethane; organic compounds are plentiful. The atmospheric pressure on the surface is 1.6 times that of the Earth's air at sea level; the surface temperature is −178°C.

Table 6 Space-craft to Saturn

Name	*Launch date*	*Encounter date*	*Closest approach, km*	
Pioneer 11	5 Apr. 1973	11 Sept. 1979	20 880	Preliminary results
Voyager 1	5 Sept. 1977	12 Nov. 1980	124 200	Images included Titan
Voyager 2	20 Aug. 1977	25 Aug. 1981	101 300	Went on to Uranus and Neptune
Cassini/Huygens	15 Oct. 1997	2004		Scheduled orbiter and landing on Titan (Huygens)

Table 7 Satellites of Saturn

Name	*Mean distance from Saturn, km*	*Orbital period, d*	*Longest diameter, km*	*Density, water-1*	*Escape velocity, km/s*
Pan	133 583	0.525	19	?	Low
Atlas	137 640	0.602	37	?	Low
Prometheus	139 350	0.613	145	0.27	0.02
Pandora	141 700	0.625	114	0.70	0.28
Epimetheus	151 422	0.694	144	0.63	0.32
Janus	151 172	0.694	196	0.67	0.35
Mimas	185 520	0.942	421	1.17	0.16
Enceladus	238 020	1.370	512	1.24	0.21
Tethys	294 600	1.888	1038	0.98	0.44
Telesto	294 600	1.888	34	?	Low
Calypso	294 600	1.888	30	?	Low
Dione	377 400	2.737	1120	1.49	0.50
Helene	377 400	2.737	36	?	Low
Rhea	527 040	4.517	1528	1.33	0.66
Titan	1 221 850	15.945	5150	1.88	2.65
Hyperion	1 481 100	21.278	410	1.47	0.11
Iapetus	3 561 300	79.330	1460	1.21	0.59
Phoebe	12 952 000	550.48 (ret)	220	0.77	0.07

The remaining outer satellites are below 50 km in diameter.

Nothing definite is known about the nature of the surface. The Voyagers could do no more than send back images of the top of a layer of orange haze; infrared images taken with the Hubble Space Telescope and, the Keck II telescope in Hawaii have shown bright patches and darker regions. It has been suggested that there could be frozen landmasses and frigid hydrocarbon seas or lakes, but we must await the arrival of the Huygens probe, scheduled to land on Titan in early 2005. Certainly Titan is unlike any other body in the Solar System. It does not seem to have an internal magnetic field, but orbits near the outer edge of Saturn's vast magnetosphere. As with most of the other satellites, its rotation is synchronous, so that the same hemisphere always faces Saturn.

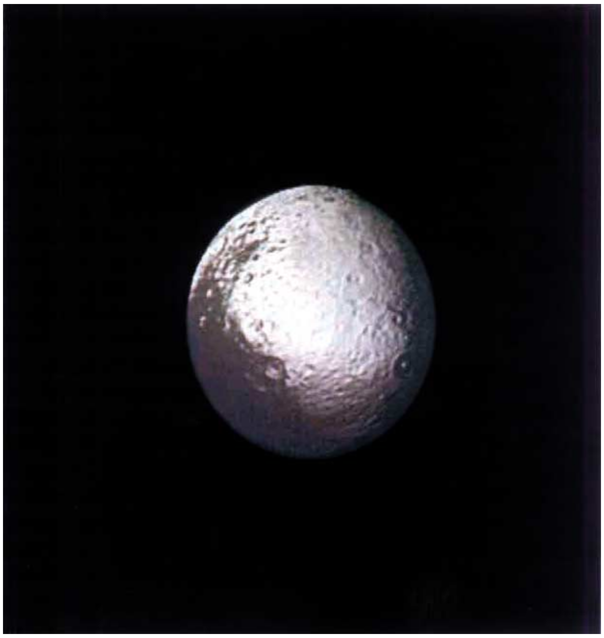


Figure 7 Iapetus, imaged by Voyager 2, 1981. The dark area is well-defined; like the bright areas, it is cratered. Copyright Cawin J. Hamilton 1999.

Medium-sized icy satellites These seem to form 'pairs'; Rhea/Iapetus (Figure 7), Tethys/Dione, and Mimas/Enceladus. Two very small satellites, Telesto and Calypso, move in the same orbit as Tethys, while Helene shares the orbit of Dione. Hyperion, moving between the orbits of Titan and Iapetus, is irregular in shape, and has a darkish surface often regarded as 'dirty ice'; the rotation is not synchronous, and on average the rotation period is around 13 days, There are several craters, and one long ridge or scarp. It is possible that Hyperion is part of a larger body which broke up. Phœoebe, outermost of the named satellites, is much further from Saturn, and has retrograde motion, so that it must be a captured asteroid. Little is known about its surface, but it was imaged in 2004 by the Cassini probe and found to be crater scarred. It was the first satellite to be discovered photographically (by WH Pickering in 1898).

Mimas Mimas is only slightly denser than water, and consists largely of ice, though there may well be a small rocky core. The surface is dominated by a huge, deep crater named Herschel in honour of the discoverer of Mimas. The 130 km crater is one-third the diameter of Mimas itself, so that the impact which formed it must have come close to disrupting the entire satellite. Parallel grooves indicate that the surface must have been subjects to considerable strain.

Enceladus Enceladus is the most interesting of all the satellites from a geological point of view, because there are at least five different types of terrain. Craters exist in many areas, and give the impression of being young and sharp; there is also an extensive grooved plain which is crater-free. Surprisingly, Enceladus may be active, with a liquid interior; if so, we are seeing what is termed cryovulcanism, the icy equivalent of volcanic action. The interior of Enceladus is presumably being tidally flexed by the gravitational pulls of Saturn and the more massive satellite Dione, the orbital period of which is twice that of Enceladus. Re-surfacing has led to the obliteration of old craters.

Tethys Tethys has a very low density, and is probably composed almost entirely of ice. One crater, Odysseus, has a diameter of 400 km larger than Mimas. The main surface feature is a huge trench, running from near the north pole across the equator and along to the south polar region.

Dione Dione is only slightly larger than Tethys, but much denser and more massive. The surface is not uniform. The trailing hemisphere is relatively dark and heavily cratered, the leading hemisphere is much lighter. One prominent feature, named Amata, is associated with a system of bright wispy features which extend over the trailing hemisphere and are accompanied by narrow linear troughs and ridges. Geologically, Dione seems to have been much more active than Tethys.

Rhea Rhea is heavily cratered; as with Dione the trailing hemisphere is the darker of the two, and there are not many really large formations. There are two distinct types of terrain, the first contains craters over 40 km across, while the second, covering parts of the polar and equatorial regions, is characterized by craters of much smaller size. Rhea seems to have a rocky core, around which most of the material is ice.

Iapetus Iapetus is unusual in many ways. The trailing hemisphere is bright and icy, but the leading hemisphere is as dark as a blackboard. The demarcation line is not abrupt; there is a 200–300 km transition zone. The low density of Iapetus shows that the dark areas are due to surface materials, which have welled up from below, but their thickness is not known. It is also notable that many of the craters in the bright areas have dark floors. There have been suggestions that the dark material has been wafted on to Iapetus from the outer satellite Phœoebe, but this seems unlikely, partly because Phœoebe is so small and remote but mainly because the colours do not match. We know little about the dark area, but the bright regions contain craters of the usual type. As seen from Earth, Iapetus is very variable; it is brightest when west of Saturn, with the bright hemisphere facing us.

Minor satellites Pan, discovered on a photograph taken by the Voyager 2 space-craft, moves within the outer division of Saturn's ring system (the Encke Division). Atlas, moves near the edge of Ring A: nothing is known about its surface details. Prometheus and Pandora, the F-ring 'shepherds' have been imaged; both consist mainly of ice, and both are cratered, Prometheus shows ridges and valleys, while Pandora has two 30 km craters. Janus and Epimetheus have the same mean distance from Saturn; every four years they approach each other, and actually exchange orbits. They are irregular in shape, and may well be the remnants of a larger body which has broken up. Predictably, both are cratered.

Outer minor satellites All these are small, though one, known at present as S/2000 S3, may be almost 50 km in diameter. By May 2003 the total number of satellites had risen to 31. The outer minor satellites tend to form clusters, some with direct motion and others retrograde; most of them seem to be fragments of larger satellites which have broken up. Nothing is known about their surface features. No doubt many more tiny satellites await discovery.

Future space probes, and more powerful Earth-based telescopes, will add greatly to our knowledge. Certainly there is no doubt that to the geologist, the giant planets and their satellites are of surpassing interest.

See Also

Solar System: Asteroids, Comets and Space Dust; Neptune, Pluto and Uranus.

Further Reading

Alexander AFOB (1952) *The Planet Jupiter*. London: Faber and Faber.

Alexander AFOB (1958) *The Planet Saturn*. London: Faber and Faber.

Asimov I (1995) *The Ringed Planet Saturn*. Milwaukee: Gareth Stevens.

Beebe H (1991) *Jupiter.* Washington: Smithsonian Institution Press.

Burns M (ed.) (1986) *Satellites.* Arizona: University of Arizona Press.

Clustenis A and Taylor F (1999) *Titan: the Earthlike Satellite.* Singapore: World Scientific.

Greenberg R and Ehahic A (1984) *Planetary Rings.* University of Arizona Press.

Gehres T (ed.) (1986) *Jupiter.* Arizona: University of Arizona Press.

Gehres T (ed.) (1984) *Saturn.* Arizona: University of Arizona Press.

Hunt G and Moore P (1980) *Atlas of Jupiter.* London: Mitchell Beazley.

Hunt G and Moore P (1982) *Atlas of Saturn.* London: Mitchell Beazley.

Morrison B and Samz J (1982) *Voyager to Jupiter.* NASA.

Morrison B (1982) *Voyager to Saturn.* NASA.

Moore P (2001) *Astronomy Data Book.* London: Institute of Physics Publishing.

Peek BM (1981) *The Planet Jupiter.* London: Faber and Faber.

Rugers J (1990) *The Giant Planet Jupiter.* Cambridge: Cambridge University Press.

Rothery D (1992) *Satellites of the Outer Planets.* Oxford: Clarendon Press.

Lang K (2003) *Cambridge Guide to the Solar System.* Cambridge: Cambridge University Press.

Neptune, Pluto and Uranus

P Moore, Selsey, UK

Introduction

Five planets have been known since ancient times: Mercury, Venus, Mars, Jupiter, and Saturn, all of which are prominent naked-eye objects. Since the invention of the telescope, three more planets have been discovered beyond the orbit of Saturn: Uranus in 1781, Neptune in 1846, and Pluto in 1930, though the planetary status of Pluto is a matter for debate. This article describes the outer members of the Solar System, together with their numerous satellites.

Uranus

Uranus was discovered in March 1781 by William Herschel, one of the greatest of all observers. He was Hanoverian by birth, but spent most of his life in England; by profession he was a musician, but his main interest was in astronomy, and he built excellent reflecting telescopes. With one of these he began systematic 'reviews of the heavens', and came across an object which was certainly not a star. It showed a perceptible disk, and it moved from night to night. Herschel believed it to be a comet, but before long its planetary nature became evident.

Uranus is a giant world, but it is not of the same type as Jupiter and Saturn. Rather than being described as a gas-giant, it is better referred to as an 'ice-giant' (Figure 1). The outer atmosphere is made up chiefly of hydrogen (probably about 83% by number of molecules) and helium (15%); methane accounts for 2%, so that there are only traces of other substances. Methane freezes out at a very low temperature, and forms a thick cloud layer above which is the predominantly hydrogen atmosphere. Methane absorbs red light, which is why Uranus appears bluish-green. Minor constituents such as ethane C_2H_6 and acetylene (C_2H_2) play a role in forming 'hazes'.

Below the atmosphere come the 'ices', a mixture of water, methane, and ammonia with traces of other substances. These materials behave as liquids under the temperature and pressure conditions inside the globe. However, Uranus differs from the other giant

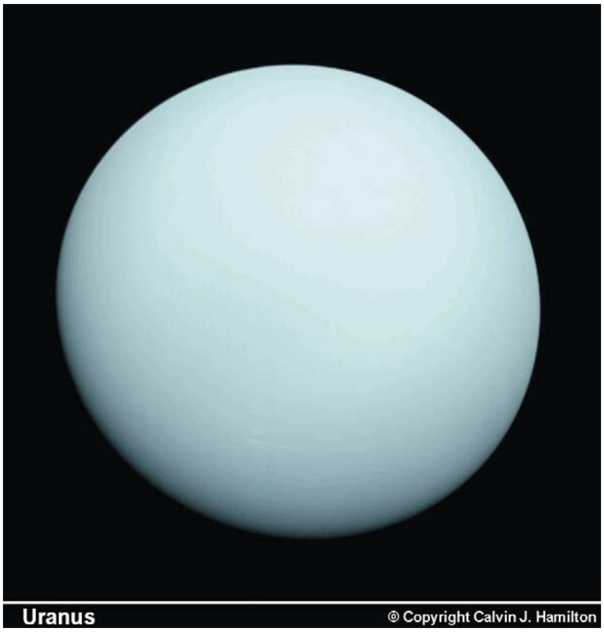


Figure 1 Uranus, imaged from Voyager 2 in January 1986. Compared with the other giants, Uranus is bland in appearance, and there are no well-marked features.

Table 1 Data for Uranus and Neptune

	Uranus	*Neptune*
Distance from Sun	Max 3 005 200 000 km (20.088 a.u.)	4 347 000 000 km (30.316 a.u.)
	Mean 2 869 600 000 km (19.181 a.u.)	4 496 700 000 km (30.058 a.u.)
	Min 2 734 000 000 km (18.275 a.u.)	4 456 000 000 km (29.800 a.u.)
Orbital period	84.01 years.	164.8 years
Synodic period	369.66 days.	367.5 days
Rotation period	17.24 h (17 h 14.4 m)	16.4 h (16 h 7 m)
Mean orbital velocity	6.82 km s^{-1}	5.43 km s^{-1}
Axial inclination, degrees	97.86	28.48
Orbital inclination, degrees	0.773	1.769
Orbital eccentricity	0.0462	0.009
Diameter, km	Equatorial 51 118	50 538
	Polar 49 946	49 600
Mass, Earth-1	14.6	17.2
Volume, Earth-1	64	57
Escape velocity, km s^{-1}	21.1	23.9
Surface gravity, Earth-1	1.17	1.2
Density, water-1	1.27	1.77
Oblateness	0.023	0.02
Albedo	0.51	0.35
Mean surface temperature, °C	−214	−220

planets in that it seems to have at best a very weak internal heat-source. This means that the temperatures of the surfaces of Uranus and Neptune are almost equal, even though Neptune is so much farther from the Sun.

Uranus is unusual in one respect; its axial inclination makes more than a right angle to its orbit, so that the rotation is technically retrograde, though not usually classed as such (Table 1). From Earth the equator of Uranus is regularly presented, as was the case in 1923 and 1966; at other times a pole may lie in the centre of the disk – the south pole in 1985, for instance, and the north pole in 2030. The rotation period is only 17.2 h, so that the Uranian calendar is strange. During one orbit each pole has a 'night', lasting for 21 Earth years, with corresponding daylight at the opposite pole. The reason for the extreme axial inclination is not known.

Uranus has a magnetic field, the polarity of which is opposite to that of the Earth, and the magnetic axis is displaced from the axis of rotation by 57.6°, again for reasons which are unknown; neither does the magnetic axis pass through the centre of the globe – it is offset by 8000 km.

Our main knowledge of Uranus has been drawn from Voyager 2, the only probe to have encountered the planet. Launched in 1977, Voyager 2 encountered Jupiter in 1979 and Saturn in 1981, and on 24 January 1986, it passed over the north pole of Uranus, only about 80 000 km above the cloud-tops, before going on to rendezvous with Neptune in 1989. Clouds were recorded on the globe, and images were obtained of all the major satellites, as well as the ring system.

Table 2 Rings of Uranus

	Distance from Uranus, km	*Width, km*	*Eccentricity*	*Period, h*
6	41 837	1.5	0.0010	6.1988
5	42 235	2	0.0019	6.2875
4	42 571	2.5	0.0010	6.3628
Alpha	44 718	4–10	0.0008	6.5808
Beta	45 661	5–11	0.0004	7.0688
Eta	47 176	1.6	0.004	7.4239
Gamma	47 626	1–4	0.0001	7.5307
Delta	48 303	3–7	0.0	7.6911
Lambda	50 024	2	0.0	8.1069
Epsilon	51 149	20–96	0.0079	8.3823

The rings of Uranus are quite unlike those of Saturn; they are thin and dark, so that they are not easy to study with Earth-based telescopes. Details of the system are given in Table 2. Ten rings are known; their thickness is between 0.1 and 1 km, and only one, the outermost (the Epsilon ring), is of considerable width. This ring is appreciably eccentric; the tiny satellites Cordelia and Ophelia, discovered by Voyager 2, act as 'shepherds' to it. The rings seem to be made up mainly of particles a few metres in diameter.

Uranus has an extensive satellite system; data are given in Table 3. Only Miranda, Ariel, Umbriel, Titania, and Oberon were known before the Voyager mission (the names come from Shakespeare and Pope's poem *Rape of the Lock* – a strange departure from the usual tradition of mythology). All these were imaged by Voyager 2. The largest satellites, Titania and Oberon, are heavily cratered, but are not alike. Titania has clearly been the site of much past tectonic

Table 3 Satellites of Uranus

	Mean distance from Uranus, km	*Orbital period, d*	*Diameter, km*	*Magnitude*	*Escape velocity, km s^{-1}*	*Density, water-1*
Cordelia	49 471	0.330	26	24.2	V. low	?
Ophelia	53 796	0.372	32	23.9	V. low	?
Bianca	59 173	0.433	42	23.3	V. low	?
Cressida	54 777	0.463	62	22.3	V. low	?
Desdemona	62 676	0.473	54	22.5	V. low	?
Julia	64 352	0.493	84	21.7	V. low	?
Portia	66 085	0.513	105	21.1	V. low	?
Rosalind	69 941	0.558	54	22.1	V. low	?
Belinda	75 258	0.663	65	22.3	V. low	?
Puck	86 000	0.762	164	20.4	V. low	?
Miranda	129 400	1.414	481	16.3	0.5	1.3
Ariel	191 000	2.520	1158	14.2	1.2	1.6
Umbriel	256 300	4.144	1169	14.8	1.2	1.4
Titania	435 000	8.206	1578	13.7	1.6	1.6
Oberon	583 500	13.463	1523	13.9	1.5	1.5
Caliban	7 230 000	579	60	22.3	V. low	?
Stephano	8 002 000	676	39	24.0	V. low	?
S/2001 UI	8 571 000	769	20	25	V. low	?
Sycorax	12 179 000	1 283	120	20.7	V. low	?
Prospero	16 418 000	1 992	40	23	V. low	?
Setebos	17 459 000	2 202	40	23	V. low	?

activity; there are ice cliffs, fault valleys, and trench-like features, one of which extends for over 1450 km. Oberon has a brownish surface; some of the craters are dark-floored, and there are several systems of bright rays. Ariel is also cratered, but the dominant features are broad, branching, smooth-floored valleys which look as though they have been cut by fluid, though water is not a likely candidate because of Ariel's small size and low temperature. Umbriel has a darker and more subdued surface, with one prominent feature of uncertain nature near the edge of the best image (it must be remembered that Voyager could cover only half of the total surface). Miranda, passed by Voyager at a mere 3000 km, has an amazingly varied landscape; there are several distinct types of terrain – old, cratered plains, brighter areas with cliffs and scarps, and large, trapezoidal regions known as corona. Large craters are lacking, but there are fault valleys, parallel ridges, and graben up to 15 km across. Miranda presents real problems of interpretation, particularly because the various types of terrain seem to have been formed at different periods. It has been suggested that the satellite may have been shattered and re-formed several times, but this would involve considerable heating, which, in view of Miranda's small size, and icy nature, does not sound probable.

All the remaining satellites are small, only Puck is as much as 100 km in diameter. The five outer satellites have retrograde motion, and are presumably captured asteroids.

Neptune

Neptune was discovered in 1846 by J Galle and H D'Arrest, from the Berlin observatory. Irregularities in the movements of Uranus had enabled the French mathematician UJJ Le Verrier to calculate the position of the body responsible for them (similar calculations by JC Adams in England had given much the same result). Neptune is a twin of Uranus, but the two worlds are by no means identical. Neptune is very slightly the smaller of the two, but it is appreciably denser and more massive, and is a much more dynamic world (Table 1). Unlike Uranus, it has a strong internal heat source, and sends out 2.6% more energy than it would do if it depended entirely upon what it receives from the Sun. It does not have an exceptional axial inclination; at the time of the Voyager 2 pass it was the south pole which was in sunlight. There is a magnetic field, but the magnetic axis is displaced by 47° from the axis of rotation and does not pass through the centre of the globe, so that in this respect Neptune really does resemble Uranus. The interior of Neptune is presumably not unlike that of Uranus, apart from the greater internal heat source.

The atmosphere consists mainly of hydrogen, with an appreciable amount of helium and some methane together with traces of other substances such as hydrogen cyanide, acetylene, and ethane. Voyager identified various cloud layers, at a level where the pressure is 3.3 bars there is a layer which seems to be of hydrogen sulphide, above which are layers of hydrocarbons, with a methane layer and an upper methane haze. Above the hydrogen sulphide layer there are discrete clouds with diameters up to 100 km, casting shadows on the cloud deck 50 to 75 km below. These clouds may be described as 'methane cirrus'. Temperature measurements from Voyager show that there is a cold mid-latitude region, with a warmer equator and pole (we know little about the north pole, which was in darkness during the Voyager encounter). There are strong winds; most of them blow in a westerly direction (that is to say, opposite to the planet's rotation), and are distinctively zonal. At the equator they blow westward at up to 450 m s^{-1}. Further south they slacken, and beyond latitude −50° they become eastward (prograde) up to 300 m s^{-1}, decreasing once more near the south pole. There is, in fact, a broad equatorial retrograde jet between latitudes +45° and −50°, with a relatively narrow prograde jet at around latitude −70°.

At the time of the Voyager encounter the most conspicuous feature on the disk was the Great Dark Spot, a huge oval with a longer axis of 10 000 km, drifting westward relative to the adjacent clouds, it was a high-pressure area, rotating counter-clockwise and showing all the signs of an atmospheric vortex. Hanging above it were methane cirrus clouds, and between these and the main cloud deck there was a clear region 50 km deep. Other, smaller spots were seen at different latitudes, and the whole disk was extremely active. Later images obtained with the Hubble Space Telescope show that the Great Dark Spot has disappeared, so that the surface shows marked changes over relatively short periods.

Neptune has an obscure ring system (Table 4). The outer ring, named after Adams, is 'clumpy', with three brighter arcs, while the Lassell ring is a diffuse band of material containing a high percentage of very small particles. There may be 'dust' extending from the inner Galle ring almost down to the cloud-tops.

Eleven satellites are known (Table 5), but of these only two, Triton and Nereid, were discovered before the Voyager 2 fly-by. Triton is one of the most remarkable bodies in the entire Solar System. It was found by the English astronomer W Lassell a few weeks after the discovery of Neptune itself, and is brighter than any of the satellites of Uranus; it is also more reflective, with an albedo in places of 0.8, and it is the coldest world ever encountered by a space-craft – the temperature is −235°C, a mere 18° above absolute zero. The globe seems to be made up of a mixture of rock and ice. There is an extensive though very tenuous atmosphere, made up almost entirely of nitrogen with a trace of methane. Triton has retrograde motion, and there seems no doubt that it is a captured body rather than a bona fide satellite.

The surface is very varied. There is a general coating of ice, presumably water ice overlain by nitrogen ice; there is little surface relief, and there are few craters. The area surveyed by Voyager 2 was divided into three parts: polar (Uhlanga Regio), eastern equatorial (Monad) and western equatorial (Bubembe Regio). The polar area is covered with a pink cap of nitrogen snow and ice, and there are geysers, completely unexpected before the Voyager mission. Apparently there is a sub-surface layer of

Table 4 Rings of Neptune

	Distance from centre of Neptune, km	*Width, km*
Galle	41 900	2000
Le Verrier	53 200	110
Lassell	53 200	4000
Arago	57 200	100
–	61 950	(indistinct)
Adams	62 933	50

Table 5 Satellites of Neptune

	Mean distance from Neptune, km	*Orbital period, d*	*Orbital eccentricity*	*Orbital inclination*	*Diameter, km*	*Magnitude*
Naiad	48 227	0.29	0.0003	4.74	58	25
Thalassa	50 075	0.31	0.0002	0.21	80	24
Despina	52 526	0.33	0.0001	0.07	148	23
Galatea	61 953	0.43	0.0001	0.05	158	23
Larissa	71 548	0.55	0.0014	0.20	208	21
Proteus	1 17 647	4.12	0.0004	0.04	436	20
Triton	33 476	5.87	0.0000016	157.34	2705	13.6
Nereid	5513	360.14	0.7512	7.23	340	18.7
S/2002 N2	20 200	2525	0.17	57	40	25
S/2002 N3	21 390	2751	0.47	43	40	25
S/2002 N1	21 990	2868	0.43	121	40	25

liquid nitrogen. If any of this migrates upward, the pressure is relaxed and the nitrogen explodes in a shower of ice and vapour, travelling quickly up the nozzle of the geyser-like vent – fast enough to make it rise to several kilometres before falling back; the outrush sweeps dark débris along it, blown by winds in the tenuous atmosphere. The edge of the cap is well-defined, and north of it there is a darker, redder region. Monad Regio is part smooth, part hummocky, with rimless pits (paterae), mushroom-like features (guttae) and low-walled plains; Bubembe Regio is characterized by the so-called cantaloupe terrain – a nickname given to it because of its superficial resemblance to a melon-skin! Fissures cross it, meeting at elevated X or Y junctions; this is probably the oldest part of Triton's surface. Of course, we have no information about the hemisphere which was in darkness at the time of the Voyager encounter.

Nereid, the other satellite known before the Voyager mission, was discovered by G Kuiper in 1949. It is small, only 340 km in diameter, and has a very eccentric orbit, so that its distance from Neptune ranges between 1.35 million km and 9.62 million km. The orbital period is just over 360 days, but it is unlikely that its rotation period is synchronous; as with Hyperion in Saturn's system, the rotation period may be chaotic. It has direct motion, and although not well imaged from Voyager 2, it seems to be fairly regular in shape; there were vague indications of a few large craters. Of the other satellites, Proteus is actually larger than Nereid, but its closeness to Neptune means that it is not observable with Earth-based telescopes. The rotation is synchronous, and the albedo is low; it has been said that Proteus is 'as dark as soot'. The area imaged by Voyager is dominated by a circular depression, Pharos, 225 km in diameter and up to 15 km deep. The remaining inner satellites are small, icy, and presumably cratered; Voyager sent back one image of Larissa. The orbit of the 158 km satellite Galatea is very close to the Adams Ring. Three outer asteroidal satellites were discovered in 2001.

Pluto

Pluto, discovered by Clyde Tombaugh in 1930, is an enigma; it is smaller than the Moon or even Triton, and has an orbit which is both eccentric and inclined. It has one satellite, Charon. Data for Pluto and Charon are given in Table 6.

Table 6 Pluto and Charon

	Pluto	*Charon*
Distance from Sun, km	Max 7 381 200 000 Mean 5 906 400 000 Min 4 445 800 000	
Orbital period, days	90 465 (247.7 years)	6 d 9 h 17 m (round Pluto)
Rotation periods	6 d 9 h 17 m	6 d 9 h 17 m
Mean orbital velocity	4.75 km s^{-1}	0.23
Axial inclination, degrees	122.46	–
Orbital inclination, degrees	17.14	(to Pluto) c-9
Orbital eccentricity	0.2488	0.0076
Diameter, km	2 324	1270
Density, water-1	2.05	2
Volume, Earth-1	0.006	
Mass, Earth-1	0.0022	
Max. surface temperature, °C	about−233	
Escape velocity, km s^{-1}	1.18	0.58
Surface gravity, Earth-1	0.06	0.21
Albedo	0.55	0.36

The calculations leading to the discovery of Pluto were made from 1905 by Percival Lowell, and were based upon perturbations of Neptune and (particularly) Uranus. Searches carried out with the Lowell refractor at Flagstaff in Arizona were unsuccessful, but the planet was finally identified at Flagstaff by Tombaugh, not too far from the position predicted by Lowell. After some discussion it was named Pluto, after the God of the Underworld. When near perihelion, it is closer-in than Neptune, as was the case between 1979 and 1999, but there is no danger of collision, because Pluto's orbit is inclined at an angle of 17°. After the discovery it was found that Pluto had been photographed at Flagstaff in 1915 and Mount Wilson in 1919, but had been missed because it was so much fainter than had been expected. A telescope of fair size is needed to show it at all; Voyager went nowhere near it, but some surface features have been recorded by the Hubble Space Telescope (**Figure 2**). There are indications of a dark equatorial band and brighter polar areas. The axial inclination is 122°, so that, as with Uranus, the rotation is technically retrograde; the rotation period is 6.34 days.

Pluto's density, twice that of water, indicates that the globe contains more rock than in the icy satellites of the giant planets, but our knowledge of the internal structure is decidedly meagre. Pluto may or may not be differentiated, but the gravitational pressure may not be adequate to increase the rock density deep inside the globe to a marked degree. The surface is coated with methane ice at least in some areas. There is a tenuous but surprisingly extensive atmosphere, made up chiefly of nitrogen together with methane. As Pluto moves out toward aphelion, due in the year 2114, the temperature will drop and the atmosphere may freeze out, so that for part of its long 'year' there may be no atmosphere at all.

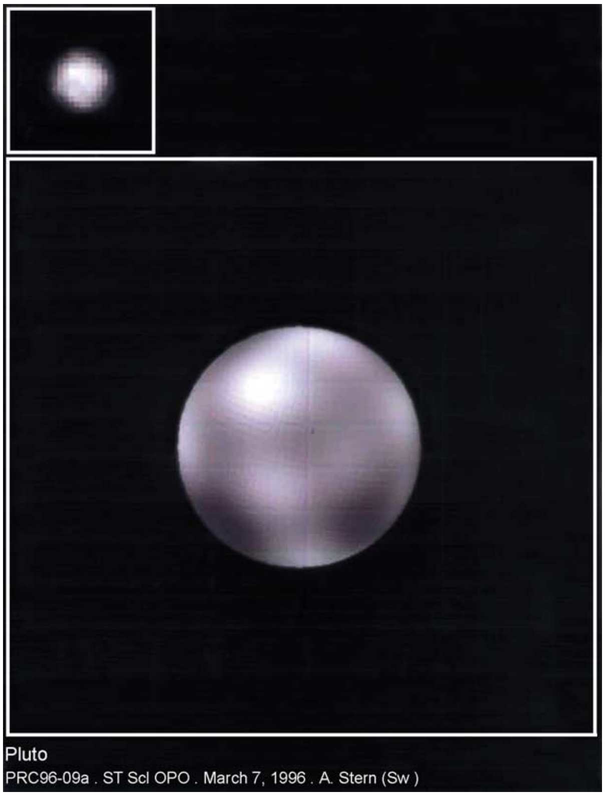


Figure 2 Pluto, imaged with the Hubble Space Telescope. Regions of different brightness are shown, but of course no definite surface features.

Pluto has a companion, Charon, discovered in 1978. Its diameter is more than half that of Pluto, so that it cannot be regarded as a conventional satellite. Its orbital period is the same as Pluto's rotation period, and the surface-to-surface distance between the two bodies is no more than 18 000 km. To an observer on Pluto, Charon would remain 'fixed' in the sky. The density of Charon is lower than that of Pluto; the surface is probably coated with water ice, and there is no trace of atmosphere. Mutual transit and occultation events for Pluto and Charon were observed between 1985 and 1990, and were very informative. When Charon passed behind Pluto it was completely hidden, and Pluto's spectrum could be seen alone, when Charon passed in front of Pluto the two spectra were seen together, so that of Pluto could be subtracted.

From 1992 many small bodies have been found orbiting the Sun beyond Neptune. These make up the Kuiper Belt, named after the Dutch astronomer GP Kuiper, who suggested its existence (*see* **Solar System:** Asteroids, Comets and Space Dust. One of these, Quaoar, may be as large as Charon, and others are around 1000 km in diameter. There are grounds for proposing that Pluto and Charon should be regarded simply as the largest known Kuiper Belt objects, though some astrnonomers are reluctant to deprive Pluto of its planetary status!

Is there another large planet moving far beyond Neptunes and Pluto? Slight irregularities in the movements of Uranus and Neptune indicated that this might be the case, and periodical searches for 'Planet X' have been carried out, but with no success. Recently it has been claimed that improved values for the masses of Uranus and Neptune show that no unexplained perturbations occur, and there is no need for Planet X, but one thing is certain; Pluto's mass is too small to cause the effects which led Lowell to calculate a position for it. Therefore, either Lowell's reasonably correct result was purely fortuitous, or else the planet for which he was searching has yet to be discovered. If Planet X really exists, it will no doubt eventually be found.

Sedna, discovered in 2004, may be 1300 km in diameter, its orbital period is over 12 000 years and it may recede almost as far as the Oort Cloud.

See Also

Solar System: Asteroids, Comets and Space Dust; Meteorites; Jupiter, Saturn and Their Moons.

Further Reading

Chuikshank D (ed.) (1995) *Neptune and Triton.* Arizona: University of Arizona Press.

Elliott and Kehh (1984) *Rings.* Cambridge Massachusetts: MIT Press.

Miner E (1998) *Uranus.* New York: Wiley/Praxis.

Hunt G and Moore P (1990) *Atlas of Uranus.* Cambridge: Cambridge University Press.

Hunt G and Moore P (1995) *Atlas of Neptune.* Cambridge: Cambridge University Press.

Miner ED and Wessen RR (2002) *Neptune.* Berlin, Heidelberg, New York: Springer-Verlag.

Stern A and Mitton J (1999) *Pluto and Charon.* New York: Wiley.

Tombaugh C and Moore P (1980) *Out of the Darkness: the Planet Pluto.* London: Stackpole and Lutterworth.

Bergstralh, *et al.* (eds.) (1991) *Uranus.* Arizona: University of Arizona Press.

Moore P (1996) *The Planet Neptune.* New York: Wiley.

SPACE DUST

See **SOLAR SYSTEM: Asteroids, Comets and Space Dust**

STRATIGRAPHICAL PRINCIPLES

N MacLeod, The Natural History Museum, London, UK

Introduction

Stratigraphy is the branch of geology that deals with the formation, composition, sequence, and correlation of stratified rocks. Since the whole Earth is stratified, at least in a broad sense, bodies of all the different types of rock – igneous, sedimentary, and metamorphic – are subject to stratigraphic study and analysis. In most cases, however, stratigraphy focuses on the evaluation of sedimentary rock strata. Modern principles of stratigraphic analysis were developed in the eighteenth and nineteenth centuries by geologists such as Niels Stensen, James Hutton, Georges Cuvier, William Smith, and Charles Lyell. By 1900 all the intellectual tools needed for the description, sequence, and correlation of strata were in place. Shortly after 1900, the tools needed to establish the absolute ages of minerals containing unstable radioisotopes also became available, giving stratigraphers a physical basis for making chronostratigraphic correlations, at least in certain favourable stratigraphic situations. Since the 1950s efforts have been made to establish international standards for stratigraphic nomenclature and the usage of stratigraphic terms and the internationally agreed designation of 'type-sections' or stratotypes for various sorts of stratigraphic unit, especially those relating to chronostratigraphy.

First Principles

The study of stratigraphy began with attempts to understand common observations, such as what the rocks we call fossils are and how the rocks that comprise mountains came to be elevated above the land surface. Of course, both fossils and mountains were well known to ancient Greek natural historians, such as Plato, Hereodotus, Aristotle, Xenophanes, and Pliny. Although a variety of explanations were offered for these phenomena, no systematic investigations of modern aspect were carried out by these classical scholars, according to the intellectual style of their time. The organic nature of fossils was recognized by a number of Renaissance scholars, including Leonardo da Vinci (1452–1519) and Conrad Gesner (1516–1565). Da Vinci's writings were particularly prescient in that he recognized that fossil mollusc shells from the tops of mountains were similar to the shells of modern molluscs and that this similarity implied that sediments occupying the mountain tops must originally have been deposited beneath marine waters. These were isolated musings, however.

The first modern treatment of a stratigraphic problem was published by Niels Stensen (1638–1686, also known by his anglicized literary name, Nicholas Steno) in 1669 (*see* **Famous Geologists:** Steno). Most scholars mark Steno's *De solido intra solidum naturaliter contento disseratiinis prodomus* as the first stratigraphic treatise. In this short work–which was presented to Steno's patron, the Grand Duke Ferdinand II of Tuscany – Steno establishes three cardinal principles of stratigraphic analysis and then uses these to reconstruct the geological history of Tuscany. Steno's principles are as follows.

1. Original horizontality: unconsolidated sediments deposited on a solid base must have originally formed horizontal layers, since the sediment particles would have 'slithered' to the lowest point. Thus, consolidated strata inclined at an angle must have become tilted after consolidation.
2. Original continuity: layers of unconsolidated sediments deposited on a solid base would have formed continuous sheets of material. Thus, bands of consolidated sediments whose ends have been broken must have experienced this breakage and erosion after consolidation.
3. Superposition: since each layer of unconsolidated sediment deposited on a solid base must have formed after the basal layer had been deposited, overlying layers of sediment are younger than underlying layers.

Using these principles, Steno argued that Tuscan geology, and especially the stratified sediment layers forming its mountains, represented the remains of a series of subterranean-erosion and land-surface-collapse events (Figure 1). Not only did this model reconcile the cyclic and directional aspects of the Tuscan stratigraphic record, it also established the principal of stratigraphic correlation as the matching of stratigraphic observations from distant outcrops in order to obtain a sense of a rock body's geometric structure (Figure 2).

The next significant contribution to stratigraphic principles was made in 1785 by the Scottish lawyer–gentlemen farmer James Hutton (1726–1797) (*see* **Famous Geologists:** Hutton), who stressed the cyclic aspects of the stratigraphic record in his doctrine of

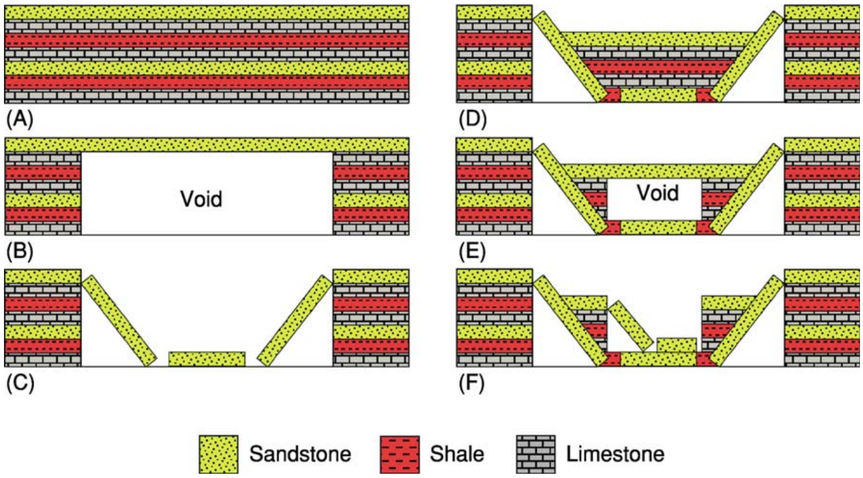


Figure 1 Steno's conceptual interpretation of the stratigraphic history of Tuscany. (A) Flat-lying continuous sediments were deposited beneath marine waters. (B) Lithified sediments were uplifted, and subterranean voids or caverns developed through the erosive action of subsurface waters. (C) When the subterranean voids grew sufficiently large, the roofing layers collapsed, elevating the cavern walls, down-dropping flat-laying layers that remained intact, and tilting blocks adjacent to the elevated areas. (D) Submergence of the entire land surface once again caused flat-lying continuous sediments to be deposited. (E) These new sediments were lithified and uplifted, after which new cavernous voids developed. (F) A new round of erosional collapse further modified the landscape. Note how Steno's model encompasses both the apparently directional nature and the cyclic nature of stratigraphic deposits and landscape formation. (Redrawn from Steno's diagram in *De solido intra solidum naturaliter contento disseratiinis prodomus*.)

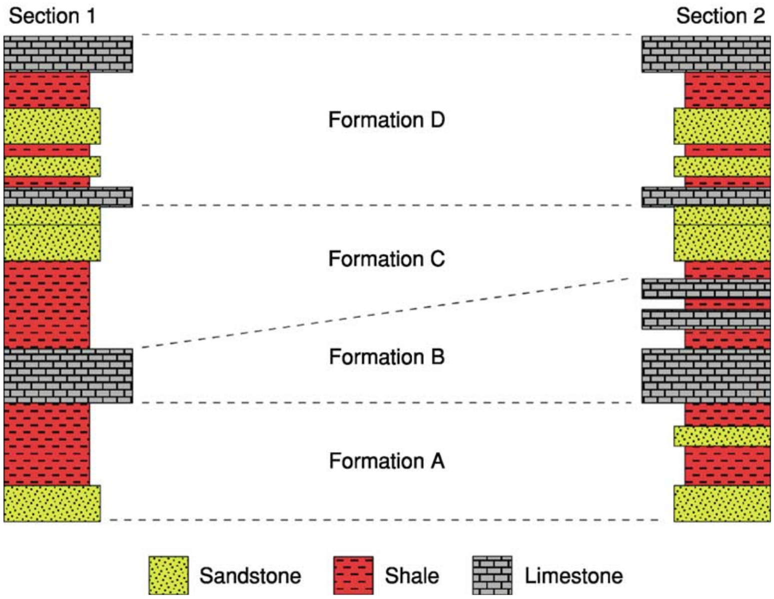


Figure 2 In addition to developing his theory of landscape formation, Steno stressed the importance of stratigraphic correlation – the matching of stratigraphic sequences between outcrops. In this illustration two hypothetical outcrop sections have been correlated based on rock type and subdivided into lithologically unified packages of strata.

uniformitarianism. Citing evidence from the angular unconformities exposed at such Scottish localities as Jedburgh and Siccar Point, Hutton reasoned that the originally horizontal marine sediments of the lower succession must have been consolidated, then tilted as they were raised up above the water's surface, planed off by erosion, submerged, and buried by additional horizontally deposited sediments, which were then consolidated, before the entire sequence was lifted again to become the rock bodies we see before us at these, and other, localities. Hutton believed that these erosion–deposition–uplift cycles had been repeated endlessly in Earth history, implying that: first, the Earth itself is very old; second, the processes we see working today (e.g. erosion, deposition, gradual uplift) operated in the past; third, the power for uplift came from the heat generated by compaction, supplemented by heat at depth left over from the Earth's initial formation; and, fourth, the ultimate purpose of this system was to produce a self-renewing Earth that was 'adapted to the purposes of man'. In particular, Hutton denied that fossils provided any evidence for the directional passage of time because each uniformitarian cycle's biota was 'perfect'.

Slightly later (in 1812), the French Baron Georges Cuvier (1769–1832) (*see* **Famous Geologists:** Cuvier) published a summary of his palaeontological studies in the Paris Basin in his book *Recherches sur les Ossemens Fossils*, the first chapter of which took issue with Hutton's uniformitarian approach to stratigraphic analysis. Cuvier argued that the abrupt disappearances of entire fossil marine faunas that characterize several horizons within this basin, and the equally abrupt appearances of new terrestrial faunas in strata lying just above these marine beds, were evidence for the repeated, sudden, and, in ecological terms, catastrophic elevation of the land. In contrast to Hutton, who believed in endless cycles, Cuvier and his colleagues – who came to be known as 'catastrophists' – envisioned an Earth whose internal core was undergoing constant thermal contraction. As this core pulled away from the hard crust gaps opened up. It was these gaps that were responsible for the catastrophes. In a manner analogous to that proposed in Steno's model, crustal failure occurred when the subterranean gaps become too large to support the burden of the overlying crust. It was supposed that these failures happened suddenly, down-dropping entire regions, the surrounding parts of which would appear to be thrust up (in relative terms) as mountains. Unlike Hutton's endless uniformitarian cycles, Cuvier's hypothesis of Earth history was resolutely directional and finite. The Earth would eventually cool to the point where no more contraction would take place, thus bringing the catastrophes to an end. Also, unlike Hutton, the catastrophists saw extinction as a real phenomenon, with new biotas responding to the changed environment in unique ways.

The next major contribution to stratigraphy was made by the English canal surveyor and geologist William Smith (1769–1839) (*see* **Famous Geologists:** Smith). Smith was the first to recognize the difference between lithostratigraphy (the characterization of rock strata by the kind and/or arrangement of their mineralogical constituents) and biostratigraphy (the characterization of rock strata by their biological constituents). Before Smith, the remains of once-living creatures and the mineral particles of which sedimentary rocks are made were considered to be of equal value in recognizing strata. Smith made a conceptual distinction between lithological and palaeontological sources of stratigraphic information and, by careful analysis of the fossils contained in stratigraphic bodies, demonstrated that strata with very similar lithological constituents could be distinguished on the basis of their fossil content. Even more importantly, Smith showed that the successive biotas preserved in the sedimentary strata of the British Midlands always occurred in the same sequence, regardless of the character of the local lithological sequences. This key stratigraphic principle later became known as the 'principle of faunal succession' (Figure 3). By applying the principle of faunal succession to his biostratigraphical observations, Smith was not only able to predict more accurately the types of rock that would be encountered during canal construction, but also able in 1815 to produce the first modern geological map. While William Smith was not given to abstract theorizing, his commitment to field observations, his willingness to accept those observations at face value, and his use of fossil extinction events as a basis on which to recognize the directional passage of time were far more in line with the philosophical tenets of catastrophism than with those of uniformitarianism.

Uniformitarianism's champion was Charles Lyell (1797–1875) (*see* **Famous Geologists:** Lyell). Lyell accepted the cyclic nature of Huttonian uniformitarianism to the extent that he denied the possibility of both extinction and evolution (though, to be fair, it must be said that the latter was denied by Cuvier as well, albeit on different grounds). Lyell also emphasized and greatly developed Hutton's idea of a mechanistic uniformitarianism in which known natural laws and processes operated at rates comparable to those observed today. Lyell believed that these mechanisms were responsible for all features of the geological record. It is interesting to note that Cuvier, Agassiz, (*see* **Famous Geologists:** Agassiz) and the other

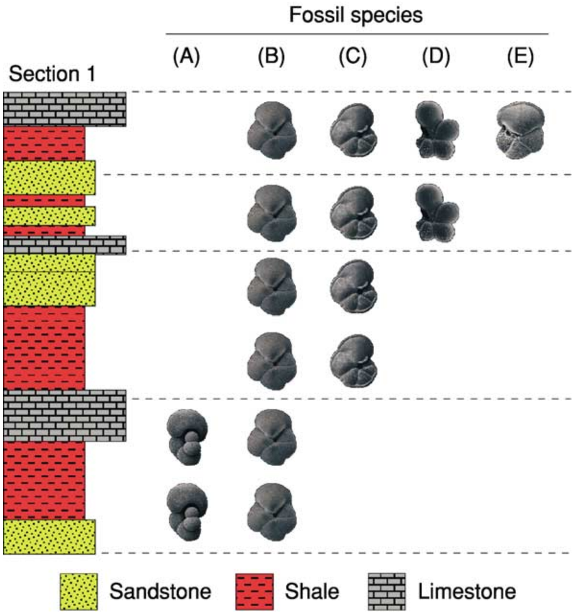


Figure 3 Organization of stratigraphic sequences into units based on their fossil content using the principle of faunal succession. In this illustration the distributions of five fossil planktonic foraminiferal species have been used to recognize stratigraphic units on the basis of unique associations of species. The principle of faunal succession can be used to recognize stratigraphic units because fossil species are individualized in the sense that they have definite and unique starting (speciation) and ending (global extinction) points. Thus, the stratigraphic range of a fossil species encompasses a distinct time interval. Note also that palaeontologically defined stratigraphic units differ in both number and kind from lithologically defined units (compare with Figure 2).

scientific catastrophists were aligned with Lyell in accepting the principle of mechanistic uniformitarianism. Lyell summarized his arguments, and supported them with examples drawn from his geological travels throughout Europe, in a massive three-volume work *Principles of Geology* published in 1830–1833.

While the uniformitarian–catastrophist debate has often been portrayed as a triumph of dispassionate scientific reason over theologically driven special pleading, with the Lyell uniformitarians founding the sciences of stratigraphy and sedimentary geology as we know them today, a more faithful description of the historical record reveals a far more interesting story. Lyellian uniformitarianism did indeed triumph, but not so much over the scientific catastrophism of Cuvier, Brongniart, d'Orbigny, and Agassiz as over the theological catastrophism embraced by the school of Natural Theology (especially in England) and sheer scientific fantasy. Lyell's reasoned approach, which emphasized modern processes working over long periods of time, appealed to many, not least Charles Darwin (*see* **Famous Geologists:** Darwin) who read Lyell's treatise during the *Beagle* voyage and used Lyellian principles as a basis for his geological explorations. Lyell's commitment to the basic uniformitarian doctrine of endless and ahistorical cyclicity, however, was not accepted even among Lyell's contemporaries. Lyell was caricatured for his position by Henry de la Beche in a famous cartoon (Figure 4) and was forced to retract from it by stages in subsequent editions of his *Principles* volumes. Neither was Lyell's view of the value of fossils in stratigraphic correlation – at least for higher taxonomic groups – accepted by his contemporaries, much less by contemporary stratigraphers. Modern uniformitarianism is a combination of the Huttonian–Lyellian emphasis on modern observable processes operating over long periods of time, but nevertheless allowing for the incorporation of processes that have no modern counterpart (e.g. Louis Agassiz's continental glaciations, enormous flood-basalt volcanic eruptions, asteroid impacts), and a thoroughly catastrophist emphasis on extinction, the existence of intervals of (geologically) rapid and widespread global change, and the directional nature of geological time.

Following Smith's demonstration of the power of biostratigraphy, the forefront of stratigraphic research turned to the identification of biostratigraphic zones that could be used to facilitate long-range stratigraphic correlation (e.g. intrabasinal, interbasinal, and intercontinental). This immediately raised a further conceptual problem. Did the identification of the same biozone in different localities mean that the resulting correlation located the two sections in terms of their position in the sequence of biotas preserved over geological time (homotaxis) or in terms of geological time itself (homochrony)? These concepts are distinct because the same sequence of events could be preserved at different localities without the individual events having taken place at the same time.

Until 1900 stratigraphers had been forced to couch their observations in terms of relative time (e.g. event A took place before or after event B) because there was no way to measure absolute time in stratigraphic successions. Of course, attempts to estimate absolute time were made, usually based on modern rates of sediment accumulation and estimates of compaction ratios for different types of sedimentary rock. Nevertheless, since these rates and ratios vary widely, and since there was no way of confirming that any given estimate was correct, such calculations were approximate at best.

This situation changed in the early 1900s, however, with the discovery of natural radioactivity and unstable radioisotopes of naturally occurring elements. Radioisotopes have unstable nuclei that spontaneously decay

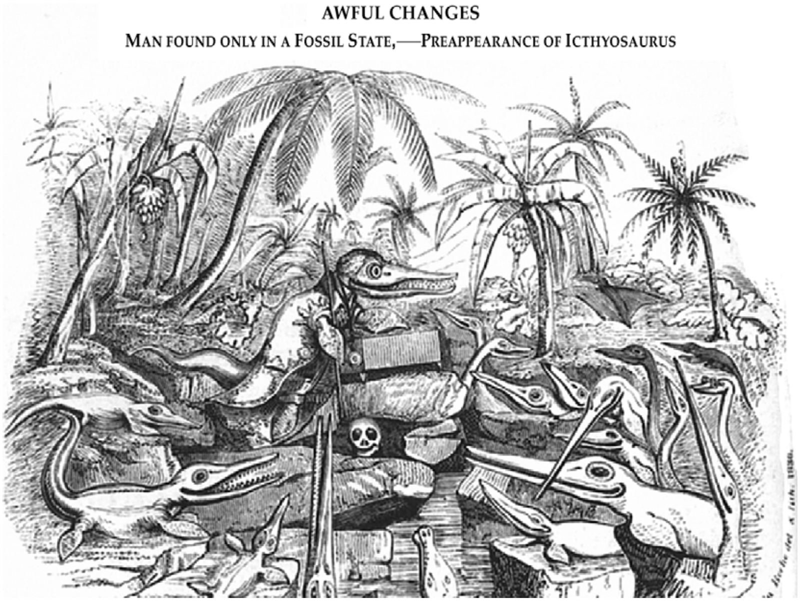

AWFUL CHANGES
MAN FOUND ONLY IN A FOSSIL STATE,—PREAPPEARANCE OF ICTHYOSAURUS

Figure 4 Henry de la Beche's caricature of Charles Lyell as 'Professor Icthyosaurus' lecturing to an eager audience of saurian students at some time in the future on the topic of an 'insignificant' and 'lower-order' fossil animal from Earth's distant past (the representation of a human skull in the drawing's centre). This image, which was originally published as the frontispiece to Frank Buckland's *Curiosities of Natural History*, poked fun at Lyell's belief in the cyclic, or uniformitarian, nature of geological processes, which predicted the re-emergence of extinct fossil forms when future environmental conditions matched those of the past. Modern uniformitarianism no longer embraces this aspect of Hutton and Lyell's original formulation but represents a dynamic amalgam of nineteenth century uniformitarian and catastrophist theory. See text for discussion.

through the emission of subatomic particles from the isotope's nucleus at a fixed and measurable rate. Daughter isotopes are produced as the products of this decay process, along with various types of radiation. If the amounts of original radioisotopic material of a specific type in a particular mineral species and the daughter-product isotope are known, the absolute age of the mineral can be calculated, subject, of course, to several assumptions (e.g. a correct value of the decay constant, accurate measurements, no loss of daughter-product isotope) (*see* **Time Scale**).

Unfortunately, accurate isotopic dating cannot usually be carried out on sediments directly. Most sedimentary rocks are composed of mineral grains whose origin predates that of the sedimentary rock body by a substantial time interval. In some instances, though, a layer of volcanic material (e.g. an ash-fall tuff) with newly formed mineral crystals can become interbedded in a suite of sedimentary rock. In such cases, the age obtained from the volcanic deposit can be used to constrain the ages of the immediately overlying and underlying sediments, subject, once again, to assumptions. By using isotopically datable materials located stratigraphically near major biostratigraphically defined boundaries in the stratigraphic record (*see* below and **Time Scale**), it is possible to estimate absolute ages for these boundaries.

Stratigraphic Classification

As stratigraphers combined the principles of stratigraphic analysis set down by Steno, Hutton, Cuvier, Smith, Lyell, and others with lithostratigraphic, biostratigraphic, and geochronological observations during the first half of the twentieth century, the true geometric relations between observed lithostratigraphic and biostratigraphic units emerged, along with their mutual relations to an entirely conceptual 'chronostratigraphy' (the characterization of rock strata by their temporal relations). These concepts are illustrated in Figure 5, and are usually discussed in terms of the distinction between rock stratigraphic units (that are distinguished by physical or biotic criteria that can be observed at the outcrop, core, well-log, etc.) and time stratigraphic units (that are in all cases inferences based on stratigraphic observations, but have the advantage of being referable to a common geological time-scale). There has been, and continues to be, much confusion over the use of these terms, primarily because of the genuine subtlety of the distinction, but also because of problems stemming from the definition of certain sorts of stratigraphic unit (e.g. biostratigraphic Oppel zones, which are defined on rock-stratigraphic criteria chosen for their supposed ability to achieve time-stratigraphic

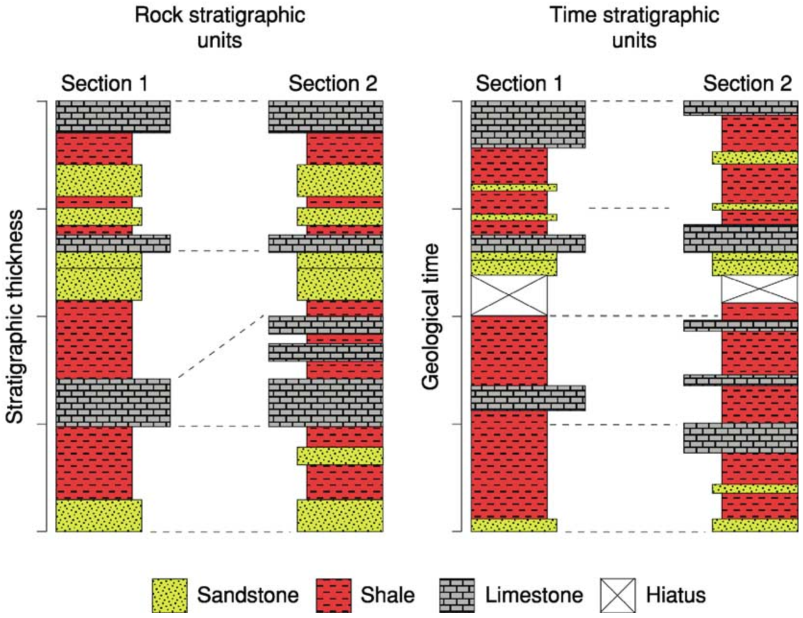


Figure 5 The difference between rock stratigraphic units (left) and time stratigraphic units (right). In this illustration the rock stratigraphic units, along with their lithostratigraphic correlation, are scaled to stratigraphic thickness, as they would be observed in a field study. When these same sections are portrayed as time stratigraphic units (and organized according to the time intervals during which they were deposited), however, the character of their comparative relations (both inter-sectional and intra-sectional, as well as their inter-section correlations) changes.

correlations) and the fact that many stratigraphers prefer to report their rock-stratigraphic observations (e.g. position in a measured section or core) in terms of time-stratigraphic inferences.

In order to stabilize stratigraphic classification and nomenclature, the International Subcommisson on Stratigraphic Classification (ISSC) (originally the International Subcommission on Stratigraphic Terminology (ISST)) was created in 1952 at the 19th International Geological Congress (Algiers). From 1952 to 1965 the ISSC operated as a standing committee under successive international geological congresses. In 1965 responsibility for the ISSC was transferred to the International Union of Geological Sciences (IUGS), where it remains. The ISSC maintains a web site at http://www.geocities.com/issc_arg/. The ISSC has several purposes. Among these is to publish and maintain the International Stratigraphic Guide, whose purpose is to promote international agreement on principles of stratigraphic classification and to develop a common internationally acceptable stratigraphic terminology and rules of stratigraphic procedure. The various stratigraphic-unit concepts and definitions currently recognized by the ISSC are summarized briefly below.

Lithostratigraphic Units

The basic unit of lithostratigraphy is the formation, which is the smallest mappable rock unit possessing a suite of lithologic characteristics that allow it to be distinguished from other such units. Formations need not be lithologically homogeneous, but the entire interval of strata should be diagnosable. Moving up the lithostratigraphic hierarchy to more inclusive units, a set of contiguous formations may be combined to form a group (e.g. the Lias Group), membership of which is usually identified based on common lithological characteristics (e.g. dominantly argillaceous facies) or genetic characteristics (e.g. a suite of formations bounded by two basin-wide unconformities). Occasionally, contiguous groups will themselves be placed into subgroups or supergroups (e.g. the Newark Supergroup, the Wealden Supergroup) based on genetic characteristics. Subgroups and supergroups may also include formations not previously assigned to a group. The most inclusive lithostratigraphic unit is a complex, which is distinguished by its diverse lithological composition – including sedimentary, metamorphic, and/or igneous rocks – and its intricate structure.

Moving down the lithostratigraphic hierarchy to more exclusive units, a member is a subdivision of a formation, recognized on lithologic criteria (e.g. the sandy member of a formation representing a suite of deltaic strata). Typically, members consist of more than a single bed, although some massive bodies with no internal stratification are recognized as members. The smallest formal lithostratigraphic unit is a bed, which is a thin lithostratigraphically monotonous sequence with some locally unique lithological character (e.g. the *Hypsilophodon* Bed). A hypothetical example of this lithostratigraphic hierarchy is presented in Figure 6.

Igneous and/or metamorphic rock bodies of tabular form and stratified nature may be admitted within this lithostratigraphic classification, either in themselves or in combination with adjacent sedimentary units. Igneous rock bodies that cut across stratified rocks of any type can be handled within this scheme under the informal designation of being associated with (in the sense of 'bounded by' or 'included within') a larger, formal, lithostratigraphic unit.

Biostratigraphic Units

The basic unit of biostratigraphy is the biozone, which is any unit of rock distinguished from other such units on the basis of its fossil content. Unlike formations, biozones do not need to be mappable units and so can vary greatly in thickness and geographical extent. Biozones may be defined on the basis of a wide variety of criteria (*see* **Biozones**). Intervals of strata between biozones that lack fossils are referred to as barren interzones, while barren intervals within biozones may be termed barren intrazones. Moving up the biostratigraphic hierarchy, a set of contiguous biozones may be grouped into superbiozones. Superbiozones do not need to be genetically linked in the same way as do the higher-level lithostratigraphic units, but some justification for the designation should be made at the time of the superbiozone's proposal. Biozones may also be subdivided into sub-biozones in order to express finer levels of biostratigraphic detail or identify a biotically distinctive regional grouping of strata. The term zonule is used to refer to a biostratigraphically diagnosable unit that is subordinate to a sub-biozone. Finally, individual stratigraphic surfaces characterized by a distinctive biotic component are referred to as biohorizons. A hypothetical example of this biostratigraphic hierarchy is presented in Figure 7.

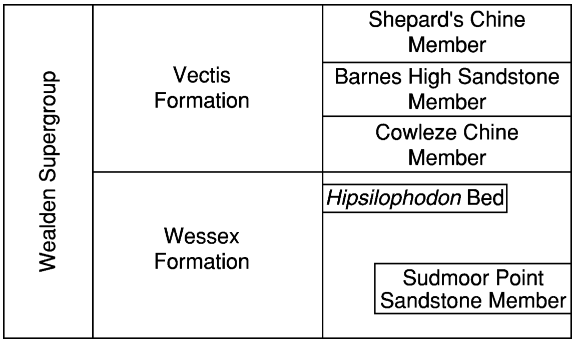


Figure 6 The use of lithostratigraphic units to subdivide a classic Lower Cretaceous suite of non-marine sediments in the Wessex Basin of Great Britain. See text for discussion.

Chronostratigraphic and Geochronological Units

Chronostratigraphic units comprise groups of strata that are recognized as having formed during a specific interval of geological time. While chronostratigraphic terms are conceptual rock stratigraphic units, their classification is mirrored by the geochronological or time-stratigraphic classification scheme. To understand the difference between these two scales, consider an hourglass. Sand falling through the neck of the hourglass is deposited in the lower reservoir over a certain time interval (say 1 h). A chronostratigraphic unit is equivalent to the sand deposit, while the associated geochronological unit is equivalent to the amount of time over which the sand deposit accumulated (1 h). The chronostratigraphic unit accumulated over the time interval and can be said to represent that interval in terms of the deposit's thickness and

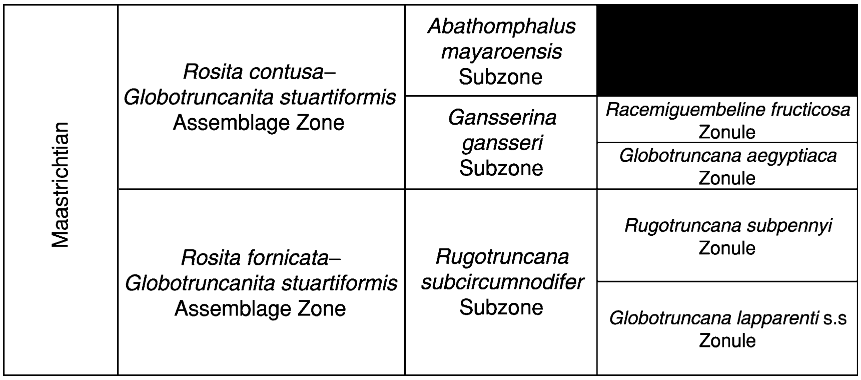

Maastrichtian
Rosita contusa–Globotruncanita stuartiformis Assemblage Zone
Rosita fornicata–Globotruncanita stuartiformis Assemblage Zone
Abathomphalus mayaroensis Subzone
Gansserina gansseri Subzone
Rugotruncana subcircumnodifer Subzone
Racemiguembeline fructicosa Zonule
Globotruncana aegyptiaca Zonule
Rugotruncana subpennyi Zonule
Globotruncana lapparenti s.s Zonule

Figure 7 The use of biostratigraphic units to zone a classic Upper Cretaceous suite of deep-marine sediments in north-central Texas on the basis of their planktonic foraminiferal content. Note the chronostratigraphic series unit (Maastrichtian) and that not all sub-biozones are divided into zonules. See text for discussion.

extent. But the sand deposit itself cannot be said to be time. Table 1 lists the chronostratigraphic and geochronometric unit equivalents.

The application of chronostratigraphic unit classification may be illustrated by the chronozone, which is equivalent to a geochronological chron. All stratigraphic intervals represent potential chronozones and chrons, as do all lithostratigraphic and biostratigraphic units. For example, the (hypothetical) *Exus alphus* biozone represents a chronozone that begins with the stratigraphic horizon deposited at the time of the speciation of this (hypothetical) species and ends with the stratigraphic horizon deposited at the time of its global extinction (Figure 8). This chronozone corresponds to the chron, which is defined as the time interval between this species' global speciation and extinction events. Both the chronozones and the chrons are worldwide in extent, though it may not be possible to recognize either in localities remote from the geographical range of the species. The chronozones and chrons will also be estimates (at least for biostratigraphic zones) and are subject to revision.

Stages (equivalent to geochronological 'ages') are the most common chronostratigraphic unit and are usually defined on the basis of the chronozones of a series of biozones (e.g. the Maastrichtian Stage/Age). Note that biozone boundaries themselves cannot be used to achieve a true chronostratigraphic system because they are inherently diachronous (see Figure 8). Stages

Table 1 Nomenclatural equivalents with examples

Chronostratigraphic units	*Geochronological units*	*Example*
Eonathem	Eon	Phanerozoic
Erathem	Era	Mesozoic
System	Period	Cretaceous
Series	Epoch	Upper Cretaceous
Stage	Age	Maastrichtian
Chronozone	Chron	*Belemnella occidentalis* Zone

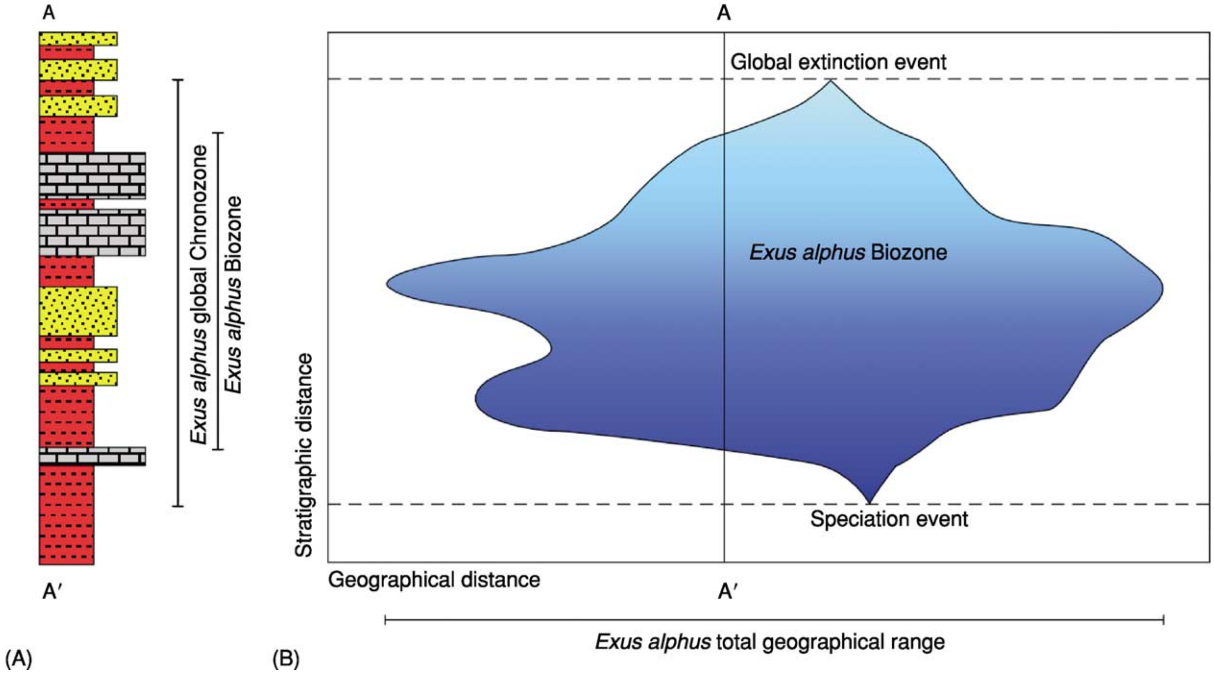


Figure 8 Relation between the (hypothetical) *Exus alphus* Biozone and its corresponding chronozone. (A) Zone expression at the level of a local stratigraphic sequence A–A′. (B) Two-dimensional slice through the *Exus alphus* Biozone containing the A–A′ sequence. Note that the rock-stratigraphic expression of the local biozone (which also represents a local chronozone) underestimates the extent of the global chronozones based on the interval between the species' speciation and global extinction horizons. See text for discussion.

may be subdivided into substages. Systems (equivalent to geochronological 'periods') are composed of a sequence of stages. For example, the Induan, Olenekian, Anisian, Laningian, Carnian, Norian, and Rhaetian stages/ages, all of which are defined on the basis of biochronozones, combine to form the Triassic System/ Period. Similarly, erathems (equivalent to geochronological 'eras') are composed of a sequence of systems. Three erathems/eras are currently recognized, the Palaeozoic, Mesozoic, and Cenozoic. Finally, eonathems (equivalent to geochronological 'eons') are composed of a sequence of erathems. Thus, the Palaeozoic, Mesozoic, and Cenozoic combine to form the Phanerozoic Eonathem/Eon. This was preceded in geological time successively by the Proterozoic and Archaean eonathems/eons.

There have been several recent proposals (*see* **Time Scale**) to dispense with the dual chronostratigraphic/ geochronological classifications of rock units and time units in favour of a single scheme based on the current geochronological classification. Under this scheme, little-used terms such as 'eonathem', 'erathem', 'age' and 'chron' would be considered redundant in favour of their more frequently seen geochronological equivalents. The reasoning behind this proposal does not challenge the logical distinction between the rock-stratigraphic and time-stratigraphic concepts of the dual classification, but instead appeals to the advantages of simplicity, greater ease of imparting the relevant concepts to students, editorial efficiency, and the fact that acceptance of the concept of global stratotype sections and points (GSSP) (see below) by most stratigraphers has rendered – at least to some – the dual rock–time system unnecessary. On the other side of this argument is the simple fact that, by blurring the distinction between rock-stratigraphic and time-stratigraphic units, one is, to some extent, blurring the distinction between observation (rock) and interpretation (time) in chronologically orientated stratigraphic investigations and, in doing so, losing the ability to distinguish between the two. What will become of these proposals remains to be seen.

Other Types of Stratigraphic Unit

With the advent of geophysical methods of analysis, several special types of lithostratigraphic classification have been developed to take advantage of the chronostratigraphic implications of such methods. Perhaps the best example is the study of rock magnetism, which can be used in some lithologies to determine the ancient polarity of the Earth's magnetic field. Based on such observations, a magnetozone can be defined as an interval of strata possessing a characteristic magnetic polarity, either normal or reversed (*see* **Magnetostratigraphy**). These can then be related to time through the use of the chronostratigraphic equivalent of the magnetozone, the magnetochron.

Magnetozones are particularly useful for chronostratigraphic analysis because the time interval over which the Earth's magnetic field changes polarity is short compared with the duration of the magnetozones, biozones, and formations. However, magnetozones can rarely be recognized on the basis of their magnetic properties alone, necessitating the use of other types of stratigraphic analysis – usually biostratigraphy – to achieve the identifications. This increases the complexity of the analysis (and the corresponding chance of error) significantly. Nevertheless, combined magneto-bio-chronostratigraphic analysis has resulted in marked improvements in our understanding of the stratigraphic record. Other types of lithostratigraphic observation that have proven useful in this context include chemical stratigraphy, isotope stratigraphy, seismic stratigraphy, climate stratigraphy, cycle stratigraphy, and orbital stratigraphy.

Stratotypes

With recognition of the distinction between rock stratigraphic units and time stratigraphic units, the ISSP recognized the need to designate 'type-sections' or stratotypes that would constitute standards of reference for various sorts of stratigraphic unit. There are two primary kinds of stratotype: unit stratotypes, which serve as the standard of definition for a stratigraphic unit, and boundary stratotypes, which serve as the standard of definition for a stratigraphic boundary. Unit stratotypes can be either single sections or suites of sections that, when taken together, form a composite unit stratotype. The primary requirement for a stratotype is that it adequately represents the concept of the stratigraphic unit or boundary in all essential particulars. This ideal, however, is rarely met in practice. All real stratigraphic sections exhibit a collection of generalized and idiosyncratic characteristics, and no stratigraphic section can be regarded as truly representative of all other sections and cores worldwide. In addition, disagreements over which section to select as an official ISSP-recognized stratotype have tended to incorporate appeals to historical precedent, priority, and even nationalism, as well as more objective scientific criteria. There is also the ever-present danger that new discoveries might render a designated stratotype incorrect. For example, the boundary at the base of the Cambrian System is defined as the level of the first occurrence of the trace fossil *Treptichnus pedum*, which was thought to occur

2.4 m above the base of Member 2 of the Chapel Island Formation at Fortune Head, Newfoundland, but which subsequent investigations have shown occurs at least 4 m below that horizon in the same section. (Note that, in recognition of the inherently provisional nature of stratigraphic boundary definitions, the ISSP now provides a procedural means for updating boundary stratotype definitions.) Despite these practical deficiencies, the stratotype concept has proven to be popular and has undoubtedly contributed to stabilizing the definitions of stratigraphic units.

One recent modification of the boundary-stratotype concept that has proven to be particularly useful is the 'topless' mode of boundary-stratotype designation. Under this convention, a boundary stratotype designated to serve as the reference for the base of one stage is automatically regarded as defining the top of the underlying stage. This convention elegantly solves the problem of designating unit stratotypes for two successive stages and then finding that the upper boundary of the lower unit and the lower boundary of the upper unit have been placed at different horizons, leading to the artificial production of a stratigraphic gap or overlap.

Conclusion

The principles of stratigraphic analysis were worked out during the nineteenth century. During the twentieth century they were applied at an intercontinental scale and modified to accommodate technological developments that allowed more and different types of geological observations to be employed in stratigraphic correlation. No doubt the former trend will be further refined, and the latter extended, during the twenty-first century. New developments will involve the creation of databases that summarize stratigraphic observations over the Earth's surface (and extending into its subsurface), the development of automated algorithms for comparing the data included in such databases and resolving conflicts between alternative sources of information, and the training of stratigraphers to better appreciate the proper use, strengths, and weaknesses of each source of stratigraphic information so that they may apply the age-old principles of stratigraphy to optimal effect.

Glossary

Angular unconformity A surface of erosion separating lower strata that dip at a different angle from the overlying younger strata.

Biochronozone The associated chronozone of a biozone.

Biostratigraphy The characterization of rock strata by their biological constituents.

Chronostratigraphy The characterization of rock strata by their temporal relations.

Depositional hiatus A horizon within a body of sedimentary rock that represents a gap in time due to the non-deposition of sediment, active erosion, or structural complications.

Diachrony The condition of taking place at different times.

Facies A stratigraphic body distinguished from other such bodies by a difference in appearance or composition.

Geochronology The geological study of absolute time.

Homochrony The condition of taking place at the same time.

Homotaxis The condition of occupying the same position in a sequence.

Isochrony The condition of being created at the same time.

Lithostratigraphy The characterization of rock strata by the kind and/or arrangement of their mineralogical constituents.

Radioisotope An isotope of an element that is capable of changing spontaneously into an isotope of another element by emitting a charged particle from its nucleus.

Stratotype The original or subsequently designated type of a named stratigraphic unit (unit stratotype) or stratigraphic boundary (boundary stratotype).

Stratum (plural strata) A tabular section of a rock body that consists throughout of the same type of rock material.

See Also

Biozones. **Famous Geologists:** Agassiz; Cuvier; Darwin; Hutton; Lyell; Smith; Steno. **Magnetostratigraphy**. **Sequence Stratigraphy**. **Time Scale**. **Unconformities**.

Further Reading

Adams FD (1938) *The Birth and Development of the Geological Sciences*. London: Williams & Wilkins.

Ager DV (1993) *The Nature of the Stratigraphical Record*, 3rd edn. New York: John Wiley & Sons.

Gould SJ (1987) *Time's Arrow, Time's Cycle: Myth and Metaphor in the Discovery of Geological Time*. Cambridge: Harvard University Press.

Hedberg HD (1976) *International Stratigraphic Guide: A Guide to Stratigraphic Classification, Terminology, and Procedure*. New York: John Wiley & Sons.

Rawson PF, Allen PM, Brenchley PJ, *et al.* (2002) *Stratigraphical Procedure*. London: The Geological Society.

Rudwick MJS (1972) *The Meaning of Fossils: Episodes in the History of Palaeontology.* London: MacDonald.

Salvador A (1994) *International Stratigraphic Guide.* Trondheim: International Union of Geological Sciences.

Shaw A (1964) *Time in Stratigraphy.* New York: McGraw-Hill.

Zalasiewicz JA, Smith A, Brenchley PJ, *et al.* (2004) Simplifying the stratigraphy of time. *Geology* 32: 1–4.

STROMATOLITES

See **BIOSEDIMENTS AND BIOFILMS**

SUN

See **SOLAR SYSTEM: The Sun**

TECTONICS

Contents

Convergent Plate Boundaries and Accretionary Wedges

G K Westbrook, University of Birmingham, Birmingham, UK

Introduction

At subduction zones, some of the material on the subducting plate is scraped off and added to the leading edge of the overriding plate to form an accretionary wedge (prism, complex), so-called because it is predominantly wedge-shaped in cross-section. The material removed from the subducting plate is primarily sedimentary. Occasionally, igneous crust is transferred to the overriding plate, and when this process occurs at a large scale, it is usually called obduction. Accretionary wedges are analogous to the foreland fold-and-thrust belts of mountain ranges, but they are predominantly submarine and the rates of convergence are typically hundreds of times greater. Where wedges grow large ($\geq$200 km wide and >20 km thick), parts of them emerge as islands; examples include Barbados, which is off the Lesser Antilles island arc, and Nias, off Sumatra. Large areas of south-eastern Iran and south-western Pakistan form the Makran accretionary complex, which occupies a belt 240 km across and extends a further 160 km offshore. In the deeper parts of accretionary wedges, the accreted rocks are metamorphosed in a low-temperature/high-pressure environment to produce blueschists, which may be brought to the surface by exhumation following continental collision or other events that halt subduction. Accretionary wedges are not ubiquitous at subduction zones. Their presence is favoured by thick sequences of sediment on the subducting plate and by low rates of subduction. At most subduction zones, low sediment supply, sediment subduction, and tectonic erosion conspire to suppress the formation of accretionary wedges, which are absent or small and transitory.

Wedge Geometry and Fluid Pressure

The geometry of an accretionary wedge is controlled by the shape of the bounding basement surfaces and by the shear stress on the slip surface between the wedge and the subducting plate. If the surface of the crystalline crust of the overriding plate dips seaward, then the landward part of the accretionary wedge overlies it and forms the seaward margin of the fore-arc basin (Figure 1). If the surface of the crystalline crust of the overriding plate dips landward, then the landward part of the accretionary wedge lies beneath it, and a leaf of crystalline basement separates the accretionary wedge from the fore-arc basin.

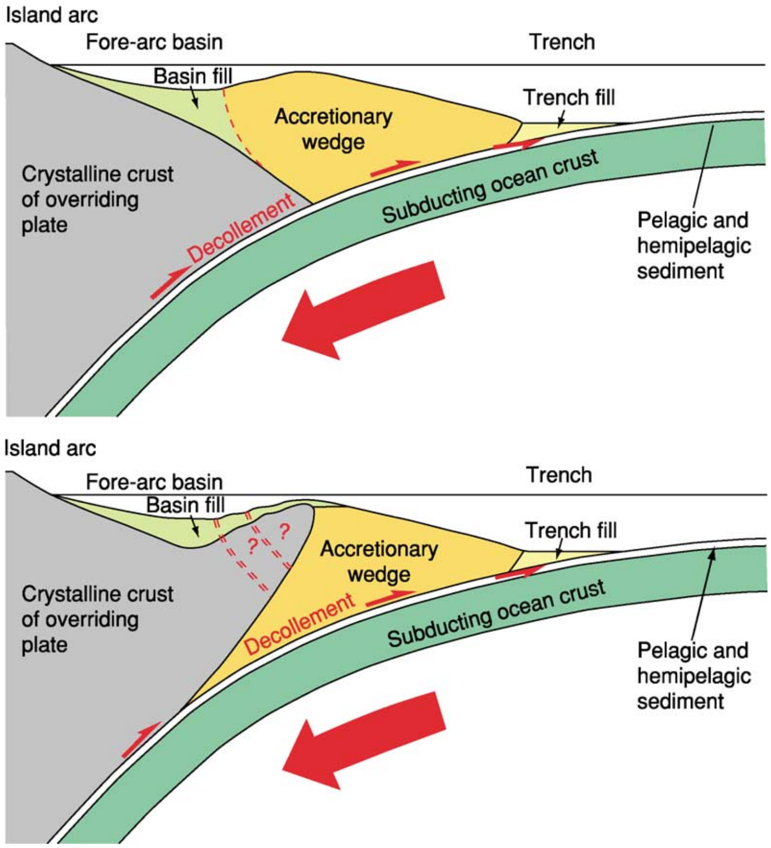


Figure 1 Accretionary wedges are formed at the leading edge of the overriding plate at a subduction zone. They occupy the region between the trench and the fore-arc basin, filling the tectonic depression created by subduction, of which the trench is its bathymetric expression. The boundary with the fore-arc basin may be provided by a leaf of crystalline crust of the overriding plate (lower), or may be a dynamic boundary (upper) at which fore-arc basin sediments lap onto the deformed sediments of the accretionary wedge and may be progressively incorporated into the wedge by deformation as the wedge grows. Copyright Graham Westbrook.

The angle of taper of the wedge (the angle between the dip of the surface and the dip of the basement) depends on the shear stress along the base of the wedge and the strength of the wedge (Figure 2). There is a critical taper at which the force imparted by the basal shear stress is matched by the gravitational spreading force produced by the weight of the wedge. If the basal shear stress is reduced, then the critical taper is reduced, and vice versa. The major factor controlling the strength of the wedge and the shear stress along its base is the frictional resistance, which is the product of the coefficient of friction and the stress normal to any plane of potential movement. The accretion of material to the toe of the wedge lengthens the wedge and changes its taper. In response to this, the wedge deforms internally, thus maintaining its critical taper. So, during accretion, the wedge is continually deforming.

The shear strength of rocks is dependent on the pressure of fluid present in them. The ratio of fluid pressure to the lithostatic pressure (the pressure exerted by the weight of rock) is usually called λ (lambda). Shear failure is governed by the effective stress, which is the difference between the normal stress (across a potential plane of failure) and the fluid pressure. The shear stress, τ, on a plane of motion between two rock masses, such as the decollement at the base of the wedge, is $\tau = \mu\sigma(1 - \lambda)$, where μ is the coefficient of friction, σ is the stress normal to the decollement, and λ is the fluid pressure ratio. The normal stress, σ, is approximately equal to the weight of the sediment in the wedge above the decollement at any particular point. If the load acting on the rock increases more rapidly than the rock can respond by compacting and expelling water, the water bears some of the increased load and becomes overpressured (i.e., its pressure is greater than hydrostatic). If λ exceeds a value of 1, fractures in rocks can be opened by the pressure of water alone (hydrofracturing). Differences in the nature of the rocks in which the decollement is situated affect the stability of the wedge. Clays have low coefficients of friction and low

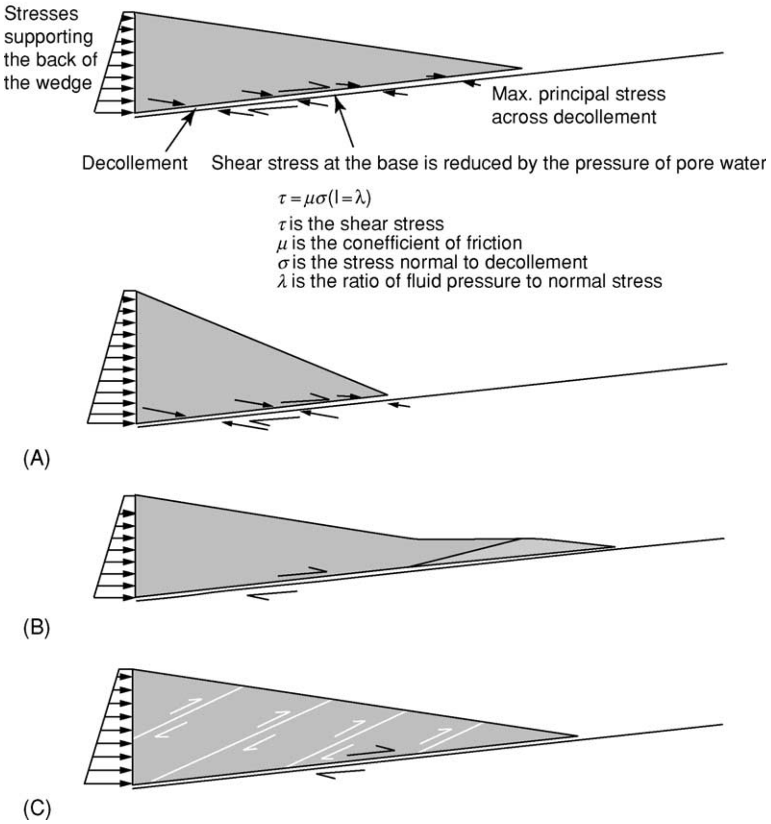

Stresses supporting the back of the wedge
Max. principal stress across decollement
Decollement
Shear stress at the base is reduced by the pressure of pore water
$\tau = \mu\sigma(\mathrm{l} = \lambda)$
τ is the shear stress
μ is the coneefficient of friction
σ is the stress normal to decollement
λ is the ratio of fluid pressure to normal stress
(A)
(B)
(C)

Figure 2 The stability of an accretionary wedge is dependent on the balance between the shear stress along its base, by the motion of the subducting plate, and the stresses generated by gravitational body forces within the wedge. For a wedge with Mohr–Coulomb rheology and constant properties, a single of angle of taper between the top and bottom surfaces, termed the 'critical taper', is established. (A) If shear stress along the base of the wedge is increased, the critical taper is increased, and vice versa. (B) Accretion of material to the toe of the wedge increases wedge width, tending to decrease the critical taper, in response to which (C) the wedge thickens by internal deformation, in order to maintain the critical taper. Sedimentation onto the wedge, erosion, and subcretion (accretion of material to the base of the wedge) similarly induce internal deformation to maintain the critical taper. Copyright Graham Westbrook.

permeability; this favours the build-up of high fluid pressure. Both of these factors reduce the shear stress, leading to low angles of taper. Sands, which are more permeable and have a higher coefficient of friction, would produce a relatively high angle of taper. Several accretionary wedges have angles of taper that are less than 5° and values of the fluid pressure ratio, λ, that are greater than 0.9. The stratigraphy of the trench and minimization of the work required to move the wedge over the subducting plate favour the formation of decollements in clay-rich formations.

At subduction zones, sediments can be become overpressured in three ways:

1. In the trench, on the subducting oceanic plate, fine-grained pelagic and hemipelagic sediments of low permeability become overpressured by the rapid overlying deposition of trench-fill sediment (Figure 3). Deposition rates in trenches, which are as much as a few kilometres per million years, are among the fastest in the world. The greatest degree of overpressure is developed near, but not at, the top of the low-permeability sediment (Figure 4).
2. Sediment accreted to the wedge becomes overpressured by the tectonic thickening of the wedge. The sediment at the base of the wedge is most prone to becoming overpressured, because it has the longest drainage paths and also because low-permeability mud predominates in the lower part of the accreted section.
3. Sediment carried beneath the wedge on the subducting plate becomes overpressured by the weight of the increasing thickness of the wedge above as it passes below. This produces the most rapid increase in load and the highest overpressure.

Wedge Growth

The accretionary wedge grows in a number of ways: by frontal accretion, subcretion, migration into a fore-arc basin, and deposition and deformation of

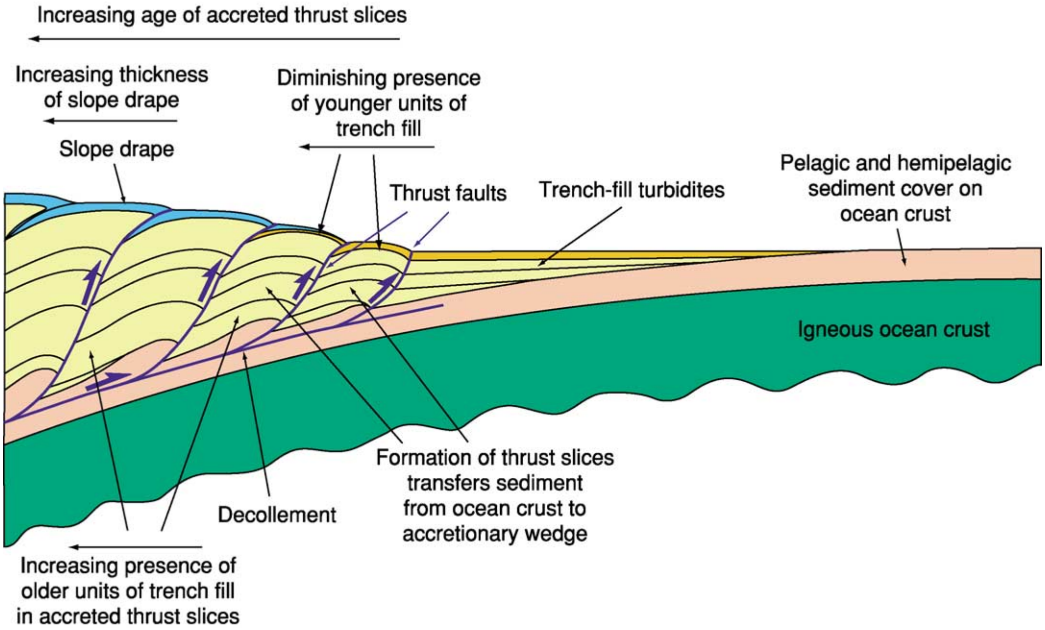

Increasing age of accreted thrust slices
Increasing thickness of slope drape
Diminishing presence of younger units of trench fill
Slope drape
Thrust faults
Trench-fill turbidites
Pelagic and hemipelagic sediment cover on ocean crust
Igneous ocean crust
Formation of thrust slices transfers sediment from ocean crust to accretionary wedge
Decollement
Increasing presence of older units of trench fill in accreted thrust slices

Figure 3 The process of accretion at the toe of an accretionary wedge. Commonly, but not invariably, the turbidites deposited in the trench form most of the accreted material. The accreted materials separate from the overlain pelagic and hemipelagic muddy sediments along a detachment surface (decollement) in the upper part of the sediment cover; the decollement is where loading from turbidite deposition has increased pore-fluid pressure to a value nearing that of the lithostatic pressure imposed by the weight of the overlying sediments, weakening them and making them prone to failure by shearing. The thrust faults that detach the individual thrust slices that form the accreted section originate in the decollement, which propagates ahead of the toe of wedge. The accreted section always includes the youngest turbidites; these are deposited in the trench but not on older accreted thrust slices, which have been uplifted out of the zone of deposition. Consequently, the stratigraphy of each successively accreted thrust slice contains younger turbidites. Within each thrust slice, the sediments upward and landward, but the overall stratigraphy of the wedge are younger becomes younger seaward. Superimposed on the accreted sediments of the wedge is a drape of hemipelagic sediment, undiluted by turbidites, and the age of the base of this drape is youngest seaward. The drape is also deformed by the deformation of the wedge as it thickens, with the oldest part of the drape sequence being more deformed than the youngest. Copyright Graham Westbrook.

slope-drape and slope-basin sediments. In frontal accretion, thrusts propagating from a decollement in a weak, overpressured horizon at the toe of the wedge divide the overlying section into thrust slices, which become added to the toe of the wedge (Figure 3). The level of the decollement is commonly in the upper part of the pelagic–hemipelagic sequence on the subducting plate, which has been overpressured by the deposition of turbidites above it in the trench. The age of the accreted sediment changes with time, giving the wedge a characteristic tectonostratigraphy. Each thrust slice is youngest upward and landward, but the sequence of successively accreted thrust slices has the youngest thrust being seaward and downward. It is this characteristic stratigraphy that can be used to identify ancient accretionary wedges, such as the Ordovician–Silurian wedge of the Southern Uplands of Scotland.

In wedge growth by subcretion, sediment is added to the base of the accretionary wedge by the formation of duplexes at a ramp where the decollement changes level. These propagate successively forward because the work required to continue to move the wedge up a ramp becomes greater than that needed to propagate displacement along the lower decollement and generate a new ramp. In the process, the energy in the sediment between the ramps is transferred from the subducting plate to the accretionary wedge (Figure 5). It has also been suggested that the formation of a zone of tectonic melange along the decollement enables material from the subducting plate to be added to the accretionary wedge, but this can also operate in the opposite sense.

Accretionary wedge growth can occur when landward force imparted by the subducting lithosphere increases with increases in wedge width, pushing the wedge backward into the fore-arc basin and forming thrusts that incorporate fore-arc basin sediment into the wedge (Figure 5). In the mechanism involving slope-drape and slope-basin sediments, deposits directly onto the wedge are deformed by the continual deformation of the wedge beneath as it strives to maintain its critical taper. The sediment forming the slope drape is usually hemipelagic, but in some cases, turbidites derived from erosion of the fore-arc basin or upper part of the wedge are deposited in synclinal troughs to form slope basins, so-called because they occur on the inner trench slope. In foreland fold-and-thrust belts, basins such as these are termed 'piggyback' basins. In the southern part of the Barbados accretionary complex, distributary channels of the Orinoco submarine fan run along synclinal valleys in the accretionary complex, as well as on the ocean floor before it (Figure 5).

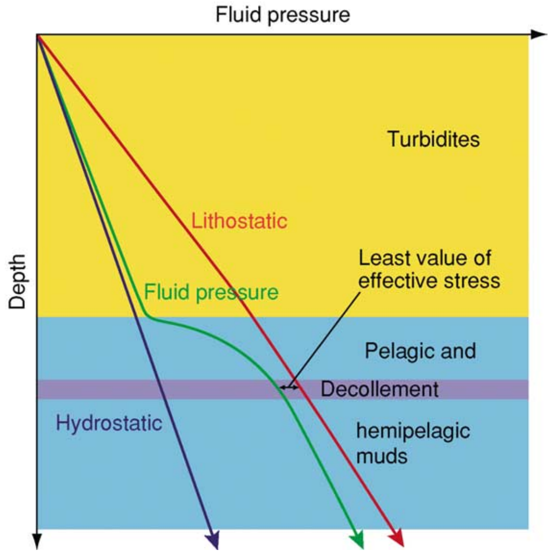


Figure 4 Fluid pressure varies with depth within the sediment in a trench, prior to accretion or subduction. Within the relatively permeable, rapidly deposited, turbidite fill of the trench, fluid pressure is only a little above hydrostatic. Within the low-permeability pelagic and hemipelagic sediments, fluid pressure increases towards lithostatic pressure in response to the loading produced by the rapid deposition of the turbidites above the sediments. The depth at which fluid pressure is nearest to lithostatic pressure is favoured for development of the decollement, although local lithologically mediated variations in the coefficient of friction and cohesion may control its actual position. Copyright Graham Westbrook.

Thickening of an accretionary wedge in its frontal part is brought about by continued displacement on the thrusts by which sediment was accreted and by general horizontal shortening, which produces folding of the thrust slices and landward rotation and steepening of structures. The amount of thickening that this can produce is limited by the rotation of the thrust faults away from the optimum angle for thrusting. Consequently, thickening in the more landward parts of the wedge is produced by motion on new, out-of-sequence, thrusts (termed 'out of sequence' because they do not follow the normal sequence of the youngest thrust being the most seaward). Thickening is also produced by subcretion, adding material from below.

The size of an accretionary wedge might be expected to be simply a product of the thickness of sediment on the subducting oceanic crust, the rate of subduction, and the period over which subduction has occurred. It is certainly the case that the really large wedges occur where major submarine fan systems of considerable thickness are being subducted (Figure 6), but the correlation between wedge size and subduction rate is negative. The very large accretionary wedges occur at subduction zones with low rates of convergence. The frictional resistance to motion along the bases of these wide wedges (which, even with a fluid pressure ratio of $\lambda = 0.9$, is a few teranewtons per square metre, per unit length of the

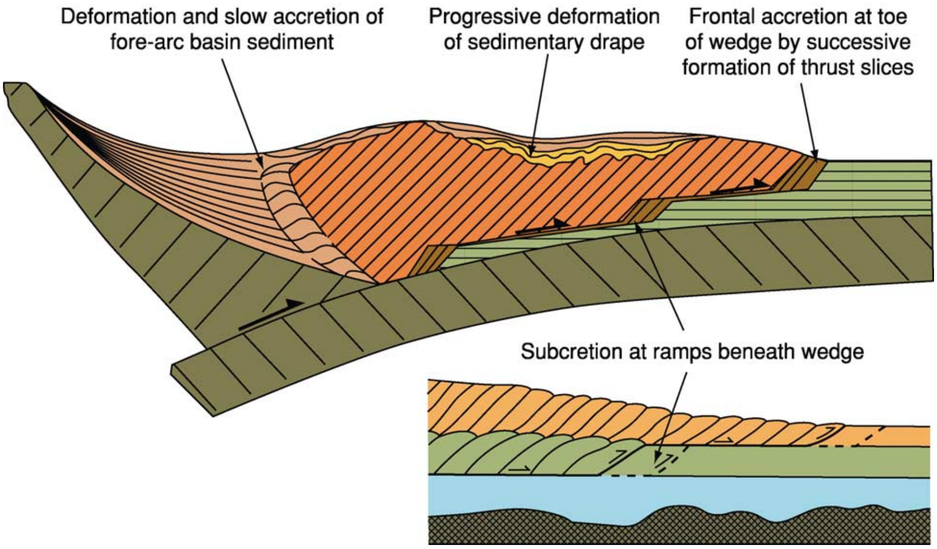

Deformation and slow accretion of fore-arc basin sediment
Progressive deformation of sedimentary drape
Frontal accretion at toe of wedge by successive formation of thrust slices
Subcretion at ramps beneath wedge

Figure 5 Different modes of accretion of sediment to an accretionary wedge. Copyright Graham Westbrook.

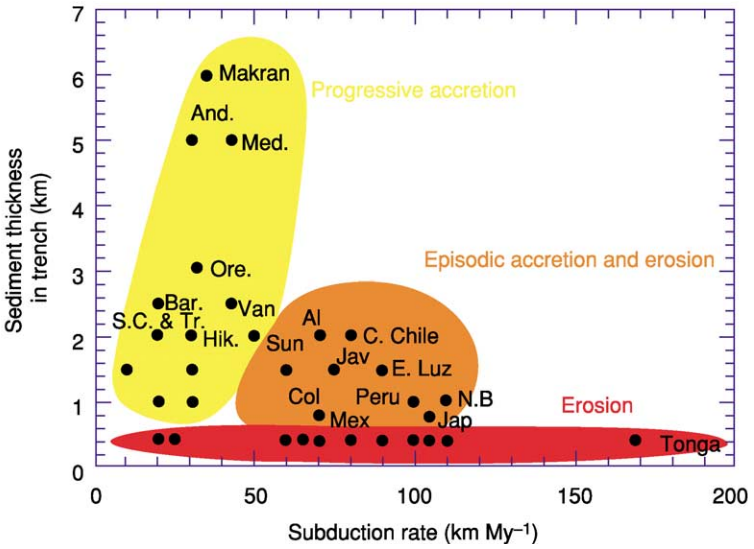


Figure 6 Thickness of sediment on a subducting plate in the trench at a subduction zone, plotted against the rate of subduction. Three discrete groups of subduction zones are identified worldwide: those that have large accretionary wedges with a history of near-continuous growth (yellow), those that have accretionary wedges of small or moderate size in relation to their budget of sediment input and history and that have undergone episodes of tectonic erosion (orange shading), and those that have no, or insignificant, accretionary wedges and are dominated by tectonic erosion (red shading). The last category is typified by very thin sediment fill in the trench, irrespective of subduction rate, and tectonic erosion by basement relief is predominant. Copyright Graham Westbrook.

plate boundary, comparable to the net plate driving forces from ridge push and subduction pull) may be a contributory cause of the low convergence rates. A crucial aspect of the accretion process is the separation of accreted from subducted sediment. Off the coast of Costa Rica, all of the sediment in the trench, which has no cover of turbidites, is subducted. Lower permeability and a higher rate of loading, which correlates with faster subduction, develop higher fluid pressure in the sediments, which favours their subduction because the decollement surface can form at a shallow level. Studies of the global budget of material accreted at convergent plate boundaries, in comparison with material supplied by subducting plate and sedimentation in the trench, have shown that there is a net loss of material. Globally, sediment subduction and tectonic erosion (see below) predominate over accretion.

Fluid Flow, Seeps, and Methane Hydrate

As the sediment incorporated into an accretionary wedge and subducted beneath the wedge compacts, it expels water. Also, the dehydration of minerals in the sediment with increasing temperature and pressure, such the transformation of smectite to illite, releases water. This water flows through the wedge and subducting crust as it escapes to the ocean, carrying with it heat, solutes, and gases (Figure 7). The permeability of the mud-rich sediments that are subducted is generally too low for the water to escape along the subduction zone. The rate of flow of the water through the sediment is less than the rate of subduction. So, the water migrates into and then through high-permeability pathways, such as faults and permeable sediments such as sands, or even the igneous ocean crust, which is about a thousand times more permeable than compacted clay-rich mud deposited on the ocean floor. The warm water expelled beneath the trench increases the heat flow from the trench. Water driven into trench sediments or expelled along faults through the wedge can contain methane that was generated by methanogenic bacteria and other hydrocarbon gases that were created in conditions of higher temperature beneath the wedge. The methane and hydrogen sulphide expelled at seeps and vents sustain chemosynthetic communities of biota. (In this setting, the hydrogen sulphide is a product of anaerobic oxidation of methane by symbiotic communities of aquatic bacteria and archaea.) The seeps and vents are usually located along faults or in mud volcanoes, which occur on the accretionary wedge and on the ocean floor in front of the wedge. Mud volcanoes are created by mud diapirism, in which the tensional stresses at the top of a body of low-density mud are sufficient to create a pathway for the mud to rise buoyantly to the seabed, where it erupts. The diapiric bodies are initially created by deformation of underconsolidated

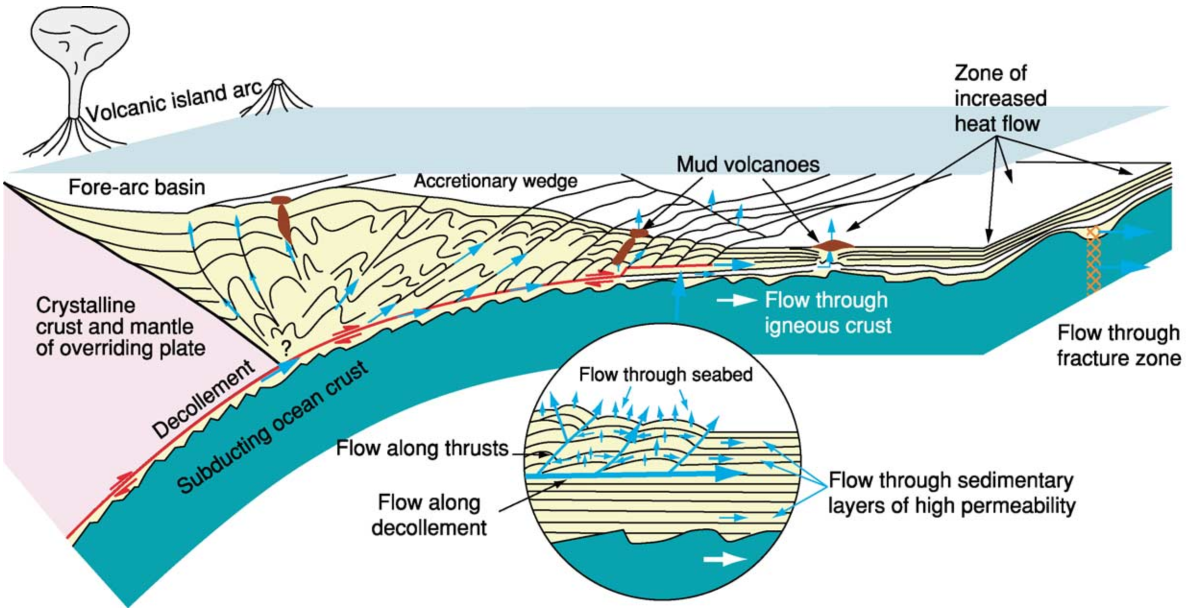

Volcanic island arc
Zone of increased heat flow
Fore-arc basin
Accretionary wedge
Mud volcanoes
Crystalline crust and mantle of overriding plate
?
Flow through igneous crust
Flow through fracture zone
Decollement
Subducting ocean crust
Flow through seabed
Flow along thrusts
Flow along decollement
Flow through sedimentary layers of high permeability

Figure 7 Fluid flow and modes of expulsion of fluid from an accretionary wedge. The sources of fluid are pore water expelled by compaction from sediment subducted beneath the accretionary wedge and sediment accreted to the wedge, and dehydration of hydrous minerals such as smectite, as temperature increases with increasing depth of subduction. Copyright Graham Westbrook.

mud-rich layers that are accreted into the wedge or subcreted beneath it. Mud volcanoes also appear to be created by the fluidization and entrainment of mud by water driven to the surface by tectonic expulsion, and this mode of formation is characteristic of mud volcanoes created in front of accretionary wedges on the ocean floor, or behind them in fore-arc basins.

The migration of methane-containing pore water through the sediments of an accretionary wedge as it compacts creates methane hydrate in the sediments occupying the first few hundred metres depth range beneath the seabed (*see* **Petroleum Geology:** Gas Hydrates). This occurs because this region lies within the stability field for methane hydrate (which is a solid clathrate formed from water and methane, in which the methane molecules are held within a cage of water molecules in an approximately 1:6 ratio). The hydrate stability field generally exists in Earth's major oceans in water depths greater than about 300 m, and is favoured by increasing pressure and decreasing temperature. Consequently, most of the sediments beneath continental margins are in the hydrate stability field down to the depth at which, because of the increase of temperature with depth, the geotherm crosses the stability boundary for hydrate. Beneath this boundary, methane can be present as free gas, in which case the boundary creates a seismic reflection because the presence of only a very small amount of free gas (less than 1% of the pore space is enough) reduces the seismic velocity of P waves significantly. The polarity of the reflection is negative, opposite to that of the seabed, because of the decrease in velocity beneath it, and is most clearly visible on seismic reflection sections where it cuts across the reflections produced by sedimentary bedding (Figure 8). This reflection is widespread in accretionary wedges, and is usually termed a bottom-simulating reflection (BSR) because its shape, to the first order, mimics that of the seabed, which, because of its nearly uniform temperature, controls the shape of the isotherms beneath it.

Uplift of the seabed produced by the thickening of the wedge continually moves the base of the zone containing hydrate upward out of hydrate stability field, causing hydrate to dissociate and release free gas that produces the BSR. Because the depth of the BSR below the seabed is controlled by the geothermal gradient, mapping the depth of the BSR has been used to map variations in heat flow from accretionary wedges, which is influenced by tectonic thickening and fluid flow. The tectonic expulsion of methane-rich pore water and the dissociation of hydrate to free gas caused by uplift results in methane hydrate and BSRs being widespread in accretionary wedges, whereas they occur only rarely in the sediments of passive continental margins.

Tectonic Erosion at Subduction Zones

The inner walls of trenches of arcs (e.g., the Mariana arc, Tonga arc, and South Sandwich arc) do not have significant accretionary wedges. Those that do occur

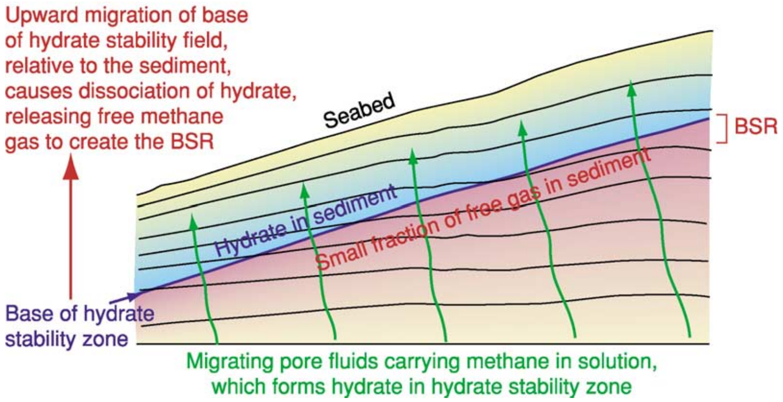

Upward migration of base
of hydrate stability field,
relative to the sediment,
causes dissociation of hydrate,
releasing free methane
gas to create the BSR
Seabed
BSR
Hydrate in sediment
Small fraction of free gas in sediment
Base of hydrate
stability zone
Migrating pore fluids carrying methane in solution,
which forms hydrate in hydrate stability zone

Figure 8 Model for the growth of methane hydrate within sediments in the uppermost part of an accretionary wedge, and for the formation of a bottom-simulating seismic reflection (BSR). Methane, dissolved in pore water expelled from sediment by compaction within and beneath the wedge, is carried upward into the hydrate stability field, where it forms hydrate. Uplift of the seabed, caused by the thickening of the wedge with continuing growth, destabilizes the base of the hydrate stability zone, releasing methane gas and water from dissociated hydrate. The presence of free gas is the principal cause of the BSR. Copyright Graham Westbrook.

are very small and transitory. Most of the evidence suggests that material from the inner trench wall is being removed and subducted. The budgets of sediment available to be added to accretionary wedges in comparison with the mass of sediment in wedges indicate that many have lost sediment or have undergone periods when none was accreted. For example, the accretionary wedge off the coast of Honshu, Japan, is composed mainly of Cretaceous sediment, with only a very little Neogene sediment and no Paleogene sediment. This suggests that there is a process that removes material from the overriding plate at subduction zones. This process, or group of processes, is referred to collectively as tectonic erosion or subduction erosion.

What is the evidence for tectonic erosion? The absence of accretionary wedges from arcs such as Tonga might be explained if all of the sediment in the trench was subducted. Fore-arc subsidence such as that off the coast of Peru, where there has been 4 km of subsidence over a 100-km width of fore arc since the late Miocene, is difficult to explain other than by the removal of material from the base of the fore arc. The landward migration of the volcanic arc across the overriding plate, as exemplified by the Andes and by island arcs such as the South Sandwich or the northern end of the Lesser Antilles, although explicable in specific instances by a reduction in the angle of dip of the subducted lithosphere, can only be generally explained by removal of crust from the leading edge of the overriding plate. The mechanisms proposed for tectonic erosion are of two principal types: erosion by basement topography and erosion from the effects of high fluid pressure.

Erosion by Basement Topography

Tectonic erosion may occur when the basement topography physically breaks off and displaces parts of the fore arc, carrying it deeper into the subduction zone; this steepens the trench slope locally, causing submarine slides into the trench. The material from the slides may then also be subducted (see Figure 9). The mechanism depends on the basement of the subducting plate being stronger than the material in the tip of the overriding plate. Where an accretionary wedge, composed of sediment or metamorphosed sediment, forms the leading edge of the overriding plate, this is normally the case, but where the edge of the overriding plate is composed of igneous or high-grade metamorphic rocks, the situation can be reversed, resulting in the accretion/obduction of the basement feature on the subducting plate. There are two broad categories of basement topography. The first is general basement relief inherited from a mid-ocean ridge and accentuated by normal faulting in the outer trench slope. This is ineffective if sediment cover is more than several hundred metres thick, because the decollement forms well above the basement and the accretionary wedge rides over the relief, undisturbed. The second category comprises discrete features of high relief, such as seamounts, transform ridges and troughs, and hotspot ridges. These are more severe in their effects, but are not present everywhere, and so they produce spatially and temporally limited episodes of tectonic erosion, of which a good example is provided by the subduction of seamounts on the Cocos plate beneath the convergent margin of Costa Rica.

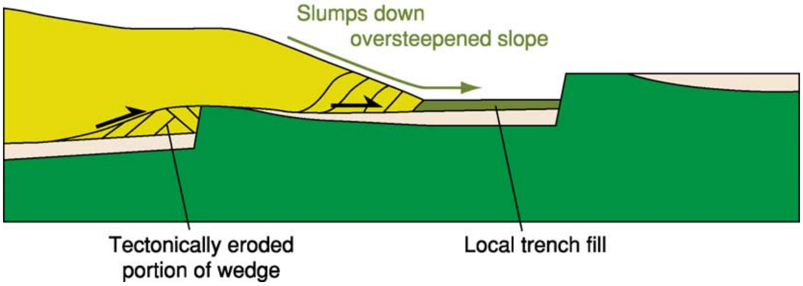

Slumps down oversteepened slope
Tectonically eroded portion of wedge
Local trench fill

Figure 9 Tectonic erosion by the relief of the oceanic basement is a common process where sediment cover on the subducting plate is thin. Scarps originally formed by normal faulting at the mid-ocean ridge are enhanced, or a new one is created by normal faulting in the outer part of the trench as the oceanic lithosphere flexes downward into the subduction zone. The tops of these scarps collide with the toe of the accretionary wedge or with the leading edge of the crystalline fore-arc of the overriding plate, forcing the decollement to jump upward to the level of the top of the scarp, thereby transferring some of the material that was in the wedge to the subducting plate. The surface of the wedge is steepened by the removal of material from its toe, inducing slumps into the trench of material that is subsequently accreted and swept into the subduction zone by the basement scarps. The subduction of seamounts produces a similar, although more severe, process of tectonic erosion that is effective in thicker sediment cover, but is geographically more localized. Copyright Graham Westbrook.

Erosion from the Effects of High Fluid Pressure

The need to invoke this second type of mechanism is presented by those subduction zones where it is clear that tectonic erosion is or has been active, yet no features of basement topography appear to be responsible. The fluid pressure ratio, λ, is greatest at the top of any body of rock in which pore water is connected. This makes the uppermost rocks weakest and most liable to failure and displacement. As zones of high pressure are driven to migrate upward, progressive failure and displacement remove material from (tectonically erode) the section through which the high-pressure pore fluid migrates, until it dissipates or escapes to the surface. As subduction is continually feeding sediments with a high water content beneath wedges of accreted sediment or the basement of the overriding plate, the potential for this type of mechanism to operate is always present, if pore water expelled from the sediment cannot escape easily through the overlying wedge. There is evidence for two possible variants of this process, the first more general, the second more specific:

1. In ‘stoping’, high-pressure fluid causes disaggregation of the base of the wedge. The disaggregated rock is incorporated into a shear zone melange and is subducted (Figure 10). Shear zone melanges are exhibited by the exhumed deeper parts of old accretionary wedges that can be observed on land, such as the Franciscan wedge in California.
2. In the process of reactivation and upward migration of detachment surfaces, high-pressure fluid weakens the upper surfaces more than the lower ones. Rock beneath the reactivated surfaces is subducted (Figure 10). Seismic images of duplexes within the subducting section have provided the indication that this process occurs.

Obduction

Sometimes, features of oceanic basement are not subducted, but are sheared off against the crystalline crust of the overriding plate, leading to obduction of part of the oceanic crust. This appears to occur most commonly where asperities exist in the oceanic crust that is being subducted, such as seamounts or the ridges flanking transform faults. The Tres Montes Peninsula on the coast of southern Chile, close to the Chile triple junction, was probably brought about by this process. Very large-scale obduction of ocean crust to create ophiolite complexes is associated with continent/continent collision or continent/island-arc collision and may also be a consequence of the closing of back-arc basins following a change in plate motions, with young buoyant ocean crust thrust onto the adjacent plate, as exemplified by the Rocas Verdes complex of southern Chile.

Oblique Subduction

There are several convergent margins, including the eastern end of the Aleutian island arc, Sumatra, the north-west United States (Cascadia), and the northern Pacific margin of Colombia, where the direction of convergence is oblique; in these cases, as well as convergence between the two plates in the direction orthogonal to the plate boundary, there is a component of motion parallel to the plate boundary. At these margins, the directions of thrust-fault outcrops and of fold axes, which form at right angles to the direction of compression, run subparallel to the margin, not at right angles to the direction of convergence of the two plates. The reason for this is that the displacement between the accretionary wedge and the subducting plate is partitioned into a component orthogonal to

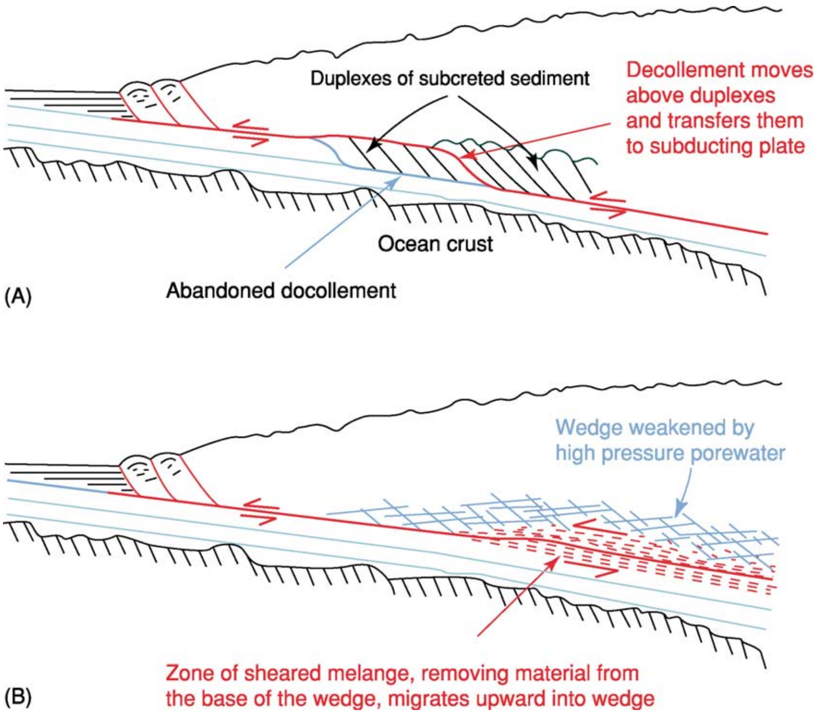

Duplexes of subcreted sediment
Decollement moves above duplexes and transfers them to subducting plate
Ocean crust
Abandoned docollement
(A)
Wedge weakened by high pressure porewater
Zone of sheared melange, removing material from the base of the wedge, migrates upward into wedge
(B)

Figure 10 The hypothesis of tectonic erosion from the effect of high pore-fluid pressure. (A) Loss of subcreted material from the accretionary wedge by relocation of the active decollement to the upper of the two levels separating the subcreted duplexes, because of an increase in drainage from the lower decollement and/or a restriction in the drainage from the upper level. (B) Material above the main decollement has been weakened by high fluid pressure; a melange of this disaggregated material is incorporated into a shear zone, which transports the material deeper into the subduction zone. Removal of this material makes it easier for high-pressure fluid in the shear zone to infiltrate the zone and propagate the process of tectonic erosion upward into the overlying wedge. Copyright Graham Westbrook.

the margin and a component parallel to it. A strike–slip fault, or series of faults, separates the wedge, and often the fore-arc basin as well, from the volcanic arc and the rest of the plate. The horizontal motion between the fore-arc region and the remainder of the plate along this fault compensates the obliquity of plate convergence, so that the relative motion between the subducting plate and the accretionary wedge is nearly orthogonal to the margin (Figure 11). It has been demonstrated theoretically and by experiments with sand-box models that, when the direction of convergence is greater than about 15° from orthogonal to the margin, displacement can be partitioned in this way. This is because the work required to move the subducting plate the shortest distance, orthogonal to the margin, against the frictional resistance of the decollement surface, plus the work to move the fore-arc along the strike–slip fault, is less than the work required to move the plate obliquely a greater distance along the decollement in the direction of convergence between the two plates. The angle of obliquity at which partitioning occurs depends on the shear stress at the base of the wedge, and if this is very low, because of fluid overpressure, the deviation from orthogonal convergence required to produce partitioning can approach 40°. Where partitioning occurs, the fore-arc can be translated large distances along the edge of the overriding plate

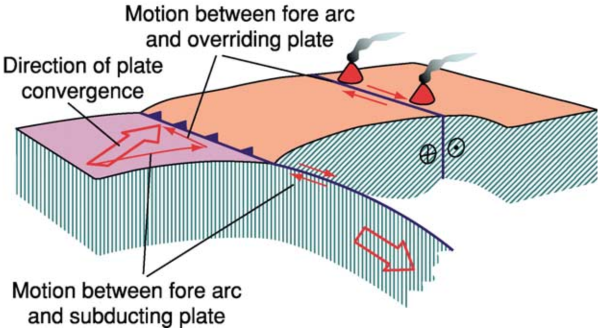

Motion between fore arc and overriding plate
Direction of plate convergence
Motion between fore arc and subducting plate

Figure 11 Where subduction is oblique to the margin, motion is partitioned into orthogonal underthrusting of the subducting plate beneath the fore arc (accretionary wedge plus or minus the fore-arc basin) and strike–slip movement between the fore arc and the overriding plate. The direction of plate motion is shown by the large arrows; the small red arrows on the subducting plate indicate components of motion normal and parallel to the margin. Copyright Graham Westbrook.

Table 1 Summary of controls on accretion, subduction, and tectonic erosion of sediment

Separation of accreted and subducted sediment	The level of detachment between accreted and subducted sediment is governed by Ratio of shear stress to effective normal stress Presence of a weak horizon, produced by Low intrinsic strength (low coefficient of friction and low cohesion) High fluid pressure from Low permeability High rate of loading, produced by Sufficient sediment supply to trench High subduction rate High angle of dip of subducting plate
Tectonic erosion	By basement topography General relief inherited from mid-ocean ridge, and accentuated by normal faulting in outer trench slope; ineffective if sediment cover is thicker than several hundred metres Discrete features of high relief Seamounts Transform ridges and troughs Hotspot ridges (swells) From the effects of high fluid pressure 'Stoping' (disaggregation of base of wedge and incorporation into a shear-zone melange) Reactivation and upward migration of detachment surfaces
Accretion vs. tectonic erosion	Progressive accretion is associated with high sediment thickness and low subduction rate Tectonic erosion is associated with low sediment thickness Episodic accretion and tectonic erosion is associated with high sediment thickness and high subduction rate

to become a displaced, or exotic, terrain. A summary of the controls on accretion, subduction, and tectonic erosion of sediment is given in Table 1.

See Also

Andes. **Europe:** Caledonides of Britain and Ireland; Mediterranean Tectonics. **Japan**. **Metamorphic Rocks:** Facies and Zones. **New Zealand**. **North America:** Northern Cordillera. **Oceania (Including Fiji, PNG and Solomons)**. **Petroleum Geology:** Gas Hydrates. **Seamounts**. **Sedimentary Processes:** Deep Water Processes and Deposits. **Seismic Surveys**. **Tectonics:** Mountain Building and Orogeny; Ocean Trenches. **Ultra High Pressure Metamorphism**.

Further Reading

Bebout GE, Scholl DW, Kirby SH, and Platt JP (eds.) (1996) *Subduction Top to Bottom, Geophysical Monograph 96*. Washington, DC: American Geophysical Union.

Carson B and Screaton EJ (1998) Fluid flow in accretionary prisms: evidence for focused, time-variable discharge. *Reviews of Geophysics* 36: 329–351.

Davis DJ, Suppe J, and Dahlen FA (1983) Mechanics of fold-and-thrust belts and accretionary wedges. *Journal of Geophysical Research* 88: 1153–1172.

Fryer P (1996) Evolution of the Mariana convergent plate margin system. *Reviews of Geophysics* 34: 89–125.

Hyndman RD and Davis EE (1992) A mechanism for the formation of methane hydrate and sea-floor bottom-simulating reflectors by vertical fluid expulsion. *Journal of Geophysical Research* 97: 7025–7041.

Hyndman RD, Spence GD, Chapman R, Riedel M, and Edwards RN (2001) Geophysical studies of marine gas hydrate in northern Cascadia. In: *Geophysical Monograph 124, Natural Gas Hydrates: Occurrence, Distribution, and Detection*, pp. 273–295. Washington, DC: American Geophysical Union.

Lallemand SE, Schnurle P, and Malavieille J (1994) Coulomb theory applied to accretionary and nonaccretionary wedges – possible causes for tectonic erosion and or frontal accretion. *Journal of Geophysical Research* 99: 12033–12055.

Moore JC and Vrolijk P (1992) Fluids in accretionary prisms. *Reviews of Geophysics* 30: 113–135.

Ranero CR and von Huene R (2000) Subduction erosion along the Middle America convergent margin. *Nature* 404: 748–752.

Stern RJ (2002) Subduction Zones. *Reviews of Geophysics* 40(4): 3.1–3.38.

Tarney J, Pickering KT, Knipe RJ, and Dewey JF (eds.) (1991) *The Behaviour and Influence of Fluids in Subduction Zones*. London: The Royal Society (From *Philosophical Transactions of the Royal Society, Series A* 335: 225–418.)

von Huene R and Scholl DW (1991) Observations at convergent margins concerning sediment subduction, subduction erosion, and the growth of continental crust. *Reviews of Geophysics* 29: 279–316.

Westbrook GK, Ladd JW, Buhl P, Tiley GJ, and Bangs N (1988) Cross section of an accretionary wedge: Barbados Ridge Complex. *Geology* 16: 631–635.

Earthquakes

G J H McCall, Cirencester, Gloucester, UK

Introduction

Earthquakes have many and diverse relationships with other Earth processes, and their study has a wide range of possible applications. Earthquakes will be considered here under the following headings:

1. The nature of earthquakes,
2. The importance of seismological records,
3. The global distribution of earthquakes, and
4. Earthquakes as a hazard – tectonic, volcanic, and man-made earthquakes.

The Nature of Earthquakes

An earthquake is a sudden movement of the Earth's surface, caused by a release of strain built up over long periods on faults. The rocks are elastic and can store energy in the same way as a compressed spring. Earthquakes are focused on faults in the rock mass. Most have foci within the crust but a few, in plate boundary zones and beneath stable cratonic areas (where they are related to events in nearby subduction zones), have foci at great depths, down to about 700 km, in the mantle; beyond this depth the rock mass is insufficiently rigid to rupture. The very deep earthquakes are not well understood. Most earthquakes have foci less than 30 km deep. Not all the built-up strain is relieved by earthquakes; much of it is relieved by continual small adjustments, a process of creep. However, where friction prevents such accommodation, the strain builds up until something has to give, and there is a sudden rupture of the weakest part of the solid rock, the forces being accommodated by sudden dislocation of the rocks on either side of the fault plane (Figure 1). This process can happen on all three types of fault: normal, reverse, and transcurrent. The point directly above the focus is the called the epicentre; here, the effects of the earthquake will be greatest. If the focus is shallow, the effects will be greater than if it is deep – the 1960 Agadir earthquake, Morocco, was not of great magnitude, but it was very shallow and the epicentre was right under the city.

The magnitude is a measure of the amount of energy released by the earthquake. It is calculated from the size or amplitude of the waves traced by the pen of a seismograph, an instrument that picks up the waves at some distance from the epicentre and records them in the form of a wavy trace on a rotating drum (Figure 2) coupled to a clock. The principle of the seismograph is illustrated by a chandelier that swung in the great Lisbon earthquake of 1755 – a freely pivoting horizontal strut is attached to an upright support (Figure 3). A heavy mass at the strut end is attached to the pen, which traces a continuous line on the paper. Most of the time there are no ground movements, so the trace is horizontal. All seismographs have to be standardized so that valid comparisons can be made between their traces. Waves weaken as they travel outwards from the earthquake focus, so allowance has to be made for distance between the focus and the instrument (this can be calculated by measuring the difference between the arrival times of P and S waves). Three seismographs are required to measure north–south and east–west horizontal movements and vertical movements. The Worldwide Standardized Seismograph Network was established in 1962. All the instruments are standardized as if they were situated 100 km away from the focus.

Waves produced by earthquakes spread through the Earth (Figure 4). They comprise body waves and surface waves. Body waves travel through the Earth and are of two types, primary (P) and secondary (S)

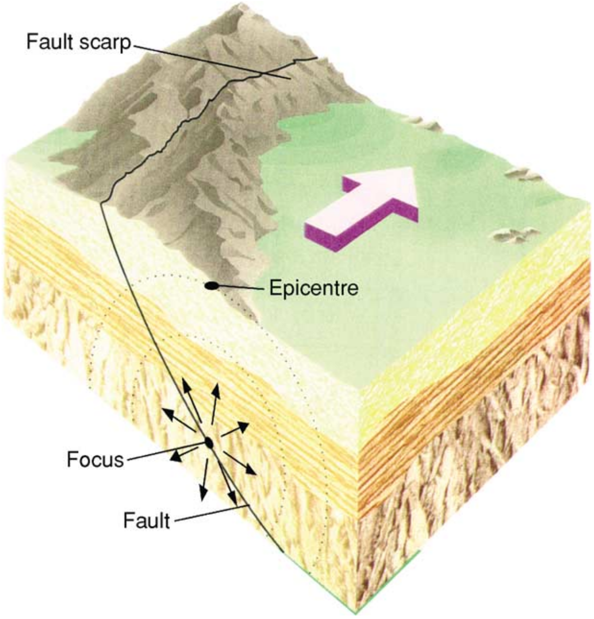


Figure 1 Block diagram showing the relationship between an earthquake focus, epicentre, and fault.

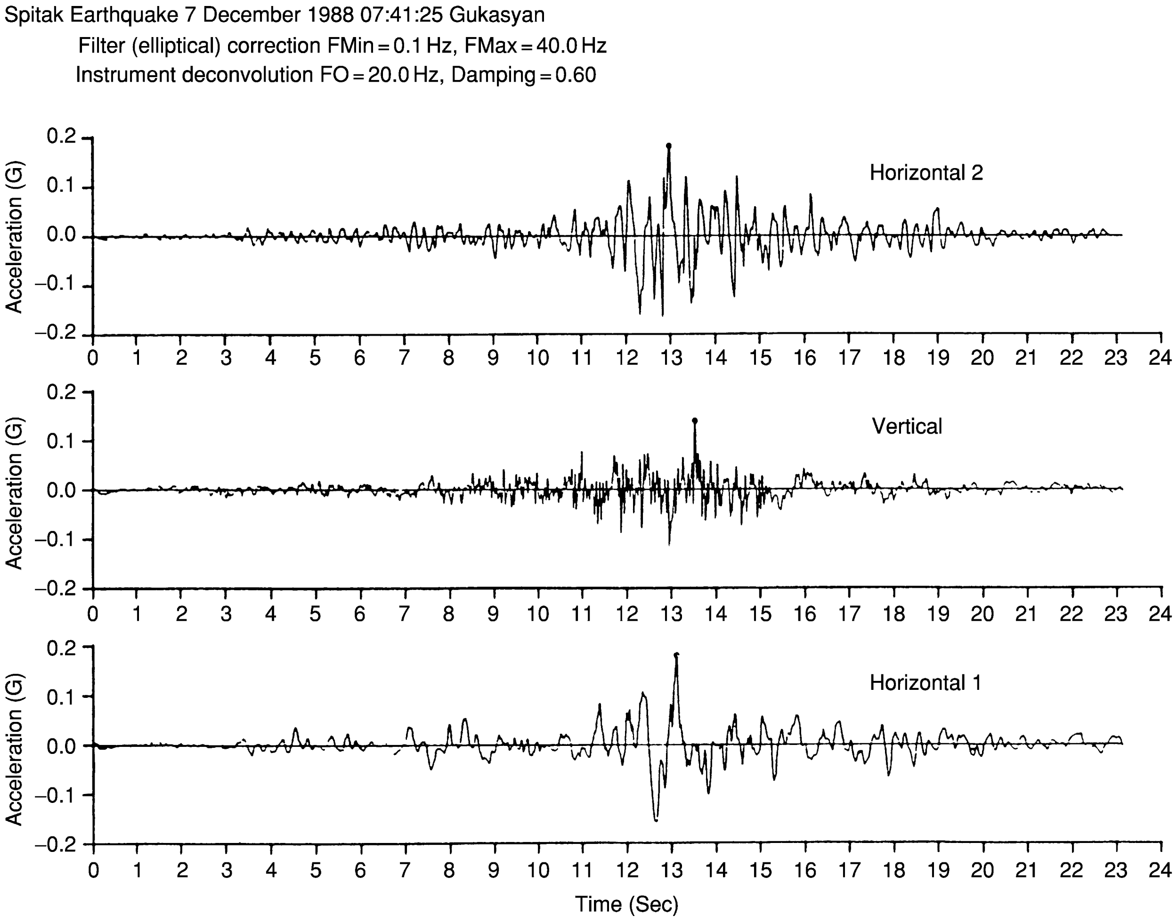


Figure 2 Record on three seismographs of the Spitak 1988 main shock (magnitude 6.9). Reproduced with permission from Rommer and Ambraseys (1989) *Earthquake Engineering and Dynamics*. Chichester: John Wiley & Sons. © John Wiley & Sons Limited.

waves. P waves can pass through both solid and molten material within the Earth's interior; they travel fastest and are the first to arrive at a given location, and they are also the first to be felt by the man in the street. They are longitudinal or compressional waves, vibrating forwards and backwards in the direction of travel. They travel at about 6 km s^{-1} through continental crust and 8 km s^{-1} through oceanic upper mantle. They may produce booming sound waves in the atmosphere. S waves travel about half as fast as P waves (3.6 km s^{-1} and 4.7 km s^{-1} in continental crust and oceanic upper mantle, respectively); they cannot pass through fluids and thus do not penetrate the liquid outer core. They are shear or transverse waves: as they pass through the rock they move particles both from side to side and up and down, at right angles to the direction of travel. Two kinds of surface waves, which travel just below the surface, are called Love and Rayleigh waves. They arrive shortly after the body waves. Love waves travel faster than Rayleigh waves and push the rock particles sideways, at right angles to the direction of travel. Like S waves, they shear buildings and constructions sideways, causing immense damage, but have no vertical motion. The slowest waves, Rayleigh waves, push particles upwards and backwards; the particles move in the vertical plane, following an elliptical path as the wave passes by.

Charles Richter in 1935, working in California, devised the Richter Scale of magnitude, in which the absolute strength at the focus can be calculated on a logarithmic scale: a rise of one unit of magnitude represents a 10-fold increase in absolute strength (i.e. a magnitude 5 earthquake is 10 times a strong as a magnitude 4 earthquake). The difference in energy release is even greater – an increase of one unit of magnitude represents 30–32 times as much energy being released. Theoretically, earthquakes with magnitudes of more than 10 could occur, but the greatest magnitude so far measured for any earthquake is about 9.5. The Richter scale is given in Table 1. It has been superseded as a scale for measuring the

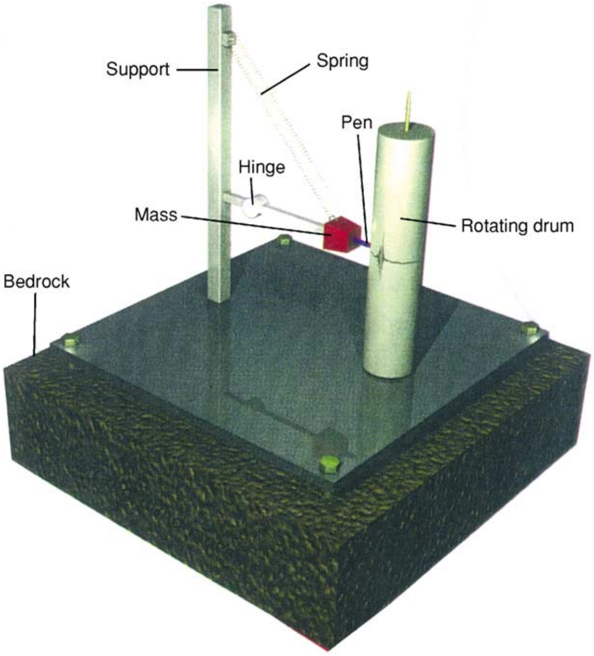


Figure 3 The components of a seismograph designed to record vertical ground movement.

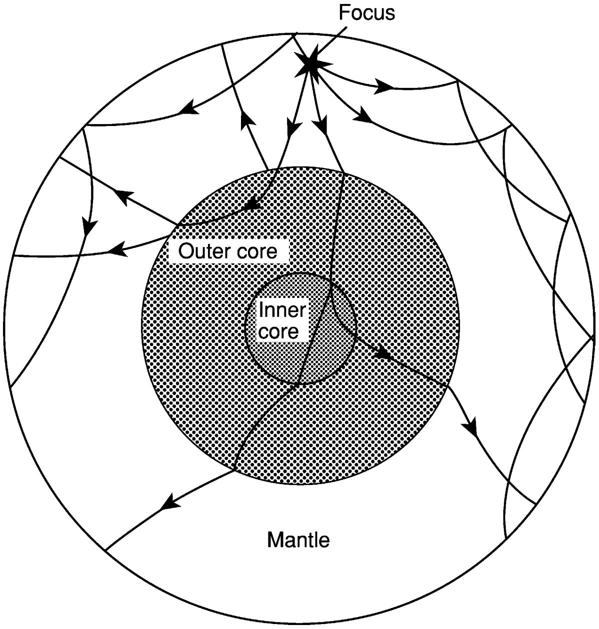


Figure 4 The way in which earthquake waves spread through the globe and are reflected at boundaries, returning to the surface. Measurement of the speed of such return is used to delineate the materials of the inner Earth according to density and physical state. Reproduced from Van Andel TJ (1994) *New Views on an Old Planet.* Cambridge: Cambridge University Press.

comparative intensity at the focus by the moment magnitude scale, but the principles are the same – the moment magnitude scale allows more refined methods of comparison.

Table 1 The Richter scale

Magnitude	*Qualitative description*	*Average number per year*	*Average intensity equivalent close to epicentre*
0–1.9		700 000	I–V; recorded but not felt
2–2.9		300 000	I–V; recorded but not felt
3–3.9	Minor	40 000	I–V; felt by some
4–4.9	Light	6200	I–V; felt by many
5–5.9	Moderate	800	V–VII; slight damage
6–6.9	Strong	120	VII; damaging
7–7.9	Major	18	IX–XI; destructive
8–8.9	Great	1 every 10–20 years	XII; widely devastating

The Importance of Seismological Records

We cannot directly study the rocks of the crust below the limits of borehole drilling (a few kilometres), though ancient rock systems do expose sections of the ancient deep crust (as in the Kapuskasing Belt, Ontario, Canada) and perhaps even the crust–mantle contact (as in Oman and Western Newfoundland). The behaviour of earthquake waves, however, provides us with invaluable evidence about the nature of the lower crust, mantle, and core because the velocities of P and S waves are functions of the density of the material through which they pass. Knowledge of rock density can tell us much about the physical state of the materials deep within the Earth, and the behaviour of S waves tells us that the outer core is molten. In Figure 5, the different densities of common rock materials are plotted against the P-wave velocity.

It is fair to say that earthquake waves form the basis of our knowledge of the mantle and core. Artificially produced seisms can also be picked up by seismographs, and, thus, nuclear explosions can be globally monitored. The explosion in the submarine Kursk in 2000 was picked up by distant seismograph stations in Africa, and this provided valuable evidence of what happened. Tomographic methods have recently been developed, producing three-dimensional images of the deep Earth, including subducted slabs of crust, using a technique akin to the use of tomography in medicine.

Earthquakes occur in sequences: slight foreshocks may give warning of a major earthquake, and aftershocks occur for long after the main shock. Foreshocks and aftershocks are generally of lower intensity than the main shock, but sometimes very large shocks occur, as in the 1999 Izmit earthquake in Turkey. In the 1988 Spitak earthquake (magnitude 6.9) an aftershock of magnitude 6.2 occurred 4 min after the main shock.

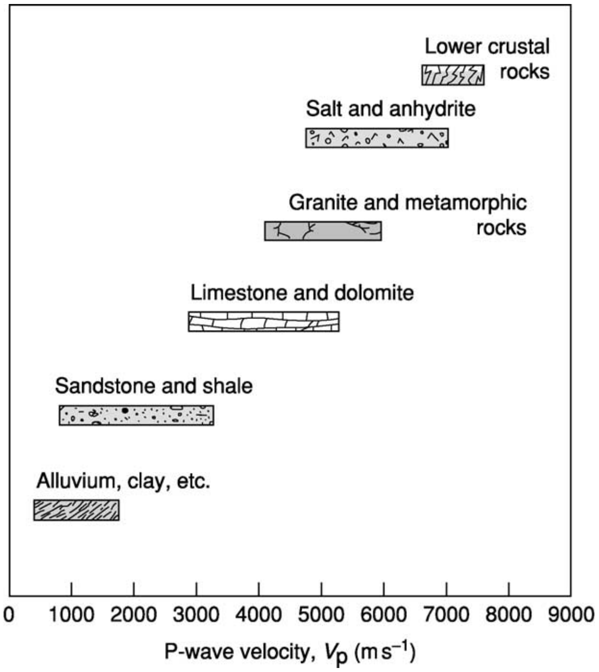


Figure 5 The different densities of common rock types plotted against P-wave velocity.

Earthquakes may produce a trace of the rupture on the land surface, dislocating the land for many kilometres. In Figure 6 the trace produced across wheat fields by the Meckering 1968 (magnitude 6.9) earthquake is shown. Such traces are invaluable in studying the sense of the movement and displacement.

The Global Distribution of Earthquakes

Earthquakes do not occur to the same extent all over the globe. The major events are largely concentrated at the boundaries of tectonic plates, and the concentration and magnitude are greater in zones of plate convergence (subduction and collision) than in zones of plate divergence (mid-ocean ridges and rift valleys). This distribution is clearly shown in Figure 7.

Not all earthquakes occur on plate boundaries, however: the destructive Killari earthquake in India in 1993 occurred within a stable cratonic area. The immensely destructive Lisbon earthquake of 1755 was also nowhere near a plate boundary.

Earthquakes as a Major Hazard: Tectonic, Volcanic, and Man-Made Earthquakes

The most damaging earthquakes are not necessarily of high magnitude. The 1994 Kobe earthquake, one

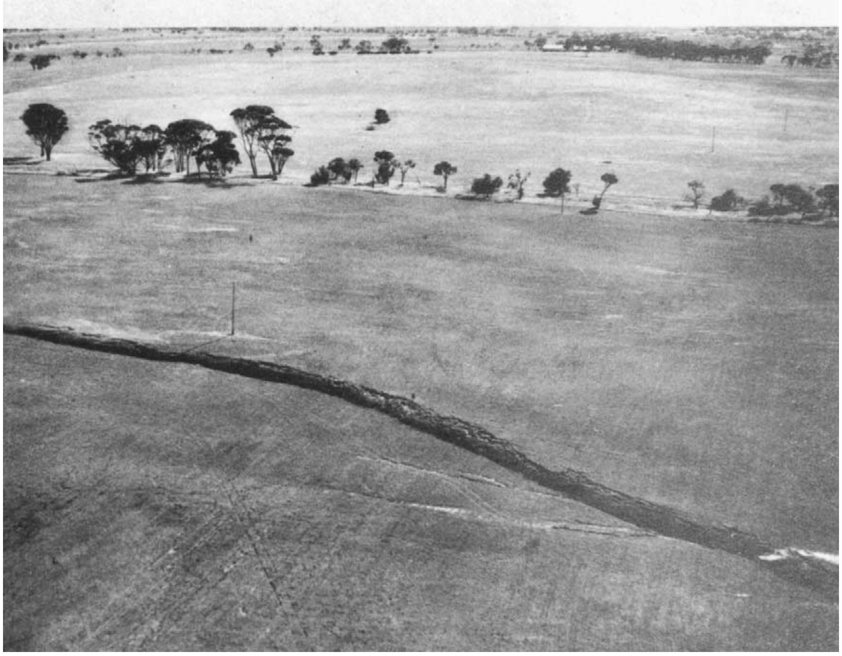


Figure 6 Surface trace of the Meckering 1968 earthquake (magnitude 6.9) in Western Australia. Reproduced from Everingham I (1968) *Preliminary report on the 14th October 1968 earthquake at Meckering, Western Australia*. Record 1968/142. Canberra: Bureau of Mineral Resources, Geology and Geophysics.

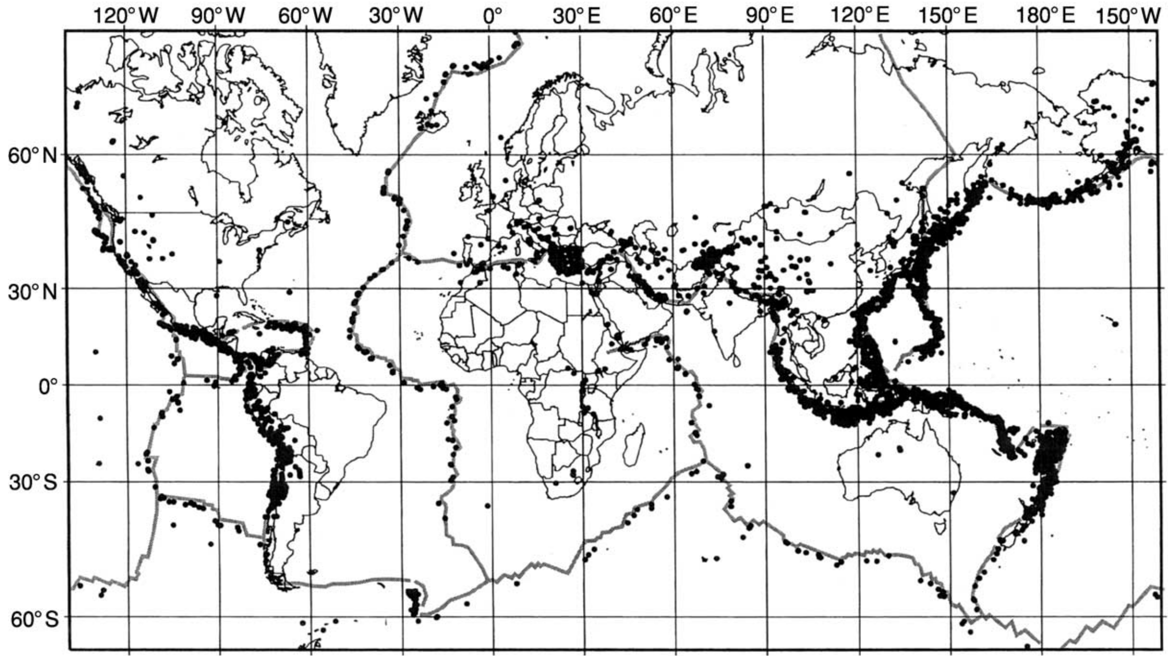


Figure 7 The global distribution of earthquakes that occurred in 1994. Reproduced from US National Earthquake Information Center.

of the most destructive and costly in living memory (55 000 houses destroyed as well as freeway, rail, and port installations), had its epicentre 20 km from the city and had a magnitude of only 6.8. The 1960 Agadir earthquake had a magnitude of only 5.8, but the focus was shallow and right beneath the city.

Thus, magnitude, though a valuable absolute measurement, tells us little about the degree of damage and the loss of property and life, even at the epicentre. The nature of the subsurface rocks can have a significant effect, especially if shock-induced liquefaction occurs, and thus we need another measurement scale. The Mercalli intensity scale measures the relative intensity of the effects felt at a specific site (the intensity will commonly decrease away from the epicentre, but secondary effects such as subsurface variation and shock liquefaction complicate this relationship). In Europe, a modification of the Mercalli scale, the MSK scale (named after Medvedev, Sponheuer, and Karnik), is now used. This scale is given in Table 2.

Tectonic Earthquakes

Earthquakes may be divided into tectonic earthquakes, volcanic earthquakes, and man-made earthquakes. In considering the natural-hazard aspect, it is the tectonic earthquakes that are by far the most destructive natural hazards. This hazard largely affects urban populations, and human design and construction has a unique role in mitigating this hazard.

The actual physical process of ground motion presents little threat to humans in the open: most casualties (other than the casualties of secondary tsunamis) occur inside buildings that partially or totally collapse. The correct design of buildings and constructions such as bridges and viaducts can thus greatly mitigate the damage and casualties resulting from an earthquake.

The vulnerability of a building to earthquake damage varies according to many factors. Vertical ground motion is the principal damaging component causing collapse, burial of people, and death. Lateral ground motion breaks or deforms power lines, pipelines, water pipes and sewers, roadways, railways (Figure 8), and bridges. Quite small lateral offsets can be very damaging. In the Mexico City earthquake of 1985, much damage was caused by adjacent high-rise buildings swaying with different wave motions and knocking each other down. It is noticeable that in Beijing, an earthquake-prone city, the high-rise buildings are widely spaced, with intervening areas of low-rise buildings, so that they cannot interact in this way. In the case of Kobe, the sixth floor of one high-rise building pancaked (Figure 9). The lower floors were built of steel-encased reinforced concrete and the upper floors of pure reinforced concrete; the junction on the sixth floor acted as an element of weakness.

Though earthquakes are mainly an urban hazard, catastrophic earthquakes may strike village populations where low-rise housing is substandard – as in

Table 2 The modified Mercalli (MSK) intensity scale

Intensity	*Effects*
I	Felt rarely. Sometimes dizzyness and nausea. Birds and animals uneasy. Trees, structures, liquids sway.
II	Felt indoors by a few, especially on upper floors. Delicately suspended objects swing.
III	Felt indoors, especially on upper floors by several people. Usually rapid vibration as if a lightly loaded lorry passing. Hanging objects and standing motor cars rock slightly.
IV	Felt indoors by many and outside by a few. Some awakened. No-one usually frightened. Sensation of heavy object striking building. Vibration as of heavy lorries passing. Crockery, windows, doors rattle. Walls and frames creak. Hanging objects and standing motor cars sway.
V	Felt indoors by almost everyone, outdoors by most people. Many awakened, a few frightened and run outdoors. Buildings tremble. Crockery and windows sometimes break. Pictures and doors clatter. Small objects move. Some liquids spilt. Clocks stopped. Trees shaken. Animals anxious.
VI	Felt by all indoors and outdoors. Many frightened, some alarmed. All awakened. People, trees, bushes shaken. Liquids set in motion. Small bells ringing. Crockery broken. Plaster cracks and falls. Books and vases fall over. Some furniture moved. Domestic animals try to escape. Minor landslides on steep slopes.
VII	All frightened, run out of doors, general alarm. Some people thrown to ground. Trees shaken quite strongly. Waves and mud stirred up in lakes. Sandbanks collapse. Large bells ring. Suspended objects quiver. Much damage to badly constructed buildings and old walls. Slight damage to well-built buildings. Chimneys crack. Much plaster, tiles, loose bricks fall. Heavy furniture overturned. Concrete ditches damaged.
VIII	Alarm approaches panic. Vehicle drivers disturbed. Trees broken and shaken. Sand and mud spurt from the ground. Marked changes to springs and wells. Much damage to ordinary and older buildings. Walls, pillars, chimneys, towers, statues, gravestones crack and fall. Very heavy furniture overturned.
IX	General panic. Ground cracked open (10 cm). Much damage to structures built to withstand earthquakes. Frequent partial or total collapse of other buildings. In reservoirs, underground pipes broken. Buildings dislodged from foundations, rock falls.
X	Widely cracked ground, fissures up to 1 m wide. Frequent river bank and coastal landslides and shifted sands. Water levels change. Water thrown onto riversides. Serious damage to dams, embankments, bridges. Severe damage to well-built wooden structures. Masonry structures destroyed along with their foundations. Rails bent. Open cracks or waves on roads. Pipes torn apart.
XI	Widespread serious ground disturbance, broad fissures, landslips, landslides. Muddy water spurts upward. Tsunamis develop. Severe damage to all wooden structures. Great damage to dams. Few masonry structures remain upright. Pillars of bridges and viaducts wrecked. All pipelines wrecked. Rails badly bent.
XII	Total damage to all constructions. Great disturbance to ground with many shearing cracks. Many landslides on slopes, rockfalls common, rock masses dislocated, water channels altered and dammed. Ground surface waves like water and ground remains undulating. Objects thrown into the air.

the case of the Cairo earthquake in 1992 and the Killari, central India, earthquake in 1993. In the case of the Cairo earthquake, poorly constructed extra storeys had been added to moderate-rise housing. In Killari, stone-built low-rise houses were poorly constructed (Figure 10). In the catastrophic Bam earthquake of December 2003, the mud bricks of the low-rise dwellings crumbled and collapsed, leaving few air spaces to allow buried victims to breathe; another factor responsible for the scale of the fatalities was the fact that all the dwellings in southern Iran have basements cooled by wind towers designed for the sweltering summer heat, and many victims would have been asleep in them.

Some of the most devastating historic and recent earthquakes are listed in Table 3.

Where cities are situated in plate-boundary zones the effects are most disastrous. The San Francisco earthquake of 1906 (Figure 11) provides an example of this. The Kobe earthquake of 1995, in which the financial loss was US $200 billion, occurred in a city that experiences a tremor every few days. An analysis of the locations of 100 of the largest cities in the world, which accommodate 10% of the global population, shows that they can expect to experience an earthquake of intensity VI or more on the MSK scale within 50 years.

The earthquake hazard extends beyond high-risk cities such as those sited on plate boundaries. Entire countries may be at quite low risk, yet have some vulnerability. The UK is a low-risk country, and earthquakes of more than magnitude 5.5 are extremely unlikely (Figure 12). Charles Davidson published a list of 1191 recorded shocks in Britain between AD 974 and AD 1924. In Lincoln in 1185, “great stones were rent; houses of stone fell; the metropolitan church of Lincoln was rent from top to bottom” and there is a similar report from the cathedral city of Wells in 1248. Two apprentices were killed in London in 1580 as a result of an earthquake in the Dover Straits. The Colchester earthquake in 1884 (magnitude 4.7) peaked in intensity near the epicentre (between Pelden and Langenhoe) (Figure 13) at MSK VIII and caused widespread damage, which was

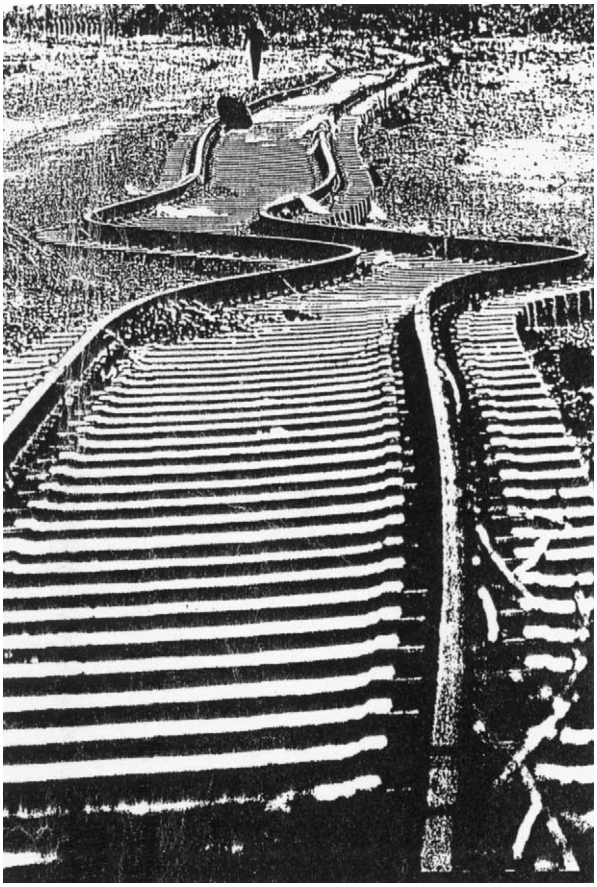


Figure 8 A railway track in the western USA twisted and shortened by lateral motion during an earthquake.

compensated by a Mansion House Fund that paid out £9000 (equivalent to £500 000 today). The Roermond earthquake in the Netherlands in 1998 (magnitude 5.8) caused £30–40 million of damage, despite the single fatality. It is predicted that a magnitude 5.7 earthquake focused at a depth of 5 km directly under Manchester would cause havoc. The increasing size of conurbations and cities increases their vulnerability: the Colchester area would suffer much more nowadays from a comparable earthquake to the 1884 event, because the population and industry are now much denser than at that time.

Earthquakes can occur in areas that are not considered to be at risk. The Spitak earthquake in Armenia in 1988 is such a case. The region was not considered to be at high risk, and a nuclear power station was planned for the Spitak area. This earthquake caused the whole process of earthquake risk assessment in the Soviet Union to be revised.

The New Madrid, Missouri, earthquakes of 1811–1812 are even more surprising. There is a reliable historical record of three earthquakes spaced over two months with magnitudes 8.2, 8.1, and 8.3. They rang the bells of Boston and rattled Quebec and provide a remarkable example of major interplate seismicity.

Secondary effects The secondary effects of earthquakes can be as destructive and lethal as the primary effects, or more so.

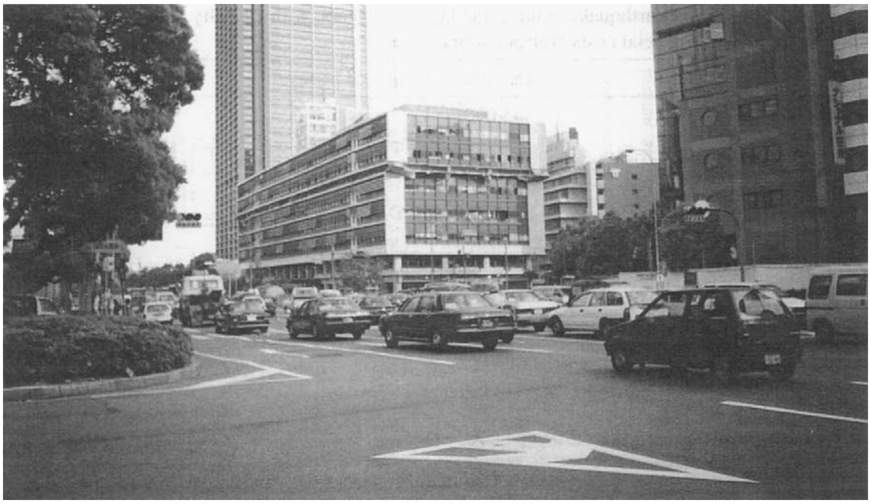


Figure 9 High-rise buildings in Kobe after the 1995 earthquake, showing the sixth floor pancaked by vertical motion. Reproduced from Esper P and Tachibana E (1998) The Kobe earthquake. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology*, pp. 105–116. Engineering Geology Special Publication 15. London: Geological Society.

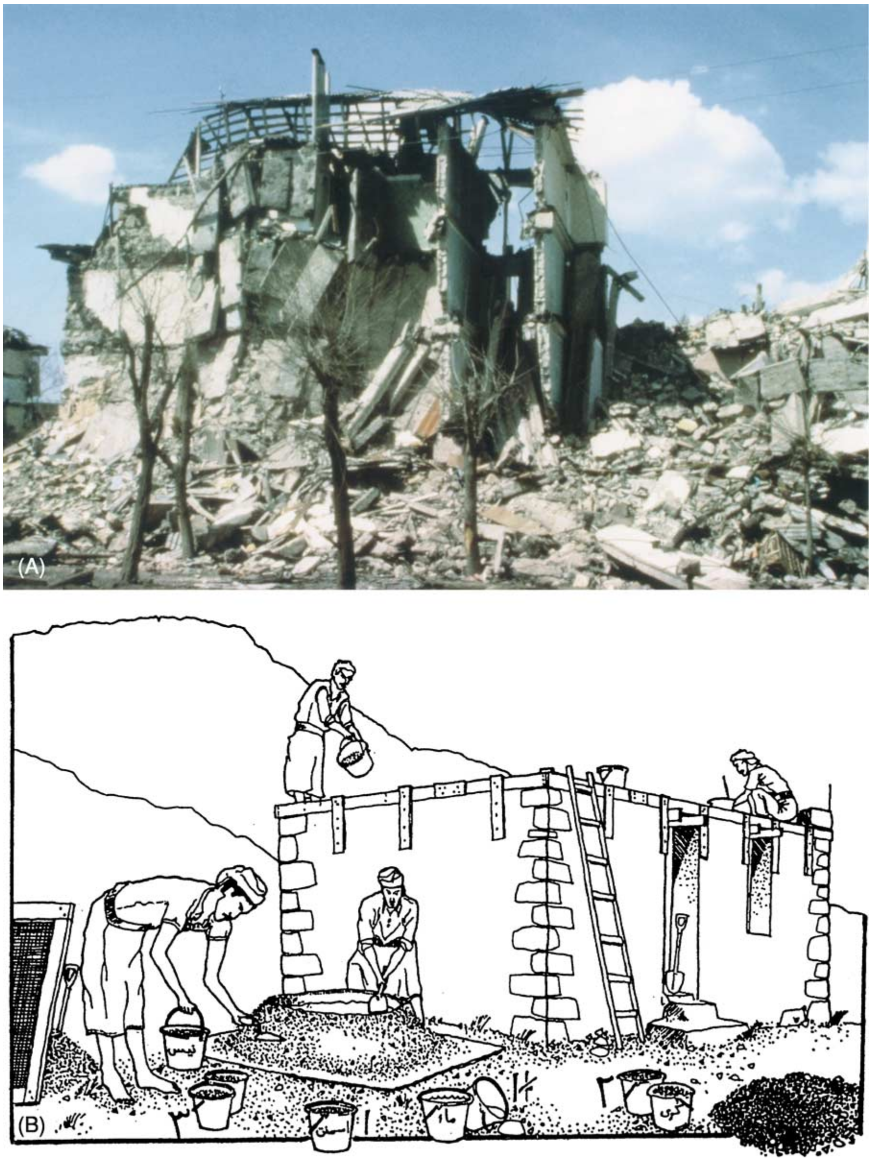


Figure 10 (A) Damage to a low-rise poorly constructed stone building of the type affected by the Killari, India, 1991 earthquake (photograph National Geophysical Data Center USA). (B) Improved training in building similar low-rise buildings in the Yemen. Reproduced from Degg MR (1995) Earthquakes, volcanoes and tsunamis: tectonic hazards in the built environment of southern Europe. *Built Environment* 21: 94–113. Courtesy of the Geological Society, London.

Tsunami The so-called tidal wave generated by earthquakes is probably the most lethal secondary effect: during the twentieth century earthquakes in Chile caused fatalities in Hawaii and Japan. The effects of tsunamis may be felt thousands of kilometres from the earthquake epicentre, but can be mitigated by systematic warning systems.

Fire Tokyo in 1923, San Francisco in 1906, and Kobe in 1995 all suffered from the secondary effects of fire. This was exacerbated by the fact that water supplies were cut off. In Tokyo there was a firestorm. In San Francisco, 70% of the damage was due to fire.

Liquefaction of sands, silts, and clays Another important secondary effect is that thixotropic sands and silts, which liquefy on shock, greatly increase the damage: examples of this are the waterfront area in the Messina earthquake, Sicily, of 1908, in which 98% of the houses were ruined and 160 000 died; Mexico City in 1985, where the worst damage was in building developments founded on old lake deposits (the wave amplitude was magnified 8–50 times here); and Anchorage, Alaska, where a magnitude 8.4 earthquake with an epicentre 130 km away caused devastation in a housing development founded on the thixotropic Bootlegger Clay

Table 3 Some important earthquakes in the last 2000 years (various sources): note that magnitudes are on various scales. An earthquake at Gujarat, India, in 2001, which killed more than 50 000 people has been omitted from the table

Year AD	*Place*	*Casualties*	*Estimated loss*
342	Antioch	40 000	
454	Sparta	20 000	
565	Antioch	30 000	
856	Corinth	45 000	
1170	Sicily	15 000	
1290	Chihli, China	100 000	
1456	Naples	60 000	
1556	Shensi, China (M 8.3?)	830 000	
1716	Algiers	20 000	
1737	Calcutta	300 000	
1755	North Persia	40 000	
1755	Lisbon (M 6.9?)	60 000	
1759	Baalbek	20 000	
1783–1786	Calabria	50 000	
1797	Quito	41 000	
1822	Aleppo	22 000	
1828	Honshu	34 000	
1896	Sanriku, Japan	28 000	
1897	Assam (M 8.7)	1542	6 major towns and all villages in 30 000 sq. miles leveled
1906	San Francisco (M 7.9?)	700	US$ 400 million
1908	Messina, Reggio (M 7.5)	160 000	98% of houses ruined
1915	Avezzano, Italy (M 7.0)	30 000	
1920	Kansu, China (M 8.5)	180 000	Vast landslides
1923	Tokyo (M 8.2)	143 000	Firestorm killed 38 000, 25 000 houses destroyed
1932	Kansu, China (M 8.5)	70 000	
1933	Long Beach, CA	120	US$ 50 million
1933	Sanriku, Japan	3000	8800 houses destroyed by tsunami
1935	Quetta (M 7.5)	60 000	
1939	Concepcion, Chile (M 8.5)	30 000	
1939	Erzincian, Turkey (M 8.0?)	40 000	30 000 dwellings destroyed
1948	Soviet–Iran border	19 000	
1960	Agadir (M 9.5?)	60 000	
1960	Chile (M 9.5?)	10 000	58 600 houses destroyed
1964	Anchorage (M 8.4)	114	
1970	Peru (M 7.9)	80 000	Devastating rock and ice falls
1971	San Fernando, CA	64	US$ 1 billion
1975	Haicheng, China (M 7.5)	1328	
1976	Guatemala (M 7.5)	22 000	
1976	Tangshan, China (M 7.7)	240 000 (some estimates are as high as 850 000)	Vast damage
1985	Mexico City (M 8.1)	10 000	US$ 4 billion
1988	Spitak (M 6.9)	30 000	US$ 14 billion; accompanied by landslides and rockfalls
1989	Loma Prieta, CA (M 7.1)	63	US$ 7 billion
1989	Newcastle, Australia	10	US$ 7 billion
1990	Northwest Iran (7.7)	40 000	US$ 8 billion
1991	Killari, India	10 000	Immense destruction of village housing
1992	Cairo (M 5.5–5.9)	<500	40 000 homeless; may have been due to construction of Aswan Dam
1994	Northridge, CA (M 6.7)	60	US$ 20 billion
1995	Kobe, Japan (M 7.2)	5429	US$ 200 billion
1999	Izmit, Turkey (M 7.6)	>17 000	Immense destruction
2003	Bam, Iran	~45 000	Immense destruction of modern city and ancient citadel destroyed

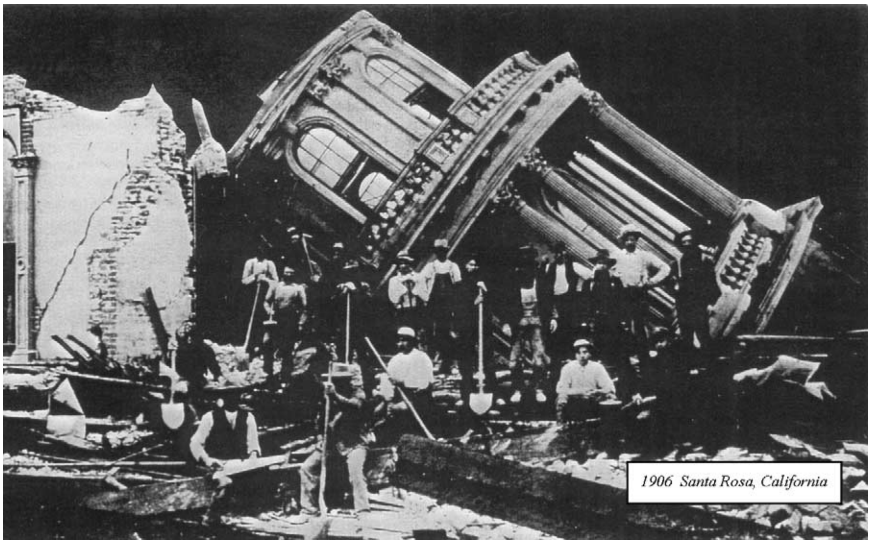


Figure 11 Gross displacement of a large building in the San Francisco 1906 earthquake.

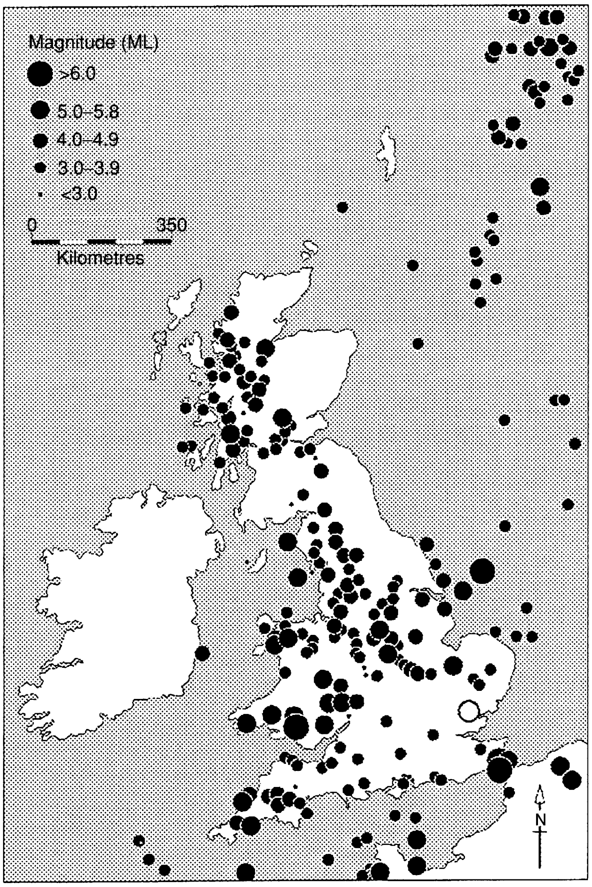


Figure 12 The magnitudes and locations of earthquakes in Great Britain greater than magnitude 3 after 1700 and greater than magnitude 4 before 1700. Reproduced from Musson R (1996) British earthquakes and the seismicity of the UK. *Geoscientist* 16: 24–25.

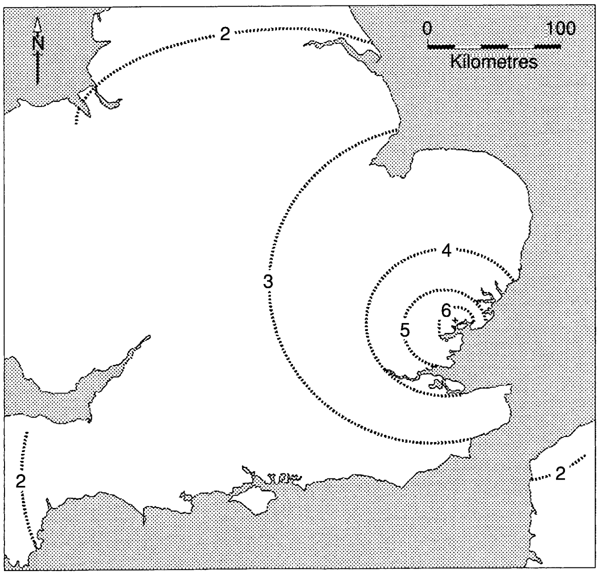


Figure 13 Intensity plot of the Great Colchester Earthquake of 1884 using the MSK scale. Reproduced from Musson R, Neislon G, and Burton PW (1990) *Microseismic Reports on Historic British Earthquakes XIV: 22 April 1984 Colchester*. BGS Seismology Report W1/90/33. Edinburgh: British Geological Survey.

Formation (Figure 14). Here, the risk was well known but there was a lack of communication between the geologists and the planners.

Landslides and rock falls Very damaging landslides or rock falls can be triggered by earthquakes and may occur some time after the main shock or aftershocks. In Montana in 1985, 30 million tonnes of rock were

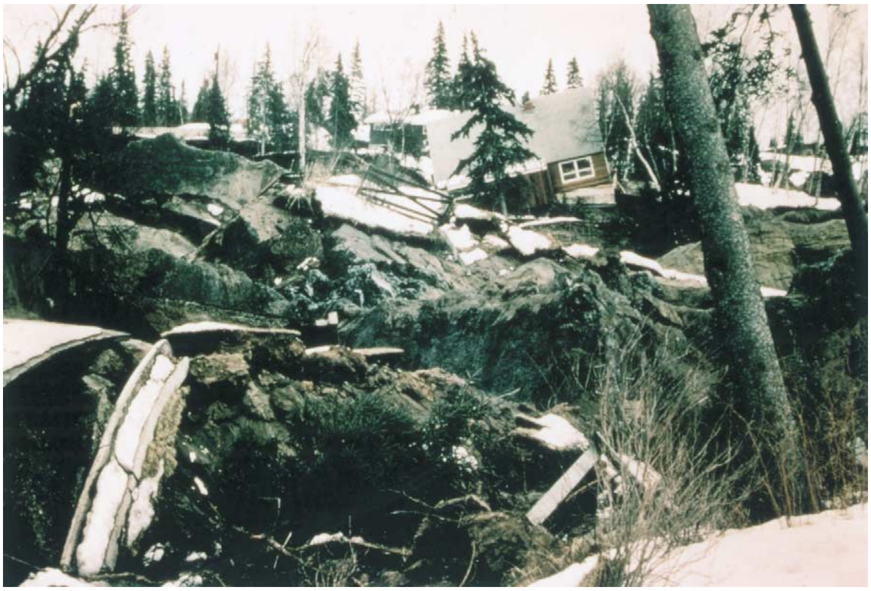


Figure 14 Destruction of the housing development at Turnagain Point above the Bootlegger Clay, Anchorage, 1994. The bluff moved 606 m towards the bay, and 75 homes were destroyed. Sand lenses in the clay lost strength (photograph National Geophysical Data Center, USA).

set in motion, fatally burying 36 people at a camp site. In Peru in 1970, rock and ice slides triggered by an earthquake killed 20 000 people.

Disease The risk of disease is a major concern after earthquakes, particularly in hot climates. Water supplies may be cut off and the populace may resort to using polluted supplies of water mixed with sewage and drainage effluents, which may be contaminated by corpses.

Starvation Normal food supplies may be cut off and transport arteries may be blocked, so it is important to bring in food supplies immediately from the world outside.

Exposure The 1988 Spitak earthquake in Armenia illustrates the problem of exposure. Many people lost their dwellings and were living in the open in very cold December climatic conditions. The need for tents, warm clothing, and blankets was urgent.

Looting Looting is prevalent after earthquakes.

Mitigation The potential for mitigation of the earthquake hazard is limited. The main ways of mitigating the hazard are through good building and constructional design, planning development away from at-risk areas, and warning. However, warning is a very difficult matter. Research into earthquakes is at an interim stage, and the scientific community is at present by no means in consensus about the physical processes involved. There is the problem of how threatened populations react to warnings: if the event fails to occur, especially more than once, the population may not heed future warnings; alternatively, giving a warning to evacuate may engender panic. In countries with controlled political systems, such as China, warning and evacuation may be easier than in a democratic Western country. The Chinese did evacuate the city of Haicheng twice in 1974 and 1975, on 4 December and again on 4 February, based on seismology, community monitoring levels, radon gas in water, water temperature, tiltmeters, magnetometer readings, and patterns of animal behaviour. An earthquake of magnitude 7.3 struck at 7.36 AM on 4 February. However, the great Tangshan earthquake of 1976 struck without any prediction or warning and killed at least 240 000 people (possibly many more).

Millions of US dollars have been spent on research into earthquakes in California, and some improvement in prediction has been achieved, yet the existence of the Northridge Fault, the site of the 1994 earthquake, which killed 60 people and caused 20 billion dollars worth of damage, was not even known before the event.

The best mitigation procedure would be to have international teams ready with emergency supplies and equipment, trained personnel, and sniffer dogs, at a distance from earthquake-prone regions, ready to been flown in by plane and helicopter.

Research into earthquakes The most important research into earthquakes has involved statistical, geographical, geological, theoretical, and mathematical studies of seismicity. An example is a study by Lya Tuliani in Russia, which addressed problems of geodynamics and seismology, tectonosphere layering, and lithostructure in seismically active regions in order to develop risk estimates. The procedure involved mathematical data processing. It was claimed that this study allowed highly accurate prediction of the coordinates of high-risk sites. This statistical, mathematical, and office-based approach contrasts with research in the USA (which involves actually drilling down to earthquake foci on faults), research into the Boothiel lineament, the site of the New Madrid earthquakes (which has revealed sand boils caused by the earthquake), and excavations in the alluvium of the rice paddies west of Beijing, China (where the actual earthquake fault of a seventeenth century event has been exposed cutting the clayey alluvium in open pits). Research has been carried out into much older earthquakes in Iran, based on the dislocation of qanats (horizontal wells). All these approaches and many more are invaluable, but the problem of predicting earthquakes is extremely complex and may never be completely solved.

Volcanic Earthquakes

There can be a connection between major tectonic earthquakes and volcanic eruptions. In Chile in 1960, a major earthquake triggered the eruptions of several volcanoes, and in Sicily there is a record of Etna erupting a day or two before a major tectonic earthquake. However, the swarms of small seisms that usually precede volcanic eruptions (though there may be no such prelude) pose little threat to life and property. They do, however, provide valuable warnings of forthcoming eruption, and arrays of instruments are mounted on dangerous volcanoes for this purpose.

Man-Made Earthquakes

Small seismic disturbances can be triggered by human activity. In the USA, the Boulder Dam and Lake Mead are constructed in a region that is highly strained; many small shocks have been correlated with changes in water depth. In Colorado, the injection of liquid wastes down boreholes has also been shown to trigger small seisms.

Moonquakes and Seisms on Other Planets

Very small earthquakes do occur on the largely quiescent Moon when it undergoes maximum tidal stresses resulting from the attraction of the Earth and Sun. Similar stresses must operate on the Earth and cause minor seisms, but the effect is of no importance in such a dynamic body. Moonquakes and artificial seisms produced on the surfaces of other extraterrestrial bodies – Mars, Venus, and Mercury – can provide a valuable insight into their internal make up. A fascinating project would be to site an instrument from an unmanned spacecraft on Io, Jupiter's volcanically active satellite, to obtain detail of its interior configuration – Io must be seismically active.

See Also

Earth: Mantle; Crust. **Earth Structure and Origins**. **Engineering Geology:** Aspects of Earthquakes; Natural and Anthropogenic Geohazards; Liquefaction. **Plate Tectonics**. **Tectonics:** Faults.

Further Reading

Bolt BA (1999) *Earthquakes*. New York: Freeman.

Bommer JJ and Ambraseys NN (1989) The Spitak, Armenia, USSR earthquake of 7 December 1988: a summary engineering geology report. *Earthquake Engineering and Structural Dynamics* 18: 921–925.

Chen Y, Tsoi KL, Chen F, *et al.* (1988) *The Great Tangshan Earthquake of 1976*. Oxford: Pergamon.

Degg MR (1992) Some implications of the 1985 Maxican earthquake for hazard assessment. In: McCall GJH, Laming DJC, and Scott SC (eds.) *Geohazards – Natural and Man-Made,* pp. 93–114. London: Chapman and Hall.

Degg MR (1995) Earthquakes, volcanoes and tsunamis: tectonic hazards in the built environment of southern Europe. *Built Environment* 21: 94–113.

Degg MR (1998) Hazard mitigation in the urban environment. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology,* pp. 329–337. Engineering Geology Special Publication 15. London: Geological Society.

Esper P and Tachibana E (1998) The Kobe earthquake. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology,* pp. 105–116. Engineering Geology Special Publication 15. London: Geological Society.

Everingham I (1968) *Preliminary report on the 14th October 1968 earthquake at Meckering, Western Australia*. Record 1968/142. Canberra: Bureau of Mineral Resources, Geology and Geophysics.

Keller GR (2000) Seismic properties of rocks. In: Hancock PL and Skinner BJ (eds.) *The Oxford Companion to the Earth*. Oxford: Oxford University Press.

McCall GJH (1996) Natural hazards. In: McCall GJH, de Mulder EFJ, and Marker BR (eds.) *Urban Geoscience,* pp. 81–125. Rotterdam: Balkema.

McCall GJH (2000) The great Colchester earthquake of 1884 revisited. *Geoscientist* 10: 4–6.

McCall GJH (2004) Remembering Bam. *Geoscientist* 14: 8–9.

Menard HW (1974) *Geology, Resources and Society.* San Francisco: WH Freeman and Co.

Musson R (1996) British earthquakes and the seismicity of the UK. *Geoscientist* 16: 24–25.

Musson R, Neislon G, and Burton PW (1990) *Microseismic Reports on Historic British Earthquakes XIV: 22 April 1984 Colchester.* BGS Seismology Report W1/90/33. Edinburgh: British Geological Survey.

Scarth A (1997) *Savage Earth.* London: HarperCollins.

Tuliani LI (1999) *Seismicity and Earthquake Risk: On the Basis of Thermodynamic and Rheological Parameters of the Tectonosphere.* Moscow: Scientific World.

Van Andel TJ (1994) *New Views on an Old Planet.* Cambridge: Cambridge University Press.

Wong IG (2000) Earthquake mechanisms and plate tectonics. In: Hancock PL and Skinner BJ (eds.) *The Oxford Companion to the Earth,* pp. 287–289. Oxford: Oxford University Press.

Faults

S Stein, Northwestern University, Evanston, IL, USA

Introduction

Faults are surfaces in the Earth along which one side moves or has moved with respect to the other. They are identified either when an earthquake occurs or by geological mapping showing that motion across the fault has occurred in the past. Many faults are inactive, in the sense that there has been no motion across them within some defined time interval, typically the past million years or less. Other faults are active, in the sense that recent motion has occurred and hence motion might be expected in the future. Faults, and the earthquakes on them, are studied to understand both the regional tectonics and the mechanics of faulting.

Typically, earthquakes occur on previously identified faults, demonstrate that the fault is active, and provide information on the fault's geometry and the motion on it. For example, in the famous 1906 San Francisco earthquake, one of the first earthquakes to be carefully studied, several metres of relative motion occurred along several hundred kilometres of the San Andreas Fault. Hence, H Reid proposed the elastic-rebound theory of earthquakes, in which materials on opposite sides of the fault move relative to each other, but friction 'locks' the fault and prevents it from slipping (Figure 1). Eventually more strain accumulates than the fault rocks can withstand, and the fault slips in an earthquake. The motion is sometimes revealed after earthquakes by linear features, including roads and rows of trees (Figure 2). Those who study earthquakes seek to understand both the geological processes causing earthquakes and the physics of faulting. These issues are important for society because knowing where and when earthquakes are likely and the expected ground motion during them can help to mitigate the risk that they pose.

The largest earthquakes occur at plate boundaries. We view them as the most dramatic part of the seismic cycle, which takes place on segments of the plate boundary over hundreds or thousands of years. During the interseismic stage, which makes up most of the cycle, steady motion occurs at a distance from the locked fault. Immediately prior to rupture there is the preseismic stage, during which small earthquakes (foreshocks) and other possible precursory effects may occur. The earthquake is the coseismic phase, during which rapid fault slip generates seismic waves. During these few seconds, metres of slip on the fault 'catch up' with the few millimetres per year of motion that have occurred over hundreds of years at a distance from the fault. Finally, a postseismic phase occurs after the earthquake, during which aftershocks and transient afterslip occur for a period of years before the fault resumes steady interseismic behaviour.

Because this cycle extends over hundreds of years, we do not have observations of it in any one place. Instead, our view of the seismic cycle is based on a combination of observations from different places. It is far from clear how good this view is and how well our models represent the complexity of the seismic cycle. As a result, earthquake and fault studies remain active research areas that integrate a variety of techniques. Seismology is used to study the motion during earthquakes. Historical records often provide data on the earthquake cycle for a given fault segment. Field studies provide information about the location, geometry, and history of faults. Geodetic measurements are used to study ground deformation before, during, and after earthquakes and thus provide information about the processes associated with fault locking and afterslip. Results for individual earthquakes are combined with those from other analyses, including laboratory studies of rock deformation, to understand how the earthquakes in a region reflect the large-scale tectonic processes causing them and to study the physics of faulting.

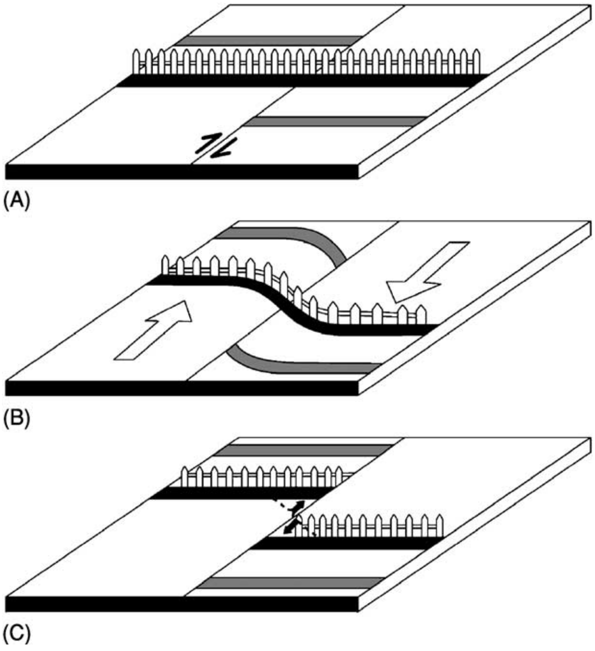


Figure 1 The elastic-rebound model of earthquakes assumes that between earthquakes material at a distance from the fault undergoes relative motion. Because the fault is locked, features across it that were originally linear (A), such as a fence, are slowly deformed with time (B). Finally, the strain becomes so great that the fault breaks in an earthquake, offsetting the features (C). (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

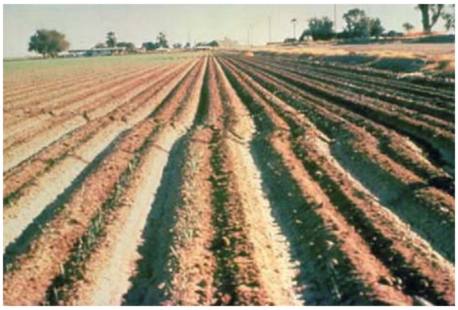


Figure 2 Displacement of crop rows resulting from an earthquake on the Imperial fault, El Centro, California on 15 October 1979. (Courtesy of the National Geophysical Data Center.)

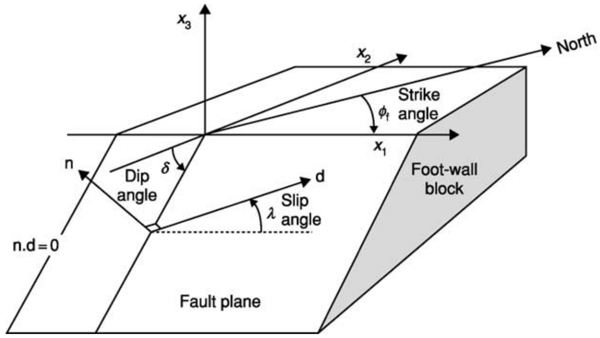


Figure 3 Fault geometry used in earthquake studies. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

Fault Geometry

We treat faults as planar surfaces across which relative motion occurs during earthquakes. Geological observations of faults that reach the surface show that this is often approximately the case, although complexities are common. This assumption is usually also consistent with seismic data.

As shown in Figure 3, the fault plane is characterized by **n**, its normal vector. The direction of motion is given by **d**, the slip vector in the fault plane. The slip vector indicates the direction in which the upper side of the fault, known as the hanging-wall block, moves with respect to the lower side of the fault, known as the foot-wall block. Because the slip vector is in the fault plane, it is perpendicular to the normal vector.

A coordinate system for studying faults has the x_1 axis in the fault strike direction, the intersection of the fault plane with the Earth's surface. The x_3 axis points upwards, and x_2 is perpendicular to the other two axes. The dip angle, δ, gives the orientation of the fault plane with respect to the surface. The slip angle, λ, gives the motion of the hanging wall with respect to the foot wall. Fault geometries are described by the slip angle (Figure 4). When the sides slide by each other, pure strike-slip motion occurs. When $\lambda = 0°$, the hanging wall moves to the right, and the motion is called left-lateral. Similarly, for $\lambda = 180°$, right-lateral motion occurs. To tell which is which, look across the fault and see which way the other side moves. The other basic fault geometries describe dip-slip motion. When $\lambda = 270°$, the hanging wall slides downwards, causing normal faulting. In the opposite case, $\lambda = 90°$ and the hanging wall goes upwards, yielding reverse or thrust faulting. Seismologists often use the terms reverse fault and thrust fault interchangeably, whereas structural geologists reserve the term thrust fault for a shallow-dipping reverse fault. Most earthquakes consist of some combination of these motions and have slip angles between these values.

If we treat a fault as rectangular, the dimension along strike is called the fault length and the dimension in the dip direction is known as the fault width. Actual fault geometries can be much more complicated than a rectangle. The fault may curve, requiring a three-dimensional description. Rupture may occur over a long period and consist of several subevents on different parts of the fault with different

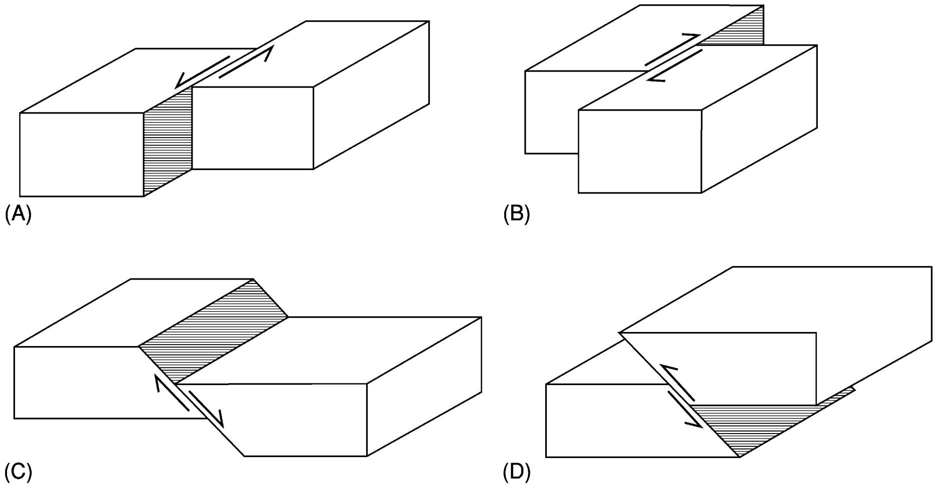


Figure 4 Basic types of faulting. (A) Left-lateral strike-slip fault, $\lambda = 0°$. (B) Right-lateral strike-slip fault, $\lambda = 180°$. (C) Normal dip-slip fault, $\lambda = -90°$. (D) Reverse dip-slip fault, $\lambda = 90°$. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

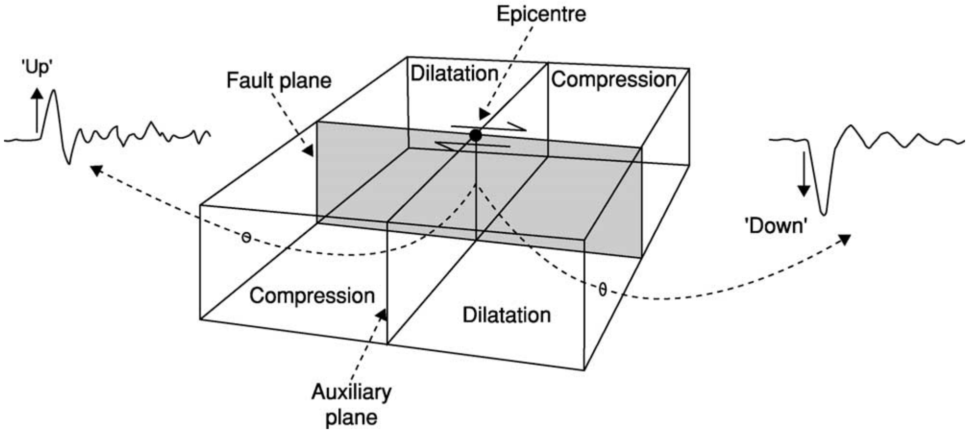


Figure 5 First motions of P waves observed at seismometers located in various directions about the earthquake provide information about the fault orientation. The nodal planes separate compressional ('up') and dilatational ('down') first arrivals. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

orientations. Such complicated seismic events, however, can be treated as a combination of simple events.

Seismological and Geodetic Studies

Seismological studies provide much of our information about earthquakes and the faults on which they occur. The arrival times of seismic waves at seismometers at different sites are first used to find the location of an earthquake, known as the focus or hypocentre. This location is often shown by the epicentre, the point on the Earth's surface directly above the earthquake. Next, the amplitudes and shapes of the radiated seismic waves are used to study the size of the earthquake, the geometry of the fault on which it occurred, and the direction and amount of slip.

The geometry of faulting during an earthquake, known as the focal mechanism, is found by using the fact that the pattern of radiated seismic waves depends on the fault geometry. Seismic waves are divided into P or compressional waves, in which material moves back-and-forth in the direction of wave propagation, and S or shear waves, where material moves at right angles to the propagation direction. P waves travel faster than S waves, so the first pulse to arrive is a P wave. The simplest method uses the first motion, or polarity, of the first-arriving P wave, which varies between seismic stations at different directions from an earthquake. As illustrated (Figure 5) for a strike-slip earthquake on a vertical fault, the first motion is either compression, for stations located such that material near the fault moves 'towards' the station, or dilatation, where the motion is 'away from' the station. The first motions define compressional and dilatational quadrants, divided by two perpendicular nodal planes – the fault plane and the auxiliary plane, which is perpendicular to the fault plane. If these planes can be found, the fault geometry is known, and can be plotted using the familiar stereographic or 'beachball' representation.

Because the first motions from slip on the actual fault plane and the auxiliary plane would be the same, first motions alone cannot resolve which is the actual fault plane. However, additional information can often settle the question. Sometimes geological or geodetic information, such as the trend of a known fault or observations of ground motion, indicates the fault plane. Often, smaller aftershocks occur on and thus delineate the fault plane. If the earthquake is large enough, the finite time required for slip to progress along the fault causes variations in the waveforms observed at different directions from the fault; such directivity effects can be used to infer the fault plane.

More sophisticated techniques use the amplitudes and shapes of the seismic waves. These waves can be body waves, which travel through the Earth's interior, or surface waves, which propagate along paths close to the Earth's surface. The approach is to compare the observed body and surface waves with theoretical, or synthetic, waveforms computed for various source parameters, and find a model that best fits the data. Such analysis also gives information about the earthquake depths and rupture processes, which cannot be extracted from the first motions.

Figure 6 shows how body waves can be used to check the mechanism and estimate the depth. Synthetic seismograms were computed for various focal depths. The left-hand panel shows the expected timing and amplitudes of various arriving phases, and the right-hand panel shows the synthetic seismogram resulting from including the effect of the earthquake source and seismometer. The data are fitted well by a source at a depth of about 30 km.

Modelling surface waves can also help to resolve earthquake focal mechanisms and depths. Depending on the fault geometry, more energy is radiated in some directions than others. Figure 7 shows theoretical radiation patterns for the two kinds of surface waves, Love and Rayleigh, corresponding to several focal mechanisms with a fault plane striking north. The patterns are distinctive: a vertical strike-slip fault has two four-lobed patterns, such that Love-wave amplitude has maxima in the north, east, south, and west directions, whereas the Rayleigh-wave amplitude has maxima in the north-east, south-east, south-west, and north-west directions. In contrast, a dip-slip fault dipping at 45° has a four-lobed Love-wave pattern and a two-lobed Rayleigh-wave pattern. Such patterns can be generated for any fault geometry and compared with observations to find the best-fitting source geometry.

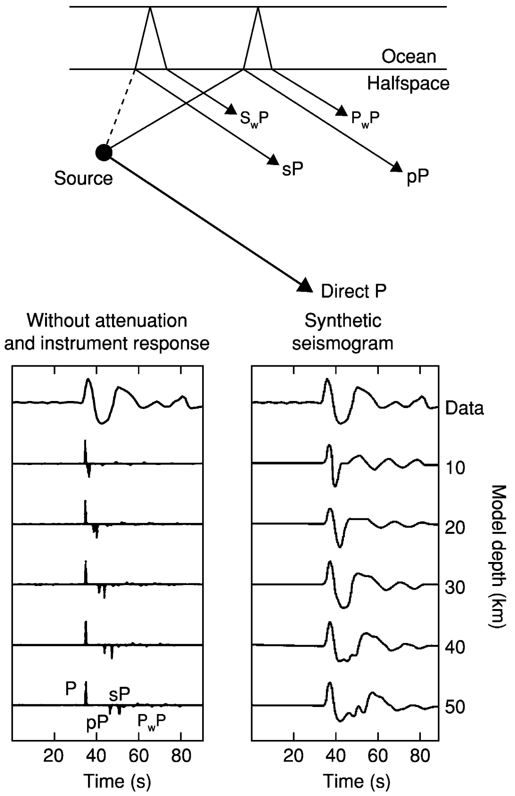


Figure 6 Body-wave modelling for depth determination. Synthetic seismograms for an assumed fault geometry, including the effects of the seismometer and attenuation, are calculated for various depths. The data are best fitted by a depth of approximately 30 km. (Reproduced from Stein S and Wiens D (1986) Depth determination for shallow teleseismic earthquakes: methods and results. *Reviews of Geophysics and Space Physics* 24: 806–832.)

For earthquakes on land, additional information is derived by measuring the deformation of the Earth resulting from the earthquake using geodesy, the science of the Earth's shape. Most such techniques rely on detecting the motion of geological or manmade features or of geodetic monuments, which are markers in the ground. Until recently, these measurements were typically made by triangulation, which measures the angles between monuments using a theodolite, or trilateration, which measures distances with a laser. Vertical motion was measured using a precise level to sight on a distant measuring rod. However, the advent of geodetic methods using signals from space permits all three components of position to be measured with sub-centimetre precision. As a result, geodetic measurements before and after earthquakes can now determine

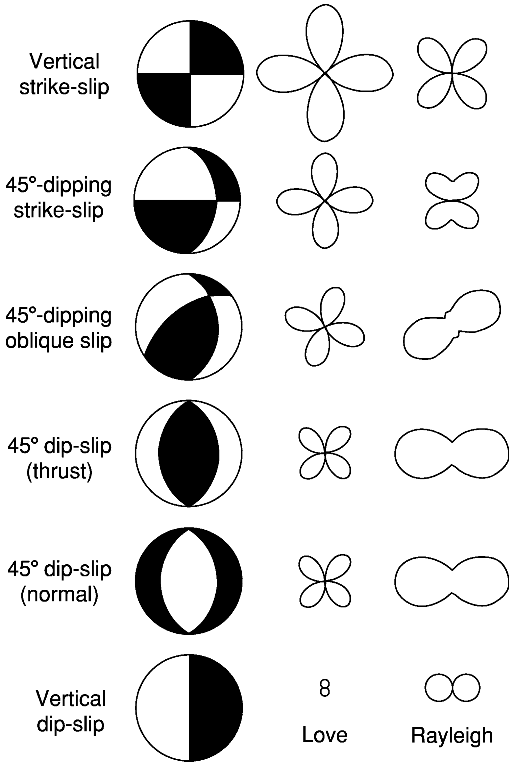


Figure 7 Focal mechanisms and surface-wave amplitude radiation patterns for six fault geometries. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

coseismic motion with high precision much more easily than was previously possible.

Faults and Stresses

The simplest theory for fracture predicts that faulting will occur on the plane on which the shear stress is highest. Although this is not exactly true, it provides an insight into the relation between fault orientations and regional tectonics. Consider an experiment in which a rock is compressed (Figure 8) with principal stresses $|\sigma_1| \geq |\sigma_2| \geq |\sigma_3|$. We expect fracture on the planes of maximum shear stress, which are 45° from the maximum and minimum principal stress axes and include the intermediate principal stress axis. Alternatively, if the experiment involves a situation known as uniaxial compression, where $|\sigma_1| \geq |\sigma_2| = |\sigma_3|$, failure should occur on any plane at 45° to the maximum principal stress (σ_1) axis. Experiments support the idea that fracture is controlled by shear stress, but in a more complicated way, so the fracture plane is often about 25° rather than 45° from the maximum principal stress direction.

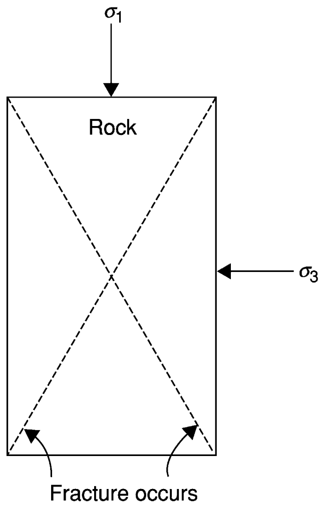


Figure 8 Schematic illustration of an experiment in which a rock sample is compressed along the direction of the maximum principal stress until fracture occurs. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

For simplicity, however, we often assume that faults in the Earth form on the planes of maximum shear stress. The three basic fault geometries – strike-slip, normal, and thrust – are related to the stress axes (Figure 9). If the vertical principal stress is the most compressive, the fault dips at 45°, and normal faulting occurs. If, instead, the vertical principal stress is the least compressive, the fault geometry is the same but reverse or thrust faulting occurs. When the vertical principal stress is the intermediate principal stress, strike-slip motion occurs on a fault plane 45° from the maximum principal stress. Thus, the geometry of faults, which can be mapped geologically or inferred from seismograms of earthquakes, can be used to study stress orientations.

Fault Strength

Using seismic waves alone limits what we can learn about the earthquake process. The seismic waves radiated from an earthquake reflect the geometry of the fault and the motion on it, and so can give an excellent picture of the kinematics of faulting. However, they contain much less information about the actual physics, or dynamics, of faulting. Hence seismological results are combined with experimental and theoretical studies of rock friction and fracture to explore the physics of faulting.

Consider the strain that results from compressing a rock specimen. For small stresses the resulting strain is proportional to the applied stress, so the material is purely elastic (Figure 10A). Once the stress reaches the rock's fracture strength, σ_f, the rock breaks. Such brittle fracture is the simplest model for an earthquake on a fault. Other materials show a change in the stress–strain curve for increasing stresses (Figure 10B). For stresses less than the yield stress, σ_o, the material acts elastically. If the stress is released, the strain returns to zero. However, for stresses greater than the yield stress, releasing the stress relieves the elastic portion of the strain but leaves a permanent deformation (Figure 10C). If the material is restressed, the stress–strain curve now includes the point of the permanent strain. The portion of the curve corresponding to stress above the yield stress is called plastic deformation, in contrast to the elastic region where no permanent deformation occurs. Materials showing significant plasticity are called ductile.

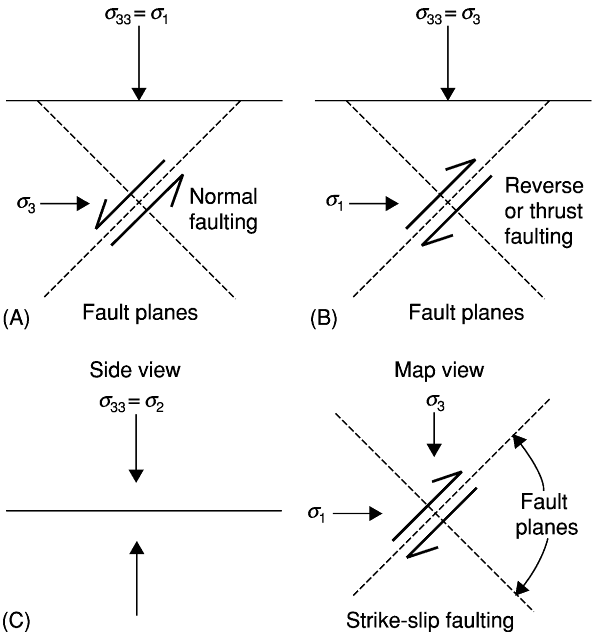


Figure 9 Stress fields associated with the three types of faulting, assuming that the earthquake occurred on a plane of maximum shear stress. (A) Normal, (B) reverse, (C) and strike-slip faulting involve different orientations of the principal stresses. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

At low pressures rocks are brittle, but at high pressures they behave ductilely, or flow. Figure 11 shows experiments where a rock is subjected to a compressive stress that exceeds a confining pressure. For confining pressures less than about 4 Kb the material behaves brittlely – it reaches the yield strength and then fails. For higher confining pressures the material flows ductilely. These pressures occur not far below the Earth's surface – each 3 km increase in depth corresponds to a 1 Kb increase in pressure, so 8 Kb is reached at about 24 km.

How fault behaviour varies with depth is often discussed in terms of the strength – the maximum difference between the horizontal and vertical stresses that the rock can support. At shallow depths rocks fail either by brittle fracture or by frictional sliding on pre-existing faults, and strength increases with depth. However, at greater depths, rocks deform ductilely, as described by flow laws, which show that strength decreases as the temperature increases. This temperature-dependent behaviour is the reason that the cold lithosphere forms the planet's strong outer layer and that earthquakes occur only to a given depth. These variations are described by strength plots known as strength envelopes.

Figure 12 shows strength envelopes appropriate for old oceanic lithosphere and a stable continental interior. In the frictional region, curves are shown for various values of the ratio of pore pressure to lithostatic pressure, because higher pore pressures result in lower strengths. Ductile flow laws are shown for quartz and olivine, minerals often used as models

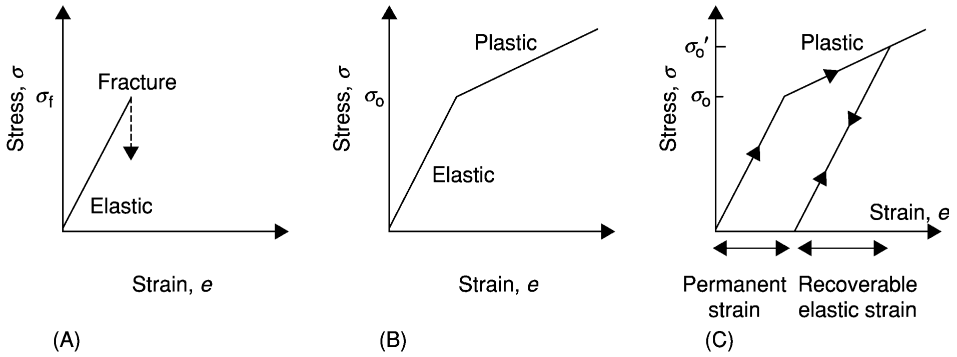


Figure 10 (A) Stress–strain curve for a material that is perfectly elastic until it fractures when the applied stress reaches σ_f. (B) Stress–strain curve for a material that undergoes plastic deformation when the stress exceeds a yield stress, σ_o. (C) Permanent strain results from plastic deformation when the stress is raised to σ_o' and released. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

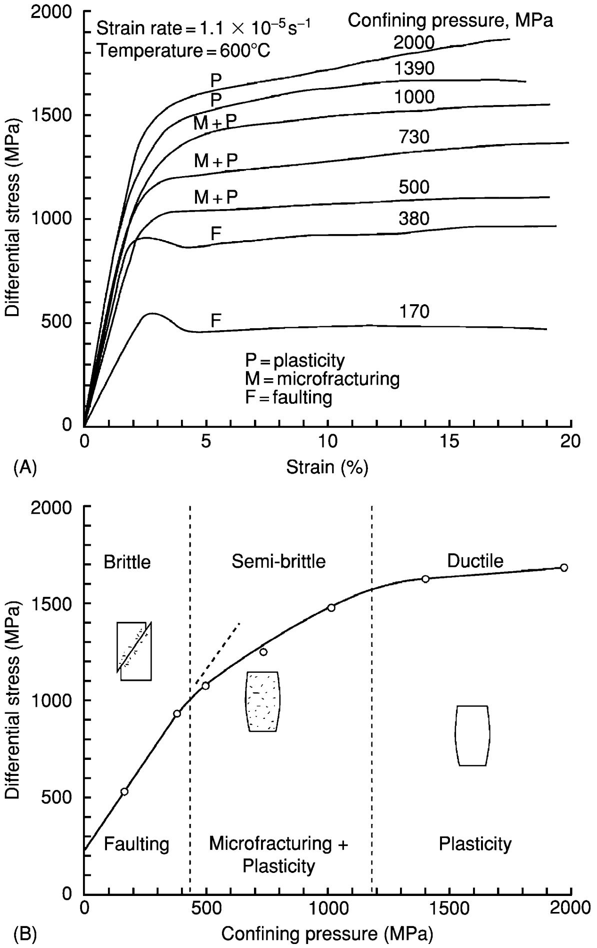


Figure 11 Results of an experiment in which rocks were subjected to a compressive stress greater than the confining pressure. (A) Differential stress–strain curves (c.f. Figure 10) for various confining pressures. (B) At low (less than 4 Kb) confining pressures the material fractures and its strength increases with pressure. At higher pressures the material is ductile and its strength increases only slowly with pressure. A semi-brittle transition regime, in which both microfractures and plasticity occur, separates the brittle and ductile regimes. (Reproduced from Kirby SH (1980) Tectonic stresses in the lithosphere: constraints provided by the experimental deformation of rocks. *Journal of Geophysical Research* 85: 6353–6363.)

for continental and oceanic rheologies, respectively. Strength increases with depth in the brittle region owing to the increasing normal stress, and then decreases with depth in the ductile region owing to increasing temperature. Hence strength is highest at the brittle–ductile transition. Strength decreases rapidly below this transition, so the lithosphere should have little strength at depths greater than about 25 km in the continents and 50 km in the oceans. As a result, the limiting temperature for continental seismicity is lower than for oceanic seismicity.

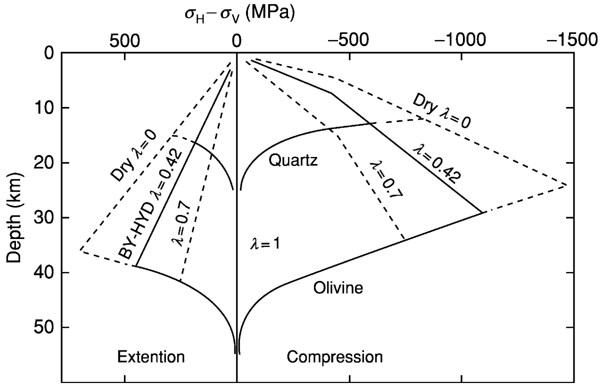


Figure 12 Strength envelopes as a function of depth for various values of λ, the ratio of pore pressure to lithostatic pressure. At shallow depths, strength is controlled by brittle fracture; at greater depths ductile flow laws predict rapid weakening. In the ductile flow regime, quartz is weaker than olivine. In the brittle regime, the lithosphere is stronger in compression (right side) than in extension (left side). (Reproduced from Brace WF and Kohlstedt DL (1980) Limits on lithospheric stress imposed by laboratory experiments. *Journal of Geophysical Research* 85: 6248–6252.)

Faulting and Rock Friction

It is natural to assume that earthquakes occur when tectonic stress exceeds the strength, so steady motion across a plate boundary would give a series of successive earthquakes at regular intervals (Figure 13). However, the time between earthquakes on plate boundaries varies even though the plate motion causing the earthquakes is steady. Some of the variation may be due to the intrinsic randomness of the failure process, such that some small ruptures cascade into large earthquakes whereas others do not. Another cause of the variation may be features of rock friction.

Interesting insight emerges from considering an experiment in which stress is applied to a faulted rock, where motion occurs once the stress reaches a certain level. As stress is reapplied, this pattern of jerky sliding and stress release continues. This stick–slip pattern looks like a laboratory version of earthquakes on a fault: as the fault is loaded by tectonic stress, occasional earthquakes occur. The analogy is strengthened by the fact that at higher temperatures (about 300°C for granite) stick–slip does not occur. Instead, stable sliding occurs, in much the same way as earthquakes do not occur at depths where the temperature exceeds a certain value. Stick–slip results from a familiar phenomenon: it is harder to start an object sliding against friction than to keep it sliding. This is because the static friction stopping sliding exceeds the dynamic friction that opposes motion once sliding starts.

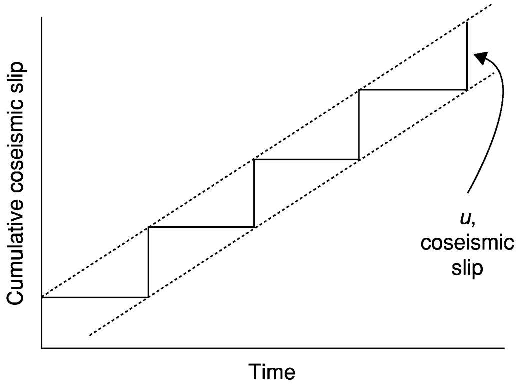


Figure 13 Slip history for an idealized earthquake cycle on a plate boundary, in which all earthquakes have the same coseismic slip. (Reproduced from Shimazaki K and Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophysical Research Letters* 7: 279–282.)

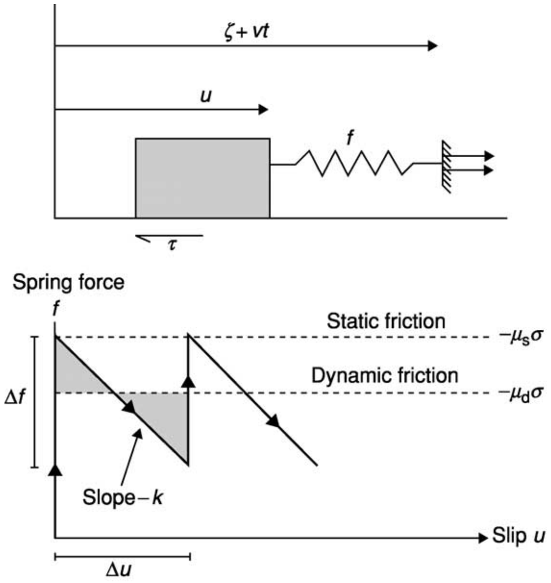


Figure 14 A simple spring and block analogue illustrating stick–slip as a model for earthquakes. (Reproduced from Stein S and Wysession M (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

To gain an insight into stick–slip as a model for earthquakes, consider the experiment illustrated in Figure 14. If an object is pulled across a table with a rubber band, jerky stick–slip motion occurs. This situation can be modelled by assuming that a block is loaded by a spring that applies a force, f, that is proportional to its stiffness, k, and extension. If loading results from the spring's far end moving at velocity v, the spring force increases with time. The block starts sliding once the spring force exceeds the frictional force $\mu_s\sigma$ where μ_s is the static friction coefficient and σ is the normal stress due to the block's weight. Once sliding starts, the friction drops to its dynamic value μ_d, and the driving force decreases as the spring shortens, until it becomes less than the friction force. The block slows and eventually stops once the shaded area above the spring-force line equals that below the line, or when the work done accelerating the block equals that which decelerated it. If the spring end continues to move, loading continues until the spring force again equals the static friction force and another slip event occurs.

Laboratory experiments show that the difference between static and dynamic friction is more complicated than is assumed in this simple model. We can think of the lower dynamic friction as showing either velocity weakening, decreasing as the object moves faster, or slip weakening, decreasing as the object moves further. Frictional models, called rate-dependent friction and state-dependent friction, with a variable coefficient of sliding friction are used to describe these effects. Velocity weakening permits earthquakes to occur by stick–slip, whereas for velocity strengthening stable sliding is expected. Laboratory results show that for granite the transition occurs at about 300°C, which should be the limiting temperature for earthquakes. Thus, the frictional model predicts a maximum depth for continental earthquakes that is similar to that predicted by the rock-strength arguments.

These results can be used to simulate the earthquake cycle. Figure 15 shows the slip history as a function of depth and time for a model in which a strike-slip fault is loaded by plate motion. The fault has rate- and state-dependent frictional properties such that stick–slip occurs above 11 km. From time A to time B, stable sliding occurs at depth and a little precursory slip occurs near the surface. The earthquake causes 2.5 m of sudden slip at shallow depths, as shown by the curves for times B and B′. As a result, the faulted shallow depths 'get ahead' of the material below, loading that material and causing postseismic slip from time B to time F. Once this is finished, the 93-year cycle starts again with steady stable sliding at depth.

Such models replicate many aspects of the earthquake cycle. An interesting difference, however, is that the models predict earthquakes at regular intervals, whereas earthquake histories are quite variable. Some of the variability may be due to the effects of earthquakes on other faults or other segments of the same fault. Figure 16 shows this idea schematically for the block model. Assume that after an earthquake cycle the compressive normal stress is reduced. This 'unclamping' reduces the frictional force resisting sliding, so it takes less time for the spring force to rise to the level needed for the next slip event. Conversely,

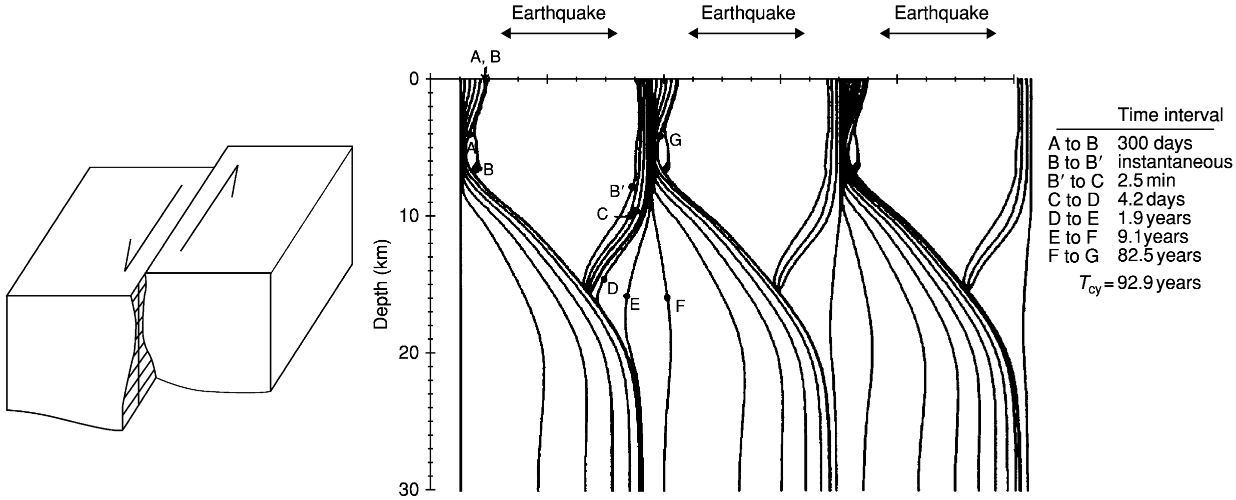


Figure 15 Earthquake cycle for a model in which a strike-slip fault with rate- and state-dependent frictional properties is loaded by plate motion. The slip history for three cycles as a function of depth and time is shown by lines representing specific times. Steady motion occurs at depth, and stick–slip occurs above 11 km. (Reproduced from Tse ST and Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *Journal of Geophysical Research* 91: 9452–9472.)

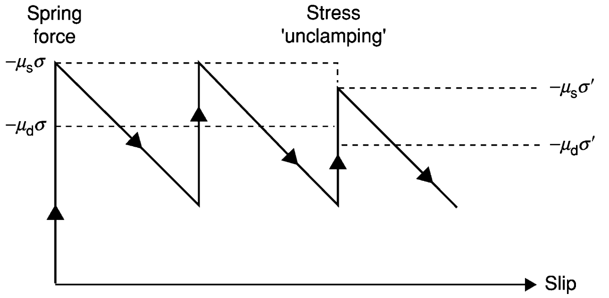


Figure 16 Modification of the block model shown in Figure 14 to include the effects of changes in normal stress. Reduced normal stress reduces the frictional force, 'unclamping' the fault and decreasing the time to the next slip event. (Reproduced from Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.)

increased compression 'clamps' the block more, increasing the time to the next slip event. This analogy implies that earthquake occurrence on a segment of a fault may reflect changes in the stress on the fault resulting from earthquakes elsewhere. Some earthquake observations provide support for this idea.

See Also

Earth: Crust. **Engineering Geology:** Seismology; Natural and Anthropogenic Geohazards. **Plate Tectonics**. **Tectonics:** Earthquakes; Neotectonics.

Further Reading

Aki K and Richards PG (2002) *Quantitative Seismology*. Sausalito: University Science.

Brace WF and Kohlstedt DL (1980) Limits on lithospheric stress imposed by laboratory experiments. *Journal of Geophysical Research* 85: 6248–6252.

Kirby SH (1980) Tectonic stresses in the lithosphere: constraints provided by the experimental deformation of rocks. *Journal of Geophysical Research* 85: 6353–6363.

Kirby SH and Kronenberg AK (1987) Rheology of the lithosphere: selected topics. *Reviews of Geophysics* 25: 1219–1244.

Lay T and Wallace TC (1995) *Modern Global Seismology*. New York: Academic Press.

Moores EM and Twiss RJ (1995) *Tectonics*. New York: W H Freeman.

Scholz CH (1990) *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge University Press.

Shearer PM (1999) *Introduction to Seismology*. Cambridge: Cambridge University Press.

Shimazaki K and Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophysical Research Letters* 7: 279–282.

Stein S and Wiens D (1986) Depth determination for shallow teleseismic earthquakes: methods and results. *Reviews of Geophysics and Space Physics* 24: 806–832.

Stein S and Wysession T (2003) *Introduction to Seismology, Earthquakes, and Earth Structure*. Malden, MA: Blackwell Publishing.

Tse ST and Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. *Journal of Geophysical Research* 91: 9452–9472.

Folding

J W Cosgrove, Imperial College London, London, UK

Introduction

A fold is defined as "…a curved arrangement of originally parallel surfaces …" and a large range of terms have been used to describe geological folds. These including folds, flexures, inflexions, bendings, plications, undulations, and crenulations. Although today most geologists use the term fold for buckle folds and many reserve flexure for layer deflections caused by bending, the terms are also used as synonyms.

Folds form on all scales, from those with a wavelength of a few mm that can only be seen in rock thin sections under the microscope, to folds with wavelengths in excess of 10 km. The mechanisms of formation are independent of size; however, for large-scale folds it is important to take into account the effect of gravity when analysing their folding behaviour.

Folds in rocks form under a wide variety of conditions. For example, folds can occur at all depths in the crust from early, near-surface folding linked to the slumping of water saturated, uncemented sediments down continental slopes, to folding of rocks under the high pressures and temperatures encountered in the lower crust. Folding of rocks occurs by various processes, the three most important being buckling, bending, and flowing. Buckling, which is the most common type of fold, requires the rock to possess a mechanical anisotropy, usually layering, and for the maximum compressive stress to act parallel or subparallel to the layering. In contrast, bending is defined as a transverse deflection of a layer or beam by a transverse couple. Flow folding occurs during the flow of a material such as lava, salt, or ice. No mechanical anisotropy is necessary for this type of folding, only passive marker bands that reveal the flow patterns within the material.

Although folds are structures that are characteristic of ductile deformation, they are often found in association with fractures (*see* **Tectonics:** Fractures (Including Joints)). The fractures result from the same stress field responsible for folding and the resulting fracture patterns, which are controlled by the mechanisms by which the folds form, can play an important role in hosting mineralization and in the storage of fluids such as water and hydrocarbons.

Fold Geometry

The main geometric parameters relating to the three-dimensional geometry of a fold are shown in Figure 1A. However, although folds are three-dimensional structures they are most commonly exposed as two-dimensional sections on joint surfaces. Consequently, their detailed classification is based on the geometry of a section. Because of the dependence of fold geometry on the orientation of the section (Figure 1B), a particular section is used to determine the geometry, namely the profile section, i.e., the section at right angles to the fold hinge. Historically, the problem of describing layer shape developed around two geometrical models, the parallel fold and the similar fold (Figure 2A and B). As the name implies, in parallel folds the orthogonal thickness of the layer remains constant. In similar folds, the layer thickness, parallel to the hinge surface remains constant. Although similar folds show considerable variations in layer thickness, this type of folding can produce folds which can extend indefinitely in the profile section, whereas with parallel folds this is not possible (Figure 2A and B).

As can be seen from Figure 2C, these two fold types are examples from a complete spectrum of possible geometries. The spectrum has been divided into five classes based on the pattern of dip isogons, i.e., lines

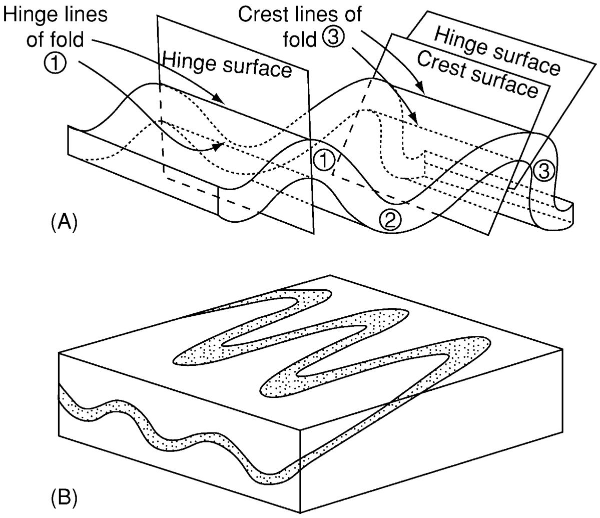


Figure 1 (A) Geometric features of a fold. Fold 1 is symmetric and folds 2 and 3 asymmetric. (B) Diagram showing the dependence of fold outcrop pattern on the orientation of the plane of exposure.

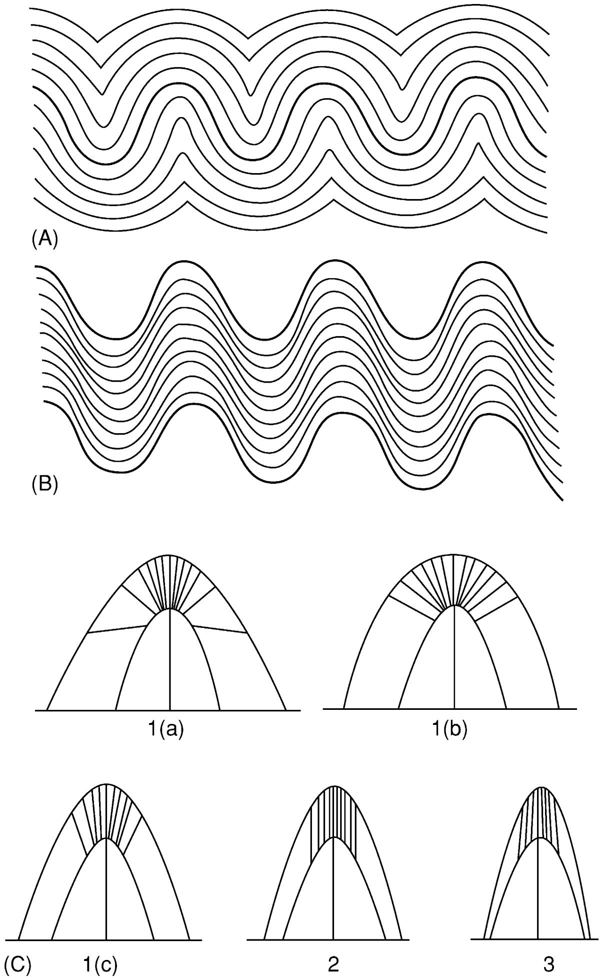


Figure 2 (A) parallel and (B) similar folds. (After Van Hise (1894).) (C) Classification of fold profiles using dip isogon patterns. 1(a) strongly convergent, 1(b) parallel, 1(c) weakly convergent, (2) similar, and (3) divergent (Ramsay (1967)).

joining points of equal dip on the two surfaces defining the folded layer. The parallel folds and similar folds are class 1b and 2, respectively.

Mechanisms of Folding

The two most important mechanisms by which folds form in rocks are buckling, the result of compression parallel or sub-parallel to the rock layering and bending, the result of compression at a high angle to the layering, (Figure 3A and B, respectively).

Buckle Folds

The most commonly formed folds in the Earth's crust are buckle folds. In order for these folds to form, the rock must possess a mechanical anisotropy. This is generally a planar mechanical anisotropy because a variety of geological processes give rise to such materials. Rocks can be intrinsically anisotropic because of the process of their formation such as, for example, a bedded sedimentary succession, or the anisotropy can be induced as a result of subsequent deformation and metamorphism during which time they can develop planar and linear mineral fabrics as they are converted to slates, phyllites, schists, and gneisses. Buckling systems can be sub-divided into four groups, namely folds formed by: (i) the buckling of a single interface, (Figure 3C); (ii) the buckling of two interfaces, which define a single layer in a matrix (Figure 3D); (iii) the buckling of several layers (Figure 3E); and (iv) the buckling of a mineral fabric such as a slate or schist (Figure 3F).

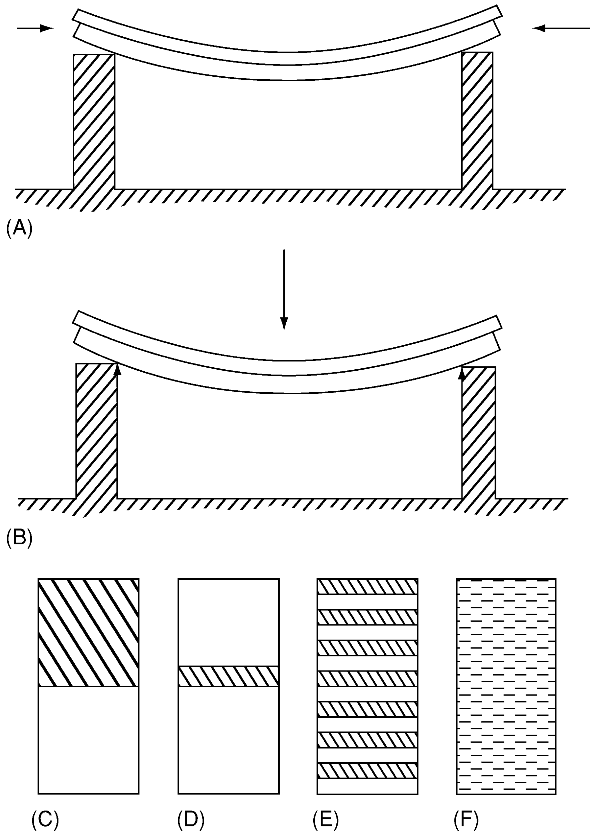


Figure 3 The orientation of the principal compression for (A) buckling and (B) bending of the layers. (C) an interface between two unlike materials; (D) a single layer; (E) a multilayer; and (F) a mechanically anisotropic material. The buckling behaviour of these systems is discussed in the text.

Interface Buckling

Interface buckling occurs on many scales, from large buckles that form at the boundary between the Mesozoic cover rocks and older basement in the French Alps as a result of the collision of the African and

European plates (Figure 4A), to small-scale examples directly observable in the field (Figure 4B) and formed in analogue models in the laboratory (Figure 4C). The folds start as symmetric sinusoidal deflections but as they amplify change their geometry into the marked cusp geometry seen in Figure 4. The cusps always point into the stronger of the two materials.

Single Layer Buckling

Single layers occur commonly in nature, for example, an isolated sandstone or limestone bed in a thick shale or marl sequence or a sheet of igneous rock intruded into an unlayered matrix (Figure 5A), and they are commonly observed to buckle. Theoretical analyses

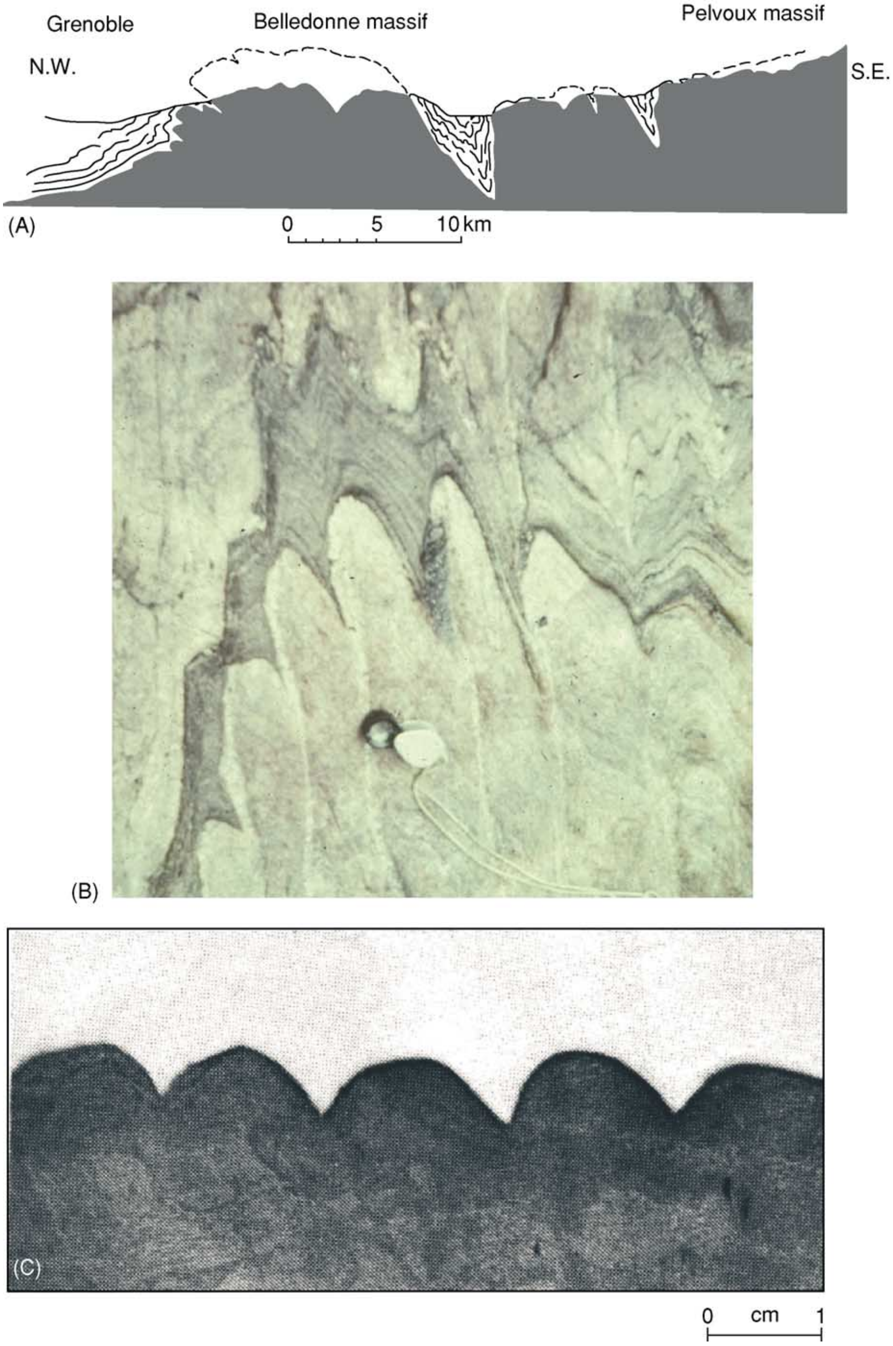


Figure 4 (A) The cuspate interface between the strong Hercynian basement rocks (black) and the weaker Mesozoic cover rocks, caused by the horizontal compression linked to the collision of the African and European plates. (B) Cusp structures formed at the interface between quartz rich (light) and mica rich (dark) bands, Loch Monar, Scotland. (C) Cusps generated by compression parallel to the interface between strong (black) and weak (light) plasticine.

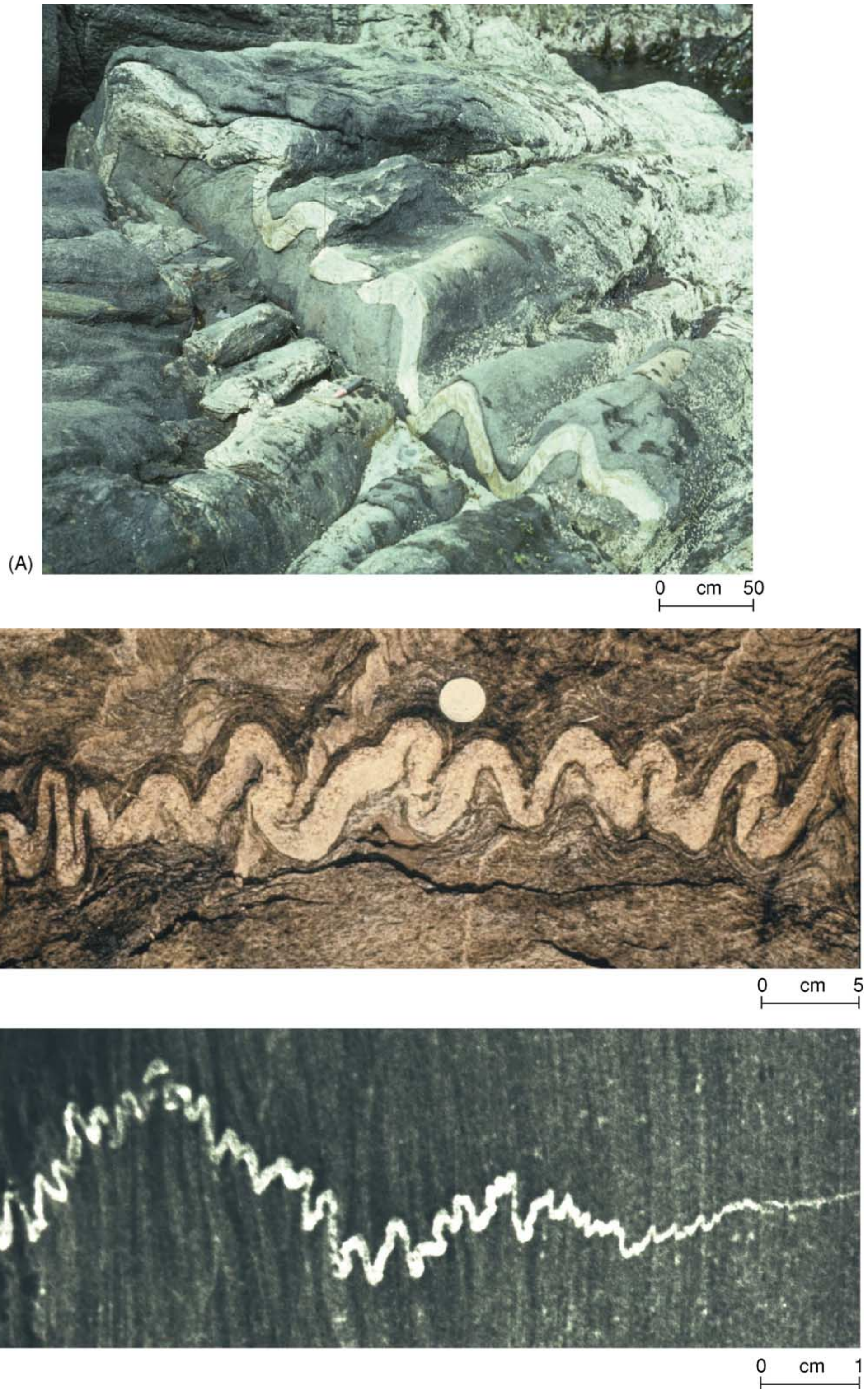


Figure 5 (A) A folded aplite dyke cutting a dolerite body, Outer Hebrides, Scotland. (B) A folded single layer, in which the folds are isoclinal. No further shortening of the layer can occur by limb rotation. Further compression may result in homogeneous flattening of the folds or in the folding of the layer into larger wavelength folds (see C). (C) A tapered quartz vein showing buckling on two scales. On both scales, the wavelength is seen to decrease with decrease in layer thickness.

predict and field observations confirm that there is a direct relationship between the wavelength (W) of folds that develop and the thickness (t) of the buckling layer (Figure 5C).

$$W = 2\pi t(\mu_1/6\mu_2)^{-1/3} \qquad [1]$$

The single layer buckling equation shows that in addition, the wavelength is also controlled by the ratio of the strength of the layer and matrix (μ_1/μ_2). The impact of the strength ratio (competence contrast) on fold style can be seen by rearranging eqn 1.

$$W/t = 2\pi(\mu_1/6\mu_2)^{-1/3} \qquad [2]$$

The wavelength/thickness ratio is determined by the relative strength of the layer and matrix. Folds with a range of W/t ratios are shown in Figure 6, which shows the control of the competence contrast between the layer and the matrix on the geometry of the buckles that form.

Once buckling has been initiated, layer shortening can continue by rotation of the fold limbs until the limbs become parallel to each other (i.e., the folds are isoclinal, Figure 5C), and no further shortening by limb rotation can occur. Shortening may continue by homogeneous flattening of the buckled layer and matrix, or alternatively continued compression may cause the buckled layer to buckle again. It has an effective thickness that is considerably greater than the original thickness of the layer and so, in accordance with the buckling equation eqn 1, will buckle with a larger wavelength (Figure 5C).

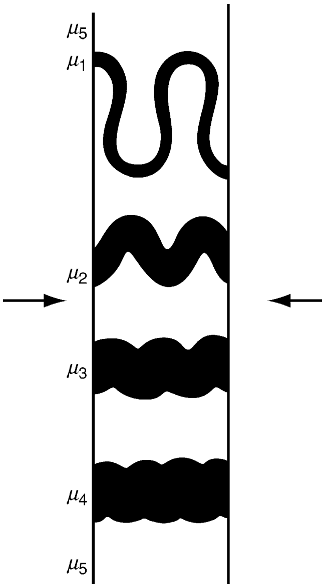


Figure 6 The effect of the ratio of the strength of the layer and the matrix on fold style. As the contrast decreases the wavelength/thickness ratio also decreases. (After Ramsay (1982).)

Multilayer Buckling

The most common type of geological layering is that of a multilayer, i.e., a succession of different layers. Often the multilayer is made up of the regular alternation between two or three rock types. The regularity of these multilayers reflects the processes by which they were formed. For example, turbidites are formed as a result of the periodic, fluid-induced collapse of a sedimentary accumulation on the edge of the continental shelf. The fluidised sediments flow as dense turbidity currents down the continental slope, depositing first the heavy sands followed by the slower settling out of the fine shale particles. In this way a sequence of alternating sandstones and shales can be built up to form a turbidite.

The buckling behaviour of multilayers is controlled by the spacing of the strong units it contains. Figure 7A and B are both physical multilayers. However, their buckling behaviours are very different. The multilayer shown in Figure 7A behaves mechanically as a series of single layers, each layers folding

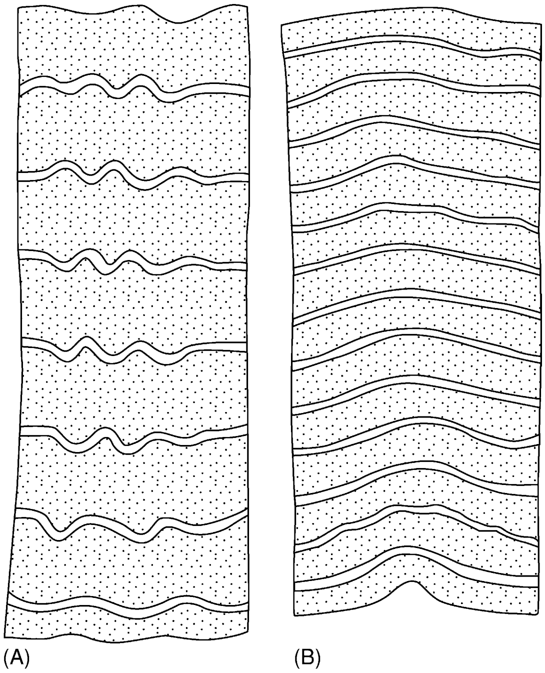


Figure 7 (A) and (B) show two compressed multilayers of alternating competent (white) and incompetent (stippled) rubber layers. The widely spaced layers (A) behave mechanically as single layers and form disharmonic folds; the more closely spaced layers behave mechanically as a multilayer and all layers form the same wavelength. (After Ramberg (1963).)

according to the single layer buckling equation, (eqn 1). This results in 'disharmonic folding'. In contrast, the multilayer shown in Figure 7B behaves mechanically as a multilayer, i.e., all the layers develop the same wavelength and amplitude. Such folding is called 'harmonic folding'.

These two types of buckling behaviour can be readily explained by considering the strain that develops in the matrix around a single, competent (i.e., strong with respect to the surrounding matrix) layer as it buckles. The zone of disturbance on each side of the buckling layer is known as the zone of contact strain. If the competent layers of a multilayer are sufficiently far apart for there to be no significant overlap of their zones of contact strain, then each layer will buckle as a single layer. If, however, the zones of contact strain of adjacent competent layers do significantly overlap, the layers can no longer buckle independently of each other. The zones of contact strain and associated zones of contact stress of adjacent layers must be compatible and, as a result, all the layers are subjected to the same stress field and develop the same wavelength.

In order to determine how close the competent layers of a multilayer must be for multilayer as opposed to single layer buckling to occur, it is necessary to know how far the zone of contact strain extends away from the layer into the matrix. For a viscous matrix it is found that the disturbance has died down to approximately 1% of its maximum value at a distance one wavelength from the layer (Figure 8).

The Buckling of Anisotropic Materials

The most complex buckling occurs in materials possessing a pervasive mechanical anisotropy. The anisotropy may be an intrinsic property of the rock resulting, for example, from the bedding parallel alignment of clay particles in a shale or induced during metamorphism when, for example, a slate or schist is formed from a mudstone. Theoretical studies and experimental work on such materials shows that there is a range of structures that can form when they are compressed parallel to the mineral fabric or layering. The two end members of this range are upright folds and box folds (Figure 9). The type of structure that forms is determined by the mechanical anisotropy of the material. As the anisotropy increases so the upright folds (Figure 9A), give way to folds with gently diverging axial planes (Figure 9B) and finally to box folds (Figure 9C).

It should be noted that if a geological multilayer has a sufficiently high mechanical anisotropy, it will buckle to form a box fold. An example of this is shown in Figure 10A, where a box fold with a wavelength of several 100 m has developed in Carboniferous turbidites from south-west England.

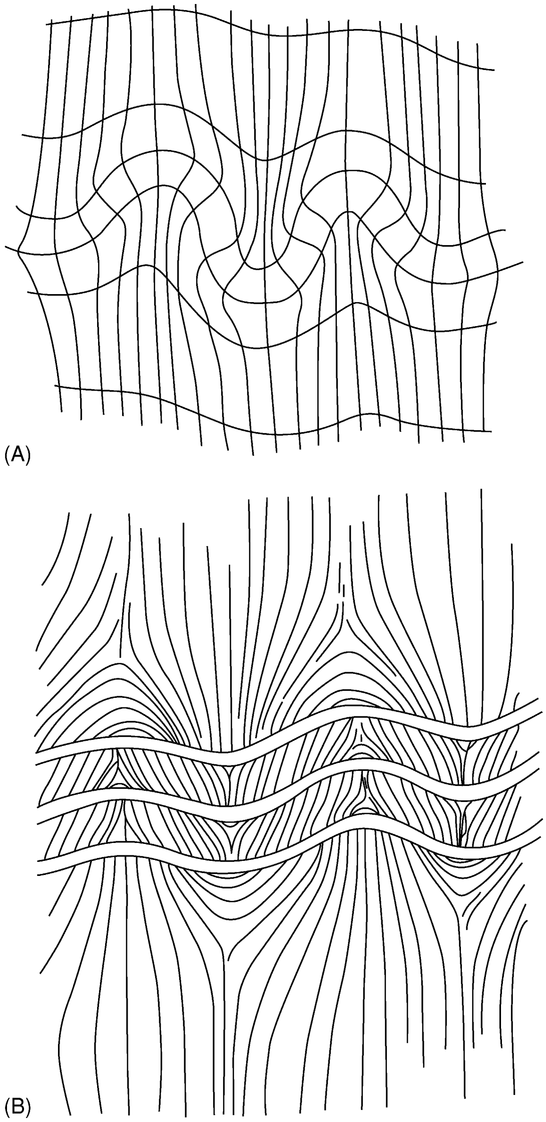


Figure 8 (A) Experimentally produced buckles in viscous materials showing the disturbance of the matrix (the zone of contact strain) by the buckling single layer (B) Folded multilayer where the proximity of the layers caused an overlap of the zones of contact stain and therefore the formation of multilayer buckles.

Experimental Work on Folds

One of the disadvantages of theoretical studies of folding is that they are often only valid for the first increment of buckling. Once buckling has been initiated, the assumptions of the analysis are violated and the theory cannot be used to predict the way in which the fold amplifies into a finite fold. In contrast, experimental work on models made from rock analogue materials such as gelatine, are ideal for the study of the amplification of folds. In addition, unlike

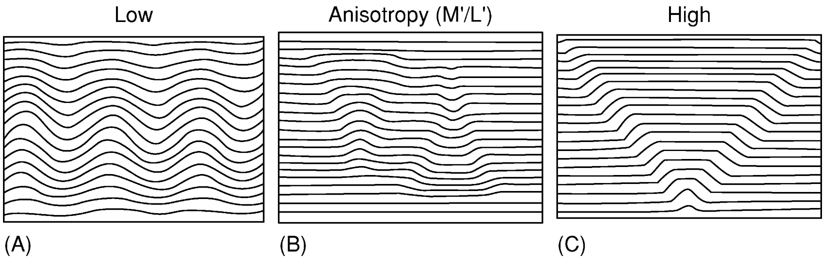


Figure 9 The structures that can form when a mechanically anisotropic material such as a mineral fabric of sedimentary layering is compressed parallel to the layering. Depending upon the anisotropy they range from upright folds with axial planes normal to the principal compression to box folds where the axial planes are inclined to the compression.

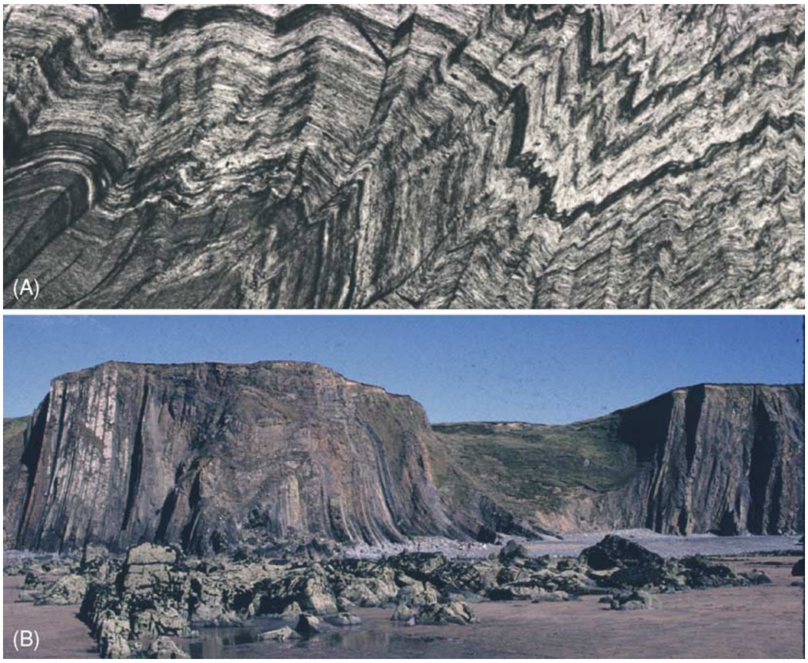


Figure 10 (A) Small-scale box fold formed in a micaceous fabric and (B) large-scale box fold formed by the buckling of a turbidite.

the theoretical treatments of folding which assume that the layered system is stressed instantaneously (i.e., that the process of stressing the material plays no part in the buckling), the experiments show that the process of stressing the layered system can influence the process of folding. Such experiments show that folds that form during the compression of the models do not generally develop synchronously. They form in a serial manner, either one after the other where the amplification of one fold stimulates the initiation and amplification of another next to it (Figure 11A), or one after the other at random positions within the model. Figure 11B shows such randomly positioned folds in the Jura mountains of Switzerland.

In the complex multilayers that occur in nature, it is often found that all types of buckling occur in the same multilayer. For example, in the multilayer shown in Figure 12, which is a finely laminated evaporite, examples of single layer buckling can be seen (layers 1 and 2) where the wavelength is determined by the thickness of the layers. The two layers are far enough apart for them to fold independently and they have done so producing an example of disharmonic folding. In area 3 of the multilayer, examples of multilayer folding where the folds have axial planes oriented at right angles to the layering can be seen (Figure 7B and 9A) and in area 4, box folds have formed with axial planes inclined to the layering (Figure 9C, 10A and B).

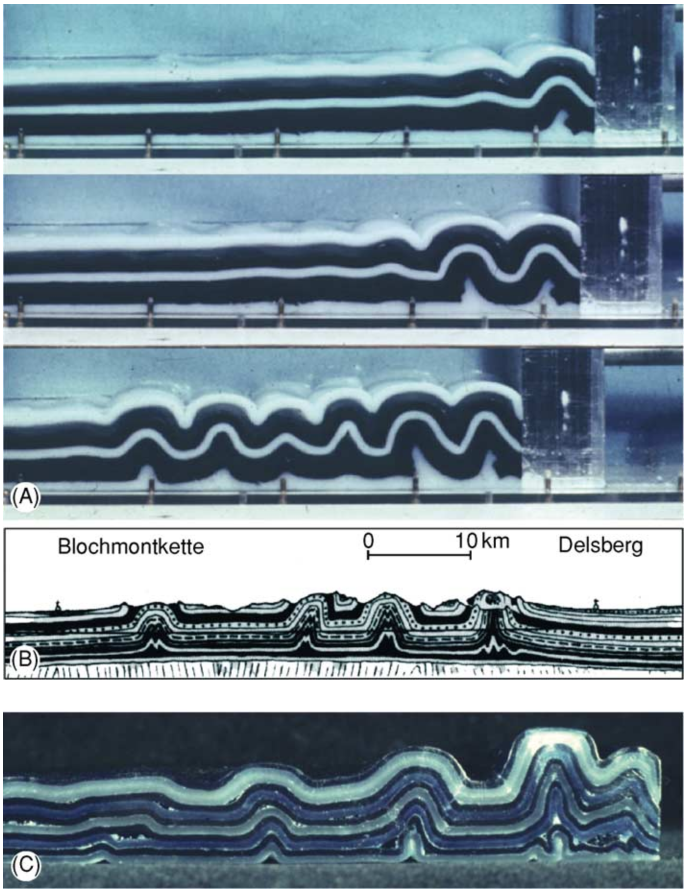


Figure 11 (A) The serial development of folds in a lubricated gelatine multilayer caused by a piston moving from right to left. (B) Isolated folds in the Jura mountains Switzerland. (C) Isolated folds in a lubricated wax multilayer.

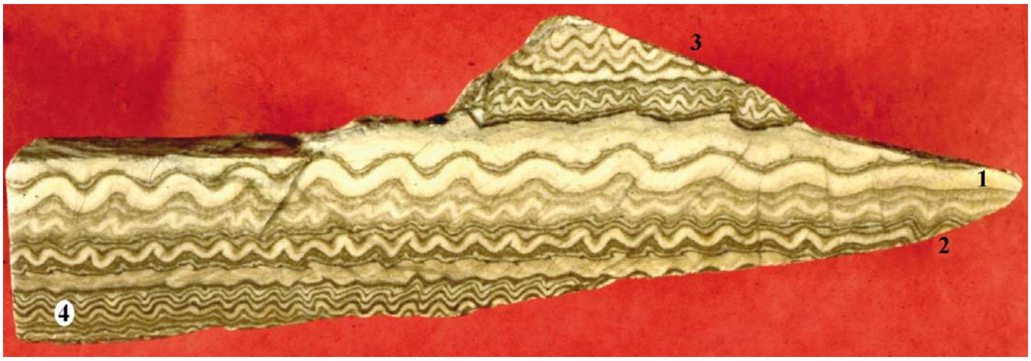


Figure 12 Harmonic and disharmonic folding in a specimen of the Castile and Todilto evaporites, New Mexico, USA (c.f. layers 1 and 2). Some of the folding (layers 1 and 2), can be described by the theory of single layer buckling, some by the theory of multilayer buckling, (3) and some (4) by the theory of buckling of an anisotropic material.

Three-Dimensional Geometry of Buckle Folds

The three-dimensional geometry and spatial organization of buckle folds has been studied primarily by field observations and analogue modelling. These studies show that buckle folds have a periclinal geometry, i.e., have the form of an elongate dome, basin, or saddle (Figure 20A). The geometry of a pericline is often given in terms of the ratio of its half wavelength and its hinge length. This is termed the aspect ratio and although it will increase as the fold amplifies, it is found that the majority of buckle folds in the upper crust have ratios of between 1:5 and 1:10. Typical geometries of geological folds are shown in Figure 13.

Bending

Bending is the term used to describe the flexuring of a layer induced by a compression acting at a high angle to the layering (Figure 1B). Geological flexures that are the result of bending are known as drape folds or forced folds and are frequently formed when sediments, which cover a more rigid basement, flex in response to components of vertical movement along basement faults (Figure 14). This may be normal movement, in which case the flexing of the layering involves layer parallel extension (Figure 14A), or reverse movement, in which case the folds that result will involve an element of buckling (Figure 14D).

A forced fold is defined as a fold in which the final overall shape and trend are dominated by the shape of some forcing member below. These are frequently fault blocks, movements of which produce linear fault scarps which in turn produce linear forced folds with an aspect ratio much higher that that associated with buckle folds and with different spatial organizations.

The two types of basement faults linked to the formation of forced folds are dip-slip faults (either normal or reverse Figure 14A and D respectively). These faults produce fault scarps at the basement–cover interface over which the forced folds form. When the third type of faulting, i.e., strike-slip or wrench faulting occurs in the basement, no fault scarp is produced and consequently no forced folds are produced in the overlying strata. However, movement on basement strike-slip faults can give rise to

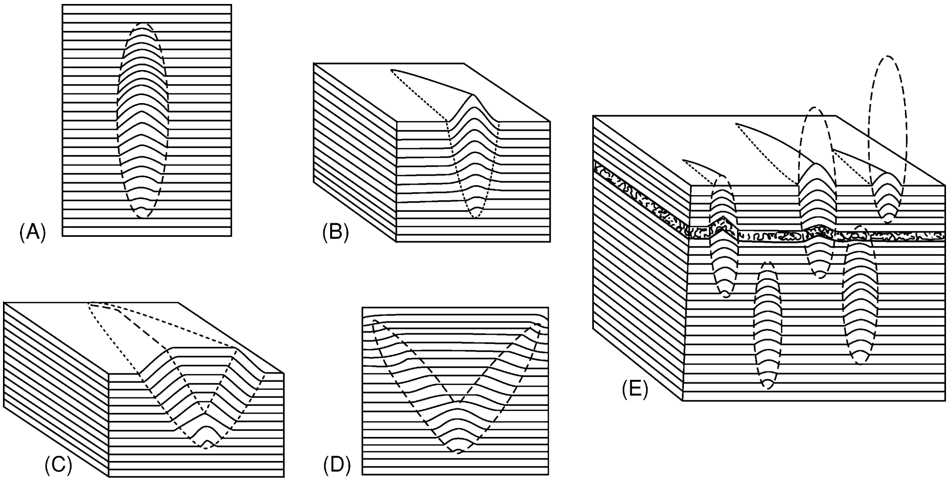


Figure 13 (A) Typical profile geometry of a fold in a multilayer. (B) Block diagram showing a fold dying out in both profile and plan view. (C) Block diagram and (D) profile geometry of a box fold. (E) The spatial organization of folds within a multilayer.

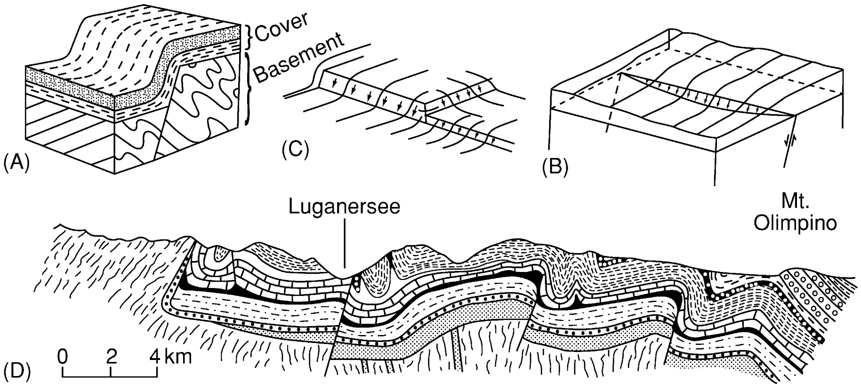


Figure 14 (A) and (B) block diagrams of drape or forced folds, the result of normal faulting in the basement. (C) shows the type of fold geometry which may be associated with block faulting in the basement. (D) Forced folds probably formed over reverse faults in the basement.

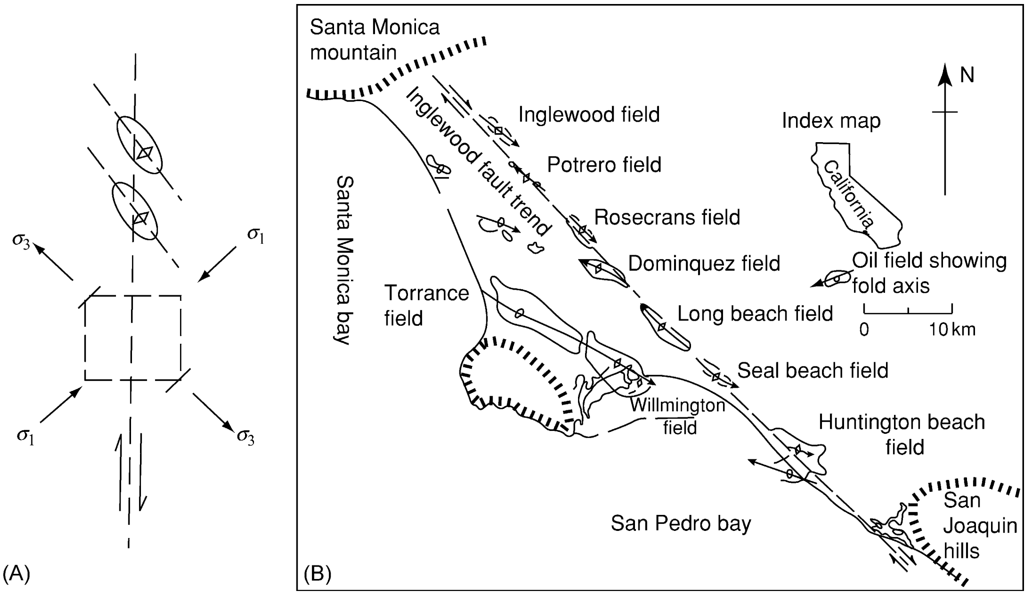


Figure 15 (A) Movement along a wrench fault in the basement produces horizontal compression in the cover rocks and the formation of a linear train of offset folds above the fault. (B) Such en echelon folds above the Inglewood fault, California.

folding in the overlying cover rocks. The stress generated in the cover rocks above a basement strikeslip fault is shown in Figure 15A. A local horizontal compression, inclined at 45° to the basement fault, is developed in the cover and this can give rise to a variety of structures, depending on the rheological condition of the cover rocks. If they behave in a ductile manner, then a series of folds may develop. They will form with their axial planes at right angles to the local maximum compression (σ_1) and will be arranged in an offset manner along the trace of the basement fault (Figure 15A). Natural examples of these linear arrays of buckle folds (see, for example, Figure 15B, which shows folding induced in the cover rocks above the San Andreas wrench fault) are excellent markers for locating major hidden faults.

Fault-Bend Folds

The forced folds shown in Figure 14 occur in cover rocks which respond in a ductile manner in response to movement on faults in a relatively rigid basement. Other well-documented relationships between faults and folds are known, one of the most familiar being fault-bend folding (Figure 16). In this type of forced folding, the folding is not the result of the movement of rigid fault blocks in the basement but rather the result of fault movement within the cover rocks. As can be seen from Figure 16, faults are not perfectly planar surfaces of slip. They generally have gentle undulations and may display substantial curvature or sharp bends. For example, the thrust fault shown in Figure 16 is made up of two horizontal portions (flats) linked by an inclined portion (ramp). As the two fault blocks slip past one another there must be deformation in at least one of them because rocks are not strong enough to support large voids. For this reason, many major folds within layered rock exist within the hanging-wall fault blocks, formed by bending the fault blocks as they slip over non-planar fault surfaces. This mechanism of folding is termed Fault-bend folding.

Flow Folding

In addition to the mechanisms of buckling and bending, folds can also be produced by flow. Impressive examples occur during the outpouring of lavas and during the slower flow of ice and salt. Salt is less dense than most rocks and therefore tends to rise diapirically though overlying strata. The resulting salt domes often emerge at the Earth's surface (Figure 17) where they can form salt glaciers (Namakiers). Figure 18 shows large flat lying flow folds formed as a result of flow within a salt glacier.

Implications of Folds Regarding the Properties of the Rock

The Earth's crust is characterized by an upper seismically active zone where brittle failure dominates, resulting in the formation of fractures and loss of continuity of the rocks, and a lower aseismic zone associated with ductile deformation, i.e., deformation that

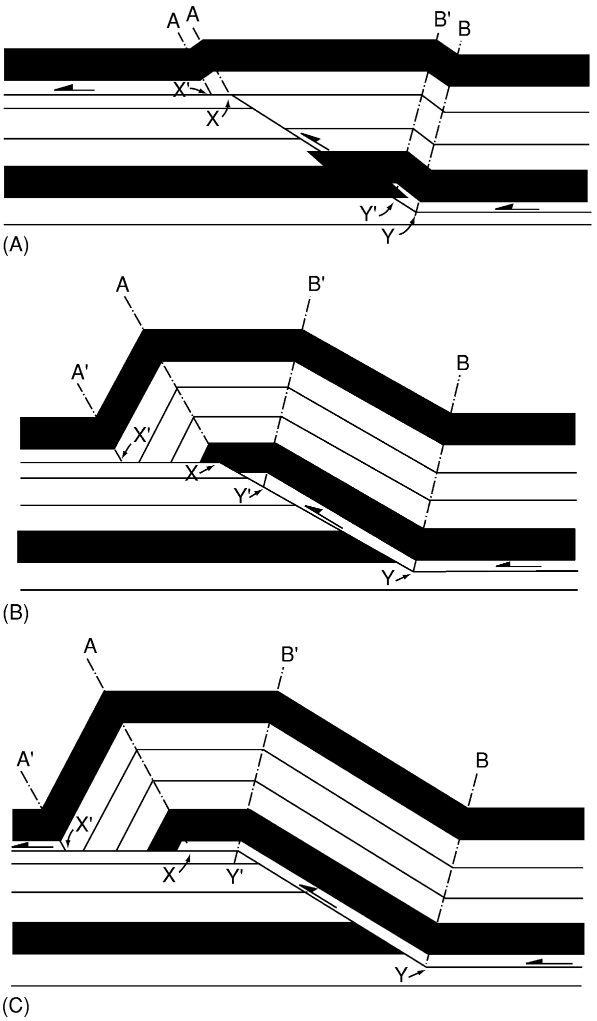


Figure 16 The development of a fault bend fold as a thrust sheet rides over a ramp in the detachment horizon. (After Suppe (1983).)

occurs without the loss of continuity of the material. Folding tends to occur within this ductile zone although, as noted earlier, folds can form at all depths in the crust. In addition, fractures can form in association with folding and the orientation of the fractures indicates clearly that they are formed by the same stress field that operated during folding (Figure 19).

Strain Within a Folded Layer and Associated Fracturing

The strain distribution (and therefore the fracture pattern) within a folded layer is dependant on the layer properties. In a homogenous isotropic layer, such as a uniform sandstone or limestone bed, the strain distribution is likely to be similar to that shown in Figure 20B, in which a layer parallel extensional field associated with the outer arc is separated from a layer parallel compression field associated with the inner arc by a neutral surface along which there is no strain. This model of strain distribution is termed Tangential Longitudinal strain folding. In contrast, a homogeneous anisotropic layer, such as a well-bedded shale, may fold by bedding parallel slip which results in the strain distribution shown in Figure 20C. This is known as flexural flow folding if the shear strain parallel to the layer boundary is uniformly distributed across the layer, and flexural slip folding if it is concentrated along distinct bedding planes. It is interesting to note that both models of folding (Figure 20B and C) produce parallel folds, i.e., folds with a constant orthogonal thickness. This illustrates the fact that the strain state within a fold cannot be deduced from the geometry of its profile section. However, the different strain patterns within the two models reflect the fact that they have very different stress fields within them which may lead to the formation of characteristic fracture patterns which enable the two fold types to be recognized in the field. For example, the extensional fractures in the outer arc of the pericline shown in Figure 20A and the shear fractures in the inner arc, indicate outer arc extension and inner arc contraction, respectively, i.e., a pattern compatible with the deformation associated with the Tangential Longitudinal Strain fold (Figure 20B).

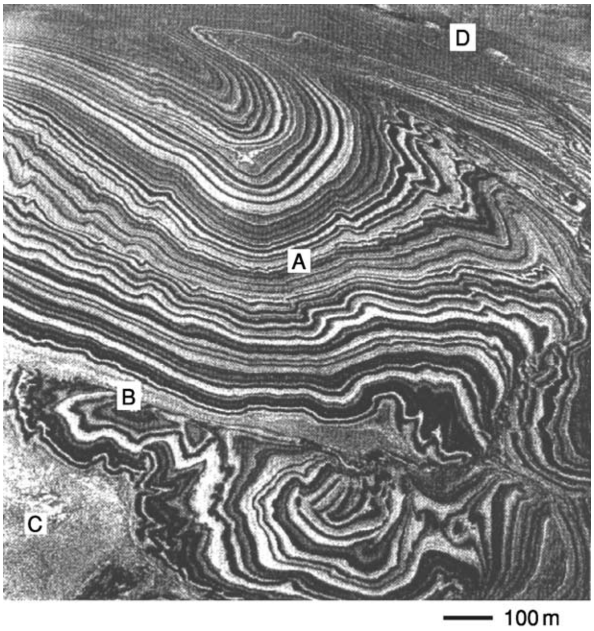


Figure 17 Oblique aerial view of a salt dome from the Great Kavir, Iran, showing the agate-like regularity of bedding in the younger salt (A) and a fault contact (B) with the more massive older salt (C).

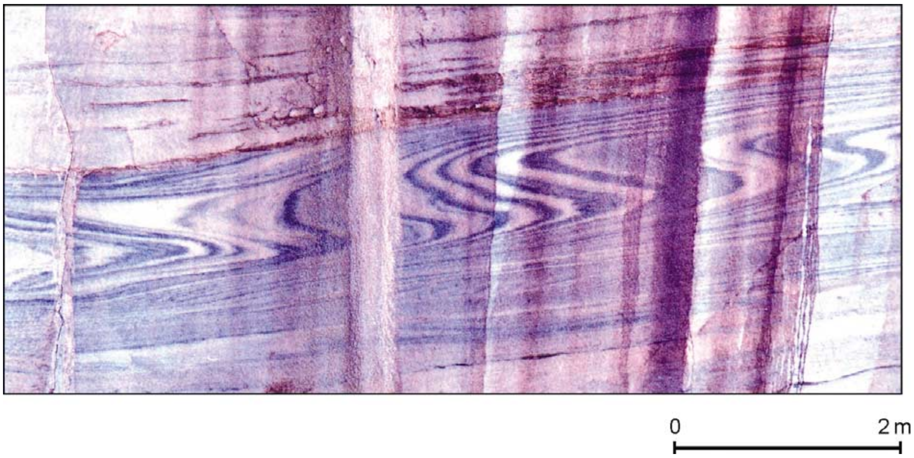


Figure 18 Large flat-lying flow folds formed as a result of flow within a salt glacier, Iran.

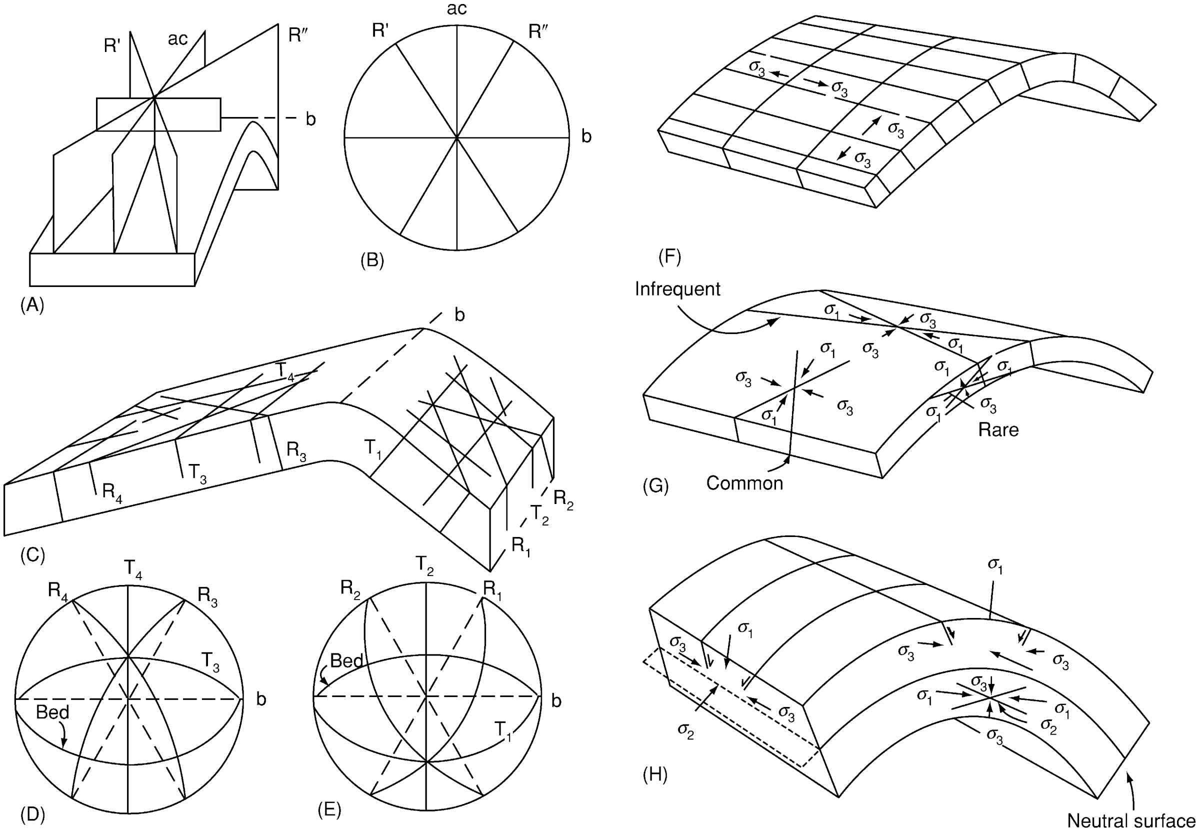


Figure 19 (A) Ideal relationship of master joints to a relatively small fold. (B) Stereographic plot of fractures shown in (A). (C) Trend of minor fractures in a folded competent unit. (D) and (E) Stereographic plots of fractures in the two limbs. R and T are shear and extension fractures, respectively. (F) Typical relationship of extensional fractures to a fold. The orientation of the least principal stress associated with each set (which are of different ages) is shown. (G) Typical orientation of shear fractures in a thin bedded layer, with associated stress systems. (H) Typical orientation of normal faults and thrusts which may develop in a thick flexed unit.

Thus, fractures formed in association with buckle folds may be the result of the local stresses generated as a result of buckling. Alternatively, they may be caused by the regional stress field. The existence of a buckle is likely to disturb the regional stress field. For example, folds are often accompanied by bedding plane slip, implying that the bedding planes cannot sustain a large shear stress. It follows, therefore, that

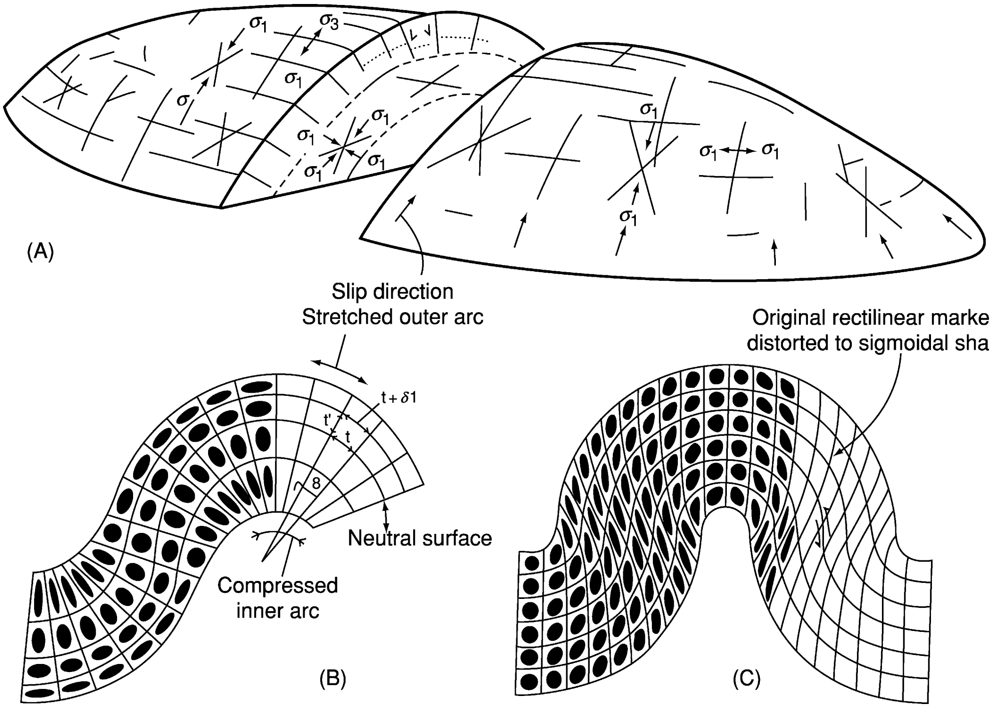


Figure 20 (A) Various fractures associated with a pericline. (After Stearns (1978).) (B) Strain distribution in a tangential-longitudinal-strain fold and (C) a flexural flow fold. (After Ramsay (1967).)

the principal stresses must be constrained to being either subparallel or subnormal to the folding layers. As a result of this stress deflection, the fractures also form normal to bedding. This is illustrated in Figure 19, which shows the predicted orientation of the shear and extensional fractures that would form in response to the regional compression generating the fold (Figure 19A), together with their projection on a stereographic plot (Figure 19B), and the frequently observed orientation of these fractures on the limbs of the fold which occurs as a result of the principal compressive stress following the layering (Figure 19C–E). The types and orientations of fracture found in association with buckle folds, which form as a result of both the regional and local stresses, are summarized in Figure 19F–H.

See Also

Tectonics: Fractures (Including Joints).

Further Reading

Cosgrove JW and Ameen MS (1999) A comparison of the geometry, spatial organization and fracture patterns associated with forced folds and buckle folds. In: Cosgrove JW and Ameen MS (eds.) *Forced Folds and Fractures*, pp. 7–21. Bath: Geological Society of London, Special Publication No. 169.

Jackson MPA, Cornelius RR, Graig CH, Stocklin J, and Talbot CJ (1990) Salt diapirs of the Great Kavir, Central Iran. *Geological Society of America Memoir 177.* pp. 139.

Price NJ and Cosgrove JW (1990) *Analysis of Geological Structures*. Cambridge: Cambridge University Press.

Ramberg H (1963) Fluid dynamics of viscous buckling applicable to folding of layered rocks. *American Association of Petrology & Geology Bulletin* 47: 485–505.

Ramsay JG (1967) *Folding and Fracturing of Rocks*. New York: McGraw Hill.

Ramsay JG and Huber MI (1987) *The Techniques of Modern Structural Geology: Volume 2: Folds and fracture*. London: Academic Press.

Stearns DW (1978) Faulting and Forced folding in the Rocky Mountain Foreland. *Geological Society of America*. Memoir 151: 1–38.

Suppe J (1983) Geometry and Kinematics of fault-bend folding. *American Journal of Science* 283: 684–721.

Van Hise CR (1894) Principals of North American pre-Cambrian geology. *US Geological Survey*, 16th annual report. pp. 571–843.

Weiss LE (1959) Structural analysis of the basement system at Turoka, Kenya. London: *Overseas Geology and Mineral Resources* 7: 3–35, 123–153.

Fractures (Including Joints)

J W Cosgrove, Imperial College London, London, UK

Introduction

Fractures are the result of brittle failure which is the general term given to failure during which continuity of the material is lost. Deformation that does not involve loss of continuity is termed ductile. Two modes of brittle failure have been recognized, namely Shear failure and Tensile failure and these can be distinguished from each other on the basis of: (i) the orientation of the fractures with respect to the principal stresses that caused them; and (ii) the relative motion of the rock on each side of the fractures (Figure 1).

If the rock moves parallel to the fracture (Figure 1A) the fracture is a shear fracture and if it only moves normal to the fracture (Figure 1B) it is a tensile fracture.

Shear fractures in rocks are called faults and tensile fractures joints.

Mechanism of Formation of Fractures

The current understanding of brittle failure in rocks is the result of a combination of field observation, experimental work, and theoretical study. Field observations show the two types of fractures, and experimental work reveals that shear fracture occurs when the principal stresses are all compressive, and tensile failure when the least principal stress is tensile and equal to or greater than the tensile strength of the rock.

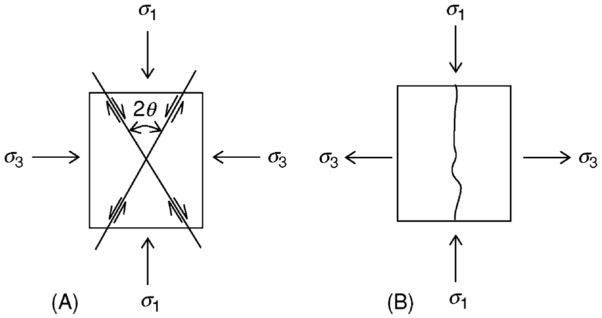


Figure 1 The two modes of brittle failure (A) Shear failure and (B) Tensile failure. They can be distinguished from each other on the basis of: (i) the orientation of the fractures with respect to the principal stresses that caused them; and (ii) the relative motion of the rock on each side of the fracture.

Shear Failure

Experiments have been performed in which cylindrical rock samples, surrounded by a hollow jacket into which a fluid can be injected to provide a confining stress (σ_3), are subjected to an axial loading (σ_1) until they fail. This results in a body of data recording the principal stress necessary for the rock to fail under a wide range of confining stress. These data can be represented graphically in two ways. The first is to plot the axial load against the confining stress (Figure 2A). For many rocks this plot produces a straight line whose intersection with the axial stress axis gives the uniaxial strength, i.e., the stress a rock can sustain with no confining stress. The graph

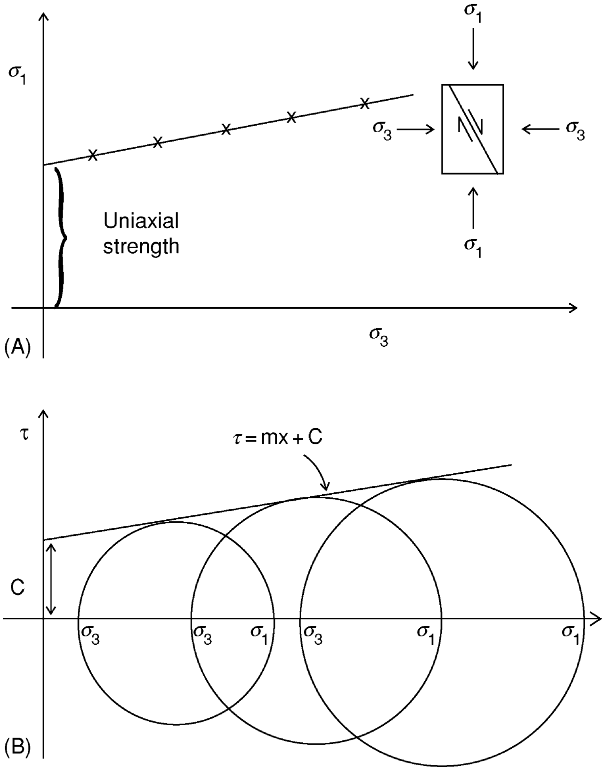


Figure 2 The graphical expression of the experimental data on shear failure. These data can be represented graphically in two ways. (A) is a plot the axial load against the confining stress at failure, (B) shows the same data plotted as a series of Mohr stress circles. The tangent to these circles represents the failure envelope for shear failure.

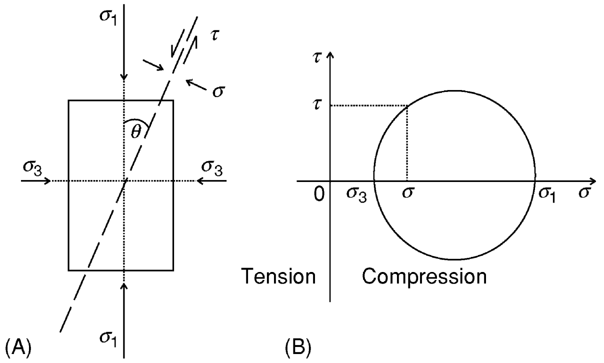


Figure 3 (A) The state of stress (σ and τ) on a plane inclined at θ° to the maximum principal compression, is given by the biaxial stress equations. (eqns [1] and [2]). The graph of σ against τ for values of θ between 0° and 180°; (B) defines a circle, the Mohr stress circle, which defines the state of stress.

indicates clearly that the strength of a rock (i.e., its ability to sustain a load without permanent deformation) is not a fixed value but depends (amongst other things) on the confining stress. The higher the confining stress, σ_3, the greater the axial load needed to cause failure. A rock's strength will, therefore, increase with increasing depth in the crust.

The second method of plotting the experimental data is to plot the stress state for each experiment (i.e., a stress state that caused the rock to fail) as a Mohr circle. The state of stress on any plane inclined at θ° to the maximum principal compression (Figure 3) is given by the biaxial stress equations:

$$\sigma = \sigma_1 \cos^2\theta + \sigma_3 \sin^2\theta \qquad [1]$$

$$\tau = (\sigma_1 - \sigma_3) \cos^2\theta \sin^2\theta \qquad [2]$$

These equations can be represented graphically by calculating σ and τ for values of θ between 0° and 180° and plotting the results on a graph of σ against τ. The resulting points define a circle, the Mohr stress circle (Figure 3B), which defines the state of stress. The diameter of the circle, which is a measure of the differential stress, $(\sigma_1 - \sigma_3)$, is determined by the values of the principal stresses, σ_1 and σ_3. As the experimental data represented in Figure 2A consists of values of the principal stresses that caused the rock to fail, these data can be plotted as a series of Mohr circles (Figure 2B). The tangent to these circles represents the failure criterion for shear failure.

For many rocks, this tangent is a straight line whose equation is:

$$\tau = m\sigma + C \qquad [3]$$

where m is the slope of the line and C its intersection with the shear stress axis.

Having established the shear failure criterion experimentally, it is important to compare it with the theoretically derived criterion. This was developed independently by Navier and Colomb, who argued that in order for a shear fracture to develop, the shear stress τ acting along the potential fracture plane (Figure 3A) must be sufficiently large to overcome the cohesion along that plane, C_0, plus the resistance to shear along the plane once it had formed. The resistance to slip is given by Amonton's law of frictional sliding which states:

$$\tau = \mu\sigma \qquad [4]$$

where τ and σ are the shear and normal stresses, respectively acting on the fracture plane and μ the coefficient of sliding friction. μ is defined as the tangent of the angle of sliding friction φ. Hence the complete criterion can be expressed in the form:

$$\tau = \sigma \tan\varphi + C_0 \qquad [5]$$

This is known as the Navier–Colomb criterion of shear failure and is identical to the criterion established experimentally (eqn [3]). The orientation of the planes where this condition is first met can be determined by substituting the biaxial stress equations (eqns [1] and [2]) into the shear failure criterion (eqn [5]) and solving for the minimum. The optimum orientations for the shear fractures are:

$$\theta = + \text{ or } - [45° - \varphi/2] \qquad [6]$$

where θ is the angle between the maximum compressive stress (σ_1) and the shear fracture (Figure 1A). Note that two fracture orientations are predicted, inclined at 45° $-\varphi/2$ each side of σ_1. They are termed conjugate shear planes and although the magnitude of the shear stress along them is the same, the sense of shear is different. Faults are the geological expression of shear failure, and conjugate small-scale faults in a sequence of alternating sandstones and shales are shown in Figure 4.

Classification of Faults

As can be seen from Figure 1A, the orientation of a fault is controlled by the orientation of the principal stresses that generate them. Field observations reveal that faults fall into three classes, normal faults, wrench (or strike slip) faults, and thrust (or reverse) faults which, as can be seen from Figure 5, correspond to stress states where σ_1, σ_2, and σ_3 are vertical.

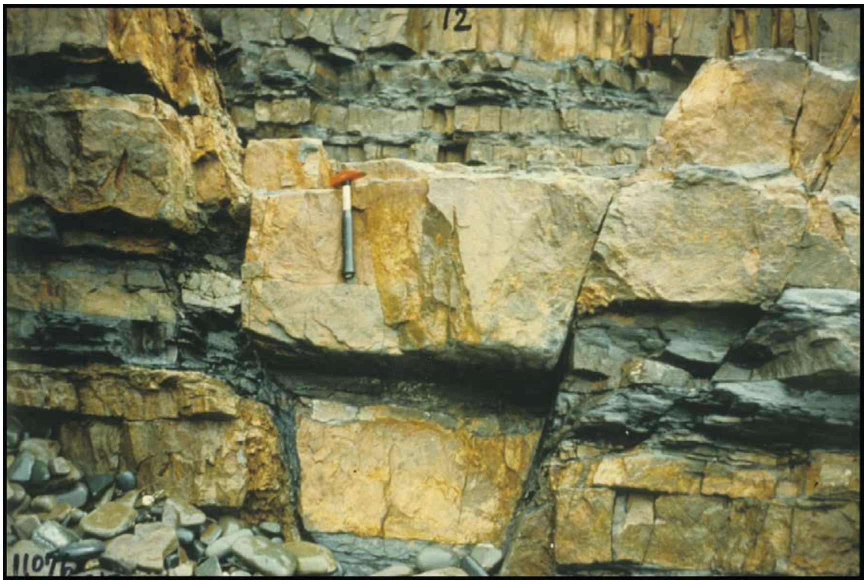


Figure 4 Conjugate normal faults in Carboniferous turbidites from Bude, Cornwall, England.

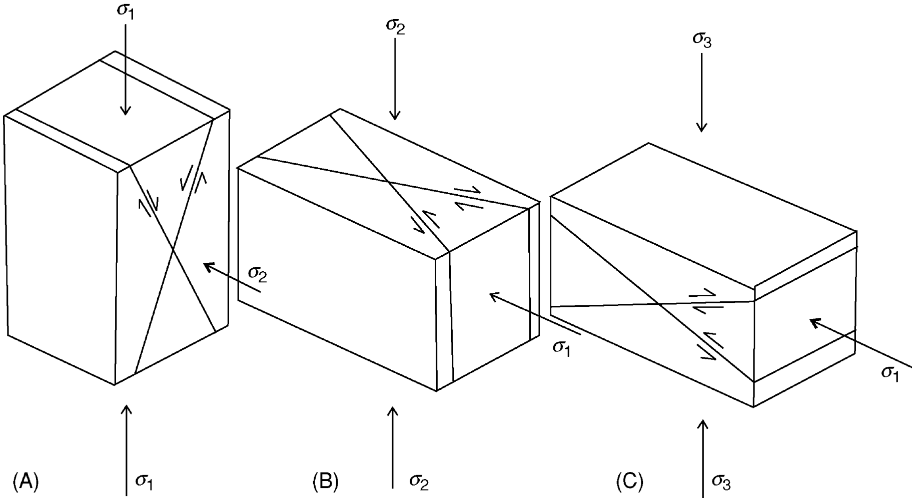


Figure 5 The orientation of shear fractures that form when (A) σ_1, (B) σ_2 and (C) σ_3 are vertical. The resulting faults are termed, normal faults, wrench (or strike slip) faults, and thrust (or reverse) faults, respectively.

This observation implies that the principal stresses tend to be oriented in one of these three orientations. In 1951, Anderson argued that this was because the Earth's surface is a free surface which cannot sustain a shear stress. Thus, in order not to generate a shear stress parallel to this surface, the principal stresses are constrained to remain either parallel or normal to it. As is discussed later, the three classes of faults characterize three different tectonic regimes.

Tensile Failure

Joints and veins are the most common geological expression of tensile failure. Experiments show that this type of failure generates fractures normal to σ_3 (Figure 1B). The theory of tensile failure was developed by Griffith (1925) who argued that in an ideal material the tensile strength of a material would be determined by the strength of the inter-atomic bonds. However, experiments revealed that the measured tensile strength of a material is usually several orders of magnitude lower than that calculated on the basis of their inter-atomic bond strength. Griffith argued that this was because the materials contain small flaws or microfractures and that these resulted in a local stress magnification at the crack tips. This is illustrated diagrammatically in Figure 6, which shows the stress configuration of a plate subjected to a horizontal extension. In the first plate (Figure 6A), the stress is evenly distributed through the plate. In the second plate, which contains a flaw represented by an elliptical crack (Figure 6B), the tensile stress is locally magnified at the fracture tips. The amount of magnification depends primarily on the orientation and eccentricity of the crack. The greater the eccentricity, the greater the magnification. Griffith argued that by this means, a relatively low applied stress could be locally amplified at fracture tips within the material to the point where it reached the stress required to break the atomic bonds within the material causing the material to fail. He developed the following failure criterion (the Griffith criterion of tensile failure) based on this model:

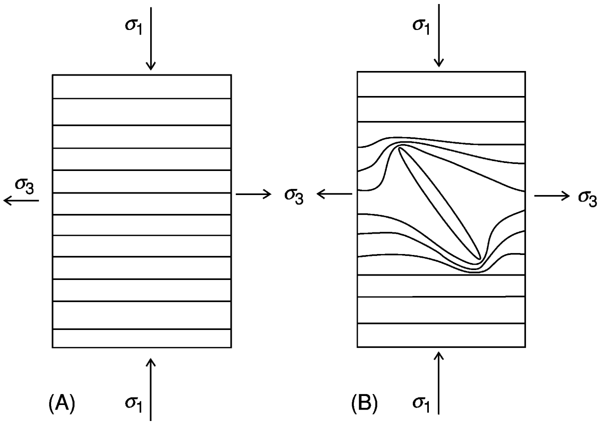


Figure 6 (A) A uniform stress field represented by uniformly spaced stress trajectories in a stretched layer. (B) The concentration of tensile stress at crack tips in a uniformly extended layer.

$$\tau^2 + 4T\sigma - 4T^2 = 0 \qquad [7]$$

where τ is the shear stress, σ the normal stress, and T the tensile strength of the material. The graphical expression of this failure criterion is shown in Figure 7A. It has the form of a parabola whose intersection with the normal stress axis gives the tensile strength and with the shear stress axis the cohesion.

The complete criteria for brittle failure, (the Griffith, Navier–Colomb criteria), is obtained by linking the two criterion (eqns [5] and [7]) at the point where their slopes are identical (Figure 7B). Any stress state can be represented on this graph as

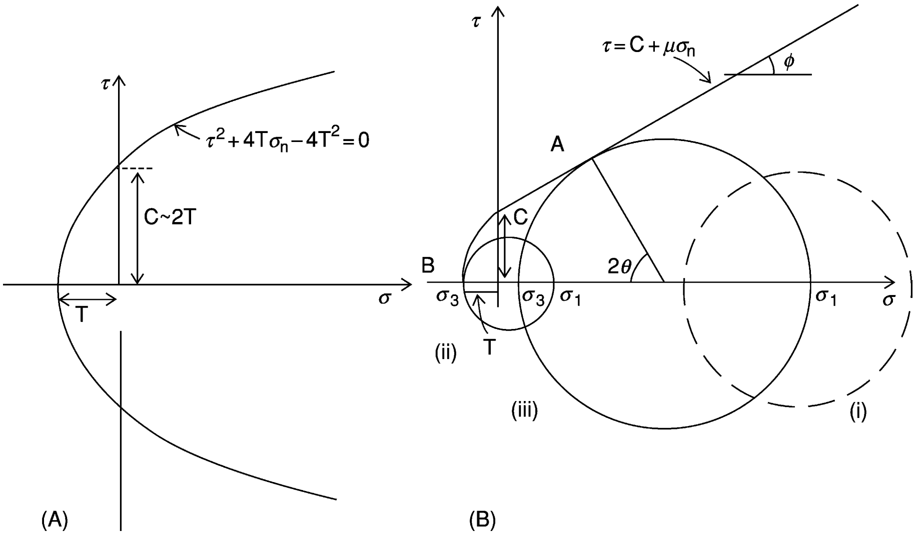


Figure 7 (A) The graphical expression of the Griffith criterion of tensile failure (eqn [7]). It has the form of a parabola whose intersection with the normal stress axis gives the tensile strength of the material and with the shear stress axis the cohesion. (B) The complete criteria for brittle failure, (the Griffith, Navier-Colomb criteria), is obtained by linking the two criteria (eqns [5] and [7]) at the point where their slopes are identical.

a Mohr circle whose position is determined by the values of the principal stresses (σ_1 and σ_3). If a stress state, when plotted on the graph, does not touch or intersect the failure envelope, the stress state is stable, i.e., will not cause the rock to fail, e.g., stress field (i) Figure 7B. If, however, it does touch the envelope, failure will occur, either by tensile fracturing if the contact is with the tensile part of the envelope (Figure 7B (ii)), or shear fracturing if it is with the shear part (Figure 7B (iii)).

What Determines Whether Tensile or Shear Fractures Form?

It can be seen from Figure 7B that shear failure is associated with a large differential stress ($\sigma_1 - \sigma_3$) i.e., the Mohr's stress circle must be large in order to intersect the shear failure envelope, and that tensile failure is associated with a low differential stress, i.e., the Mohr's stress circle must be small in order to intersect the tensile failure envelope. The precise conditions necessary for the formation of the two types of fractures are:

$$\text{For tensile failure to occur} \quad (\sigma_1 - \sigma_3) < 4\text{T} \qquad [8a]$$

$$\text{For shear failure to occur} \quad (\sigma_1 - \sigma_3) > 4\text{T} \qquad [8b]$$

where T is the tensile strength of the material.

The geometrical relationships between the principal stresses and the fractures they produce (i.e., a conjugate set of shear fractures symmetrically about σ_1 and a single set of tensile fractures at right angles to σ_3) is shown in Figure 1 and, as noted below, the understanding of these relationships provides a powerful tool in fracture analysis.

It follows therefore that the orientation of the fractures that form in response to a stress field is determined by the orientation of the principal stresses (Figure 1), and the type of fracture (shear or tensile) by the magnitude of the differential stress.

The Effect of a Fluid Pressure on Fracturing

Fluid-Induced Failure

The state of stress in the crust is dominantly compressional. For example, in a nontectonic environment, the stress at any depth is generated by the overburden which produces a compressive vertical stress which induces a compressive horizontal stress. Thus, at any depth the Mohr stress circle will plot in the compressive regime in Figure 7B and there will be no possibility of tensile failure. Geologists were, therefore, perplexed to find that large numbers of tensional fractures occur in the crust. This paradox was resolved when the importance of fluid pressures within a rock was understood. Pore fluid pressure within a rock increases as the rock is buried. (*see* **Tectonics:** Hydrothermal Activity). The stress state within the pores is hydrostatic and the pressure acts so as to appose the lithostatic stress caused by the overburden. This effect can be shown diagrammatically by representing the lithostatic stress as an ellipse with the stress acting compressively and the fluid pressure as a circle with the pressure acting outwards (Figure 8A). The fluid pressure reduces all the lithostatic stresses by an amount P_{fluid} to give an effective stress. Thus, the principal stresses σ_1 and σ_3 become ($\sigma_1 - \text{P}_{\text{fluid}}$) and ($\sigma_3 - \text{P}_{\text{fluid}}$). This new stress field can be plotted as a Mohr stress circle (Figure 8B). It can be seen that the original lithostatic stress circle is moved towards the tensile regime but that the diameter of the circle, i.e., the differential stress, remains unchanged.

The amount of migration of the stress circle is determined by the magnitude of the fluid pressure. Thus, as the fluid pressure gradually increases during burial, the stress circle is pushed inexorably towards the failure envelope. When it hits the envelope, failure occurs. Such failure is termed fluid induced or hydraulic fracturing. In this way an originally compressional stress regime can be changed so that one or more of the principal stresses becomes effectively tensile and the conditions for tensile failure can be satisfied.

The Expression of Fluid-Induced Failure

In the example shown in Figure 8B the lithostatic stress had a small differential stress (i.e., less that 4T (see eqn [8]) and as a result the induced hydraulic fractures were tensile fractures. If it had been greater than 4T, shear fractures would have formed.

The Organization of Tensile Fractures

The Mohr circles shown on Figure 9 all intersect the failure envelope in the tensile regime, i.e., the differential stresses are all less than 4T and will all therefore result in tensile failure. Their differential stresses vary from just less than 4T (circle (i) Figure 9), to zero (circle (iv), Figure 9). Note that when the stress state is hydrostatic, the Mohr circle is reduced to a point.

As noted above, tensile fractures form normal to the minimum principal compressive stress σ_3 (Figure 1B), i.e., they open against the minimum compressive stress. The stress state represented by Mohr circle (i) in Figure 9, has a relatively large differential stress and there is, therefore, a definite direction of easy opening for the fractures. The fractures would

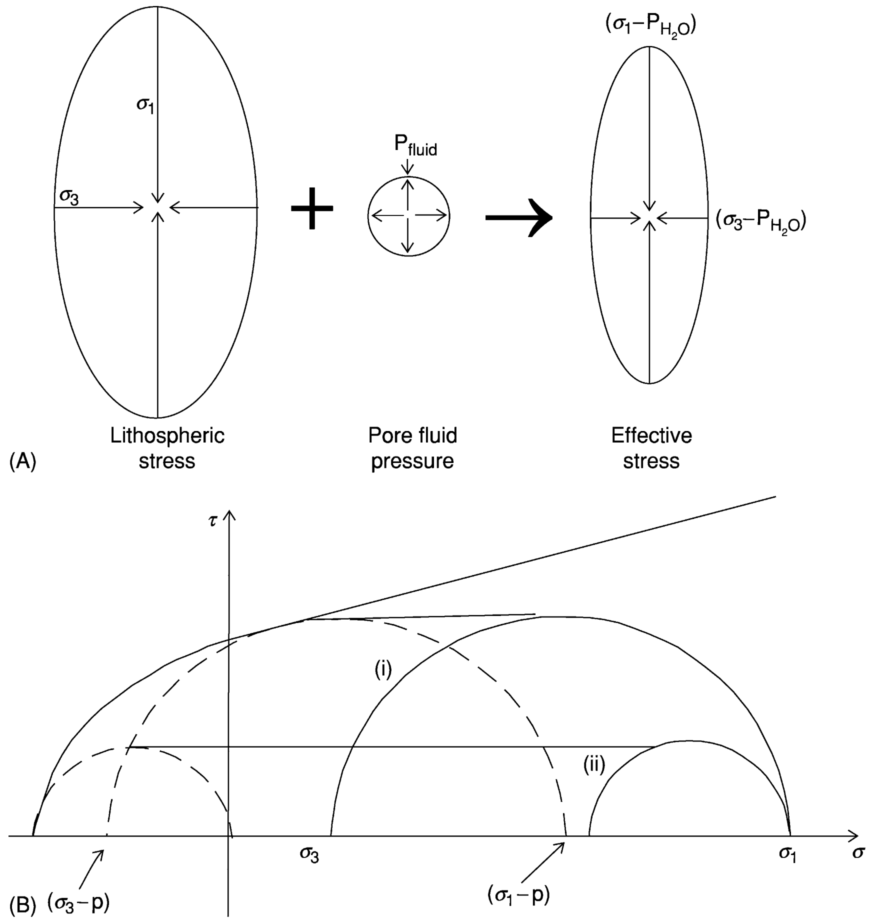


Figure 8 (A) Diagramatic representation of the effect of a fluid pressure (the circle with the outwardly acting stress) on the stress state in a rock (the ellipse with the inwardly acting stresses). All normal stresses are reduced to an effective stress ($\sigma - p_{fluid}$) but the differential stress ($\sigma_1 - \sigma_3$) remains unchanged. The effect is to cause the Mohr stress circle to move to the left by an amount equal to the fluid pressure; (B). Thus depending on the magnitude of the differential stress the induced fractures will be either shear (stress state (i)) or tensile (stress state (B)) (*see* **Tectonics:** Folding).

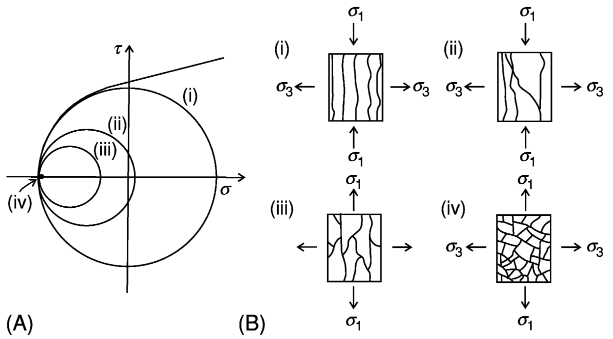


Figure 9 (A) Mohr stress circles (i)–(iv) representing a range of stress states, all of which will lead to tensile failure. NB the Mohr circle (iv), that represents hydrostatic stress is a point. (B) Patterns of tensile failure generated by the corresponding stress states shown in (A).

therefore exhibit a marked alignment normal to this direction (Figure 9B (i)). However, for the stress states represented by the Mohr circles (ii–iv), the differential stress becomes progressively smaller until, for the hydrostatic stress represented by circle (iv), the differential stress is zero. In a hydrostatic stress field the normal stress across all planes is the same and there is, therefore, no direction of relatively easy opening for the fractures. Thus, they will show no preferred orientation and, if they are sufficiently closely spaced and well developed, will produce a brecciation of the rock (Figure 9B (iv)). It can be seen that as the differential stress becomes progressively lower so the tendency for the resulting tensile fractures to form a regular array normal to σ_3 decreases. Tensile fracture systems ranging from well-aligned fractures to randomly

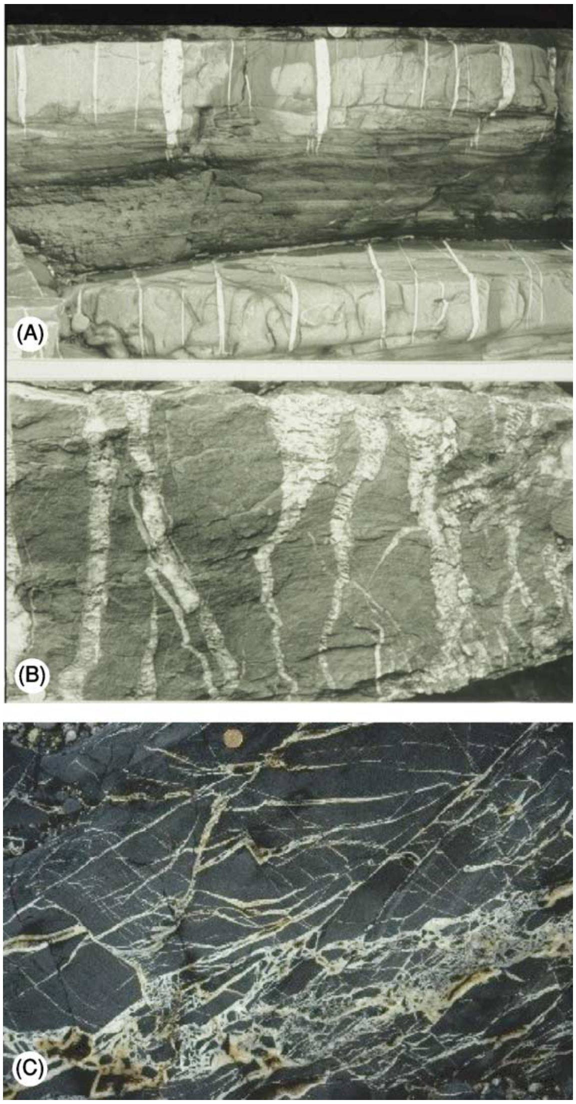


Figure 10 (A) A regular array of tensile fractures exposed on a bedding plane in Carboniferous sandstone, Millook, Cornwall, England. (B) Less well organized tensile fractures cutting Devonian sandstones, St. Anne's Head, South Wales. (C) Carboniferous sandstone cut by randomly oriented tension fractures.

oriented fractures are to be expected in rocks, and field observations support this idea, Figure 10.

As noted above, the problem of forming tensile fractures in the compressive stress field that generally characterizes the Earth's crust can be solved by appealing to high fluid pressures. However, tensile failure can occur in rocks without the aid of a high internal fluid pressure, for example, during the contraction of a layer as a result of desiccation of a sediment (Figure 11) or the cooling of an igneous body (Figure 12).

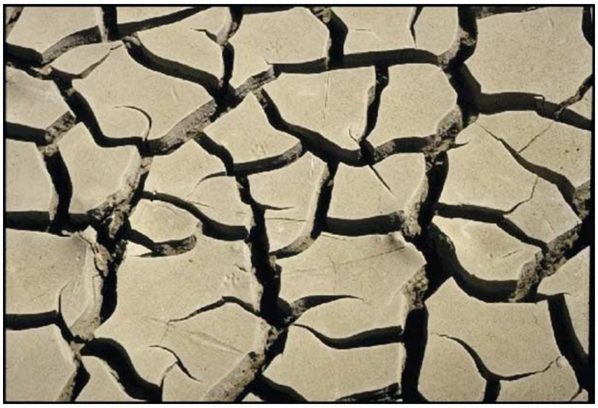


Figure 11 Polygonal arrays of tensile fractures cause by the desiccation of a silt layer.

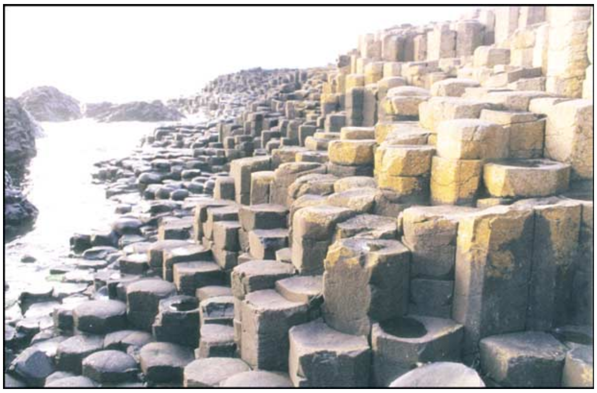


Figure 12 Polygonal arrays of tensile fractures cause by the cooling of a lava flow. (Giant's Causeway Northern Ireland).

In both these examples, the fractures are organized into polygonal arrays showing that the tensile stresses generated were the same in all directions.

Fracture Sets

Generally, the state of stress in the Earth's crust is not hydrostatic. Consequently a single episode of deformation is likely to generate a set of fractures with the same orientation. However, most rocks experience several different stress regimes during their history with the result that several fracture sets are frequently found superimposed on each other to produce a fracture network (Figure 13). The interaction of late fractures with early fractures is illustrated in Figure 14.

The effect of early fractures on later ones is to arrest their propagation and to modify their orientation. It can be seen from Figures 14A and B that one fracture often ends abruptly against another. This abutting relationship gives the relative age of the fractures, i.e., the later fracture abuts against the earlier fracture. If an early fracture is an open fracture then it will represent a free surface within the rock and, as discussed in the above section on classification of faults, will be unable to support a shear stress. Consequently, the principal stresses will reorient as they approach it into a position either normal or parallel to the fracture. This effect can be clearly seen in Figure 14, where later fractures curve into an orientation at right angles to the earlier fracture as they approach it.

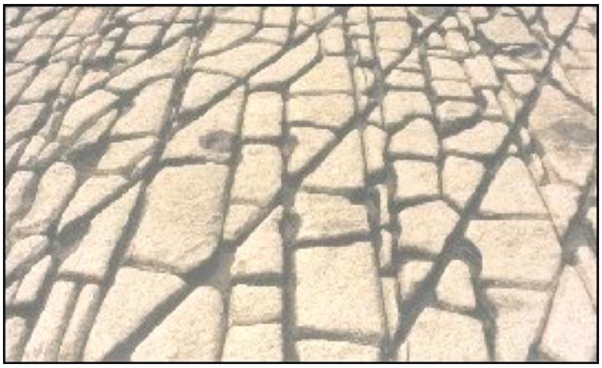


Figure 13 A fracture network in a Liassic limestone bed from Lilstock, North Somerset, England. It was produced by the superposition of individual fracture sets.

Fracture Networks

Structural geologists study the cross-cutting relationships of different fracture sets in order to determine their relative age. A variety of rules have been established to help in this task. It is found that early fractures tend to be long and relatively continuous and, as noted above, later fractures abut against these and are consequently shorter. Some of these features can be seen in Figure 15, which shows a fractured limestone pavement containing several fracture sets. The

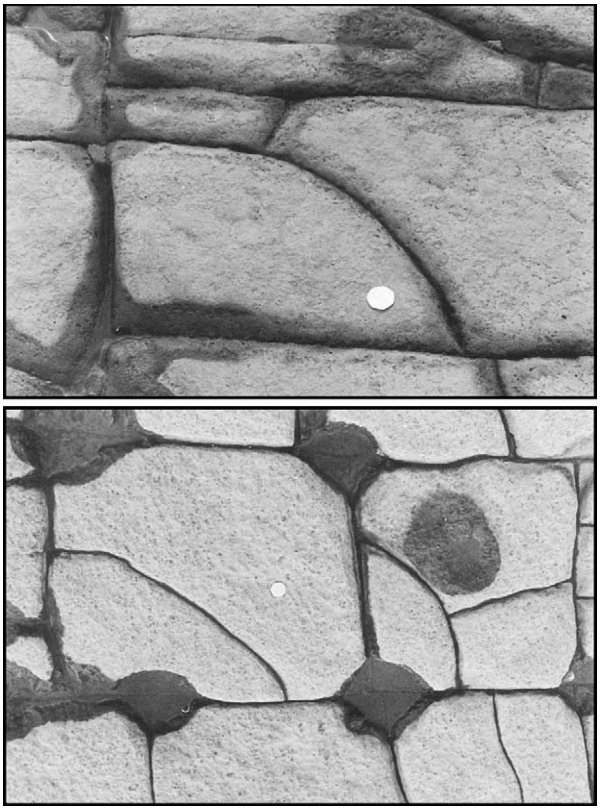


Figure 14 Details of the limestone pavement shown in Figure 13 illustrating the interaction of late fractures with early fractures. The effect of early fractures on later ones is to arrest their propagation and to modify their orientation. It can be seen that the later fractures are deflected by and abut against the earlier fractures.

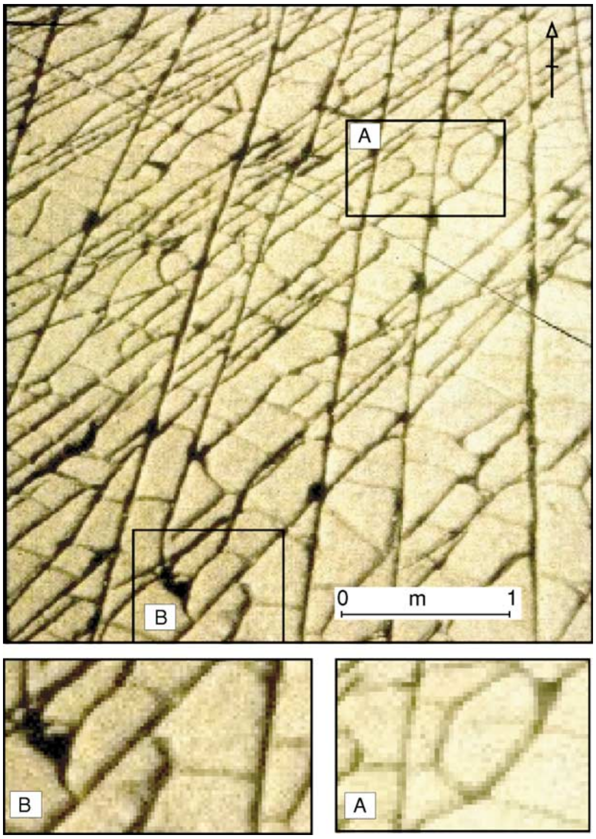


Figure 15 Fracture patterns in a limestone pavement at Lilstock, North Somerset, SW England. The older fracture sets are the most continuous and, as the sets become progressively younger, they become less continuous and less well oriented.

longest and most continuous runs approximately N–S. These are the oldest fractures and are crosscut by several younger sets which become progressively less continuous and less aligned as the regional stress fields responsible for their formation becomes progressively modified by the pre-existing fractures. The fracture set trending approximately NW–SE, the second set to form, shows a remarkable degree of continuity, being only affected by the N–S fractures; its orientation is related directly to the regional stress field.

However, as more fracture sets develop in the rock mass, modification of the stress orientation by the pre-existing fractures may result in there being a poor correlation between the fracture orientation and the regional stress field responsible for its formation. This is well illustrated in subarea A in Figure 15 which has been enlarged in the bottom left-hand corner of the figure. The influence of the pre-existing fractures on the orientation of the late fracturing is so marked that the later fractures display a polygonal organization and cannot be linked directly to the regional stress field responsible for their formation.

Fracture Analysis

A fracture analysis is the study of a fractured rock mass in order to: (i) establish the detailed geometry of the fracture network; (ii) determine the sequence of superposition of the different fracture sets that make up the fracture network; and (iii) deduce the stress regime associated with the formation of each fracture set. The reason why a detailed knowledge of the geometry of the fracture network is so important is that the bulk properties (e.g., strength, permeability) of a fractured rock mass (and most natural rocks are fractured) are generally determined by the fractures they contain rather than by the intrinsic rock properties.

Stages (ii) and (iii) of a fracture analysis are carried out using the principals outlined above relating to the interaction of fractures and the relationship between the stress field and fracture orientation (Figure 1).

Types of Faults a Plate Margins

The 'type' of plate margin is controlled by the relative motion of the two adjacent plates. They can be subdivided into three classes, convergent, divergent, and strike-slip. Convergent margins lead to compressional regimes at the plate margins which results in the formation of mountain belts. The stress regime is that appropriate for thrusts to form, namely a horizontal maximum principal compressive stress and a vertical minimum stress (Figure 5C). Divergent plate margins result in the formation of oceans and the separation of plates. The initial stage of this process is the fracturing of the lithosphere and the formation in the upper crust of major rift systems such as the East African Rift (*see* **Tectonics:** Rift Valleys). The stress regime of a horizontal minimum principal compressive stress and a vertical maximum stress is appropriate for the formation of normal faults (Figure 5A). When plates move parallel to each other at different velocities, conditions are appropriate for the formation of major wrench (strike-slip) faults (Figure 5B) such as the San Andreas Fault zone of California which separates the Pacific and North American plates.

Thus it can be seen that each of the three types of plate margins is characterized by a different types of fault.

Scale of Fracturing

Fractures occur on all scales within the Earth's crust, ranging from major faults that define plate margins, through faults that can be seen on seismic sections (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges), down to faults that can be observed directly in the field, e.g, Figure 4, to microscopic fractures only visible under the microscope. Detailed studies of the microfractures in rocks at different stages of the evolution of tensile fractures show, as predicted by Griffith's theory of stress magnification (1925) outlined above, that the microfractures grow by tensile failure at the crack tips and that suitably located microfractures link to form larger fractures oriented normal to σ_3, the minimum compressive stress (*see* **Tectonics:** Faults).

More remarkably, when the growth of shear fractures are studied in the same way, it is found that

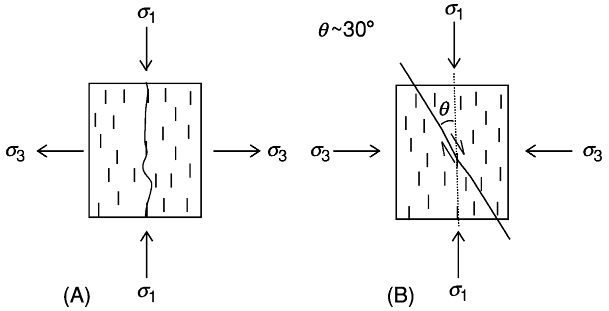


Figure 16 Randomly oriented micro-fractures within a material and their growth by tensile failure and subsequent linkage to form (A) macroscopic tensile fractures and (B) macroscopic shear fractures.

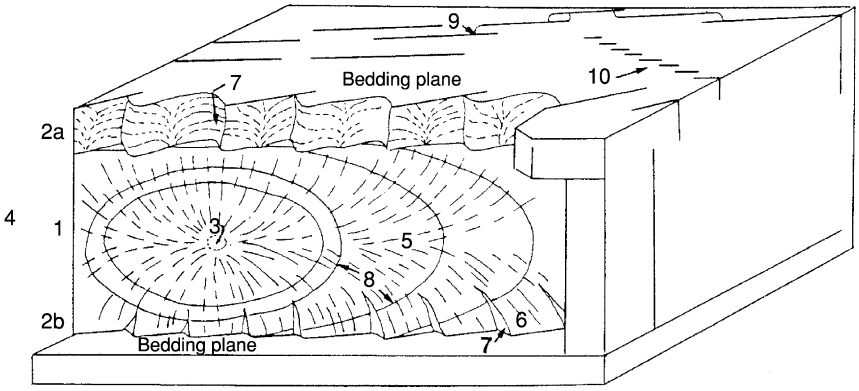


Figure 17 A block diagram illustrating the different types of surface structures (patterns) and fracture trace architecture. (Based on Kullander *et al.*, 1990.) (1) Main joint face, (2a) abrupt twist hackle fringe, (2b) Gradual twist hackle fringe, (3) Origin of fracture, (4) Hackle plume, (5) Plume axis, (6) Twist-hackle face, (7) Twist-hackle step, (8) Rig marks (front line of the fracture), (9) hooking, (10) En echelon fractures.

initially fracturing occurs by the growth of microfractures at their tips and in an orientation normal to σ_3. However, the macroscopic shear fractures are formed by the linking of offset microfractures, as shown in Figure 16. Thus, it can be seen that despite the two types of fractures having independent failure criteria and different orientations with respect to the principal stresses, they are nevertheless fundamentally linked on a microscopic scale. They are both the result of the growth of microfractures by tensile failure and differ only in the way in which these fractures are linked to a macroscopic fracture.

Surface Features of Fractures

Fractures display a variety of surface features and Figure 17 is a summary diagram showing some of these. Fractography is the science which deals with the description, analysis, and interpretation of fracture surface morphologies and links them to the causative stresses, mechanisms, and subsequent evolution of the fractures. It has been demonstrated that the diverging rays of the plumose structures (Figures 17 and 5) always remain parallel to the direction of propagation of the fracture. Thus, by constructing lines at right angles to these rays, the position and shape of the fracture front at different times of its evolution can be determined (Figures 17 and 8). When the exposures are sufficiently good, it is found that the fracture fronts form a series of concentric 'ellipses', the centre of which marks the site of fracture initiation.

See Also

Tectonics: Earthquakes; Faults; Folding; Hydrothermal Activity; Seismic Structure At Mid-Ocean Ridges; Rift Valleys.

Further Reading

Ameen MS (1995) *Fractography: fracture topography as a tool in fracture mechanics and stress analysis.* Special Publication Geological Soc. Of London, No. 92; p 240.

Anderson EM (1951) *The dynamics of faulting and dyke formation with application to Britain,* 2nd edn. Oliver & Boyd.

Griffith AA (1925) *The theory of rupture.* 1st International Conference of Applied Mechanics Proceeding Delft. P55.

Kullander BR, Dean SL, and Ward BJ (1990) *Fracture Core Analysis: Interpretation, Logging, and Use of Natural and Inducted Fractures in Core: Methods in Exploration Series, No. 8.* Tulsa, Oklahoma, USA: American Association of Petroleum Geologists.

Mandl G (1999) *Faulting in Brittle Rocks.* Springer.

Price NJ (1966) *Fault and joint development in brittle and semi-brittle rock.* Pergamon press.

Hydrothermal Activity

R P Lowell, Georgia Institute of Technology, Atlanta, GA, USA
P A Rona, Rutgers University, New Brunswick, NJ, USA

Introduction

Hydrothermal activity results from the complex interplay of heat transfer, fluid–rock chemical reactions, and fluid circulation within Earth's continental and oceanic crust. Hydrothermal circulation redistributes heat energy in the crust, often giving rise to regions of concentrated thermal output that lead to the emplacement of economically important mineral deposits and that serve as geothermal energy resources. Hydrothermal activity is thus an important component of Earth's global heat engine whereby heat transferred to the lithosphere by mantle convection is transferred to Earth's surface by thermal conduction, volcanic extrusion, and hydrothermal venting. Lithospheric plate motions, volcanic and tectonic activity, and earthquakes are manifestations of Earth's global heat engine. There is a close connection between tectonic plate boundaries and sites of hydrothermal activity (Figure 1). In the following sections of this article we compare some aspects of terrestrial and submarine hydrothermal activity, describe the basic physics and chemistry of hydrothermal circulation, briefly discuss the importance of two-phase flow, and suggest some directions for future study.

Comparison between Terrestrial and Submarine Activity

Although hydrothermal activity in terrestrial and submarine settings has many similarities there are significant distinctions. Part of the reason for this distinction stems from the manner in which heat is transported in continental and oceanic lithosphere, respectively (Table 1). In terrestrial settings, nearly 40% of the heat flux stems from radiogenic heat production in the crust, and 60% is conducted from the underlying mantle. Terrestrial hydrothermal activity accounts for less than 1% of Earth's thermal budget. On the other hand, the process of plate creation and seafloor spreading along the roughly 60 000-km ocean ridge system dominates the thermal regime of the oceanic lithosphere. Conductive heat flux from the spreading lithosphere decreases as $\tau^{-1/2}$, where τ is the lithospheric age. Hydrothermal circulation transports a significant fraction of the lithospheric heat advectively, leading to lower than expected conductive heat flow in young lithosphere (Figure 2). Nearly 25% of Earth's global heat loss and 33% of the heat loss from oceanic lithosphere result from hydrothermal activity. Most seafloor hydrothermal heat loss occurs at low temperature. High-temperature hydrothermal activity, which accounts for less than 10% of the total seafloor hydrothermal heat loss (Table 1), appears to occur only in lithosphere less than 1 My old.

Another interesting distinction between terrestrial and submarine hydrothermal activity has been their role in human endeavours. Warm and hot springs on the continents have been used for bathing and medicinal purposes since antiquity. Thermal springs were utilized throughout the Roman empire, and early descriptions of springs in Europe appear in seventeenth-century writings. Terrestrial hydrothermal systems have also long been used as an energy resource. Geothermal waters in Iceland have been used for heating for centuries, and by the 1930s a centralized heating system was established for Reykjavik. Geothermal steam has been produced at Lardarello, Italy, since the latter half of the nineteenth century, and the Geysers geothermal field in California was first exploited in the 1920s. It is remarkable that commercial development of geothermal resources occurred long before measurements of geothermal heat flux and without detailed geophysical exploration.

On the other hand, direct detection of submarine hydrothermal activity did not occur until the 1960s (Red Sea) and the 1970s at mid-ocean ridges of the Atlantic and Pacific. As a result of these discoveries the understanding of biogeochemical processes on Earth was revolutionized. It became clear that submarine hydrothermal circulation significantly impacts global geochemical cycles of both the lithosphere and the ocean. Chemical transport of certain major and trace elements to the ocean by hydrothermal discharge equals or exceeds river inputs. Moreover, hydrothermal fluids also serve as an energy resource for complex chemosynthetic biological ecosystems. The discovery of chemosynthetic ecosystems at seafloor hydrothermal vents has led to a new awareness of life in extreme environments and has stimulated the discussion of the origin of life on Earth and other planetary bodies in the solar system.

Physics and Chemistry

The fundamental components of hydrothermal activity are a heat source and a fluid circulation system

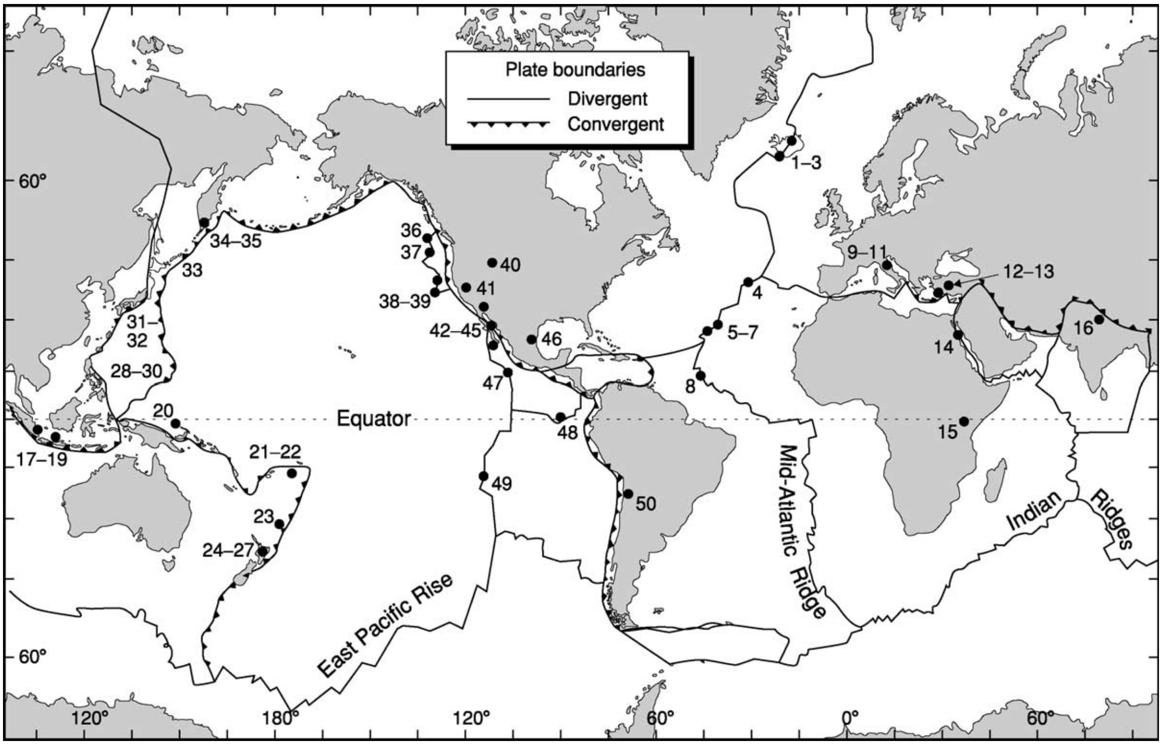


Figure 1 Plate tectonic map of the world showing locations (1–50) of selected submarine and terrestrial high-temperature hydrothermal sites as follows: (1) Krafla; (2) Namafjall; (3) Svartsengi; (4) Rainbow; (5) Lost City; (6) TAG; (7) Snake Pit; (8) Logatchev; (9) Larderello; (10) Mt. Amiata; (11) Travale; (12) Kizildere; (13) Afyon; (14) Atlantis II Deep; (15) Olkaria; (16) Puga; (17) Kawah; (18) Kamodjang; (19) Dieng; (20) PACMANUS; (21) North Fiji Basin; (22) Lau Basin; (23) Brothers; (24) Kawerau; (25) Rotorua; (26) Broadlands; (27) Wairakei; (28) Tiwi; (29) Mariana Trough; (30) Okinawa Trough; (31) Otake; (32) Sunrise; (33) Matsukawa; (34) Paratunka; (35) Pauznetska; (36) Magic Mountain; (37) Main Endeavour; (38) Sea Cliff; (39) Escanaba; (40) Yellowstone; (41) Geysers; (42) Imperial Valley; (43) Cerro Prieto; (44) Guaymas Basin; (45) East Pacific Rise 21° N;(46) Pathe; (47) East Pacific Rise 9° N; (48) Galapagos; (49) Rapa Nui; (50) El Tatio.

Table 1 Heat flux from the Earth ($\times 10^{12}$ W)[a]

Type	*Value*
Continental Crust	
(a) Crustal radiogenic heat production	4.6
(b) Conductive heat flux from the mantle	6.8
(c) Extrusion of lavas	0.03
(d) Hydrothermal flux	0.1
Total	11.5
Oceanic Crust	
(a) Conduction	20.3
(b) Extrusion of lavas	0.3
(c) Axial high-temperature hydrothermal flux	0.3
(d) Axial low-temperature hydrothermal flux	2.7
(e) Off-axis low-temperature hydrothermal flux	7
Total hydrothermal flux	10
Total	30.6
Global Heat flux	42.2

[a] Compiled from Sclater JG, Jaupart C, and Galson D (1980) and Elderfield and Schultz (1996); modified after Lowell (1991).

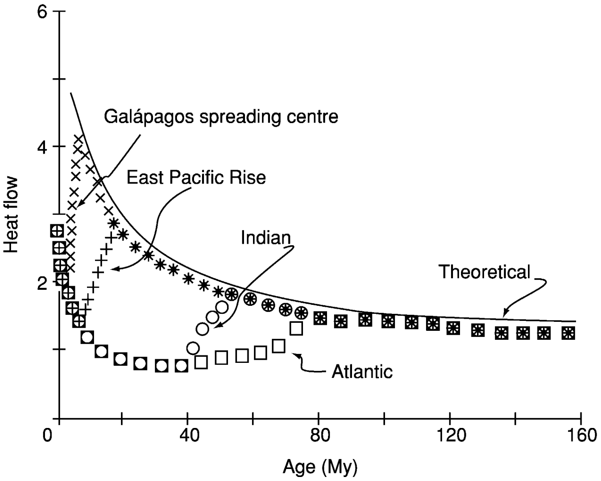


Figure 2 Observed mean heat flow for oceanic spreading centres compared with theoretical curve for conductive cooling of lithosphere. Reproduced from Anderson RN, Langseth MG Jr, and Slater JG (1977) The mechanisms of heat transfer through the floor of the Indian Ocean. *Journal of Geophysical Research* 82: 3391–3409.

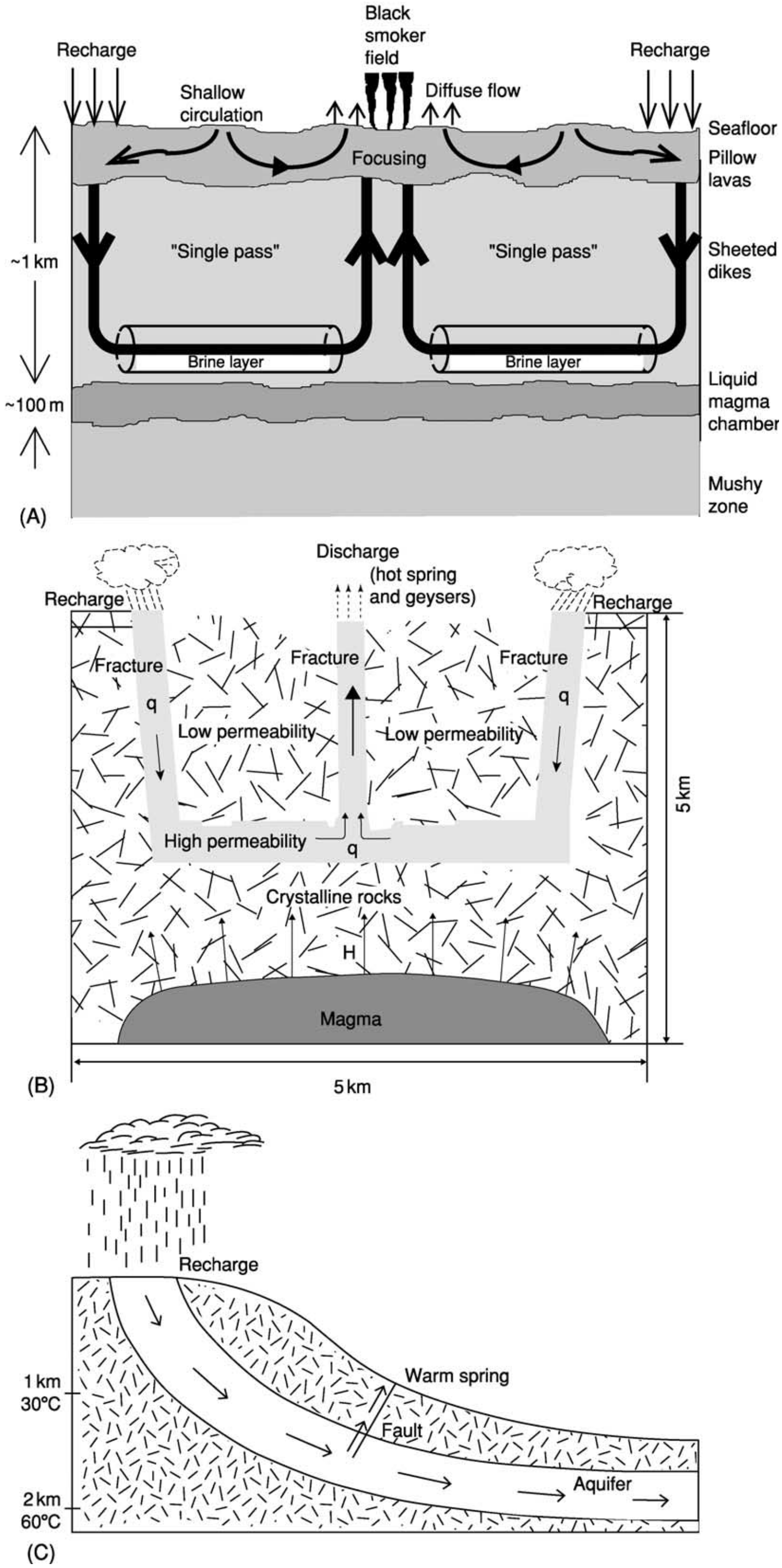


Figure 3 (A) Cartoon of single-pass hydrothermal circulation model at an ocean ridge crest. The major single-pass segment refers to a deep circulation cell in which seawater recharge penetrates through sheeted dykes to near the top of a magma body, takes up heat and undergoes water–rock chemical reactions while flowing quasihorizontally, and ascends through faults or fractures to the seafloor as high-temperature focussed black smoker flow. Mixing of the deep circulation with shallow cooler circulation in the basaltic pillow lavas may result in diffuse discharge. Reproduced from Germanovich LN, Lowell RP, and Astakhou DK (2000) Stress dependent

(*see* **Geysers and Hot Springs**). It is the nature of the heat source that generally determines whether hydrothermal activity occurs at high (>150°C) or low temperature. The circulation system consists of a recharge zone through which fluids enter the crust, a region in which the fluid takes up heat from its surroundings, and a discharge zone, through which the heated hydrothermal fluid emerges at the surface as a hot spring or hydrothermal vent. Although fluid may sometimes recirculate several times before exiting the system, it is often convenient to describe circulation in terms of a simple single-pass circulation model. Figure 3 shows cartoons of single-pass models envisioned for high-temperature terrestrial and submarine systems and a low-temperature warm spring system. In addition, all hydrothermal activity exhibits temporal variability, and chemical reactions between the circulating fluid and rock are often important.

Heat

Geothermal gradient Conductive heat flux, H, is related to the geothermal gradient by $H = \lambda \, dT/dz$, where λ is the thermal conductivity. For rocks, λ ranges from approximately 1.8 to 5 W $(\mathrm{m\,°C})^{-1}$, with most igneous and metamorphic rocks falling into a narrower range between 2.0 and 2.5 W $(\mathrm{m\,°C})^{-1}$. In older, stable continental cratons, the geothermal gradient may be as low as $10\,°\mathrm{C\,km}^{-1}$, whereas in active volcanic regions it may be more than $100\,°\mathrm{C\,km}^{-1}$. A typical geothermal gradient of $\approx 25\,°\mathrm{C\,km}^{-1}$ gives a conductive heat flux of $\approx 60\,\mathrm{mW\,m}^{-2}$.

In terrestrial low-temperature hydrothermal activity, fluids driven by a topographic head circulate to a depth of ~1–3 km in the crust where they are heated by the geothermal gradient. The fluids emerge through faults at the surface as warm or hot springs with temperatures ranging from a few tens of degrees above ambient to the local surface boiling temperature (Figure 3C). Such springs are found worldwide in areas of both normal and elevated heat flow.

Low-temperature hydrothermal circulation in oceanic crust occurs from ridge axes to a lithospheric age of ~60 My. This circulation is partially controlled seafloor topography in combination with the geothermal gradient, with discharge occurring at highs and recharge occurring at topographic lows. Type and thickness of sediment cover also influences this circulation. More than 90% of all hydrothermal heat loss from the seafloor occurs at low temperature. This circulation impacts geochemical cycles as the equivalent of an ocean volume approximately evens 10^6 years.

Magmatic heat High-temperature hydrothermal activity (typically classified as > 150°C) is associated with active volcanism. In these settings, shallow magmatic intrusions provide the heat source. Part of this heat comes from the latent heat of crystallization and part of the heat is derived from the cooling pluton. Thermal buoyancy differences between the colder and hotter parts of the system drive convective fluid motions. As volcanism is associated with ocean ridges, hot spots, and island arc systems (fore-arc, arc, and back-arc settings) at subduction zones, it is not surprising that essentially all high-temperature hydrothermal activity occurs in these regions (Figure 1).

In terrestrial settings, boiling hot springs and geysers provide the surface expressions high-temperature hydrothermal activity. Reservoir temperatures of these systems typically lie between 200 and 350°C. In oceanic settings vigorous high-temperature hydrothermal activity is exhibited as "black smoker" venting at temperatures between 300 and 400°C (Figure 4). Lower temperature "white smokers" with temperatures ~150–200°C are also common. Because of the high pressure (~250 bars) at the seafloor, these high-temperature vents lie below the boiling temperature. As discussed later, however, boiling and phase separation appear to occur in the subsurface of both terrestrial and submarine high-temperature hydrothermal systems.

Chemical heat It has long been recognized that hydration of peridotite is an exothermic reaction that produces heat, that alters the chemistry of the rocks and hydrating solutions involved, and that expands the volume of the rocks (~40%). It is only now emerging how widespread this process called the "serpentinization reaction" may be beneath ocean basins and possibly continents. The reaction involves peridotite, the characteristic ultramafic rock type of the Earth's upper mantle, and either seawater or meteoric water. Serpentinization is commonly observed in ultramafic rocks recovered from the seafloor and in slices of ancient oceanic mantle exposed on land as ophiolites. This reaction yields distinctive chemical solutions characterized by high alkalinity, high ratios of Ca to Mn and other metals, and abiogenic

permeability and the formation of seafloor event plumes. *Journal of Geophysical Research* 105: 8341–8354. (B) Analogous cartoon representing single-pass flow in a high-temperature terrestrial system. Adapted from White (1973) Characteristics of geothermal resources. In: Kruger P and Otte C (eds.) *Geothermal Energy*. Stanford, CA: Stanford University Press. (C) Cartoon of a terrestrial low-temperature warm spring system. Reproduced from Lowell (1992) Hydrothermal systems. In: *Encyclopedia of Earth System Science*, vol. 2, pp. 547–557. San Diego, CA: Academic Press.

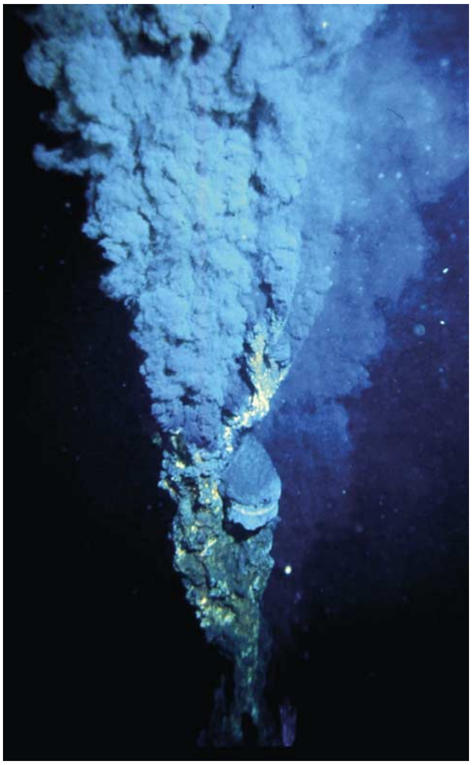


Figure 4 Black smoker vent at the East Pacific Rise 21° N hydrothermal field. (© Woods Hole Oceanographic Institution, Woods Hole, Massachusetts.)

generation of methane (CH_4) and hydrogen (H_2) gas. The heat released, depending on the volume and rate of serpentinization, may be sufficient to drive hydrothermal circulation over a range of fluid temperatures, typically low to intermediate (degrees to tens of degrees Celsius), and possibly up to several hundred degrees Celsius.

Serpentinization is favoured by conditions that facilitate access of water to large volumes of the upper mantle. In ocean basins the conditions include a low magma budget, which produces thin ocean crust, and tectonic extension and volume expansion that creates permeability through fractures and faults and that exposes rocks of the upper mantle on the seafloor. Such conditions generally occur at sections of slow-spreading ocean ridges in the Atlantic, Indian, and Arctic oceans. For example, fluids with the chemical signatures of serpentinization reactions are common along the mid-Atlantic ridge where several high-temperature (to 360°C) seafloor hydrothermal fields (Logatchev at 14°45′ N, 44°58′ W and Rainbow, 36°14′ N, 33°54′ W) at least partially situated in serpentinized ultramafic rocks of the upper mantle have been found. Only one of these fields appears to be an end member of a hydrothermal system entirely driven by serpentinization reactions (Lost City field, near 30° N, 42° W) located about 15 km west of the eastern intersection of the rift valley with the Atlantis Fracture Zone, where the field is apparently isolated from magmatic heat sources. There serpentinization-derived fluids are discharging at temperatures up to 75°C and precipitating calcium carbonate and magnesium hydroxide chimneys, which have grown up to 60 m high. Thermal and chemical fluxes from such serpentinization-driven seafloor hydrothermal systems have yet to be determined, but may be a significant fraction of global hydrothermal mass and heat budgets. Seawater and upper mantle rocks are ubiquitous in ocean basins, although the sites of serpentinization may be localized.

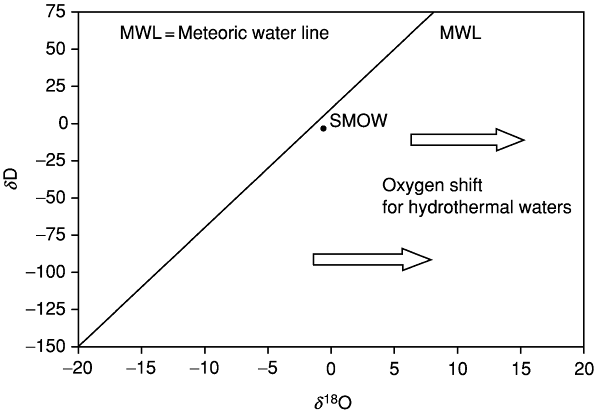


Figure 5 Schematic of δD–$\delta^{18}O$ relationship in meteoric waters. Reproduced from Craig (1961). Horizontal arrows indicate the $\delta^{18}O$ shift generally found in hydrothermal fluids that results from isotopic exchange with $\delta^{18}O$-enriched igneous and metamorphic rocks.

Fluid Sources

The aqueous fluid involved in hydrothermal activity can, in principle, have several different origins. Meteoric waters are the predominant fluid, but metamorphic or magmatic fluids may also contribute. The origin of the fluid is generally determined by examining their oxygen and hydrogen isotopic ratios (Figure 5). Meteoric waters are defined by a characteristic linear relationship between δD and $\delta^{18}O$ (MWL), whereas metamorphic and magmatic rocks and waters tend to be enriched in $\delta^{18}O$ relative to the MWL, and hence lie to the right of the curve. Ocean waters occupy a small range of δD–$\delta^{18}O$ space near the MWL (denoted by SMOW). Hydrothermal source waters are typically meteoric (or seawater) and hence lie somewhere along the MWL. At temperatures greater than 200°C, hydrothermal waters may be enriched in $\delta^{18}O$ as a result of isotopic exchange during water–rock reactions. The presence of a magmatic or metamorphic component may also move the isotopic signature of the fluid to the right of the MWL (Figure 5). Isotopic evidence of a magmatic component in active hydrothermal systems is generally inconclusive, but the presence of CO_2 in some hydrothermal fluids points to the presence of magmatic volatiles. Seawater is by far the predominant fluid in submarine hydrothermal systems.

The Circulation System

Heat transport by fluid flow through the rock requires interconnected fluid pathways and a driving force for fluid flow. The relationship between the driving force, the gradient of hydraulic head $\partial\hat{H}/\partial x_j$, and volumetric flow rate per unit area per unit time, or specific discharge q_i, is generally given by the empirical relationship called Darcy's Law.

$$q_i = (gk_{ij}/v)(\partial\hat{H}/\partial x_j) \qquad [1]$$

where g is the acceleration due to gravity, v is the kinematic viscosity of the fluid, and k_{ij} is the intrinsic rock permeability tensor, respectively. Subscripts, i, j refer to the Cartesian coordinate directions. In many applications k_{ij} is treated as a scalar k; the units of k_{ij} are m^2.

Rock permeability is the single most important physical parameter that affects hydrothermal circulation. This parameter is a measure of the interconnectivity of pore spaces and fractures; however, these features and their interconnectivity may depend upon physical and chemical processes related to the flow itself. Consequently, in any given hydrothermal environment, rock permeability may be a complex function of time and space that is difficult to determine *in situ* at the field scale. Moreover, permeability is often heterogeneous and anisotropic; it is a scale-dependent parameter that may vary over several orders of magnitude on relatively small spatial scales. Considerable research effort has been devoted to the determination of permeability and its temporal evolution during hydrothermal activity. In the following subsections we discuss some approaches to describing permeability in hydrothermal systems and its temporal evolution.

Porous medium permeability The percentage of rock volume that may be occupied by fluid is termed the porosity. To the extent that this porosity is interconnected it may give rise to permeable pathways for fluid flow. Such porosity-related permeability is termed primary permeability. There have been several mathematical models attempting to relate effective, or interconnected, porosity, ϕ, with permeability, but these have had limited success. A mathematical relationship between porosity and permeability is highly desirable for two reasons. First, porosity is scale independent and therefore laboratory-based measurements of porosity are meaningful; secondly, *in situ* porosity can be estimated from both electrical and seismic data.

Mathematical models relating effective porosity ϕ to a scalar bulk permeability k are generally of the form

$$k = Cb^2\phi^n \qquad [2]$$

where b is the average grain size of the medium and C is a numerical constant, respectively. The exponent n ranges between 2 and 3 in most formulations. The well-known Carmen-Kozeny relation is similar to eqn [2]. Equations of the form [2] often fail in practice because in most systems, even those that have significant primary permeability, field-scale permeability is controlled by fractures.

Fracture- and fault-related permeability Permeability in essentially all hydrothermally active regions is controlled by fractures and faults. Such permeability is termed secondary. In igneous and metamorphic rocks, which host most high-temperature hydrothermal activity, cracks must provide the main permeability because porosity and, hence, primary permeability is low. When permeability is controlled by fractures, large permeability can exist in the presence of very low interconnected porosity. In fracture-controlled systems, permeability is related to crack density, the abundance of crack intersections and the cube of the crack aperture.

A generalized formulation can be written as

$$k = C'\frac{l^3}{h}Na^2 \qquad [3]$$

where C' is a dimensionless coefficient describing the degree of crack interconnectivity, l is the mean crack aperture, h is the crack spacing, a is the crack length, and N is the number of cracks per unit area. As special case of [3], one may consider a set of planar parallel cracks of aperture l and spacing h. In this case the permeability is

$$k = \frac{l^3}{12h} = \frac{h^2}{12}\phi^3 \qquad [4]$$

where the porosity $\phi = l/h$. Figure 6 depicts k vs ϕ for selected values of h; the results show that large values of crack permeability can exist for $\phi \leq 1\%$. These values are several orders of magnitude greater than 10^{-18}–10^{-20} m^2, which are typical laboratory values for unfractured granite.

Estimates of permeability for hydrothermally active regions have been determined from borehole

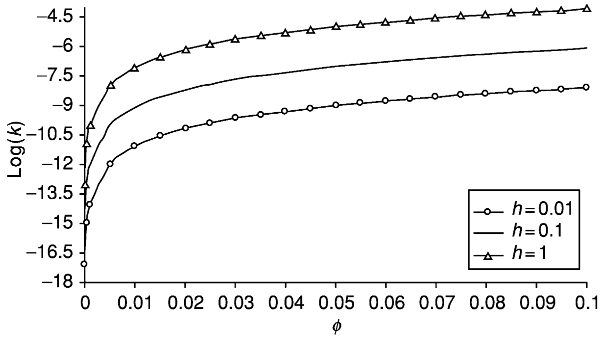


Figure 6 Graph of permeability versus porosity at a number of given fracture spacings for rock permeability resulting from planar parallel fractures. The curves show that high fracture permeability can occur in low-porosity rocks.

measurements and from field measurements of crack distributions in fossil systems. Field-scale permeability has also been estimated from mathematical modelling of hydrothermal heat output. Measurements made in Deep Sea Drilling Project and Ocean Drillng Project boreholes give permeability values ranging from 10^{-18} m^2 in sheeted dykes to as high as 10^{-10} m^2 in pillows. Values in continental hydrothermal systems often fall between 10^{-12} and 10^{-15} m^2. Crack spacing and apertures in ophiolites yield permeability values ranging between 10^{-13} and 10^{-8} m^2. Mathematical modelling studies of high-temperature venting at ocean ridge crests give a similar range. The high values of permeability and the broad range of values estimated from field and modelling studies further indicate that permeability in hydrothermal regions is fracture-controlled.

Fracture-controlled permeability is typically heterogeneous and possibly anisotropic. Fracture concentration and orientation may result from tectonic stresses as well as processes related to magma emplacement and volcanic eruptions and thermal stresses. Zones of high fracture-controlled permeability are associated with dikes. In both continental and submarine hydrothermal systems discharge zones are often focussed along tectonic faults. Recharge zones are more problematical, but faulting could be important there as well.

Although fractures and faults control hydrothermal circulation patterns, one does not often consider flow in discrete fractures. Mathematical models usually treat flow in fractured rock as an equivalent porous medium and apply Darcy's Law. It is important in this regard, however, to recognize that fracture permeability can exist on several spatial scales and that the permeability may depend on time.

Temporal variations in permeability Temporal changes in permeability can result from several mechanisms. In addition to changes in response to tectonic and magmatic processes, both thermal and chemical processes can be significant. Because the processes have not been quantified in great detail, their relative importance is uncertain.

As circulating aqueous fluids encounter different pressure and temperature environments dissolution and precipitation of chemical constituents may occur. During water–rock reactions, hydrothermal fluids reach thermodynamic equilibrium with quartz. At pressures of a few hundred bars, the solubility of quartz reaches a maximum between 350 and 400°C. Thus if the hydrothermal solution is heated above 400°C quartz will precipitate and clog fractures and pore spaces. Similarly as the hydrothermal solution ascends towards the surface, both lower pressures and temperatures in the environment will foster quartz precipitation. Both quartz and amorphous silica are common vein minerals in hydrothermal systems, and precipitation of these phases may exert a strong influence on hydrothermal circulation over time. The development of a low permeability barrier as a result of quartz precipitation is likely an important factor in the evolution of vapour-dominated hydrothermal systems.

In submarine hydrothermal systems precipitation of anhydrite may be important in both recharge and discharge zones. Because the solubility of anhydrite decreases with increasing temperature, heating of seawater during hydrothermal recharge to $T \geq 150°\mathrm{C}$ results in precipitation of anhydrite in recharge zones (Figure 3A). Sulphate is removed from seawater by precipitation of anhydrite and reaction with crustal rocks; however, upon ascent, mixing of sulphate-poor, hot hydrothermal fluid with cold sulphate-rich seawater may again result in the precipitation of anhydrite. Mixing in the subsurface may contribute to focussing seafloor venting into black smokers; whereas mixing above the seafloor contributes to the formation of chimney structures.

Thermoelastic stresses result from the passage of either cold fluids through initially hotter rock or hot fluid through cooler rocks. In the former case, cooling of rock surfaces leads to thermal contraction and the enhancement of permeability. In the latter case, heating of the rock leads to thermal expansion and reduction of permeability. The dependence of permeability on temperature can be expressed as

$$k = k_0[1 - \gamma(T - T_0)]^3 + k_{\mathrm{res}} \qquad [5]$$

where k_0 is the permeability of the main crack network, k_{res} is a finer-scale residual permeability, and γ is factor expressing the strength of the thermoelastic effect.

At temperatures exceeding 350–400°C, rocks begin to exhibit ductile behaviour. Although this behaviour depends upon the rate which stresses are applied, ductile behaviour will tend to seal cracks. Thus permeable pathways that may initially be opened by stresses resulting in brittle failure may gradually close. This process may limit the depth to which cracks remain open in the crust and the extent to which hydrothermal circulation may approach magma bodies.

Water–Rock Chemical Reactions

As aqueous fluids pass through hot subsurface rocks, chemical reactions occur. Some chemical constituents may be removed from the fluid, others may be extracted from the rock. The reactions may also involve isotopic exchange between the fluid and rock. These reactions are complex functions of temperature, pressure, lithology, permeability structure, duration of activity, and other factors. A detailed discussion of this topic is beyond the scope of this article; however, the use of geochemical thermometers and the formation of hydrothermal ore deposits are discussed briefly.

Geochemical thermometers The strong temperature dependence of solubility of certain chemical constituents in hydrothermal fluids, the temperature dependence of elemental partitioning between rock and solution, and the temperature dependence of isotopic partitioning between mineral and fluid phases have led to the development of a variety of geochemical thermometers to deduce subsurface conditions from surface samples. The quartz geothermometer utilizes the strong temperature dependence of quartz solubility and the slow kinetics of quartz precipitation at low temperature. As hydrothermal solutions in equilibrium with quartz at high temperature rise to the surface and cool, the high degree of disequilibrium in the measured quartz concentration permits a calculation of the equilibrium temperature at depth Other common geothermometers include:

1. Na/K, which makes use of the temperature dependence of partitioning of these elements between aluminosilicate rocks and hydrothermal fluid;
2. Na–K–Ca, which includes the effect of Ca in the partitioning; and
3. ratios of stable isotopes such as $\delta^{13}C$, $\delta^{18}O$, δD, and $\delta^{34}S$.

Various factors affect the resolution and reliability of each of these geothermometers, so often many independent ones are used.

Ore deposits–fossil hydrothermal systems As a result of water–rock chemical reactions at intermediate to high temperature, trace metallic ore-forming metals such as Fe, Cu, Zn, Sb, Au, Ag, and Pb are transferred from the rock to the hydrothermal solution. Because most of these metals form metallic sulphides that are highly insoluble in water, solubility is achieved by the formation of bisulphide or chloride ion complexes. Various mechanisms may cause local precipitation of these metal–ion complexes, resulting in a concentrated accumulation of metallic ore. A rapid drop in the solution temperature because of thermal conduction or mixing with cooler fluids, a change in solution pH, and boiling can all lead to ore deposition.

Many types of ore deposits in the geological record, such as porphyry ore deposits associated with silicic volcanism, Mississippi Valley-type lead-zinc deposits, and volcanically hosted massive sulphide deposits are linked to hydrothermal activity. Such ore deposits thus present an integrated fossil record of hydrothermal activity, and provide a window into subsurface heat transfer and fluid flow processes. By coupling this integrated fossil record with studies of active ore-forming processes on the seafloor and in other active hydrothermal environments, one can obtain a more complete picture of hydrothermal activity (*see* **Mining Geology:** Hydrothermal Ores).

Temporal Variability in Hydrothermal Activity

Temporal variability on a range of time-scales is a fundamental characteristic of hydrothermal activity (Table 2). Some of this variability is linked to episodicity in magmatic and tectonic activity, or climate changes. The occurrence of these processes ranges from scales of plate reorganization of $\sim 10^6$–10^7 years to magma replacement times of $\sim 10^1$–10^4 years at fast and slow spreading ridges, respectively. Temporal variability related to the fluid circulation system, mainly resulting from changes in crustal permeability, occurs on time-scales ~ 1–10^2 years. Seafloor hydrothermal activity is known to change on time-scales of hours to months following earthquakes, igneous intrusions (e.g., dykes), or volcanic eruptions. Climate changes may alter precipitation patterns, and hence fluid recharge, on time-scales of 10–10^3 years; ice ages and glaciation may affect high-altitude systems on similar time-scale.

Two-Phase Flow

Boiling and phase separation commonly occur in high-temperature hydrothermal systems. For pure water, boiling is defined by the boiling point curve as a function of pressure. Liquid phase occurs below

Table 2 Time-scale of events and processes related to hydrothermal activity[a]

Time-scale	*Activity or process*
10^6–10^7 years	Plate reorganization
10^6–10^7 years	Episodes of seafloor spreading
10^5 years	Magnetic polarity interval
10^1–10^6 years	Duration of ore formation processes
10^3–10^4 years	Eruption cycle on slow spreading ridges
10^1–10^3 years	Eruption cycle on fast spreading ridges
10^3–10^4 years	Glacial episodes
10^3–10^6 years	Duration of hydrothermal activity
10^1–10^3 years	Episodes of climate change
10^0–10^2 years	Duration of individual seafloor hydrothermal vent
10^0–10^1 years (hours to decade)	Duration of volcanic eruption
10^0–10^1 years	Residence time of hydrothermal fluid in oceanic crust
10^4–10^7 s	Transit time of upwelling hydrothermal fluid
10^5–10^7 s	Duration of earthquake swarms
10^5–10^7 s	Duration of dyke emplacement event
10^5–10^6 s	Duration of seafloor event plume
10^3–10^6 s	Period of tidal signals
0.1–3 s	Precipitation of sulphide particles during mixing of high-*T* hydrothermal fluid with ambient seawater

[a]Modified from Lowell RP, Rona PA, and Von Herzen RP (1995) and Rona (1988).

the boiling point and the vapour phase occurs above it. The two phases are in equilibrium along the curve, and the volume fractions of each phase along the curve depend upon the enthalpy of the system. The phase diagram for water (Figure 7) most readily shows these relations. A key feature of the pure water phase diagram is the critical point defined by $P_c = 218$ bars, $T_c = 374°\mathrm{C}$. Above the critical point only a single-phase "water substance" exists.

Most hydrothermal aqueous fluids contain some amount of dissolved salts; however, the presence of these dramatically alters the two-phase behaviour of water. First of all, both P_c and T_c increase as the salt content increases. For seawater, which can be represented by salinity $x \approx 3.2\%$ NaCl solution, $P_c \approx 300$ bars, $T_c \approx 405°\mathrm{C}$. Secondly, two-phase flow is defined by a region of *P*-*T*-*x* space rather than a curve. Thirdly, because fluid density is a function of salinity, the fractionation of salt between the liquid and vapour phases affects the dynamics of the phases. Figure 8 depicts part of the phase diagram for NaCl–water solution.

In terrestrial systems subsurface boiling results in either liquid- or vapour-dominated systems. In liquid-dominated systems, fluid pressures are near hydrostatic. Hot spring fluids are generally neutral to alkaline pH and chloride rich. Liquid and vapour phases are intermingled within the two-phase zone, and the two-phase zone overlies a single-phase fluid. By contrast, in vapour-dominated systems low, nearly uniform, vapour-static fluid pressure occurs over a considerable thickness. Fluid discharge occurs at low pH and low chloride. The presence of a region of underpressure implies a permeable barrier between the vapour-dominated zone and the surrounding cold recharge. Vapour-dominated systems act as a heat pipe with near zero net mass flux; heat is carried upwards by high enthalpy vapour while small amounts of low-enthalpy liquid flow downwards. Most systems are liquid-dominated, including the geyser basins of Yellowstone National Park, USA, Wairakei and Broadlands, NZ, and Ahuachapan, MX; vapour-dominated systems include Geysers, USA, Lardarello, IT, Kamodjang, IND, and Matsukawa, JP (Figure 1).

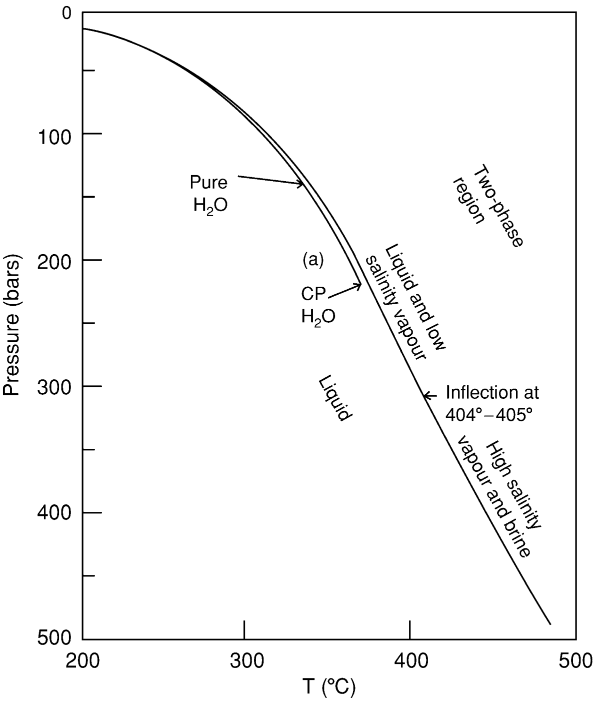


Figure 7 Boiling point curve for pure water compared with that for seawater. Note that the pure water curve ends at the critical point. The region above the pure water curve is pure vapour and the two-phases only exist along the boiling curve. Reproduced from Bischoff and Rosenbauer (1984) The critical point and two-phase boundary of seawater, 200–500°C. *Earth and Planetary Science Letters* 68: 172–180.

In submarine systems venting black smoker fluids are mostly in the liquid phase; however, the chlorinity of vent fluids seldom corresponds to seawater (≈540 mmol kg). Rather it ranges from ≈30 to ≈1200 mmol kg (Table 3). The departure of vent salinity from that of seawater is attributed to phase separation. The lowest chlorinity values often occur shortly after magmatic eruptions or diking events and thus are indicative of active phase separation.

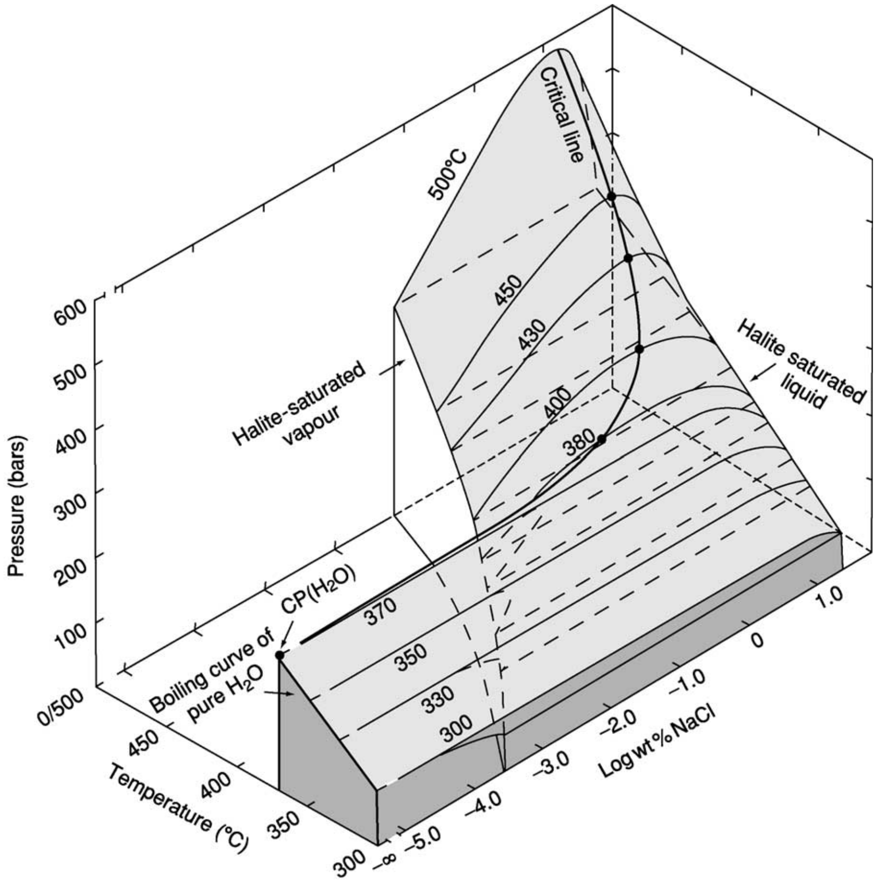


Figure 8 Three-dimensional perspective of the NaCl–H_2O phase diagram between 300 and 500°C. Reproduced from Bischoff and Pitzer (1989) Liquid-vapor relations for the system NACL-H_2O: Summary of the P-T-x surface from 300° to 500°C. *American Journal of Science* 289: 217–248.

Table 3 Chlorinity of selected high-temperature seafloor hydrothermal Vents[a]

Vent site	*Year(s) sampled*	*Value (mmol kg^{-1})*[b]
East Pacific rise		
9–10° N	1991	32–860
21° N	1979, 81, 85	489–579
Juan de Fuca ridge		
North cleft	1990–92	730–1245
South cleft	1984	896–1087
Endeavour	1984–88	253–505
Axial volcano	1986–88	176–624
Mid-Atlantic ridge		
TAG	1986	659
MARK	1986	559
Lau basin	1989	650–800

[a] Modified from Von Damm (1985).

[b] Normal seawater chlorinity = 540 mmol kg^{-1}.

Phase separation also results in the formation of saline brines that, because of their high density, sink towards the base of the system. Later mixing of these saline brines with seawater may result in vent chlorinity greater than seawater. The formation of a brine layer at the base of a hydrothermal system may act as a thermal conductive barrier between the overlying hydrothermal circulation and the magma body (Figure 3A) and be a salinity source for saline vent fluids.

Future Directions

Hydrothermal activity represents an exciting dynamic area for future research. This is particularly true for submarine systems because of their links to studies of the origin of life, life in extreme environments, and the continued discovery of novel types of hydrothermal activity. The detailed sampling and data analysis and continued exploration for serpentinization-driven hydrothermal activity will likely grow during the next decade. At ridge crest and volcanic island arc systems, advances in ocean drilling technology, remote and autonomous sensing devices, long-term monitoring, integrated interdisciplinary experiments at various well-characterized seafloor sites, and improvements in mathematical modelling techniques will stimulate the science over the next decade. Finally, we believe that analysis of seafloor hydrothermal activity over geological time, and attempts to discern the importance of hydrothermal activity elsewhere in the solar system will emerge as an important endeavour. Such studies are necessary to understand the links between hydrothermal activity and life.

In terrestrial systems continued exploitation as an energy resource will be important. Moreover, climatically induced precipitation changes, because of the link to hydrothermal recharge, may alter warm spring and geyser behaviour. Utilization of hydrothermal activity as a climate monitor has yet to receive attention.

See Also

Geysers and Hot Springs. **Igneous Processes**. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Origin of Life**. **Plate Tectonics**. **Tectonics:** Faults.

Further Reading

Anderson RN, Langseth MG Jr, and Sclater JG (1977) The mechanisms of heat transfer through the floor of the Indian Ocean. *Journal of Geophysical Research* 82: 3391–3409.

Bischoff JL and Pitzer KS (1989) Liquid-vapor relations for the system NACL-H_2O: Summary of the P-T-x surface from 300° to 500°C. *American Journal of Science* 289: 217–248.

Bischoff JL and Rosenbauer RJ (1984) The critical point and two-phase boundary of seawater, 200–500°C. *Earth and Planetary Science Letters* 68: 172–180.

Craig H (1961) Standard for reporting concentrations of deuterium and oxygen-18 in natural waters. *Science* 133: 1833–1834.

Elder J (1981) *Geothermal Systems*. San Diego, CA: Academic Press.

Elderfield H and Schultz A (1996) Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annual Reviews of Earth and Planetary Science* 24: 191–224.

Germanovich LN, Lowell RP, and Astakhov DK (2000) Stress dependent permeability and the formation of seafloor event plumes. *Journal of Geophysical Research* 105: 8341–8354.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (eds.) (1995) *Seafloor Hydrothermal Systems: Physical, Chemical, and Biological Interactions,* AGU Geophysical Monograph 91. Washington, DC: American Geophysical Union.

Kelley DS, Baross JA, and Delaney JR (2002) Volcanoes, fluids and life at mid-ocean ridge spreading centers. *Annual Review of Earth and Planetary Science* 30: 385–491.

Kruger P and Otte C (eds.) (1973) *Geothermal Energy.* Stanford, CA: Stanford University Press.

Lowell RP (1991) Modeling continental and submarine hydrothermal systems. *Reviews of Geophysics* 29: 457–476.

Lowell RP (1992) Hydrothermal systems. In: Nierenburg WA (ed.) *Encyclopedia of Earth System Science,* vol. 2, pp. 547–557. San Diego, CA: Academic Press.

Lowell RP, Rona PA, and Von Herzen RP (1995) Seafloor hydrothermal systems. *Journal of Geophysical Research* 100: 327–352.

Rona PA (1988) Hydrothermal mineralization at oceanic ridges. *Canadian Mineralogist* 26: 431–465.

Sclater JG, Jaupart C, and Galson D (1980) The heat flow through oceanic and continental crust and the heat loss of the Earth. *Review of Geophysics* 18: 269–311.

Von Damm KL (1995) Controls on the chemistry and temporal variability of seafloor hydrothermal fluids. In: Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (eds.) *Seafloor Hydrothermal Systems: Physical, Chemical, and Biological Interactions*, AGU Geophysical Monograph 91, pp. 222–247. Washington, DC: American Geophysical Union.

White DE (1973) Characteristics of geothermal resources. In: Kruger P and Otte C (eds.) *Geothermal Energy.* Stanford, CA: Stanford University Press.

Mid-Ocean Ridges

K C Macdonald, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

The mid-ocean ridge system is the largest mountain chain and the most active system of volcanoes in the solar system. In plate-tectonic theory, the ridge is located between plates of the Earth's rigid outer shell that are separating at speeds of approximately 10–170 mm year^{-1} (up to 220 mm year^{-1} in the past). The ascent of molten rock from deep within the Earth (*ca.* 30–60 km) to fill the void between the plates creates new seafloor and a volcanically active ridge. This ridge system wraps around the globe like the seam of a baseball and is approximately 70 000 km long (including the lengths of ridge offsets, such as transform faults). Yet the ridge itself is only about 5–30 km wide, very small compared with the plates, which can be thousands of kilometres across (Figure 1).

Early exploration showed that the gross morphology of spreading centres varies with the rate of plate separation. At slow spreading rates (10–40 mm year^{-1}) a rift valley 1–3 km deep marks the axis, while for fast spreading rates (more than 90 mm year^{-1}) the axis is characterized by an elevation of the seafloor of several hundred metres called an axial high (Figure 2). The rate of magma supply is a second factor that may influence the morphology of mid-ocean ridges. For example, a very high rate of magma supply can produce an axial high even where the spreading rate is slow; the Reykjanes Ridge south of Iceland is a good example. Also, for intermediate spreading rates (40–90 mm year^{-1}) the ridge crest may have either an axial high or a rift valley depending on the rate of magma supply. The depth to the seafloor increases from a global average of approximately 2600 m at the spreading centre to more than 5000 m beyond the ridge flanks. The rate of deepening is proportional to the square root of the age of the seafloor because it is caused by the thermal contraction of the lithosphere. Early mapping efforts also showed that the mid-ocean ridge is a discontinuous structure, which is offset at right angles to its length by numerous transform faults that are tens to hundreds of kilometres long.

Maps are powerful: they inform, excite, and stimulate. Just as the earliest maps of the world in the sixteenth century ushered in a vigorous age of exploration, so the first high-resolution continuous-coverage maps of the mid-ocean ridge system stimulated investigators from a wide range of fields, including petrologists, geochemists, volcanologists, seismologists, tectonicists, and practitioners of marine magnetics and gravity, as well as researchers outside the Earth sciences, including marine ecologists, chemists, and biochemists. Marine geologists have found that many of the most revealing variations are observed by exploring along the axis of the active ridge. This along-strike perspective has revealed the architecture of the global rift system. The ridge axis undulates up and down in a systematic way, defining a fundamental partitioning of the ridge into segments bounded by a variety of discontinuities. These segments behave like giant cracks in the seafloor, which can lengthen or shorten and have episodes of increased volcanic and tectonic activity. In fact, elementary fracture mechanics can be used to explain the interaction between neighbouring ridge segments.

Another important change in perspective came from the discovery of hydrothermal vents by marine geologists and geophysicists. It became clear that,

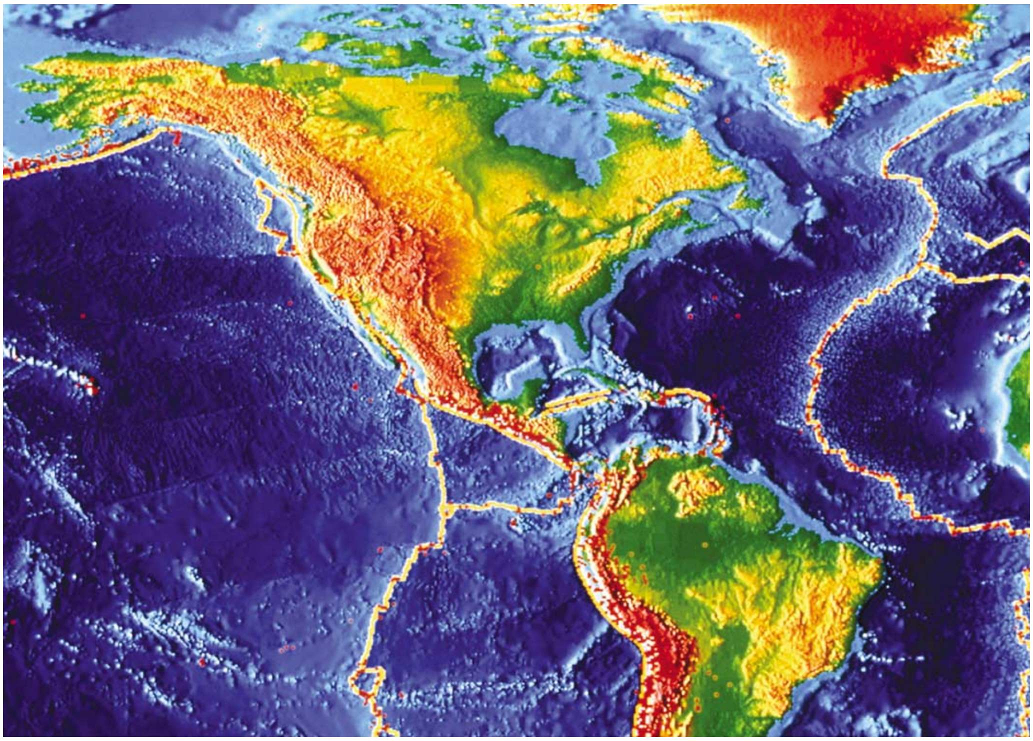


Figure 1 Shaded relief map of the seafloor showing parts of the East Pacific Rise, a fast-spreading centre, and the Mid-Atlantic Ridge, a slow spreading centre (courtesy of the National Geophysical Data Centre).

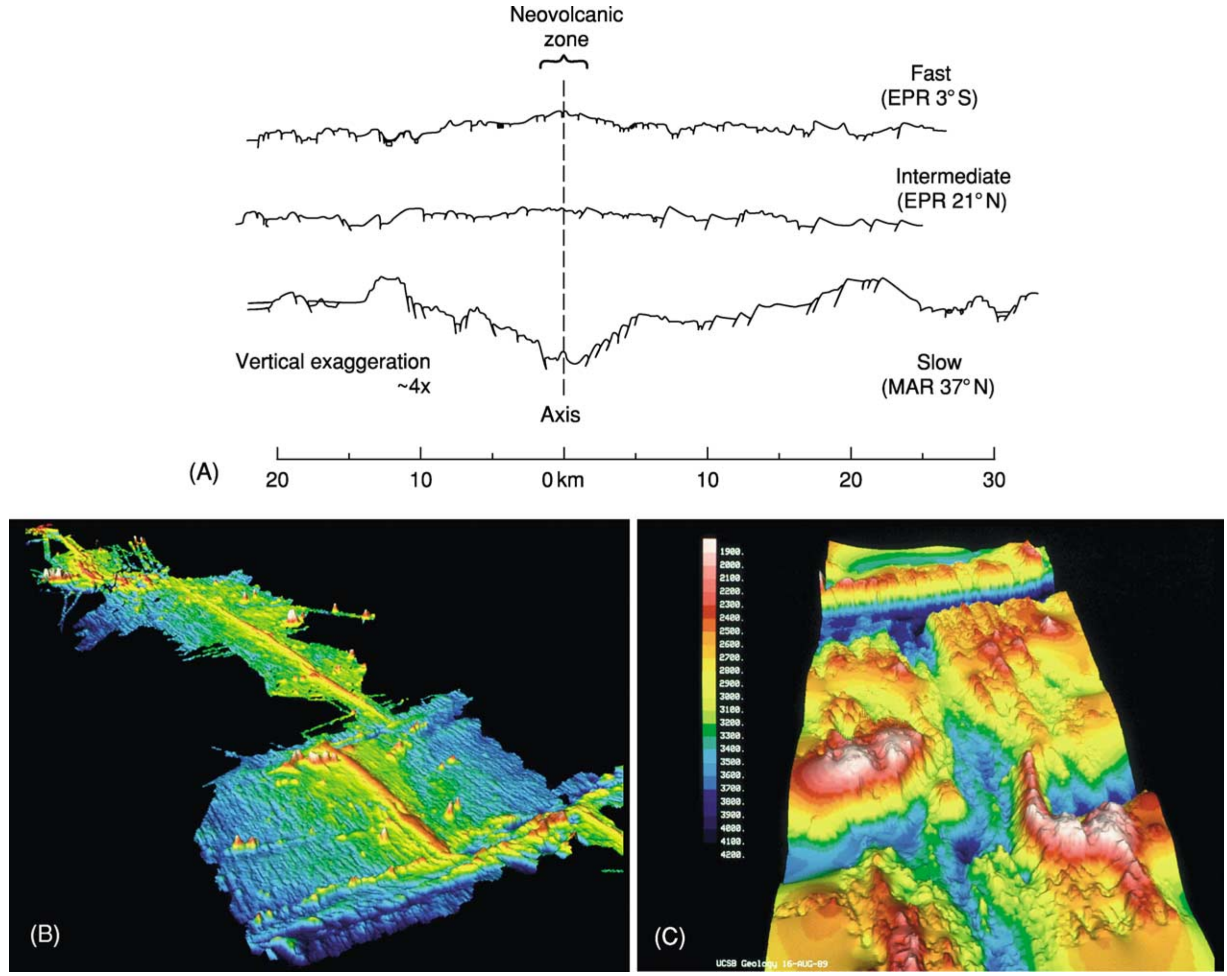


Figure 2 Topography of spreading centres. (A) Cross-sections of typical fast-intermediate- and slow-spreading ridges based on high-resolution deep-tow profiles. The neovolcanic zone (the zone of active volcanism) is indicated and is several kilometres wide; the zone of active faulting extends to the edge of the profiles and is several tens of kilometres wide. (Reproduced from Macdonald KC (1982) Mid-ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annual Review of Earth and Planetary Sciences* 10: 155–190.) EPR, East Pacific Rise; MAR, Mid-Atlantic Ridge. (B) Shaded relief map of a 1000 km stretch of the East Pacific Rise extending from 8° N to 17° N. Here, the East Pacific Rise is the boundary between the Pacific and Cocos plates, which are separating at a 'fast' rate of 120 mm year^{-1}. The map reveals two kinds of discontinuity: large offsets, about 100 km long, known as transform faults, and smaller offsets, about 10 km long, called overlapping spreading centres. Colours indicate depths of 2400 m (pink) to 3500 m (dark blue). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.) (C) Shaded relief map of the Mid-Atlantic Ridge. Here, the ridge is the plate boundary between the South American and African plates, which are spreading apart at the slow rate of approximately 35 mm year^{-1}. The axis of the ridge is marked by a 2 km deep rift valley, which is typical of most slow-spreading ridges. The map reveals a 12 km jog of the rift valley, a second-order discontinuity, and also shows a first-order discontinuity called the Cox transform fault. Colours indicate depths of 1900 m (pink) to 4200 m (dark blue). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

in studies of mid-ocean ridge tectonics, volcanism, and hydrothermal activity, the greatest excitement is in the linkages between these different fields. For example, geophysicists searched for hydrothermal activity at mid-ocean ridges for many years by towing arrays of thermistors near the seafloor. However, hydrothermal activity was eventually documented more effectively by photographing the distribution of exotic vent animals. Even now, the best indicators of the recency of volcanic eruptions and the duration of hydrothermal activity are found by studying the characteristics of benthic faunal communities. For example, during the first deep-sea mid-ocean-ridge eruption witnessed from a submersible, divers did not see a slow lumbering cascade of pillow lavas, as observed by divers off the coast of Hawaii. What they saw was completely unexpected: white bacterial matting billowing out of the seafloor, creating a scene much like a mid-winter blizzard in Iceland, covering the freshly erupted glassy black lava with a thick blanket of white bacterial ‘snow’.

Ridge Segmentation

The most recognizable segmentation of mid-ocean ridges is that defined by transform faults. These plate boundaries are usually perpendicular to the ridge segments they offset and are tens to hundreds of kilometres long, although some exceed 1000 km in length (e.g. the Romanche and San Andreas faults). In plate tectonics, a transform fault traces a small circle about the Euler pole of opening between any pair of plates. Thus the transform fault and its off-axis fracture zone traces may be used to determine the pole of opening as well as changes in the pole of opening. At a ridge–transform intersection, normal spreading processes are truncated. Normal faulting predominates on mid-ocean ridges, while strike-slip faulting dominates along transform faults. The transition can be very complex, with normal faults and strike-slip faults occurring along trends that are affected by shear stresses on the transform fault. Crustal accretionary processes are also affected by the juxtaposition of thick cold lithosphere against the end of a spreading segment. This effect increases with the age and thickness of the lithosphere that is sliding past the ridge–transform intersection. Transverse ridges occur along the length of some of the largest transform faults; some of these ridges have been elevated above sea-level for part of their history.

Between major transform faults, the axial depth profile of mid-ocean ridges undulates up and down with a wavelength of tens of kilometres and an amplitude of tens to hundreds of metres at fast-spreading and intermediate-spreading ridges. This pattern is also observed for slow-spreading ridges, but the wavelength of undulation is shorter and the amplitude is larger (Figure 3). In most cases, ridge-axis discontinuities occur at local maxima of the axial depth profile. These discontinuities include transform faults, as discussed above (first order), overlapping spreading centres (second order), and higher order (third and fourth order) discontinuities, which are increasingly short-lived, mobile, and associated with smaller offsets of the ridge (see Table 1 and Figure 4).

A much-debated hypothesis is that the axial depth profile (Figures 3 and 5) reflects the magma supply along a ridge segment. According to this idea, the magma supply is enhanced along shallow portions of ridge segments and is relatively starved at segment ends (discontinuities). In support of this hypothesis is the observation at ridges with an axial high (fast-spreading ridges) that the cross-sectional area or axial volume varies directly with depth (Figure 6). Maxima in the cross-sectional area (more than 2.5 km^2) occur at minima along the axial depth profile (generally not near ridge-axis discontinuities) and are thought to correlate with regions where magma supply is robust. Conversely, small cross-sectional areas (less than 1.5 km^2) occur at local depth maxima and are interpreted to reflect minima in the magma-supply rate along a given ridge segment. On slow-spreading ridges characterized by an axial rift valley, the cross-sectional area of the valley is at a minimum in the mid-segment regions, where the depth is at a minimum. In addition, there are more volcanoes in the shallow mid-segment area, and fewer volcanoes near the segment ends. Studies of crustal magnetization show that very highly magnetized zones occur near segment ends; these are most easily explained by a local restriction of magma supply resulting in the eruption of highly fractionated lavas that are rich in iron.

Multichannel seismic and gravity data support the axial volume–magma supply–segmentation hypothesis (Figure 6). A bright reflector, which is phase-reversed in many places, occurs commonly (>60%

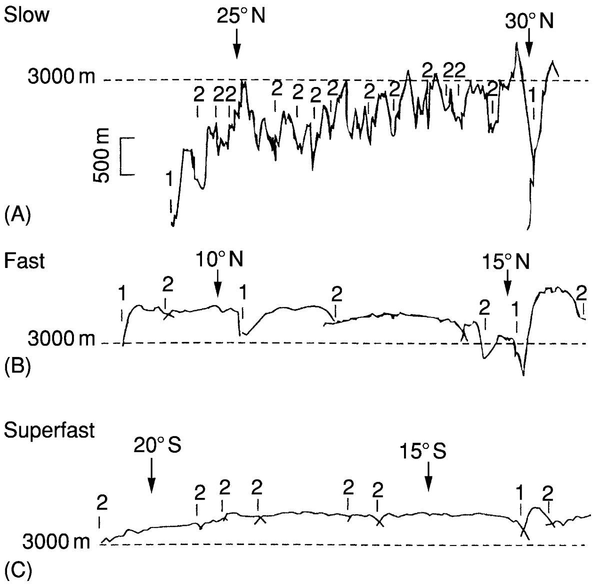


Figure 3 Axial depth profiles for (A) slow-spreading, (B) fast-spreading, and (C) superfast-spreading ridges. Discontinuities of orders 1 and 2 typically occur at local depth maxima (discontinuities of orders 3 and 4 are not labelled here). The segments at faster-spreading ridges are longer and have smoother lower-amplitude axial depth profiles. These depth variations may reflect the pattern of mantle upwelling. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

Table 1 Characteristics of segmentation, updated from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid-ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986–994 (see references therein). This four-tiered hierarchy of segmentation probably represents a continuum in segmentation

	Order 1	*Order 2*	*Order 3*	*Order 4*
Segments				
Segment length (km)	600 ± 300[a] (400 ± 200)[b]	140 ± 90 (50 ± 30)	20 ± 10 $(15 \pm 10?)$	7 ± 5 $(7 \pm 5?)$
Segment longevity (years)	$>5 \times 10^6$	$0.5-5 \times 10^6$ $(0.5-30 \times 10^6)$	$\sim 10^4-10^5$ (?)	$<10^3$ (?)
Rate of segment lengthening (long-term migration) (mm y^{-1})	0–50 (0–30)	0–1000 (0–30)	Indeterminate: no off-axis trace	Indeterminate: no off-axis trace
Rate of segment lengthening (short-term propagation) (mm y^{-1})	0–100 (?)	0–1000 (0–50)	Indeterminate: no off-axis trace	Indeterminate: no off-axis trace
Discontinuities				
Type	Transform, large propagating rifts	Overlapping spreading centres (oblique shear zones, rift-valley jogs)	Overlapping spreading centers (intervolcano gaps), devals	Devals, offsets of axial summit caldera (intravolcano gaps)
Offset (km)	>30	2–30	0.5–2.0	<1
Offset age (years)[c]	$>0.5 \times 10^6$ $(>2 \times 10^6)$	0.5×10^6 (2×10^6)	~ 0	~ 0
Depth anomaly (m)	300–600 (500–2000)	100–300 (300–1000)	30–100 (50–300)	0–50 (0–100?)
Off-axis trace	Fracture zone	V-shaped discordant zone	Faint or none	None
High-amplitude magnetization?	Yes	Yes	Rarely (?)	No? (?)
Breaks in axial magma chamber?	Always	Yes, except during OSC linkage? (NA)	Yes, except during OSC linkage? (NA)	Rarely
Breaks in axial low-velocity zone?	Yes (NA)	No, but reduction in volume (NA)	Small reduction in volume (NA)	Small reduction in volume? (NA)
Geochemical anomaly?	Yes	Yes	Usually	$\sim 50\%$
Break in high-temperature venting?	Yes	Yes	Yes (NA)	Often (NA)

[a] Values are ±1 standard deviation.

[b] Where information differs for slow- and fast-spreading ridges (<60 mm y^{-1}), it is placed in parentheses.

[c] Offset age refers to the age of the seafloor that is juxtaposed to the spreading axis at a discontinuity.

NA, not applicable; ?, not presently known as poorly constrained; OSC, overlapping spreading centre.

of ridge length) beneath the axial regions of both the northern and southern portions of the fast-spreading and ultra-fast-spreading East Pacific Rise. This reflector has been interpreted as a thin lens of magma residing at the top of a broader axial magma reservoir. The amount of melt is highly variable along strike, varying from a lens that is primarily crystal mush to one that is close to 100% melt. This 'axial magma chamber' reflector is observed where the ridge is shallow and where the axial high has a broad cross-sectional area. Conversely, it is rare where the ridge is deep and narrow, especially near ridge-axis discontinuities. A reflector may occur beneath ridge-axis discontinuities during propagation and ridge-axis realignment, as may be occurring now on the East Pacific Rise near 9° N (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges; Propagating Rifts and Microplates At Mid-Ocean Ridges).

There is evidence that major-element geochemistry correlates with axial cross-sectional area (Figure 7). On the East Pacific Rise between 13° S and 21° S there is a good correlation between MgO wt% and cross-sectional area (higher MgO indicates a higher eruption temperature and perhaps a greater local magmatic budget). The abundance of hydrothermal venting (as measured by light transmission and backscatter in the water column and geochemical tracers) also varies directly with the cross-sectional area of the East Pacific Rise. It is not often that one sees a correlation between two such different kinds of measurement. It is all the more remarkable considering that the measurements of hydrothermal activity are sensitive to changes on a time-scale of days to months, while the cross-sectional area probably reflects a time-scale of change measured in tens of thousands of years.

On slow-spreading centres, such as the Mid-Atlantic Ridge, the picture is less clear. Seismic and gravity data indicate that the oceanic crust thins significantly near many of the transform faults, even

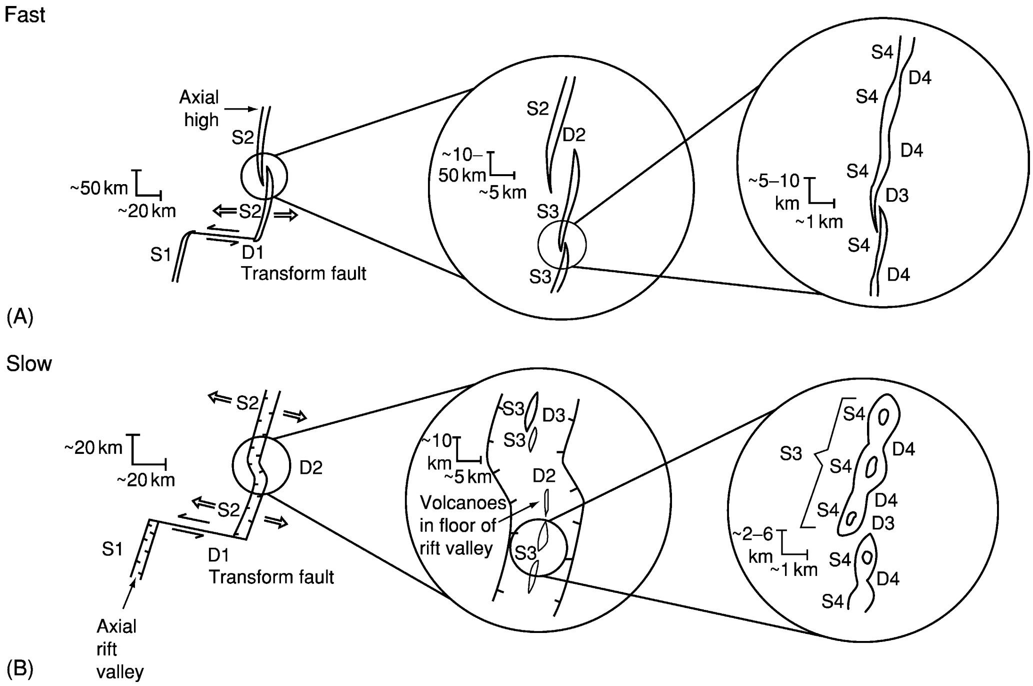


Figure 4 A possible hierarchy of ridge segmentation for (A) fast-spreading and (B) slow-spreading ridges. S1–S4 are ridge segments of orders 1–4, and D1–D4 are ridge-axis discontinuities of orders 1–4. At both fast-spreading and slow-spreading centres, first-order discontinuities are transform faults. Examples of second-order discontinuities are overlapping spreading centres on fast-spreading ridges and oblique shear zones on slow-spreading ridges. Third-order discontinuities are small overlapping spreading centres on fast-spreading ridges. Fourth-order discontinuities are slight bends or lateral offsets of the axis of less than 1 km on fast-spreading ridges. This four-tiered hierarchy of segmentation probably represents a continuum; it has been established, for example, that fourth-order segments and discontinuities can grow to become third-, second-, and even first-order features and vice versa at both slow-spreading and fast-spreading centres. (Reproduced from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid-ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986–994.)

those with small offsets. This is thought to be the result of highly focused mantle upwelling near the mid-segment regions, with very little along-axis flow of magma away from the upwelling region. Focused upwelling is inferred from 'bulls-eye'-shaped residual gravity anomalies and variations in crustal thickness that have been documented by seismic refraction and studies of microearthquakes. At slow-spreading centres, melt probably resides in small, isolated, and very short-lived pockets beneath the median valley floor (Figure 5) and beneath elongate axial volcanic ridges. An alternative view is that the observed along-strike variations in topography and crustal thickness can be accounted for by along-strike variations in mechanical thinning of the crust by faulting. There is no conflict between these models, so both focused upwelling and mechanical thinning may occur along each segment.

One might expect the same to hold at fast-spreading centres, i.e. crustal thinning adjacent to overlapping spreading centres. This does not appear to be the case at 9° N on the East Pacific Rise, where seismic data suggest a thickening of the crust towards the overlapping spreading centres and a widening of the axial magma chamber reflector. There is no indication of crustal thinning near the Clipperton transform fault either. And yet, as one approaches the 9° N overlapping spreading centres from the north, the axial depth plunges, the axial cross-sectional area decreases, the axial magma chamber reflector deepens, the average lava age increases, the MgO content of dredged basalts decreases, hydrothermal activity decreases dramatically, crustal magnetization increases significantly (suggesting eruption of more fractionated basalts in a region of decreased magma supply), crustal fracturing and inferred depth of fracturing increase (indicating a greater ratio of extensional strain to magma supply), and the throw of off-axis normal faults increases (suggesting thicker lithosphere and greater strain) (Figure 8A). How can these parameters all correlate so well, indicating a decrease in the magmatic budget and an increase in amagmatic extension, yet the seismic data suggest crustal thickening off-axis from and a wider magma lens near the overlapping spreading centres?

One possibility is that mantle upwelling and the axial magmatic budget are enhanced away from ridge-axis discontinuities even at fast-spreading

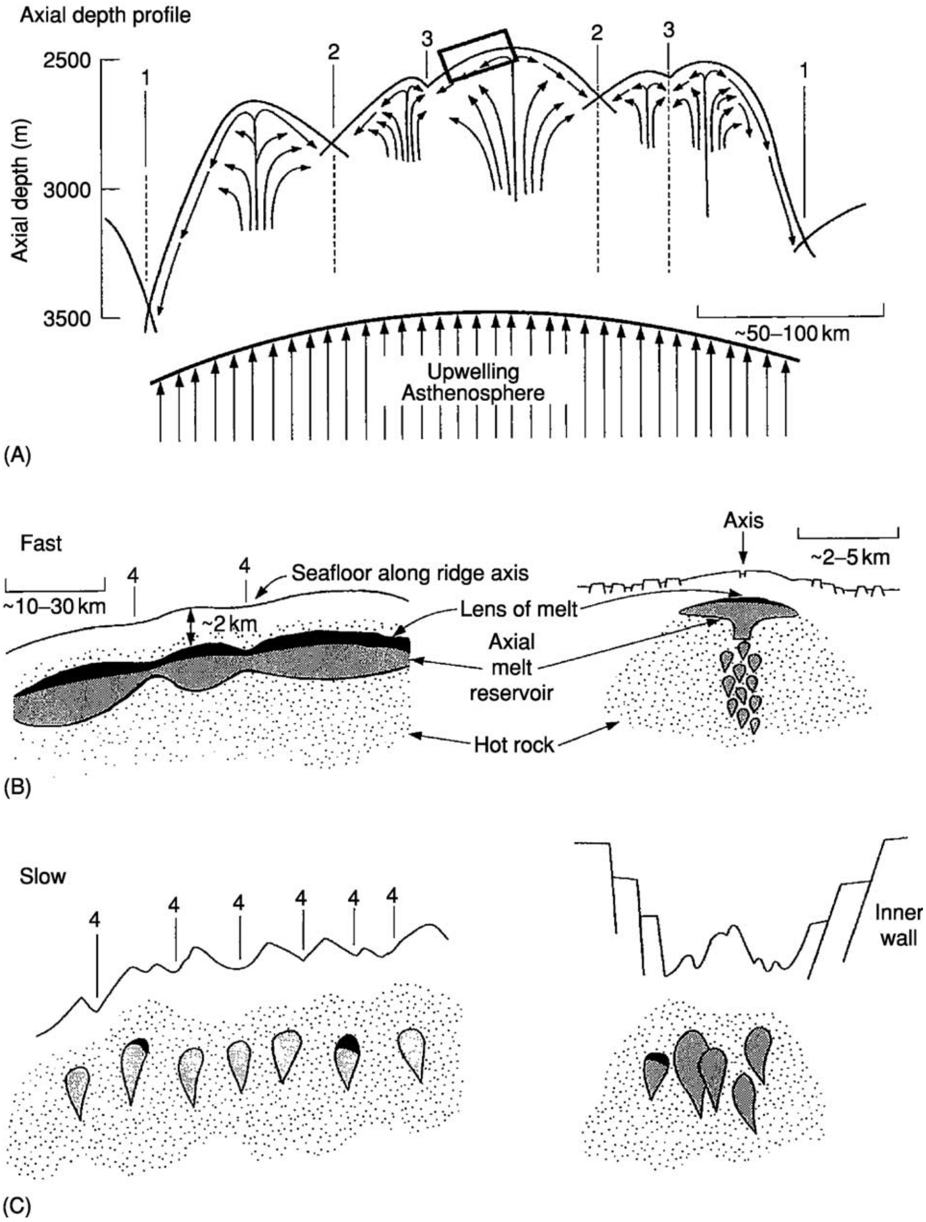


Figure 5 (A) How ridge segmentation may be related to mantle upwelling, and (B and C) the distribution of magma supply. In (A), the depth scale applies only to the axial depth profile; numbers denote discontinuities and segments of orders 1–3. Decompression partial melting in the upwelling asthenosphere occurs at depths of 30–60 km beneath the ridge. As the melt ascends through a more slowly rising solid residuum, it is partitioned at different levels to feed segments of orders 1–3. Mantle upwelling is hypothesized to be 'sheetlike' in the sense that melt is upwelling along the entire length of the ridge, but the supply of melt is thought to be enhanced beneath shallow parts of the ridge away from major discontinuities. The rectangle indicates the area enlarged to show fine-scale segmentation for (B) a fast-spreading example and (C) a slow-spreading example. In (B) and (C) along-strike cross-sections showing the hypothesized partitioning of the magma supply relative to fourth-order discontinuities (4s) and segments are shown on the left. Across-strike cross-sections for fast-spreading and slow-spreading ridges are shown on the right. (Reproduced from Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid-ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986–994.)

centres, but subaxial flow of magma 'downhill' away from the injection region redistributes the magma episodically (Figure 5). This along-strike flow and redistribution of magma may be unique to spreading centres with an axial high, such as the East Pacific Rise or Reykjanes, where the axial region is sufficiently hot at shallow depths to facilitate subaxial flow. It is well-documented in Iceland and other volcanic areas analogous to mid-ocean ridges that magma can flow in subsurface chambers and dykes

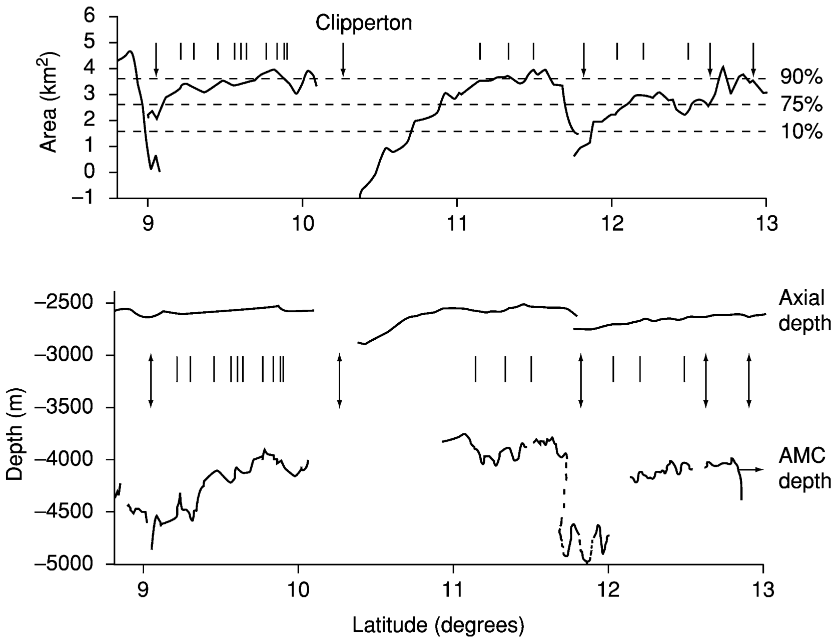


Figure 6 Profiles of the along-axis cross-sectional area, axial depth, and axial-magma-chamber (AMC) seismic-reflector depth for the East Pacific Rise 9°–13° N. The locations of first- and second-order discontinuities are denoted by vertical arrows (first-order discontinuities are named); each occurs at a local minimum of the ridge area profile and a local maximum of the ridge-axis depth smaller discontinuities are denoted by vertical bars. There is an excellent correlation between ridge-axis depth and cross-sectional area; there is a good correlation between cross-sectional area and the existence of an axial magma chamber, but detailed characteristics of the axial magma chamber (depth, width) do not correlate. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

for many tens of kilometres away from the source region before erupting. In this way, thicker crust may occur away from the mid-segment injection points, proximal to discontinuities such as overlapping spreading centres.

Based on studies of the fast-spreading East Pacific Rise, a 'magma supply' model has been proposed, which explains the intriguing correlation between over a dozen structural, geochemical, and geophysical variables within a first-, second-, or third-order segment (Figure 9). It also addresses the initially puzzling observation that the crust is sometimes thinner in the mid-segment region, where upwelling is supposedly enhanced. Intuitively, one might expect the crust to be thickest over the region where upwelling is enhanced, as is observed on the Mid-Atlantic Ridge. However, along-axis redistribution of melt may be the controlling factor on fast-spreading ridges, where the subaxial melt region may be well-connected for tens of kilometres. In this model, temporal variations in along-axis melt connectivity may result in thicker crust near the mid-segment when connectivity is low (most often at slow-spreading ridges) and thicker crust closer to the segment ends when connectivity is high (most often, but not always, at fast-spreading ridges).

The basic concepts of this magma-supply model also apply to slow-spreading ridges that are characterized by an axial rift valley. Mantle melting is enhanced beneath the mid-segment regions. However, the axial region is colder (averaged over time), and along-strike redistribution of melt is impeded. Thus, the crust tends to be thickest near the mid-segment regions and thinnest near ridge-axis discontinuities (Figures 8B and 9).

It is possible that the segmentation of mid-ocean ridges and the observations that correlate with segmentation (e.g. axial depth, geochemistry of lavas, lava morphology, etc.; Figure 8) are not related to the supply of magma to the ridge. This is still an area of active research and debate. For example, it has been suggested that the supply of melt is uniform along fast-spreading ridges and that along-strike variations are caused by differences in hydrothermal heat loss. If heat loss were enhanced near segment ends, this could cause many of the along-strike variations noted in Figure 8. So far, however, there is no indication that hydrothermal heat loss is greater near segment ends. On the contrary, hydrothermal heat loss is least evident at segment ends and is often enhanced near the shallow mid-sections of first-, second-, and third-order segments.

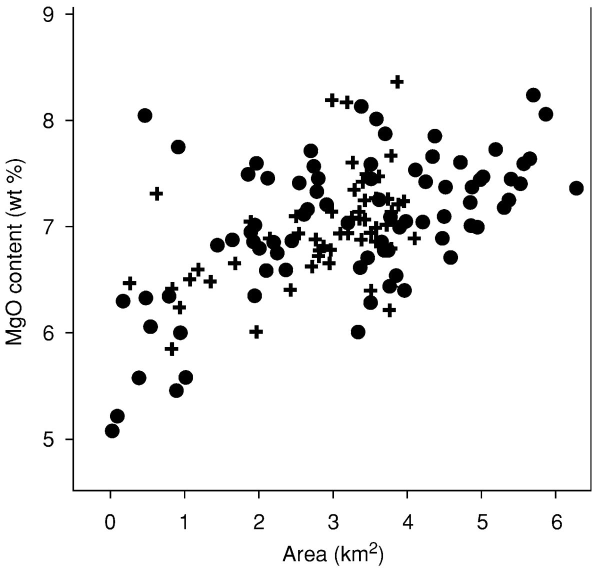


Figure 7 Cross-sectional area of the East Pacific Rise plotted against the MgO content of basalt glass (crosses from 5–14° N; circles from 13–23° S). There is a tendency for high MgO contents (interpreted as higher eruption temperatures and perhaps a higher magmatic budget) to correlate with larger cross-sectional areas. Smaller cross-sectional areas are correlated with lower levels of MgO and a greater scatter in MgO content, suggesting magma chambers that are transient and changing. Thus shallow inflated areas of the ridge tend to erupt hotter lavas. Updated from Scheirer and Macdonald (1993) and references therein. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid-ocean ridge tectonics, volcanism and geomorphology*, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

Fine-Scale Variations in Ridge Morphology Within the Axial Neovolcanic Zone

The axial neovolcanic zone occurs on or near the axis of the axial high of fast-spreading centres and within the floor of the rift valley of slow-spreading centres (Figures 2A). Studies of the widths of the polarity transitions of magnetic anomalies, including *in situ* measurements from ALVIN, document that approximately 90% of the volcanism that creates the extrusive layer of oceanic crust occurs in a region 1–5 km wide at most spreading centres. Direct qualitative estimates of lava age at spreading centres using submersibles and remotely operated vehicles tend to confirm this, as do recent high-resolution seismic measurements that show that layer 2A (interpreted to be the volcanic layer) achieves its full thickness within 1–3 km of the rise axis (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges). However, there are significant exceptions, including small-volume off-axis volcanic constructions and voluminous off-axis floods of basaltic sheet flows.

The axial high on fast-spreading and intermediate-spreading centres is usually bisected by an axial summit trough approximately 10–200 m deep, which is found along approximately 60–70% of the axis. Along the axial high of fast-spreading ridges, side-scan sonar records show that there is an excellent correlation between the presence of an axial summit trough and an axial magma chamber reflector as seen on multichannel seismic records (over more than 90% of the ridge length). Neither axial summit troughs nor axial magma chambers occur where the ridge has a very small cross-sectional area.

In rare cases, an axial summit trough is not observed where the cross-sectional area is large. In these locations, volcanic activity is presently occurring or has occurred within the last decade. For example, on the East Pacific Rise, near 9°45′–9°52′ N, a volcanic eruption documented from the submersible ALVIN was associated with a single major dyke intrusion, similar to the 1993 eruption on the Juan de Fuca Ridge. Side-scan sonar records showed that an axial trough was absent from 9°52′ N to 10°02′ N and, in subsequent dives, it was found that dyke intrusion had propagated into this area producing very recent lava flows and hydrothermal activity complete with bacterial ‘snowstorms’. A similar situation has been thoroughly documented at 17°25′–17°30′ S on the East Pacific Rise, where the axial cross-sectional area is large but the axial summit trough is partly filled. Perhaps the axial summit trough has been flooded with lava so recently that magma withdrawal and summit collapse are still occurring. Thus, the presence of an axial summit trough along the axial high of a fast-spreading ridge is a good indicator of the presence of a subaxial lens of partial melt (axial magma chamber); where an axial summit trough is not present but the cross-sectional area is large, this is a good indicator of very recent or current volcanic eruptions; where an axial summit trough is not present and the cross-sectional area is small, this is a good indicator of the absence of a magma lens (axial magma chamber).

In contrast to the along-axis continuity of the axial neovolcanic zone of fast-spreading ridges, the neovolcanic zone of slower-spreading ridges is considerably less continuous and there is a great deal of variation from segment to segment. Volcanic constructions, called axial volcanic ridges, are most common along the shallow mid-segment regions of the axial rift valley. Near the ends of segments, where the rift valley deepens, widens, and is truncated by transform faults or oblique shear zones, the gaps between axial volcanic ridges become longer. The gaps between axial volcanic ridges are regions of older crust, characterized by faulting and a lack of recent volcanism.

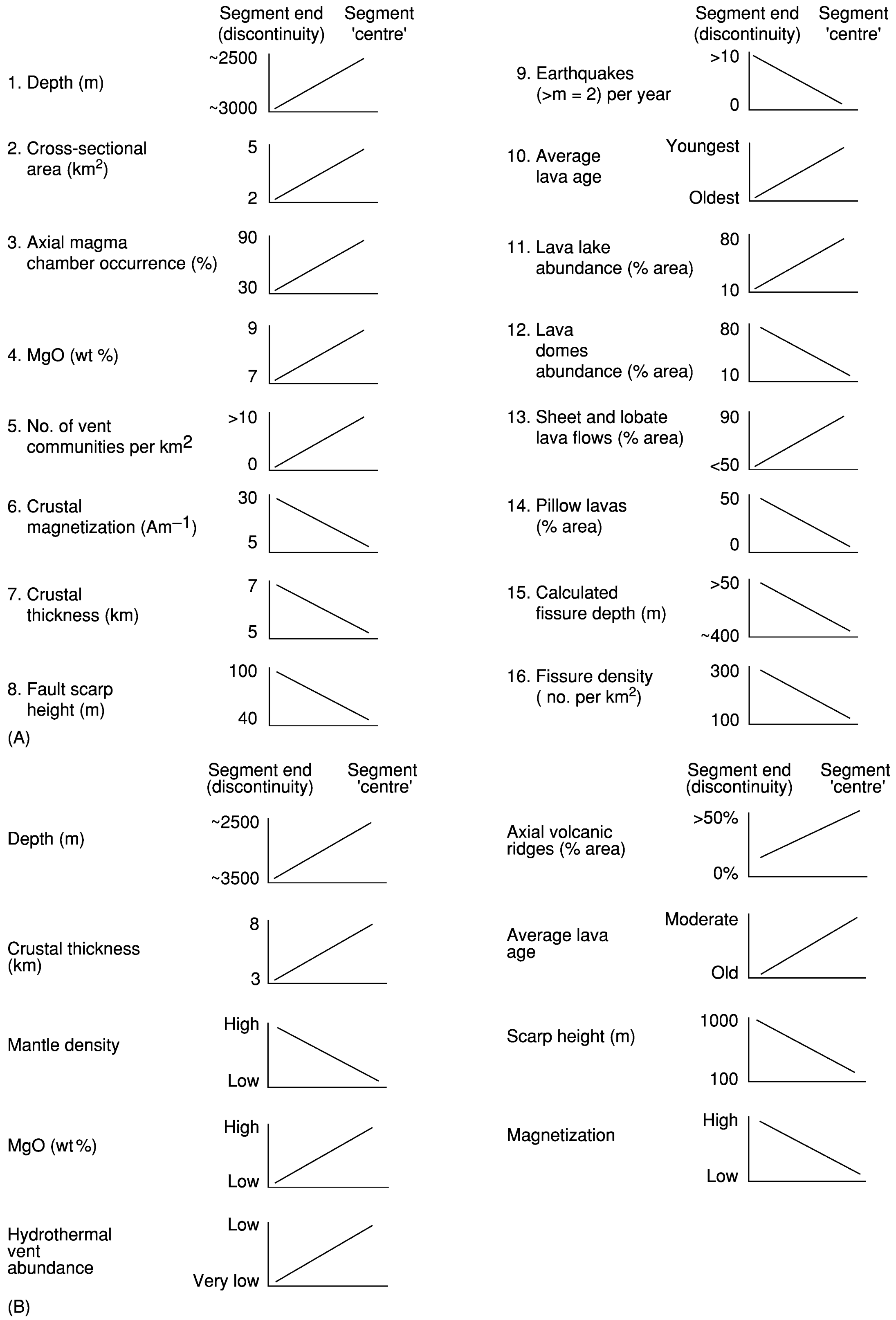


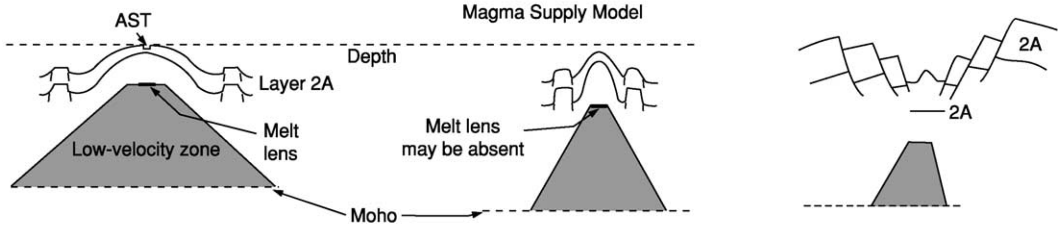
Figure 8 Summary of along-axis variations in spreading-centre properties from segment end (discontinuity of order 1, 2, or 3) to segment mid-section areas for (A) fast-spreading ridges with axial highs and (B) slow-spreading ridges with axial rift valleys. A large number of parameters correlate well with location within a given segment, indicating that segments are distinct independent units of crustal accretion and deformation. These variations may reflect a fundamental segmentation of the supply of melt beneath the ridge.

Continued

These gaps may correspond to fine-scale (third- and fourth-order) discontinuities of the ridge.

There is another important difference between volcanism on fast-spreading and slow-spreading ridges. Axial volcanic ridges represent a thickening of the volcanic layer atop a lithosphere that may be 5–10 km thick, even on the axis. In contrast, the volcanic layer is usually thinnest along the axis of the fast-spreading East Pacific Rise (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges). Thus, the axial high of fast-spreading ridges is not a thickened accumulation of lava, while the discontinuous axial volcanic ridges of slow-spreading ridges are.

On both slow-spreading and fast-spreading ridges, pillow and lobate lavas are the most common lava morphologies. Based on laboratory studies and observations of terrestrial basaltic eruptions, this indicates that the lava effusion rates are slow to moderate on most mid-ocean ridges. High volcanic effusion rates, indicated by fossil lava lakes and extensive outcrops



(A) High magma budget	(B) Moderate to low magma budget	(C) Very low, sporadic magma supply
Shallow water depth	Deeper water depth	Deep rift valley floor
Broad cross-sectional area	Small cross-sectional area	Wide valley floor
Wide low-velocity zone High MgO content High temperature Low density	Narrow (except where 2 LVZ coalesce at an OSC)* Lower MgO content Lower temperature Higher density	Narrow low-velocity zone
Melt lens ubiquitous, depth and width variable*	Melt lens less common, highly variable width, depth where present, disrupted at RADs*	Melt lens very rare
Thinner crust (melt flows 'downhill' away from mantle upwelling zone)*	Thicker crust, but highly variable*	Highly variable crustal thickness
Axial summit trough present unless in eruptive phase*	Axial summit trough rare	
Hydrothermal venting abundant*	Hydrothermal venting rare*	Hydrothermal vents rare, fault controlled
Less crustal fissuring*	More crustal fissuring*	Extensive fissuring
Younger average lava age*	Older average lava age*	Highly variable but generally older lava age, mostly pillows
Lower crustal magnetization*	Higher crustal magnetization*	
Smaller throw on flanking fault scarps	Larger throw on flanking fault scarps	Large fault scarps, especially at inside corner highs
Layer 2A thinnest along axis	Layer 2A thinnest along axis	Layer 2A thick beneath axial volcanic ridges

Figure 9 Magma-supply model for mid-ocean ridges. (A) A segment with a robust magmatic budget, generally a fast-spreading ridge away from discontinuities or a hotspot-dominated ridge with an axial high. (B) A segment with a moderate magma budget, generally a fast-spreading ridge near a discontinuity or a non-rifted intermediate-rate ridge. (C) A ridge with a sporadic and diminished magma supply, generally a rifted intermediate-to-slow-spreading centre (for along-strike variations at a slow ridge, see Figure 8B). AST, axial summit trough; LVZ, low-velocity zone; OSC, overlapping spreading centres; RAD, ridge-axis discontinuity. See references in Buck *et al.* (1998) (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

(Less than 1% of the ridge has been studied in sufficient detail to create this summary.) (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

of sheet-flow lavas, are very rare on slow-spreading ridges. High-effusion-rate eruptions are more common on fast-spreading ridges and are more likely to occur along the shallow inflated mid-segment regions of the rise, in keeping with the magma-supply model for ridges discussed earlier (Figure 9A).

Very little is known about eruption frequency. It has been estimated based on some indirect observations that, at any given place on a fast-spreading ridge, eruptions occur approximately every 50–100 years, and that on slow-spreading ridges eruptions occur approximately every 5000–10 000 years. If this is true, then the eruption frequency varies inversely with the square of the spreading rate. On intermediate- to fast-spreading centres, if one assumes a typical dyke width of around 50 cm and a spreading rate of 5–10 cm year^{-1}, then an eruption could occur approximately every 5–10 years. This estimate is in reasonable agreement with the occurrence of megaplumes and eruptions on the well-monitored Juan de Fuca Ridge. However, observations of sheeted-dyke sequences in Iceland and of ophiolites indicate that only a small percentage of the dykes reach the surface to produce eruptions.

On fast-spreading centres, the axial summit trough is so narrow (30–1000 m) and well-defined in most places that tiny offsets and discontinuities of the rise axis can be detected (Table 1 and Figure 2). This finest scale of segmentation (fourth-order segments and discontinuities) probably corresponds to individual fissure eruption events similar to the Krafla eruptions in Iceland and the Kilauea east rift zone eruptions in Hawaii. Given a magma chamber depth of 1–2 km, an average dyke ascent rate of approximately 0.1 km h^{-1}, and an average lengthening rate of approximately 1 km h^{-1}, typical diking events would give rise to segments 10–20 km long. This agrees with observations of fourth-order segmentation and the scale of recent diking events on the Juan de Fuca Ridge and in other volcanic rift zones. The duration of such segments is thought to be very short, of the order of 100–1000 years (too brief in any case to leave even the smallest detectable trace off-axis; Table 1). Yet, even at this very fine scale, excellent correlations can be seen between average lava age, density of fissuring, the average widths of fissures, and the abundance of hydrothermal vents within individual segments. In fact, there is even an excellent correlation between ridge cross-sectional area and the abundance of benthic hydrothermal communities (Figure 8).

A curious observation on the East Pacific Rise is that the widest fissures occur in the youngest lava fields. If fissures widen over time with increasing extension, one would expect the opposite: the widest fissures should be in the oldest areas. The widest fissures are approximately 5 m wide. Using simple fracture mechanics, it can be concluded that these fissures probably extend all the way through layer 2A and into the sheeted-dyke sequence. These have been interpreted as eruptive fissures, and this is where high-temperature vents (more than 300 °C) are concentrated. In contrast to the magma-rich dyke-controlled hydrothermal systems that are common on fast-spreading centres, magma-starved hydrothermal systems on slow-spreading ridges tend to be controlled more by the penetration of seawater along faults near the ridge axis.

Faulting

Extension at mid-ocean ridges causes fissuring and normal faulting. The lithosphere is sufficiently thick and strong on slow-spreading centres to support shear failure on the axis, so normal faulting along dipping fault planes can occur on or very close to the axis. These faults produce grabens that are 1–3 km deep. In contrast, normal faulting is not common on fast-spreading centres within 2 km of the axis, probably because the lithosphere is too thin and weak to support normal faulting. Instead, the new thin crust fails by simple tensional cracking.

Fault strikes tend to be perpendicular to the direction of least compressive stress; thus, they also tend to be perpendicular to the spreading direction. While there is some ‘noise’ in the fault trends, most of this noise can be accounted for by perturbations in the direction of least compressive stress due to shearing in the vicinity of active or fossil ridge-axis discontinuities. Once this is accounted for, fault trends faithfully record changes in the direction of opening to within ±3° and can be used to study plate-motion changes on a finer scale than that provided by the study of seafloor magnetic anomalies. Studies of the cumulative throw of normal faults, seismicity, and fault spacing suggest that most faulting occurs within 20–40 km of the axis irrespective of spreading rate.

The occurrence of inward- and outward-dipping faults depends on the spreading rate. Most faults (*ca.* 80%) dip towards the axis on slow-spreading centres, but there is a monotonic increase in the frequency of outward-dipping faults with spreading rate (Figure 10). Inward- and outward-facing faults are approximately equally abundant at very fast spreading rates. This can be explained by the smaller mean normal stress across a fault plane that dips towards the axis, cutting through thin lithosphere, than across a fault plane that dips away from the axis, cutting through a much thicker section of lithosphere. Given reasonable thermal models, the difference in the

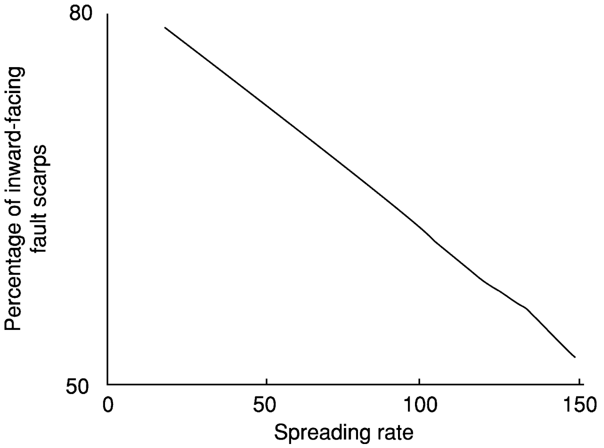


Figure 10 The effect of spreading rate on the percentage of fault scarps that are inward-facing (facing towards the spreading axis rather than away from the spreading axis). A significant increase in the percentage of inward-facing scarps occurs at slower spreading rates (mm yr). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, Mid-ocean ridge tectonics, volcanism and geomorphology, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

thickness of the lithosphere cut by planes dipping towards and away from the axis (and the mean normal stresses across those planes) decreases significantly with spreading rate, making outward-dipping faults more likely at fast spreading rates.

At all spreading rates, important along-strike variations in faulting occur within major (first-order and second-order) spreading segments. Fault throws (inferred from scarp heights) decrease in the mid-segment regions away from discontinuities (Figures 8 and 11). This may be caused by a combination of thicker crust, thinner lithosphere, greater magma supply, and less amagmatic extension away from ridge-axis discontinuities in the mid-segment region (Figure 11). Another possible explanation for along-strike variations in fault throw is along-strike variation in the degree of coupling between the mantle and crust. A ductile lower crust will tend to decouple the upper crust from extensional stresses in the mantle, and the existence of a ductile lower crust will depend on spreading rate, the supply of magma to the ridge, and proximity to major discontinuities.

Estimates of crustal strain due to normal faulting vary from 10–20% on the slow-spreading Mid-Atlantic Ridge to *ca.* 3–5% on the fast-spreading East Pacific Rise. This difference may be explained as follows. The rate of magma supply to slow-spreading ridges is relatively low compared with the rate of crustal extension and faulting, while extension and magma-supply rates are in closer balance on fast-spreading ridges. The resulting seismicity is different too. In contrast to faulting on slow-spreading ridges, where teleseismically detected earthquakes are common, faulting on fast-spreading ridges rarely produces earthquakes with magnitudes of more than 5. Nearly all of these events are associated with ridge-axis discontinuities. The level of seismicity measured at fast-spreading ridges accounts for only a very small percentage of the observed strain due to faulting, whereas fault strain at slow-spreading ridges is comparable to the observed seismic moment release. It has been suggested that faults in fast-spreading environments accumulate slip largely by stable sliding (aseismically) owing to the warm temperatures and associated thin brittle layer. At slower spreading rates, faults will extend beyond a frictional stability transition into a field where fault slip occurs unstably (seismically) because of a thicker brittle layer.

Disruption of oceanic crust by faulting may be particularly extreme on slow-spreading ridges near transform faults (Figure 12). Unusually shallow topography occurs on the active transform slip side of ridge–transform intersections; this is called the inside corner high. These highs are not volcanoes. Instead they are caused by normal faults that cut deeply into and perhaps all the way through the oceanic crust. It is thought that crustal extension may occur for 1–2 Ma on detachment faults with little magmatic activity. This results in extraordinary extension of the crust and exposure of large sections of the deep crust and upper mantle on the seafloor. Corrugated slip surfaces indicating the direction of fault slip are also evident and are called 'megamullions' by some investigators.

Disruption of oceanic crust by faulting may also reach extremes at the slowest known spreading rates of 0.5–1.0 cm year^{-1}, for example on the Gakkel Ridge in the Arctic Ocean and along parts of the Southwest Indian Ocean Ridge. There is evidence that crustal extension due to faulting exceeds 20% and that there are prolonged periods with no basaltic volcanism along the rift valley. Instead, upper-mantle peridotites upwell to fill the gap between the separating plates during the intervals between infrequent, but sometimes very large, basaltic eruptions.

At distances of several tens of kilometres from the axis, topography generated near the spreading centre is preserved on the seafloor with little subsequent change, except for the gradual accumulation of pelagic sediments at rates of approximately 0.5–20 cm per thousand years, until it is subducted. The preserved topographic highs and lows are called abyssal hills. At slow-spreading centres characterized by an axial rift valley, back-tilted fault blocks and half-grabens may be the dominant origins of abyssal hills (Figure 13),

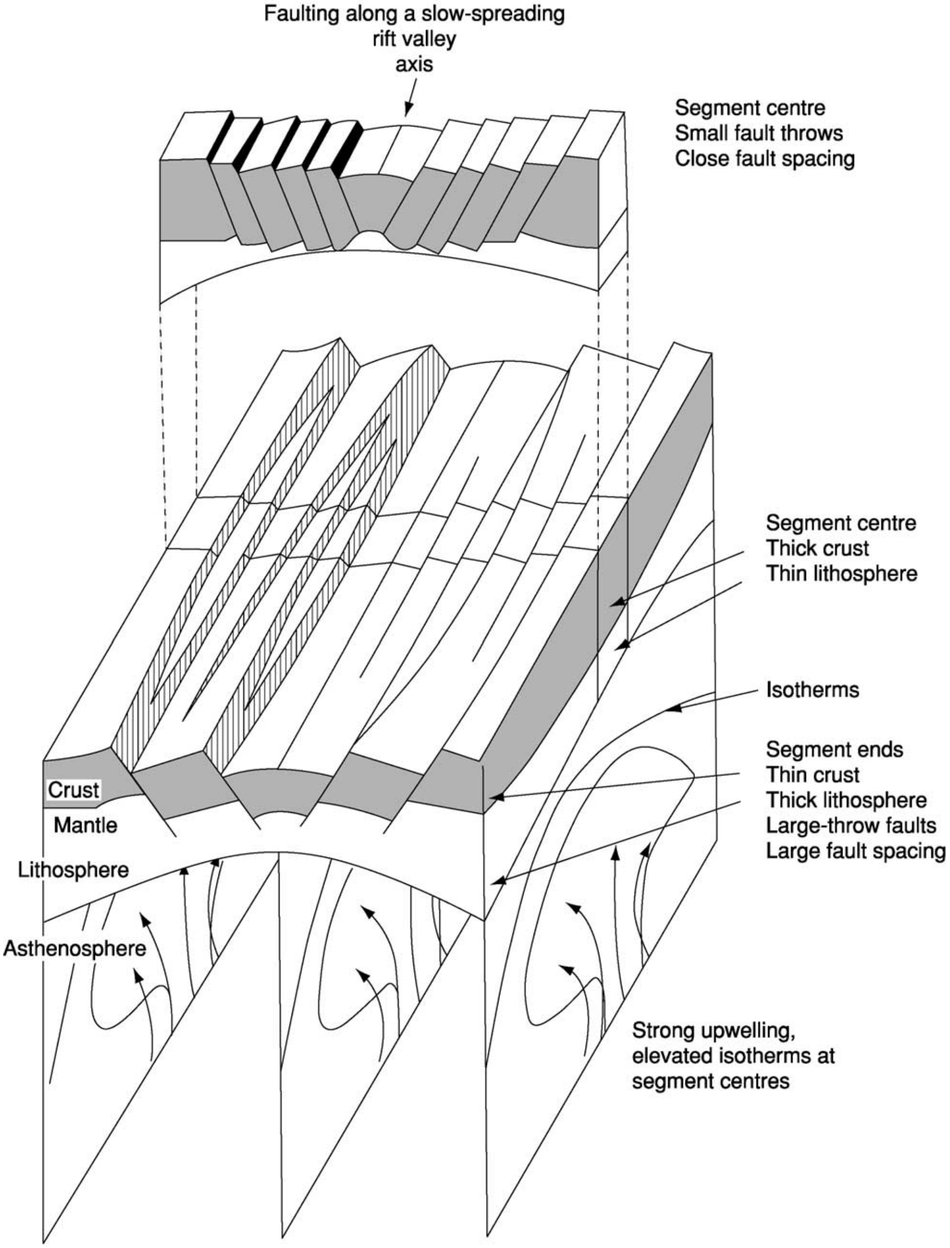


Figure 11 A geological interpretation of along-axis variations in scarp height and the occurrence of more closely spaced scarps near mid-segments on a slow-spreading ridge. The cross-section through the segment centre (top) shows more closely spaced smaller-throw faults than at the segment ends (bottom). Focused mantle upwelling near the segment centre causes this region to be hotter: the lithosphere will be thinner, while increased melt supply will create a thicker crust. In contrast to observations at fast-spreading centres, there may be very little melt redistribution along strike. Near the segment ends, the lithosphere will be thicker and magma supply will be less, creating thinner crust. Along-axis variations in scarp height and spacing reflect these along-axis variations in lithospheric thickness. Amagmatic extension across the larger faults near segment ends may also thin the crust, especially at inside corner highs. (Reproduced from Shaw PR (1992) Ridge segmentation, faulting and crustal thickness in the Atlantic Ocean. *Nature* 358: 490–493.)

although there is continued controversy over the roles of high-versus low-angle faults and listric faulting versus planar faulting and the possible role of punctuated episodes of volcanism versus amagmatic extension. At intermediate-rate spreading centres, abyssal-hill structure may vary with the local magmatic budget. Where the budget is starved and the axis is characterized by a rift valley, abyssal hills are

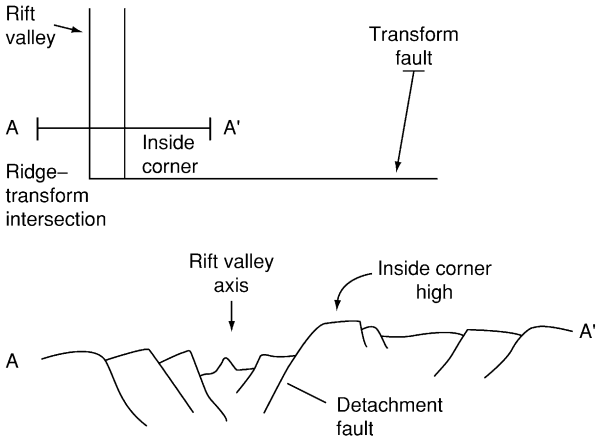


Figure 12 Inside corner high at a slow-spreading ridge–transform intersection. Extension is concentrated along a detachment fault for up to 1–2 Ma, exposing deep sections of oceanic crust and mantle. The oceanic crust is thinned by this extreme extension; crustal accretion and magmatic activity may also be diminished. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid-ocean ridge tectonics, volcanism and geomorphology*, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

generally back-tilted fault blocks. Where the magmatic budget is robust and an axial high is present, the axial lithosphere is episodically thick enough to support a volcanic construction, which may then be rafted away intact or split in two by the spreading axis, resulting in whole-volcano and split-volcano abyssal hills, respectively.

Based on observations made from the submersible ALVIN on the flanks of the East Pacific Rise, it would appear that the outward-facing slopes of the hills are neither simple outward-dipping normal faults, as would be predicted by the horst-and-graben model, nor entirely of volcanic construction, as would be predicted by the split-volcano model. Instead, the outward-facing slopes are 'volcanic growth faults' (Figure 14). Outward-facing scarps produced by episodes of normal faulting are buried near the axis by syntectonic lava flows originating along the axial high. Repeated episodes of dip-slip faulting and volcanic burial result in structures resembling growth faults, except that the faults are episodically buried by lava flows rather than being continuously buried by sediment deposition. In contrast, the inward-dipping faults act as tectonic dams to lava flows. Thus, the abyssal hills are horsts and the intervening troughs are grabens, with the important modification to the horst-and-graben model that the outward-facing slopes are created by volcanic growth faulting rather than traditional normal faulting. Thus, on fast-spreading centres, abyssal hills are asymmetric, being bounded by steeply dipping normal faults

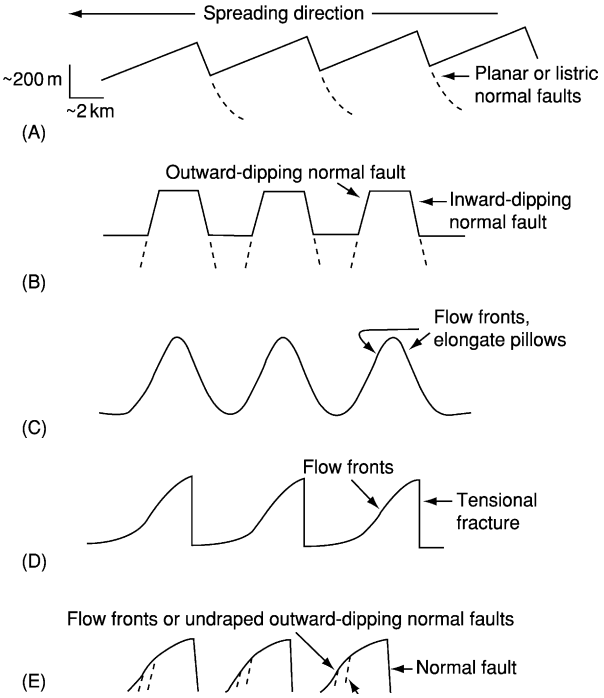

Spreading direction
~200 m
~2 km
Planar or listric normal faults
(A)
Outward-dipping normal fault
Inward-dipping normal fault
(B)
Flow fronts, elongate pillows
(C)
Flow fronts
Tensional fracture
(D)
Flow fronts or undraped outward-dipping normal faults
Normal fault
(E)

Figure 13 Five models for the development of abyssal hills on the flanks of mid-ocean ridges. (A) Back-tilted fault blocks (episodic inward-dipping normal faulting off-axis). (B) Horst and graben (episodic inward- and outward-dipping faulting off-axis). (C) Whole volcanoes (episodic volcanism on-axis). (D) Split volcanoes (episodic volcanism and splitting on-axis). (E) Horsts bounded by inward-dipping normal faults and outward-dipping volcanic growth faults (episodic faulting off-axis and episodic volcanism on-axis or near-axis). (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid-ocean ridge tectonics, volcanism and geomorphology*, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

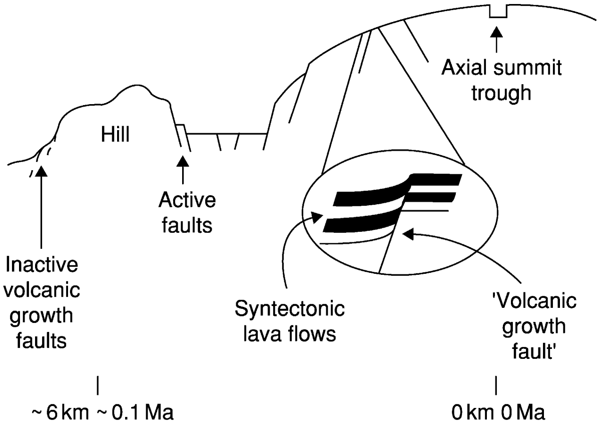


Figure 14 Cross-sectional depiction of the development of volcanic growth faults. Volcanic growth faults are common on fast-spreading centres and explain some of the differences between inward- and outward-facing scarps as well as the morphology and origin of most abyssal hills on fast-spreading centres. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid-ocean ridge tectonics, volcanism and geomorphology*, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

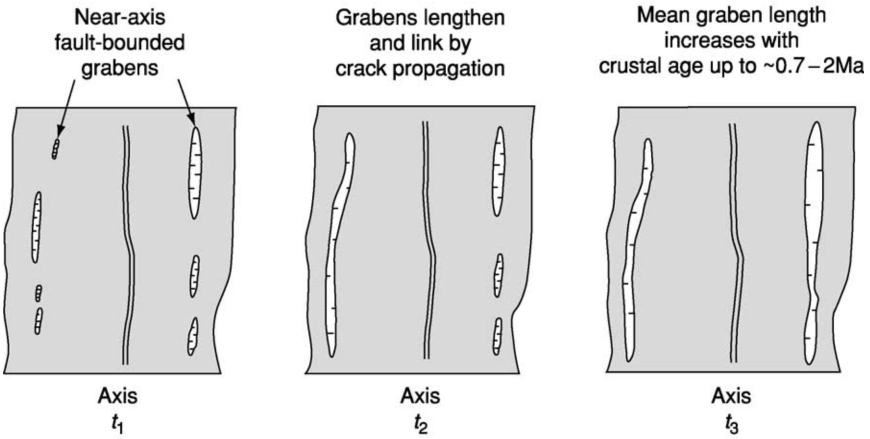


Figure 15 Proposed time sequence of along-strike propagation and linkage of near-axis faults and grabens, which define the edges of abyssal hills; time-averaged propagation rates are approximately 20–60 km Ma^{-1}. (Reprinted from *Encyclopedia of Ocean Sciences*, Steele J, Thorpe S, and Turekian K (eds.), Macdonald KC, *Mid-ocean ridge tectonics, volcanism and geomorphology*, pp. 1798–1813, Copyright (2001), with permission from Elsevier.)

facing the spreading axis and volcanic growth faults on the opposing side. These faults lengthen at a rate of 20–60 Km/Ma reaching total lengths of 10–70 km to form abyssal hills and intervening grabens (Figure 15).

See Also

Analytical Methods: Gravity. **Lava**. **Magnetostratigraphy**. **Plate Tectonics**. **Seamounts**. **Tectonics:** Hydrothermal Vents At Mid-Ocean Ridges; Seismic Structure At Mid-Ocean Ridges.

Further Reading

Buck WR, Delaney PT, Karson JA, and Lagabrielle Y (1998) *Faulting and Magmatism at Mid-Ocean Ridges*. Geophysical Monograph 106. Washington, DC: American Geophysical Union.

Fox PJ and Gallo DG (1986) The geology of North Atlantic transform plate boundaries and their aseismic extensions. In: Vogt PR and Tucholke BE (eds.) *The Western North Atlantic Region: The Geology of North America*, pp. 157–172. Boulder: Geological Society of America.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thompson RE (1995) *Seafloor hydrothermal systems: physical, chemical, biological and geological interactions*. AGU Geophysical Monograph 91. Washington, DC: American Geophysical Union.

Langmuir CH, Bender JF, and Batiza R (1986) Petrological and tectonic segmentation of the East Pacific Rise, 5°30′ N–14°30′ N. *Nature* 322: 422–429.

Macdonald KC (1982) Mid-ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annual Review of Earth and Planetary Sciences* 10: 155–190.

Macdonald KC and Fox PJ (1990) The mid-ocean ridge. *Scientific American* 262: 72–79.

Macdonald KC, Scheirer DS, and Carbotte SM (1991) Mid-ocean ridges: discontinuities, segments and giant cracks. *Science* 253: 986–994.

Menard H (1986) *Ocean of Truth*. Princeton: Princeton University Press.

Phipps-Morgan J, Blackman DK, and Sinton J (1992) *Mantle Flow and Melt Generation at Mid-Ocean Ridges*. Geophysical Monograph 71. Washington, DC: American Geophysical Union.

Shaw PR (1992) Ridge segmentation, faulting and crustal thickness in the Atlantic Ocean. *Nature* 358: 490–493.

Hydrothermal Vents At Mid-Ocean Ridges

R M Haymon, University of California–Santa Barbara, Santa Barbara, CA, USA

Introduction

On the Galapagos Rift in 1977, ~2600 m beneath the sunlit surface of the sea, scientists exploring the lightless, frigid seafloor were astonished to discover springs of warm water teeming with life (Figure 1A). It was quickly realized these remarkable ecosystems in the deep sea are supported by microbes capable of metabolizing chemicals dissolved in hydrothermal spring waters. In 1979, submersible divers exploring the crest of the East Pacific Rise came on superheated (380 ± 30°C) hot springs, where plumes of scalding fluid, blackened by minuscule mineral particles, billowed into the ocean through tall mineral conduits (Figure 1B). The minerals formed at these 'black smoker' hydrothermal vents proved to be rich in copper, iron, zinc, and other useful metals.

Hundreds of hydrothermal vents have now been located on the crest and flanks of the mid-ocean ridge, and more are found every year. The mid-ocean ridge is a globe-encircling rift where plates of ocean lithosphere are repeatedly ripped apart (*see* **Tectonics:** Mid-Ocean Ridges). New, hot seafloor freezes in the cracks between the plates and is cooled by hydrothermal circulation that entrains enormous volumes of seawater. The great magnitude of hydrothermal fluid flow through the mid-ocean ridge influences the chemistry, biology, and physical oceanography of the world ocean; the thermal structure, physical properties, and chemical composition of ocean crust; and the nature and diversity of sub-seafloor microbial environments. Many now believe that submarine hot springs may have been the biochemical crucibles for the origin of life in the solar system.

For thousands of years before mid-ocean ridge hot springs were discovered in the oceans, people mined copper from mineral deposits that were originally formed on oceanic spreading ridges. These fossil deposits are embedded in old fragments of seafloor, called 'ophiolites', that have been uplifted and emplaced onto land by fault movements. The copper-rich mineral deposits in the Troodos ophiolite of Cyprus are well-known examples of fossil ocean ridge deposits that have been mined for at least 2500 years; in fact, the word 'copper' is derived from the Latin word *cyprium*, which means 'from Cyprus'.

The mineral deposits accumulating today at hot springs along the mid-ocean ridge are habitats for a variety of remarkable organisms, ranging in size from tiny microbes to 2-m-long tubeworms. The properties of the mineral deposits are inextricably linked to the organisms that inhabit them. The mineral deposits contain important clues about the physical–chemical environments in which some of these organisms live, and also preserve fossils of some organisms, creating

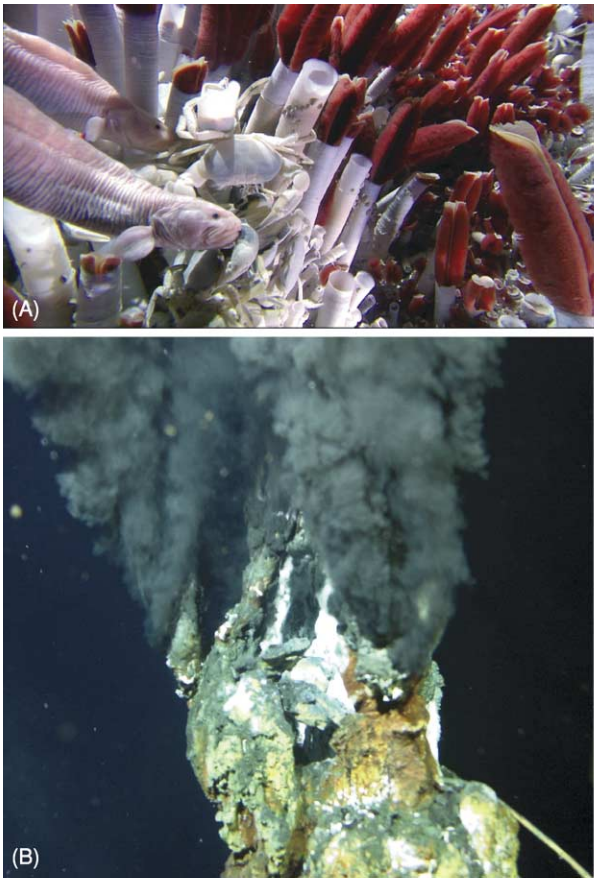


Figure 1 (A) A dense thicket of *Riftia* vestimentiferan tubeworms on the crest of the East Pacific Rise near lat. 9° 50′ N; brachyuran crabs and pink zooarcid fish are visible in the foreground. Chemosynthetic microbes, nourished by dissolved minerals in hydrothermal fluids, support many flourishing communities such as this along the mid-ocean ridge crest. Photo credit: Richard Lutz, Rutgers University, Stephen Low Productions, and Woods Hole Oceanographic Institution. (B) Vigorous plumes of mineral-blackened fluid issue from metal-enriched mineral chimneys on the northern East Pacific Rise crest. Photos were taken from the Alvin submersible (operated by the National Deep Submergence Facility at Woods Hole Oceanographic Institution), during dives funded by the National Science Foundation.

a geological record of their existence. Hydrothermal vents thus help to balance the chemical composition of the oceans, and provide; sources of energy for deep-sea ecosystems; a renewable source of metals; and a depositional record of the physical, chemical, biological, and geological processes at modern and ancient submarine vents.

Where Deep-Sea Hydrothermal Vents and Deposits Form: Geological Controls

Less than 2% of the total area of the mid-ocean ridge crest has been studied at a resolution sufficient to reveal the spatial distribution of individual seafloor hydrothermal vents, mineral deposits, and other significant small-scale geological features. Nevertheless, because study areas have been carefully selected and strategically surveyed, and because methods for remotely mapping hydrothermal plumes have been developed, much has been learned about where vents and deposits form, and about the geological controls on their distribution. The basic requirements for hydrothermal systems include heat, to drive fluid circulation, and high-permeability pathways, to facilitate fluid flow through the seafloor. These requirements frequently are met at sites of seafloor volcanism and faulting on or near plate boundaries, or at 'mid-plate' seafloor volcanoes (for example, on Loihi Seamount near Hawaii). Along the mid-ocean ridge, vents and deposits are forming at sites where ascending magma intrusions introduce heat into the permeable shallow crust, and at sites where deep cracks provide permeability and fluid access to heat sources at depth.

Fast-Spreading Ridges

Near- and on-bottom studies along the fast-spreading East Pacific Rise suggest that most hydrothermal mineral deposits form at high-temperature vents (~100–400°C) concentrated along the summit of the ridge crest within a narrow 'axial zone' less than 500 m wide. Only a few active sites of mineral deposition have been located outside this zone. However, heat flow data indicate that ~70% of the total hydrothermal heat loss from mid-ocean ridges occurs on ridge flanks, from lithosphere > 1 million years old. Clearly, more exploration of the vast 'off-axis' region beyond the axial zone is needed to identify the nature and distribution of hydrothermal vents on ridge flanks.

Within the axial zone on fast-spreading ridges, the overall spatial distribution of hydrothermal vents and mineral deposits traces the segmented configuration of cracks and magma sources along the ridge axis. Active vents and mineral deposits are concentrated along the floors and walls of axial troughs created by volcanic collapse and/or faulting along the summit of the ridge crest. A majority of the deposits are located along fissures that have opened above magmatic dyke intrusions, and along collapsed lava ponds formed above these fissures by pooling and drainage of erupted lava. Where fault-bounded troughs have formed along the summit of the ridge crest, mineral deposition is focused along the bounding faults and also along fissures and collapsed lava ponds in the trough floor. Hydrothermal vents appear to be most abundant along magmatically inflated segments of fast-spreading ridges; however, the mineral deposits precipitated on the seafloor on magmatically active segments are often buried beneath frequent eruptions of new lava flows. The greatest number of deposits, therefore, can be observed on inflated ridge segments that are surfaced by somewhat older flows, i.e., along segments where much heat is available to power hydrothermal vents and mineral deposits have had time to develop but have not yet been buried by renewed eruptions.

Intermediate- and Slow-Spreading Ridges

Most hydrothermal deposits that have been found on intermediate- and slow-spreading ridge crests are focused along faults, fissures, and volcanic structures within large rift valleys that are several kilometres wide. The fault scarps along the margins of rift valleys are common sites for hydrothermal venting and mineral deposition. Some hydrothermal systems along rift valley bounding faults are known to have been episodically active over periods of thousands of years, and have produced large mineral deposits, several hundreds of metres in length and tens of metres thick. Fault intersections are thought to be particularly favorable sites for hydrothermal mineral deposition because they are zones of high permeability that can focus fluid flow. Vents and mineral deposits on rift valley floors also are observed above dike intrusions, along eruptive fissures and volcanic collapse troughs, or on top of volcanic mounds, cones, and other constructions. On the intermediate-rate Juan de Fuca Ridge, hydrothermal vents are located above seismically detectable magma lenses. Thus, both magmatic and tectonic controls on hydrothermal processes are observed at slow and intermediate spreading rates. In general, however, faults appear to play a greater role in controlling the distribution of hydrothermal vents and mineral deposits than they do at fast-spreading ridges, where magmatic fissures are clearly a dominant geological control on where vents and deposits are forming.

Spectacular fault-controlled hydrothermal mineral deposits have formed on older seafloor (approximately 1 million years old) on the west flank of the slow-spreading Mid-Atlantic Ridge, at the intersection of the Mid-Atlantic Ridge rift valley and the Atlantic Transform Fault near 30° N. This site, known as 'Lost City', is discharging warm hydrothermal fluids (<100°C) through more than 30 calcium carbonate mineral structures, the tallest of which is 60 m in height!

Structure, Morphology, and Size of Deposits

A typical hydrothermal mineral deposit on an unsedimented mid-ocean ridge accumulates directly on top of the volcanic flows covering the ridge crest. On sedimented ridges, minerals are deposited within and on top of the sediments. Beneath the seafloor are networks of feeder cracks through which fluids travel to the seafloor. Precipitation of hydrothermal minerals in these cracks and in the surrounding rocks or sediments creates a subseafloor zone of mineralization called a 'stockwork'. In hydrothermal systems in which fluid flow is weak or unfocused, or in which the fluids mix extensively with seawater beneath the seafloor, most of the minerals will precipitate in the stockwork rather than on the seafloor.

Hydrothermal deposits on mid-ocean ridges are composed of (1) vertical structures, including individual conduits known as 'chimneys' (Figure 2) and larger structures of coalesced conduits that are often called 'edifices', (2) horizontal 'flange' structures that extend outwards from chimneys and edifices, (3) mounds of accumulated mineral precipitates (Figure 2), and (4) horizontal layers of hydrothermal sediments, debris, and encrustations. Chimneys initially are built directly on top of the seabed around focused jets of high-temperature effluents. Chimneys and edifices are physically unstable and often break or collapse into pieces that accumulate into piles of debris. The debris piles are cemented into consolidated mounds by precipitation of minerals from solutions percolating through the piles. New chimneys are constructed on top of the mounds as the mounds grow in size. Hydrothermal plume particles and particulate debris from chimneys settle around the periphery of the mounds to form layers of hydrothermal sediment. Diffuse seepage of fluids also precipitates mineral encrustations on mound surfaces, on volcanic flows and sediments, and on biological substrates such as microbial mats or the shells and tubes of sessile macrofauna.

The morphology of chimneys is highly variable and evolves as the chimneys grow, becoming more complex with time. Black smoker chimneys often are colonized by organisms and evolve into 'white smokers' that emit diffuse, diluted vent fluids through a porous carapace of wormtubes (Figure 2). Fluid compositions and temperatures, flow dynamics, and biota are all factors that influence the development of chimney morphology. Both the complexity of the interactions between these factors and the high degree of spatial–temporal heterogeneity in the physical, chemical, biological, and geological conditions influencing chimney growth account for the diversity in

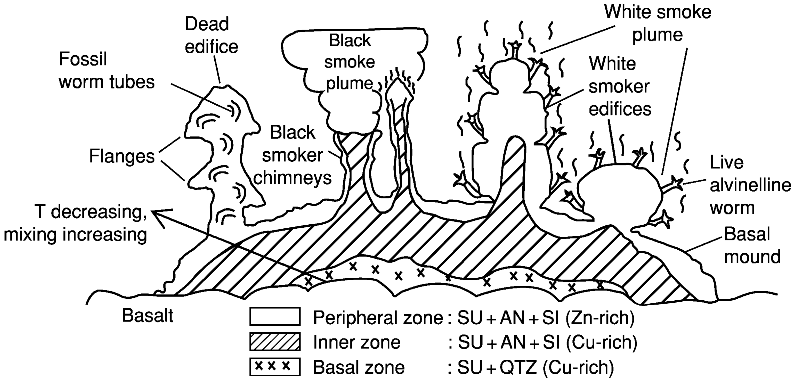


Figure 2 Composite sketch of the mineral structures and zones in hydrothermal mineral deposits on unsedimented ridge crests. Although mound interiors are seldom observed on the seafloor, the simplified sketch of mineral zoning within the mound is predicted by analogy with chimneys and massive sulphide deposits exposed in ophiolites. An outer peripheral zone of Zn-rich sulphide (SU; dominantly $ZnS + FeS_2$), anhydrite (AN), and amorphous silica (SI) is replaced in the interior by an inner zone of Cu-rich sulphide ($CuFeS_2 + FeS_2$), minor anhydrite, and amorphous silica. The inner zone may be replaced by a basal zone of Cu-rich sulphide ($CuFeS_2 + FeS_2$) and quartz (QTZ). Zones migrate as thermochemical conditions within the mound evolve. Although not shown here, it is expected that zoning around individual fractures cutting through the mound will be superimposed on the simplified zone structure depicted in this sketch. T = temperature. Modified with permission from Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America: The Eastern Pacific Ocean and Hawaii*, vol. N, pp. 125–144. Boulder: Geological Society of America.

morphology exhibited by chimneys, and presents a challenge to researchers attempting to unravel the processes producing specific morphologic features.

The sizes of hydrothermal mineral deposits on ridges also vary widely. It has been suggested that the largest deposits are accumulating on sedimented ridges, where almost all of the metals in the fluids are deposited within the sediments rather than being dispersed into the oceans by hydrothermal plumes. On unsedimented ridges, the structures deposited on the seabed at fast spreading rates typically are relatively small in dimension (mounds are typically less than a few metres in thickness and less than tens of metres in length, and vertical structures are <15 m high). On intermediate- and slow-spreading ridges, mounds are sometimes much larger (up to tens of metres in thickness, and up to 300 m in length). On the Endeavour Segment of the Juan de Fuca Ridge, vertical structures reach heights of 45 m, and the edifices are even taller at the Lost City site on the Mid-Atlantic Ridge. The size of a deposit depends on many factors, including the magnitude of the heat source, which influences the duration of venting and mineral deposition; whether deposition recurs episodically at a particular site, which depends on the nature of the heat source and plumbing system and on the rate of seafloor spreading; the frequency with which deposits are buried beneath lava flows; and the chemical compositions of vent fluids and chimney minerals, which are affected by site-specific subseafloor rock composition and water–rock reactions. The large deposits found on slower spreading ridge crests are located on faults that have moved slowly away from the ridge axis and have experienced repeated episodes of venting and accumulated mineral deposition over thousands of years, without being buried by lava flows. The tall Endeavour Segment edifices are formed because ammonia-enriched fluid compositions favour precipitation of silica in the edifice walls. The silica is strong enough to stabilize these structures so that they do not collapse as they grow taller. At Lost City, precipitation of sturdy calcium carbonate permits very tall edifices to form.

How Do Chimneys Grow?

A relatively simple two-stage inorganic growth model has been advanced to explain the basic characteristics of black smoker chimneys (Figure 3). In this model, a chimney wall composed largely of anhydrite (calcium sulphate) precipitates initially from seawater that is heated around discharging jets of hydrothermal fluid. The anhydrite-rich chimney wall precipitated during Stage I contains only a small component of metal sulphide mineral particles that crystallize from rapid chilling of the hydrothermal fluids. In Stage II, the anhydrite-rich wall continues to grow upwards and to thicken radially, protecting the fluid flowing through the chimney from very rapid chilling and dilution by seawater. This allows metal sulphide minerals to precipitate into the central conduit of the chimney from the hydrothermal fluid. The hydrothermal fluid percolates outward through the chimney wall, gradually replacing anhydrite and filling voids with metal sulphide minerals. During Stage II, the chimney increases in height, girth, and wall thickness, and both the calcium sulphate/metal sulphide ratio and permeability of the walls decrease. Equilibration of minerals with pore fluid in the walls occurs continuously along steep, time-variant temperature and chemical gradients between fluids in the central conduit and seawater surrounding the chimney. This equilibration produces sequences of concentric mineral zones across chimney walls that evolve with changes in thermal and chemical gradients and wall permeability (Figure 4).

The model of chimney growth is accurate but incomplete, because it does not include the effects on chimney development of fluid-phase separation, biological activity, or variations in fluid composition. Augmented models that address these complexities are needed to characterize fully the processes governing chimney growth.

Elemental and Mineral Compositions of Deposits

Typical ridge crest hydrothermal deposits are dominantly composed of iron, copper, and zinc sulphide minerals (*see* **Minerals:** Sulphides); calcium and barium sulphate minerals (*see* **Minerals:** Sulphates); iron oxide and iron oxyhydroxide minerals; and silicate minerals (Table 1). These minerals precipitate from diverse processes, including heating of seawater; cooling of hydrothermal fluid; mixing between seawater and hydrothermal fluid; reaction of hydrothermal minerals with fluid, seawater, or fluid–seawater mixtures; reaction between hydrothermal fluid and seafloor rocks and sediments; and reactions that are biologically mediated or catalysed. This diversity in the processes and environments of mineral precipitation results in deposition of many different minerals and elements (Tables 1 and 2). High concentrations of strategic and precious metals are found in some deposits (Table 2). The deposits are potentially valuable, if economic and environmentally safe methods of mining them can be developed.

Chimneys can be classified broadly by composition into five groups: sulphate-rich, copper-rich, zinc-rich,

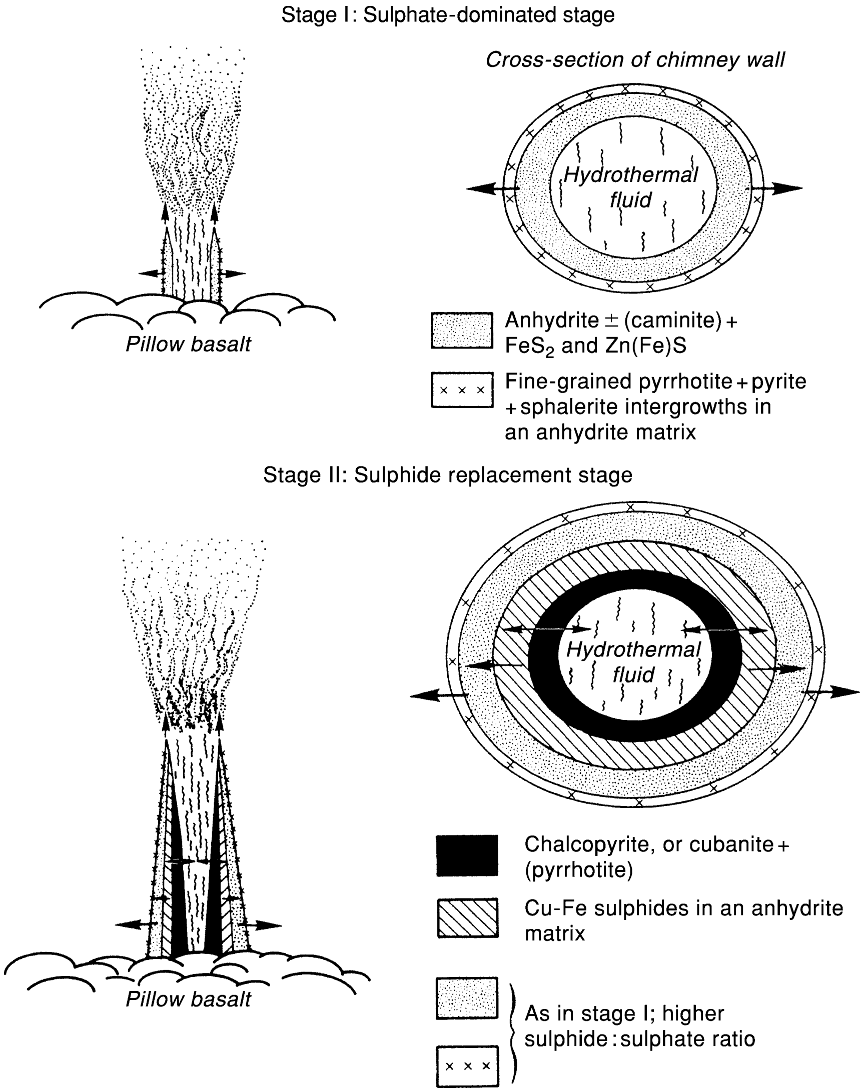

Stage I: Sulphate-dominated stage
Cross-section of chimney wall
Hydrothermal fluid
Pillow basalt
Anhydrite ± (caminite) + FeS₂ and Zn(Fe)S
Fine-grained pyrrhotite + pyrite + sphalerite intergrowths in an anhydrite matrix
Stage II: Sulphide replacement stage
Hydrothermal fluid
Pillow basalt
Chalcopyrite, or cubanite + (pyrrhotite)
Cu-Fe sulphides in an anhydrite matrix
As in stage I; higher sulphide : sulphate ratio

Figure 3 Two-stage model of black smoker chimney growth. During Stage II, several different sulphide mineral zonation sequences develop, depending on permeability and thickness of chimney walls, hydrodynamic variables, and hydrothermal fluid composition. Arrows indicate directions of growth.

silica-rich, and carbonate-rich structures. Copper-rich chimney compositions are indicative of formation at temperatures above ~300°C. Sulphate-rich compositions are characteristic of active and immature chimneys formed at temperatures above ~150°C. Carbonate-rich chimneys are found in areas where discharging hydrothermal fluids have reacted with carbonate sediments or ultramafic igneous rocks, and may form at relatively low temperatures. Many chimneys are mineralogically zoned, with hot interior regions enriched in copper and cooler exterior zones enriched in iron, zinc, and sulphate (Figures 2–4). Mounds exhibit a similar gross mineral zoning, and those that are exposed by erosion in ophiolites often have silicified (quartz-rich) interiors (Figure 2). Seafloor weathering of deposits after active venting ceases results in dissolution of anhydrite and in oxidation and dissolution of metal sulphide minerals. Small deposits that are not sealed by silicification or buried by lava flows will not be well preserved in the geological record.

Chimneys as Habitats

Chimney and mound surfaces are substrates populated by microbial colonies and sessile organisms such as vestimentiferan worms (Figure 1), polychaete worms, limpets, mussels, and clams. It is likely that pore spaces in exterior regions of chimney walls and mounds also are inhabited by microbes. All of these chemosynthesis-dependent organisms benefit from the seepage of hydrothermal fluid through active

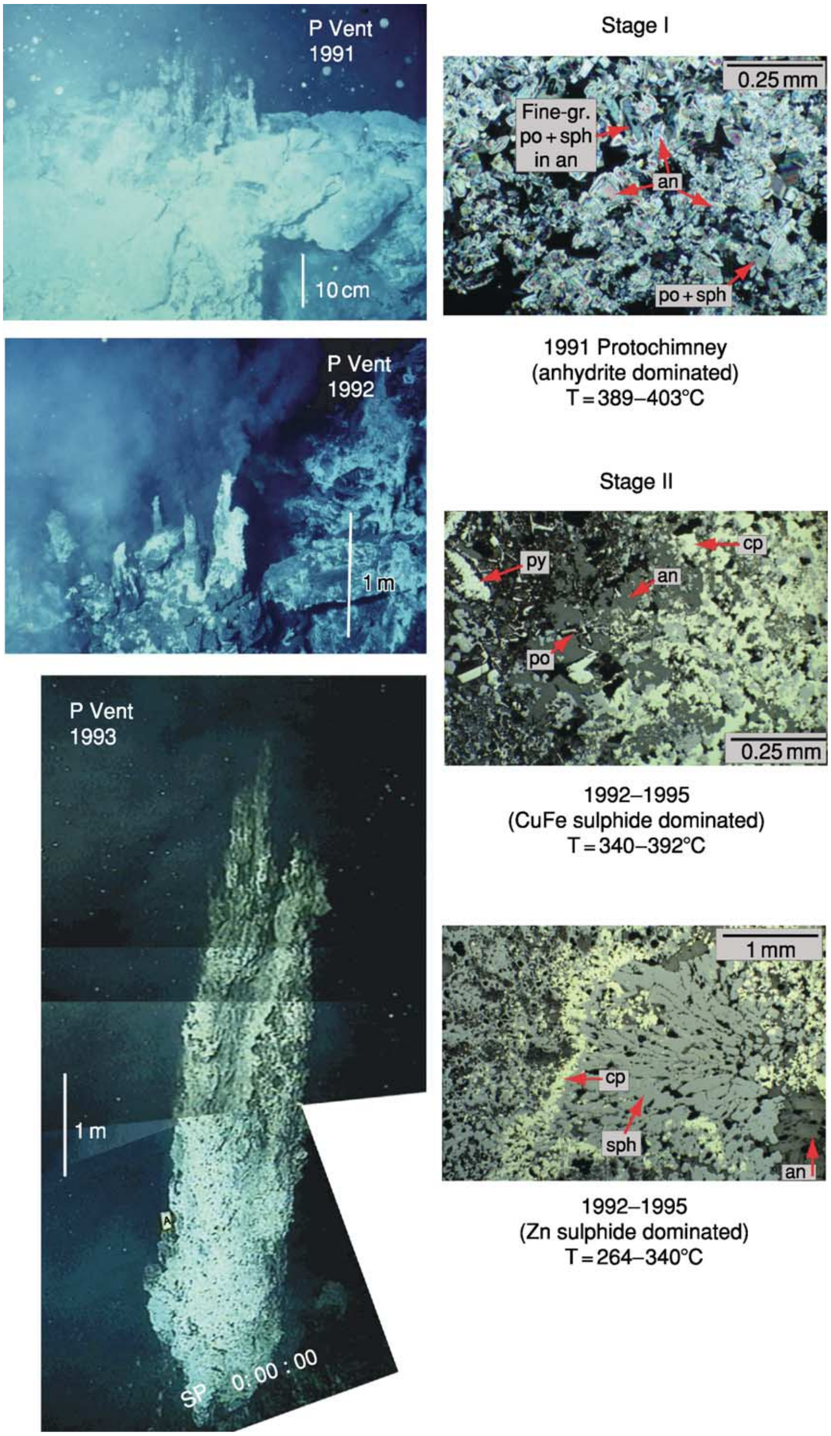


Figure 4 Morphological and mineralogical evolution of chimneys. (Left): A time-series of seafloor photographs showing the morphological development of a chimney that grew on top of lava flows that erupted in 1991 on the crest of the East Pacific Rise near 9° 50.3′ N. Within a few days to weeks after the eruption, anhydrite-rich Stage I ‘protochimneys’ a few centimetres high had formed where hot fluids emerged from volcanic outcrops that were covered with white microbial mats (top left). Eleven months later, the chimney consisted of cylindrical Stage II anhydrite–sulphide mineral spires approximately 1 m in height, and as-yet unpopulated by macrofauna (middle left). Three and a half years after the eruption, the cylindrical conduits had coalesced into a 7-m-high chimney structure that was covered with inhabited Alvinelline wormtubes (bottom left). (Right): Photomicrographs of chimney samples from the eruption area that show how the chimneys evolved from Stage I (anhydrite-dominated; top right) to Stage II (metal sulphide-dominated) mineral compositions (see text). As the fluids passing through the chimneys cooled below ~330°C during Stage II, the CuFe sulphide minerals in the chimney walls (middle right) were replaced by Zn sulphide and Fe sulphide minerals (bottom right). Abbreviations: gr, grained; po, pyrrhotite; an, anhydrite; py, pyrite; cp, chalcopyrite. Photographs reproduced from Haymon R, Fornari D, Von Damm K, *et al.* (1993) Volcanic eruption of the mid-ocean ridge along the East Pacific Rise at 9° 45–52′ N: direct submersible observation of seafloor phenomena associated with an eruption event in April, 1991. *Earth and Planetary Science Letters* 119: 85–101.

Table 1 Minerals occurring in ocean ridge hydrothermal mineral deposits

Mineral group/name	*Chemical formula*
Sulphides/sulphosalts	
Most abundant	
Sphalerite	Zn(Fe)S
Wurtzite	Zn(Fe)S
Pyrite	FeS_2
Chalcopyrite	$CuFeS_2$
Less abundant	
Iss (isocubanite)[a]	Variable ($CuFe_2S_3$)
Marcasite	FeS_2
Melnikovite	$FeS_{2\text{-}x}$
Pyrrhotite	$Fe_{1\text{-}x}S$
Bornite–chalcocite	Cu_5FeS_4–Cu_2S
Covellite	CuS
Digenite	Cu_9S_5
Idaite	$Cu_{5.5}FeS_{6.5}$
Galena	PbS
Jordanite	$Pb_9As_4S_{15}$
Tennantite	$(Cu,Ag)_{10}(Fe,Zn,Cu)_2As_4S_{23}$
Valeriite	$2(Cu,Fe)_2S_23(Mg,Al)(OH)_2$
Sulphates	
Anhydrite	$CaSO_4$
Gypsum	$CaSO_4 \cdot H_2O$
Barite	$BaSO_4$
Caminite	$MgSO_4 \cdot x Mg(OH)_2 \cdot (1\text{–}2x)H_2O$
Jarosite–natrojarosite	$(K,Na)Fe_3(SO_4)_2(OH)_6$
Chalcanthite	$CuSO_4 \cdot 5H_2O$
Carbonates	
Magnesite	$MgCO_3$
Calcite	$CaCO_3$
Aragonite	$CaCO_3$
Elements	
Sulphur	S
Oxides/hydroxides	
Goethite	FeO(OH)
Lepidocrocite	FeO(OH)
Hematite	Fe_2O_3
Magnetite	Fe_3O_4
Brucite	$Mg(OH)_2$
Amorphous Fe compounds	—
Amorphous Mn compounds	—
Psilomelane	$(Ba,H_2O)_2Mn_5O_{10}$
Silicates	
Opaline silica	$SiO_2 \cdot nH_2O$
Quartz	SiO_2
Talc	$Mg_3Si_4O_{10}(OH)_2$
Nontronite	$(Fe,Al,Mg)_2(Si_{3.66}Al_{0.34})O_{10}(OH)_2$
Illite–smectite	—
Aluminosilicate colloid	—
Hydroxychlorides	
Atacamite	$Cu_2Cl(OH)_3$

[a] Iss, Intermediate solid solution.

Table 2 Ranges of elemental compositions in bulk mid-ocean ridge hydrothermal mineral deposits

Element	*Range*[a]
Cu	0.1–15.0 wt.%
Fe	2.0–44.0 wt.%
Zn	$<$0.1–48.7 wt.%
Pb	0.003–0.6 wt.%
S	13.0–52.2 wt.%
SiO_2	$<$0.1–28.0 wt.%
Ba	$<$0.01–32.5 wt.%
Ca	$<$0.1–16.5 wt.%
Au	$<$0.1–4.6 ppm
Ag	3.0–303.0 ppm
As	7.0–918.0 ppm
Sb	2.0–375.0 ppm
Co	$<$2.0–3500.0 ppm
Se	$<$2.0–224.0 ppm
Ni	$<$1.5–226.0 ppm
Cd	$<$5–1448 ppm
Mo	1.0–290.0 ppm
Mn	36.0–1847.0 ppm
Sr	2.0–4300.0 ppm

[a] Data from Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America*: *The Eastern Pacific Ocean and Hawaii*, vol. N, pp. 125–144. Boulder: Geological Society of America. Hannington MD, Jonasson IR, Herzig PM, and Petersen S (1995) In: Humphris SE, Zeirenberg RA, Mullineaux LS, and Thomason RE (eds), *Physical and chemical processes of seafloor mineralization at mid-ocean ridges*, p. 115–157.

mineral structures and from the thermal and chemical gradients across mineral structures. The structures provide an interface between seawater and hydrothermal fluid that maintains tolerable temperatures for biota, and allows organisms simultaneous access to the chemical constituents in both seawater and hydrothermal fluid. However, organisms attached to active mineral structures must cope with changes in fluid flow across chimney walls (which sometimes occur rapidly) and with ongoing engulfment by mineral precipitation.

Some organisms actively participate in the precipitation and breakdown of minerals; for example, sulphide-oxidizing microbes mediate the crystallization of native sulphur crystals, and microbes are also thought to participate in the precipitation of marcasite and iron oxide minerals. Additionally, the surfaces of organisms provide favorable sites for nucleation and growth of amorphous silica, metal sulphide, and metal oxide crystals, and this facilitates mineral precipitation and fossilization of vent fauna.

Fossil Record of Hydrothermal Vent Organisms

Fossil moulds and casts of wormtubes, mollusc shells, and microbial filaments have been identified in both modern ridge hydrothermal deposits and in ancient deposits of Cretaceous, Jurassic, Devonian, and Silurian ages. This fossil record establishes the antiquity of vent communities and the long evolutionary history of specific faunal groups. The singular Jurassic fossil assemblage preserved in a small ophiolite-hosted deposit in central California is particularly interesting because it contains fossils of vestimentiferan worms, gastropods, and brachiopods, but no clam or mussel fossils. In contrast, modern and Palaeozoic faunal assemblages described thus far include clams, mussels, and gastropods, but no brachiopods. Does this mean that brachiopods have competed with molluscs for ecological niches at vents, and have moved in and out of the hydrothermal vent environment over time? Fossilization of organisms is a selective process that does not preserve all of the fauna that are present at vents. Identification of fossils at the species level is often difficult, especially where microbes are concerned. Notwithstanding, it is important to search for more examples of ancient fossil assemblages and to trace the fossil record of life at hydrothermal vents back as far as possible, to shed light on questions about how vent communities have evolved and about whether life on Earth might have originated at submarine hydrothermal vents.

Summary

Hydrothermal activity is an integral aspect of seafloor accretion at mid-ocean ridges. The circulating fluids and plumes facilitate thermal and chemical exchange between the oceans and the lithosphere, support life above and below the seafloor, and affect current flow and biological activity at mid-water depths. The mineral deposits are valuable for their metals, for the role that they play in fostering hydrothermal vent ecosystems, for the clues that they hold to understanding spatial–temporal variability in hydrothermal vent systems, and as geological records of how life at hydrothermal vents has evolved. Vent organisms exhibit novel biochemistry, genetics, taxonomy, physiology, symbiosis, and community dynamics. From submarine hydrothermal systems, the insights gained about biogeochemical processes at high temperatures and pressures can be applied to biotechnology and to understanding life in inaccessible realms within Earth's crust, or the potential for life on other planetary bodies. As the complexities of hydrothermal systems on the mid-ocean ridge are unravelled through ongoing exploration and interdisciplinary studies, exciting applications of this knowledge are being discovered.

See Also

Minerals: Sulphates; Sulphides. **Mining Geology:** Hydrothermal Ores; Magmatic Ores. **Origin of Life**. **Plate Tectonics**. **Seamounts**. **Tectonics:** Mid-Ocean Ridges.

Further Reading

Baker ET (1996) Geological indexes of hydrothermal venting. *Journal of Geophysical Research* 101(B6): 13 741–13 753.

Dilek Y, Moores E, Elthon D, and Nicolas A (eds.) (2000) *Ophiolites and Oceanic Crust: New Insights from Field Studies and the Ocean Drilling Program. Geological Society of America Memoir.* Boulder: Geological Society of America.

Haymon RM (1989) Hydrothermal processes and products on the Galapagos Rift and East Pacific Rise. In: Winterer EL, Hussong DM, and Decker RW (eds.) *The Geology of North America: The Eastern Pacific Ocean and Hawaii,* vol. N, pp. 125–144. Boulder: Geological Society of America.

Haymon RM (1996) The response of ridge crest hydrothermal systems to segmented, episodic magma supply. In: MacLeod CJ, Tyler P, and Walker CL (eds.) *Tectonic, Magmatic, Hydrothermal, and Biological Segmentation of Mid-Ocean Ridges, Special Publication,* vol. 118, pp. 157–168. London: Geological Society.

Haymon R, Fornari D, Von Damm K, *et al.* (1993) Volcanic eruption of the mid-ocean ridge along the East Pacific Rise at 9° 45–52′ N: direct submersible observation of seafloor phenomena associated with an eruption event in April, 1991. *Earth and Planetary Science Letters* 119: 85–101.

Humphris SE, Zierenberg RA, Mullineaux LS, and Thomson RE (eds.) (1995) *Seafloor Hydrothermal Systems: Physical, Chemical, Biological, and Geological Interactions, Geophysical Monograph.* vol. 91. Washington, DC: American Geophysical Union.

Kelley DS, Karson JA, Blackman DK, *et al.* (2001) An off-axis hydrothermal field near the Mid-Atlantic Ridge at 30° N. *Nature* 412: 145–149.

Little CTS, Herrington RJ, Haymon RM, and Danelian T (1999) Early Jurassic hydrothermal vent community from the Franciscan Complex, San Rafael Mountains, California. *Geology* 27: 167–170.

Schrenk MO, Kelley DS, Delaney JR, and Baross JA (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Applied and Environmental Microbiology* 69: 3580–3592.

Shank TM, Fornari DJ, Von Damm KL, *et al.* (1998) Temporal and spatial patterns of biological community development at nascent deep-sea hydrothermal vents (9° 50′ N, East Pacific Rise). *Deep-Sea Research II* 45: 465–515.

Tivey MK, Stakes DS, Cook TL, Hannington MD, and Petersen S (1999) A model for growth of steep-sided vent structures on the Endeavour Segment of the Juan de Fuca Ridge: results of a petrological and geochemical study. *Journal of Geophysical Research* 104: 22 859–22 883.

Von Damm KL (2000) Chemistry of hydrothermal vent fluids from 9–10° N, East Pacific Rise: "Time Zero" the immediate post-eruptive period. *Journal of Geophysical Research* 105: 11 203–11 222.

Propagating Rifts and Microplates At Mid-Ocean Ridges

R N Hey, University of Hawaii at Manoa, Honolulu, HI, USA

Introduction

Propagating rifts are extensional plate boundaries that gradually break through lithospheric plates, forming new plate boundaries and rearranging the geometries of old ones. If the rifting advances to the seafloor-spreading stage, propagating seafloor-spreading centres gradually extend through the rifted lithosphere. This evolution occurs rapidly ($\sim 10^5$ years) for oceanic propagators and much more slowly ($\sim 10^7$ years) for continental ones. The orthogonal combination of seafloor spreading and propagation produces a characteristic V-shaped wedge of lithosphere, with progressively younger and longer isochrons abutting the 'pseudofaults' that bound this wedge. Oceanic propagators generally replace pre-existing spreading centres, causing lithospheric transfer from one plate to another, and sequences of spreading-centre jumps, leaving failed rift systems in their wakes. This changes the classic plate tectonic geometry and results in asymmetric accretion of lithosphere to the plates. There is pervasive shear deformation in the overlap zone between the propagating and failing rifts, much of it accommodated by bookshelf faulting, in which, e.g., right-lateral plate motion shear produces high-angle left-lateral fault slip. When the scale or strength of the overlap zone becomes large enough, it stops deforming, and instead begins to rotate as a separate microplate between dual active spreading centres. Continental propagators break apart continents and can leave failed rifts (aulocogenes) along unsuccessful propagation paths, as well as predictable patterns of deformation ahead of the propagator tips. For continental rifting, the pseudofaults are the continental margins. Rift propagation appears to be the primary mechanism by which Earth's accretional plate boundary geometry is reorganized.

Oceanic Propagators

Figure 1 shows several variations of typical oceanic ridge propagation geometry, in which a pre-existing 'doomed rift' is replaced by the propagator. This always seems to result in at least slight spreading-centre reorientation. Whether this is because rifts propagate in response to changes in direction of seafloor spreading, or because the spreading direction changes while rifts propagate, rift propagation is the primary mechanism by which many seafloor-spreading systems have adjusted to changes in spreading direction. Propagation rates and local spreading rates are often similar in magnitude, although propagation rates as high as 1000 km My^{-1} have been discovered.

Figure 1A shows the discontinuous propagation model, in which periods of seafloor spreading alternate with periods of instantaneous propagation,

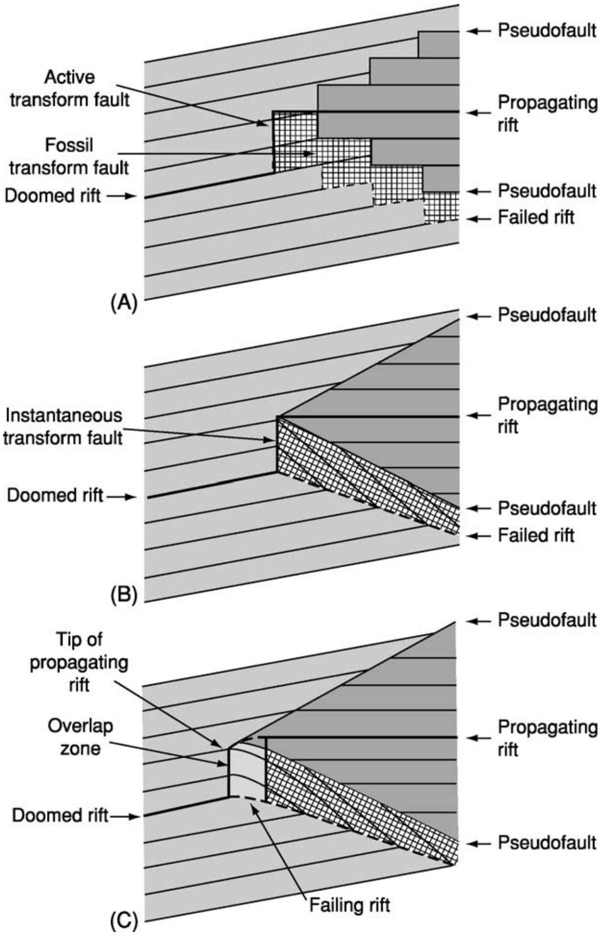


Figure 1 (A) Discontinuous, (B) continuous, and (C) non-transform-zone oceanic propagating/failing rift models. Propagating-rift lithosphere is marked by dark stipple, normal lithosphere created at the doomed rift is indicated by light stipple, and transferred lithosphere is cross-hatched. Heavy lines show active plate boundaries. In the non-transform-zone model(C), active axes with full spreading rate are shown as heavy lines; active axes with transitional rates are shown as dashed lines. The overlap zone joins these transitional spreading axes. Reproduced with permission from Hey RN, Sinton JM, and Duennebier FK (1989) Propagating rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii*, pp. 161–176. Boulder, CO: Geological Society of America.

producing *en echelon* failed rift segments, fossil transform faults, and fracture zones, and blocks of progressively younger transferred lithosphere. Figure 1B shows the pattern produced if propagation, rift failure, and lithospheric transferal are all continuous. In this idealized model, a transform fault migrates continuously with the propagator tip, never existing in one place long enough to form a fracture zone, and thus V-shaped pseudofaults are formed instead of fracture zones. Figure 1C shows a more geologically plausible model, in which the new spreading centre requires some finite time to accelerate from zero to the full rate, with concomitant decreases on the failing spreading centre, so that lithospheric transferal is not instantaneous. Instead of a transform fault, a migrating broad 'non-transform' zone of distributed shear deformation connects the overlapping propagating and failing ridges during the period of transitional spreading. Deformation occurring in this overlap zone is preserved in the zone of transferred lithosphere. This zone, left behind as the overlap zone migrates, is bounded by the failed rifts and inner (proximal) pseudofault. Even more complicated geometries occur in some places where the doomed rift, instead of failing monotonically as the propagator steadily advances, occasionally propagates in the opposite direction.

Standard plate tectonic geometry holds for the area outside the pseudofaults and zone of transferred lithosphere, but rigid plate tectonics breaks down in the overlap zone where some of the lithosphere formed on the doomed rift is progressively transferred to the other plate by the rift propagation and resulting migration of the overlap zone. Shear between the overlapping propagating and failing rifts appears to be accommodated by bookshelf faulting, probably along the pre-existing abyssal hill faults. This produces oblique seafloor fabric, with trends quite different from the ridge-parallel and-perpendicular structures expected from classic plate tectonic theory.

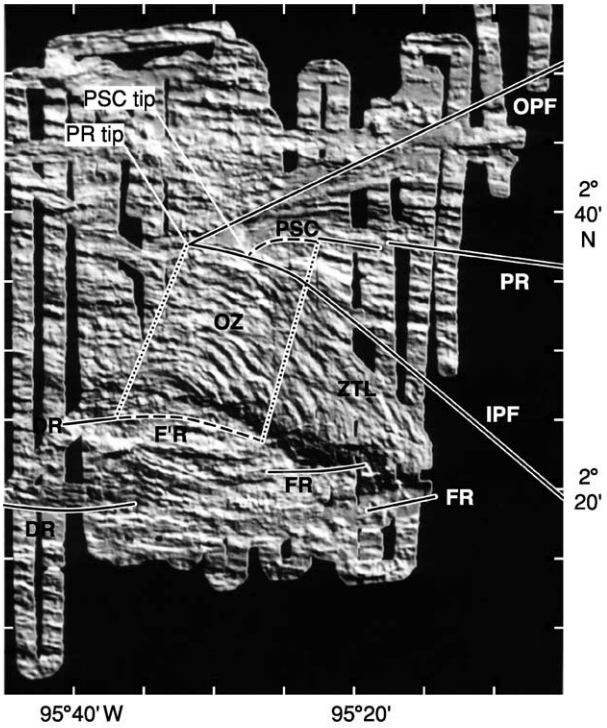


Figure 2 Shaded relief map of digital Sea Beam swath bathymetry at the Galapagos propagating rift system (95.5° W). The relative plate motion is nearly north–south; propagation is to the west. The oblique structures in the overlap zone and its wake, the zone of transferred lithosphere, are clearly evident. PR, Propagating rift; PSC, propagating spreading centre; OPF, IPF, outer and inner pseudofaults; OZ, overlap zone; ZTL, zone of transferred lithosphere; DR, doomed rift; F'R, failing rift; FR, failed rift grabens. Adapted by permission from Hey RN, Sinton JM, and Duennebier FK (1989) Propagating rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii*, pp. 161–176. Boulder, CO: Geological Society of America.

Figure 2 is a relief map of the type-example propagating rift, at 95.5° W along the Cocos–Nazca spreading centre. This area closely resembles the geometry in Figure 1C, except that rift failure occurs more discontinuously. This propagator is breaking westward away from the Galapagos hotspot through 1-My-old Cocos lithosphere at a velocity of about 50 km My^{-1}. Well-organized seafloor spreading begins about 10 km behind the faulting, fissuring, and extension at the propagating rift tip. This 200 000-year time lag between initial rifting and the rise of asthenosphere through the lithospheric cracks to form a steady-state spreading centre suggests an asthenospheric viscosity of about 10^{18} Pa s. The combination of seafloor spreading (about 60 km My^{-1}) and propagation produces a V-shaped wedge of young lithosphere bounded by pseudofaults and pre-existing lithosphere. The propagating rift lithosphere is characterized by high-amplitude magnetic anomalies and by unusual petrological diversity, including highly fractionated ferrobasalts. This propagator is replacing a pre-existing spreading system about 25 km to the south, and thus spreading-centre jumps and failed rifts result. Although propagation is continuous, segmented failing rift grabens seem to die episodically on a time-scale of about 200 000 years. This has produced a very systematic pattern of spreading-centre jumps identified from magnetic anomalies, in which each jump was younger and slightly longer than the preceding jump. The spreading-centre orientation is being changed clockwise by about 13°, and more than 10^4 km^3 My^{-1} of Cocos lithosphere is being transferred to the Nazca plate.

The active propagating and failing rift axes overlap by about 20 km and are connected by a broad and anomalously deep zone of distributed shear deformation rather than by a classic transform fault. Most of the seismic activity occurs within this 'non-transform' zone, where the pre-existing abyssal hill fabric originally created on the doomed rift is sheared and tectonically rotated into new oblique trends. Simple equations accurately describe this geometry in terms of ratios of propagation and spreading rates, together with the observed propagating and doomed rift azimuths. For example, for the simplest continuous propagation geometry, if u is the spreading half-rate and v is the propagation velocity, the pseudofaults form angles $\tan^{-1}(u/v)$ with the propagator axis, and the isochrons and abyssal hill fabric in the zone of transferred lithosphere have been rotated by an angle $\tan^{-1}(2u/v)$.

The boundaries of the Galapagos high-amplitude magnetic anomaly zone, the ferrobasalt province, and the spreading-centre jumps are all coincident with the pseudofaults bounding the propagating rift lithosphere. All of these observations can be explained as mechanical and/or thermal consequences of a new rift and spreading centre breaking through cold lithosphere, with increased viscous head loss and diminished magma supply on the propagating spreading centre close to the propagator tip. This leads to an unusually deep axial graben and unusually extensive fractional crystallization and differentiation. The 95.5° W propagator tip is also a mantle geochemistry boundary, implying that this rift propagation is associated with plume-related subaxial asthenospheric flow away from the Galapagos hotspot.

Causes of Rift Propagation

One important observation is that many rifts and spreading centres propagate down topographic gradients away from hotspots or shallow ridge axis topography. For example, all six known active Galapagos propagators are propagating away from the hotspot. Plume-related asthenospheric flow generates gravitational stresses on the shallow spreading-centre segments near the hotspot that promote crack propagation away from the hotspot. Flow of asthenosphere into these cracks produces new lithosphere at propagating seafloor-spreading centres. Regionally high deviatoric tensile stresses associated with regional uplift provide a quantitatively plausible driving mechanism. Crack growth occurs when the stress concentration at the tip, characterized in elastic fracture mechanics by a stress intensity factor, exceeds the resisting stress intensity contribution. The spreading-centre propagation rate could be limited by the viscosity of the asthenosphere flowing into the rift, producing viscous suction forces at a local tip depression, or by process zone deformation at the rift tip. The overlap/offset ratio of propagating and failing rifts tends to be ~1, close to the ratio at which the stress intensity factor is maximized (*see* **Tectonics:** Seismic Structure At Mid-Ocean Ridges).

Although many rifts appear to propagate in response to hotspot-related stresses, others appear to propagate because of stresses producing changes in plate motion. Subduction-related stresses appear to be a common mechanism for producing propagation in the North-east Pacific. This probably explains most of the massive reorganizations of the spreading geometry as the Pacific–Farallon ridge neared the Farallon–North America trench, as clearly evident even in the classic Raff–Mason magnetic anomaly data (Figure 3), although some propagation away from the Axial Seamount hotspot has also occurred in the Juan de Fuca area. Propagation may be produced in many ways over a wide range of scales, including small-scale propagation of overlapping spreading centres away from local magmatic centres. The larger reorganizations sometimes involve transient microplate formation, geometrically similar to the broad overlap zone model of Figure 1C, but on a much larger scale.

Microplates

Microplates are small, mostly rigid areas of lithosphere, located at major plate boundaries but rotating as more or less independent plates. They can form in many tectonic settings. The two main types along mid-ocean ridges, those formed at triple junctions and those formed away from triple junctions, share many similarities. Although it was once thought that stable, growing microplates could eventually grow into major oceanic plates, it now appears that these are transient phenomena resulting from large-scale rift propagation. When the overlap zone becomes too big and strong to deform by pervasive bookshelf faulting, it changes mechanical behaviour and accommodates the boundary plate motion shear stresses by beginning to rotate as a separate microplate. The most well-studied oceanic microplates are the Easter microplate along the Pacific–Nazca ridge and the Juan Fernandez microplate at the Pacific–Nazca–Antarctica triple junction (Figure 4). Despite their different tectonic settings, they show many striking similarities.

The scales of the Easter (~500 km diameter) and the Juan Fernandez (~400 km diameter) microplates are similar. The eastern and western boundaries of both microplates are active spreading centres, propagating north and south, respectively. Both microplates began forming about 5 Ma, and both

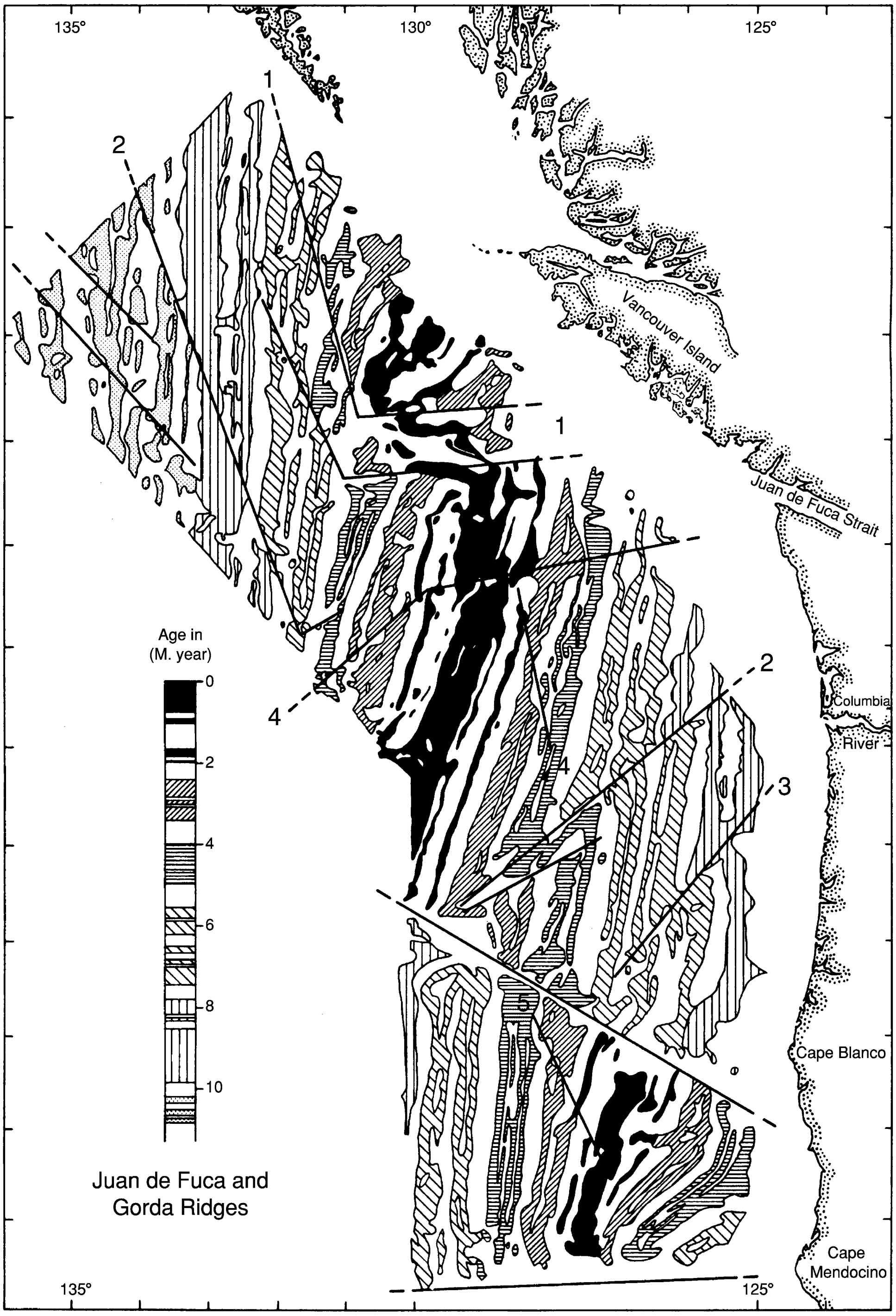


Figure 3 Raff–Mason magnetic anomalies in the Juan de Fuca area. Positive anomalies are shaded. Numbers denote major propagation sequences. Reproduced with permission from Hey RN (1977) A new class of pseudofaults and their bearing on plate tectonics: A propagating rift model. *Earth and Planetary Science Letters* 37: 321–325.

East rifts have been propagating into roughly 3-My-old Nazca lithosphere. Extremely deep axial valleys occur at their tips, ~6000 m at Pito Deep at the north-eastern Easter microplate boundary, and ~5000 m at Endeavour Deep at the north-eastern Juan Fernandez microplate boundary. The northern and southern boundaries are complicated deformation zones, with zones of shear, extension, and significant areas of compression. The dominant (East) rift of the Easter microplate, the dominant (West)

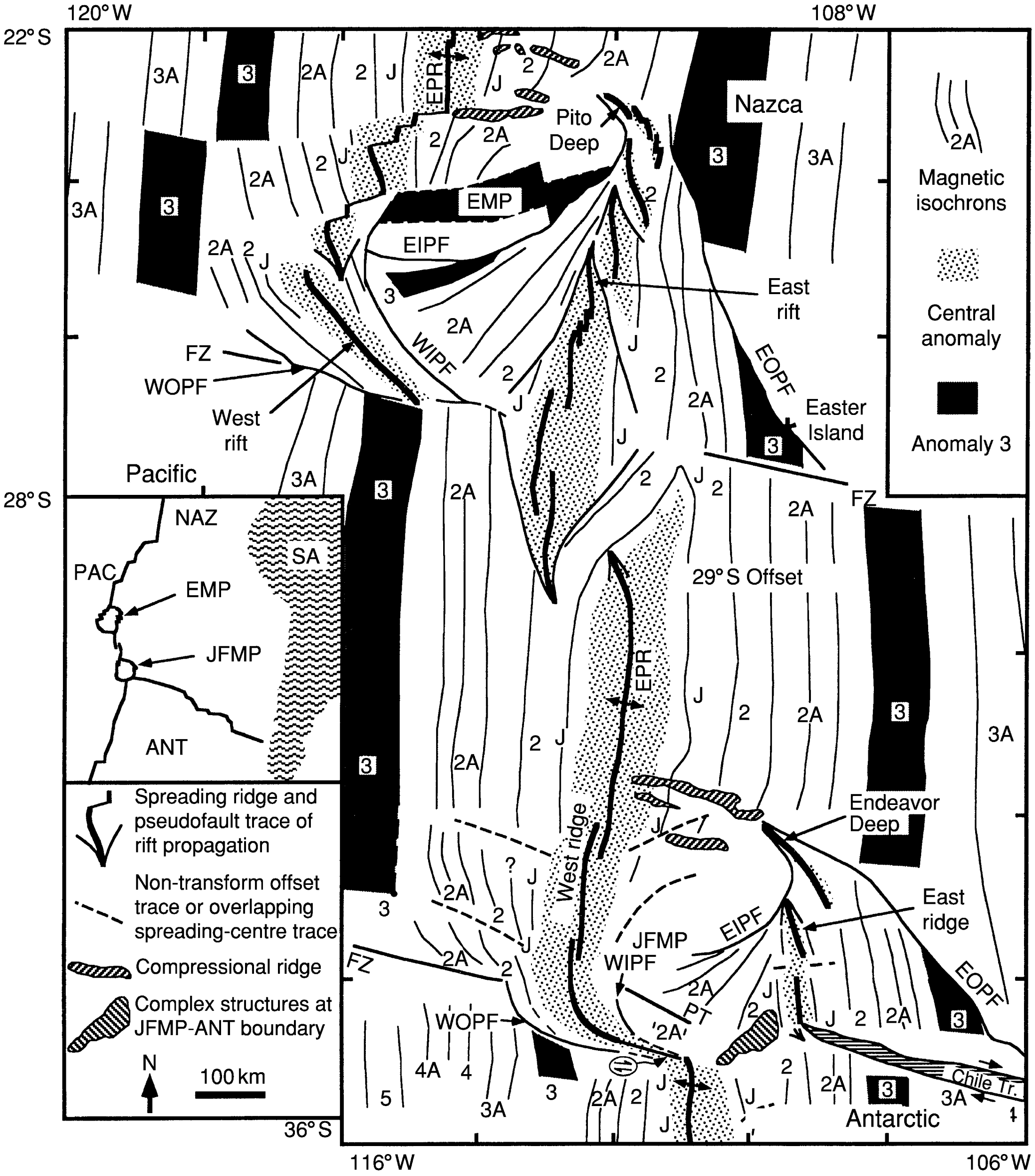


Figure 4 Tectonic boundaries, magnetic isochrons, and structures of the Easter microplate (EMP) and Juan Fernandez microplate (JFMP). EPR, East Pacific Rise; FZ, fracture zone; NAZ, Nazca; PAC, Pacific; SA, South America; ANT, Antarctic; WOPF, WIPF, EOPF, and EIPF are western outer and inner and eastern outer and inner pseudofaults, respectively. Numbers (e.g. 2, 2A) identify magnetic chrons from magnetic reversal time scale, J is Jaramillo reversal ~1 Ma, PT is paleotransform. Reproduced with permission from Bird RT and Naar DF (1994) Intratransform origins of mid-ocean ridge microplates. *Geology* 22: 987–990.

rift of the Juan Fernandez microplate, and the dominant (West) rift of the duelling propagators between the microplates, are all propagating away from the Easter mantle plume (or the intersection of this plume with the ridge axis), suggesting that microplate formation as well as rift propagation can be driven by plume-related forces.

Both microplates have large (~100 × 200 km), complex, pervasively deformed cores, which may have formed by bookshelf faulting in overlap zones during an initial large-scale propagating rift stage of evolution. Both show more recent stages of growth as independent microplates, with deformation concentrated along the plate boundaries. At present, both microplates are rotating clockwise very rapidly about poles near the microplate centres, spinning like roller-bearings caught between the major bounding plates. The Easter microplate rotation velocity is about 15° My^{-1} and the Juan Fernandez velocity is about 9° My^{-1}. The roller-bearing analogy has been quantified in an idealized edge-driven model of microplate kinematics. If microplate rotation is indeed driven by shear on the boundaries between the microplate and surrounding major plates, the rotation velocity (in radians) is $2u/d$, where $2u$ is the major plate relative velocity and d is the microplate diameter. This follows because the total spreading on the microplate boundaries must equal the major plate motion if the microplate did not exist. The rotation (Euler) poles describing the motion of the microplate relative to the major plates will lie on the microplate boundaries, at the farthest extensions of the rifts, which must continually lengthen by a different kind of rift propagation as the microplate rotates.

This idealized geometry (Figure 5) requires a circular microplate shape, yet also requires seafloor spreading on the dual active ridges, which must constantly change this shape. The more the microplate grows, the more deformation must occur as it rotates, and the less successful the rigid plate model will be. Although it would appear that this inevitable plate growth would soon invalidate the model, numerous episodes of rift propagation helping to maintain the necessary geometry are observed to have occurred at the Easter and Juan Fernandez microplates. All propagation was on the microplate interior side of the failing rifts, thus transferring microplate lithosphere to the surrounding Pacific and Nazca plates, shaving the new microplate growth at the edges and maintaining a shape circular enough for the edge-driven model to be very successful.

According to the edge-driven model, a microplate may stop rotating if one of the bounding ridge axes

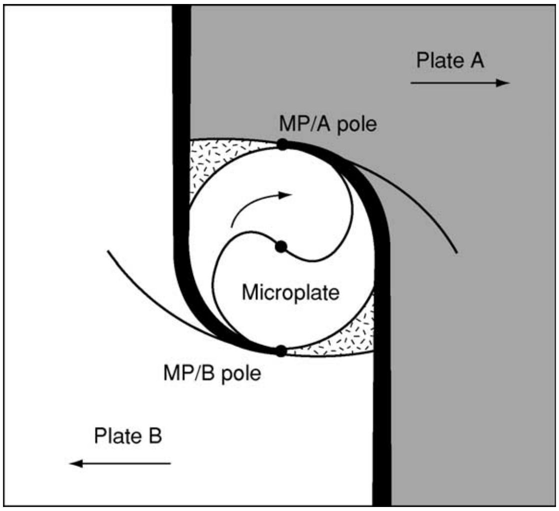


Figure 5 Roller-bearing model of microplates based on a simple, concentrically rotating bearing. The microplate is approximated by a circular plate that is caught between two major plates (MP/A and MP/B). The main contacts between microplate and major plates are also the positions of the relative rotation poles (dots). Dark shading shows major spreading centres, overlapping about the microplate. Cross-hatched corners are areas of compression. Medium curved lines are predicted pseudofaults; plate arrows show relative motions. This schematic model assumes growth from an infinitesimal point to a present circular shape; the model can be extended to take account of growth from a finite width, eccentric motions, and growth of the microplate. Reproduced with permission from Searle RC, Bird RT, Rusby RI, and Naar DF (1993) The development of two oceanic microplates: Easter and Juan Fernandez microplates, East Pacific Rise. *Journal of the Geological Society* 150: 965–976.

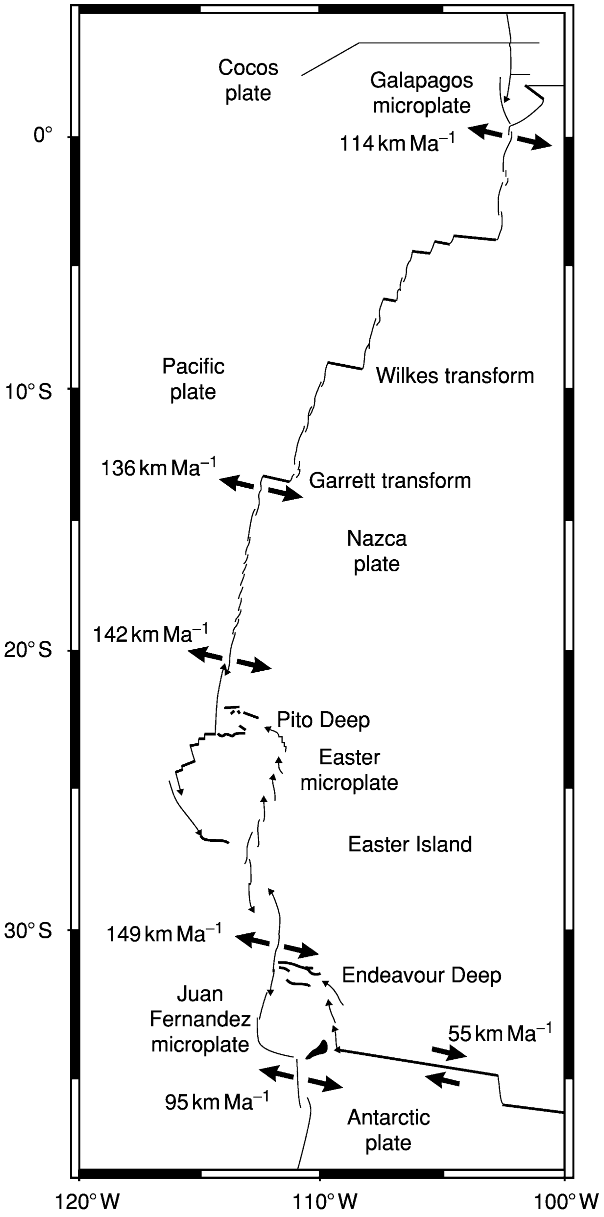


Figure 6 Plate tectonic geometry and relative plate motions along the southern East Pacific Rise. Light lines are ridges, and those with arrows are propagating. Heavy straight lines are transform faults. Reproduced with permission from Hey RN, Johnson PD, Martinez F, *et al.* (1995) Plate boundary reorganization at a large-offset, rapidly propagating rift. *Nature* 378: 167–170.

propagates through to the opposite spreading boundary, eliminating coupling to one of the bounding plates. Dual spreading would no longer occur, spreading would continue on only one bounding ridge, and 10^6–10^7 km^3 of microplate lithosphere would accrete to one of the neighbouring major plates. Active microplates are thus modern analogues for how large-scale (hundreds of kilometres) spreading centre jumps occur. There is evidence in the older seafloor record that this happened many times along the ancestral East Pacific Rise.

All large right-stepping offsets along the Pacific–Nazca spreading centre are transform faults, whereas all large left-stepping offsets are microplates or the giant duelling propagators (possible protomicroplate) between the Easter and Juan Fernandez microplates (Figure 6). The Galapagos microplate at the Pacific–Cocos–Nazca triple junction also fits this pattern. This suggests that a recent clockwise change in Pacific–Nazca plate motion could have been an important factor triggering the formation of these microplates. Earth's fastest active seafloor spreading occurs in this area, and all parts of the plate boundary presently spreading faster than 142 km My^{-1} are reorganizing by duelling propagators or microplates (Figure 6). The combination of thin lithosphere produced at these 'superfast' seafloor spreading rates, and the unusually hot asthenosphere produced by an Easter mantle plume, would reduce the forces resisting propagation and thus make these plate boundary reorganizations easier, perhaps explaining their common occurrence in this area.

Microplates also occur in convergent settings, where small pieces of lithosphere are caught between large plates. A well-studied continental convergence example is the Mediterranean, where small plates the size of Turkey and the Aegean adapt to Africa–Eurasia convergence. Oceanic microplates can also form along convergent margins, e.g., in the West Pacific, to accommodate Pacific, Australia, and Asia convergence.

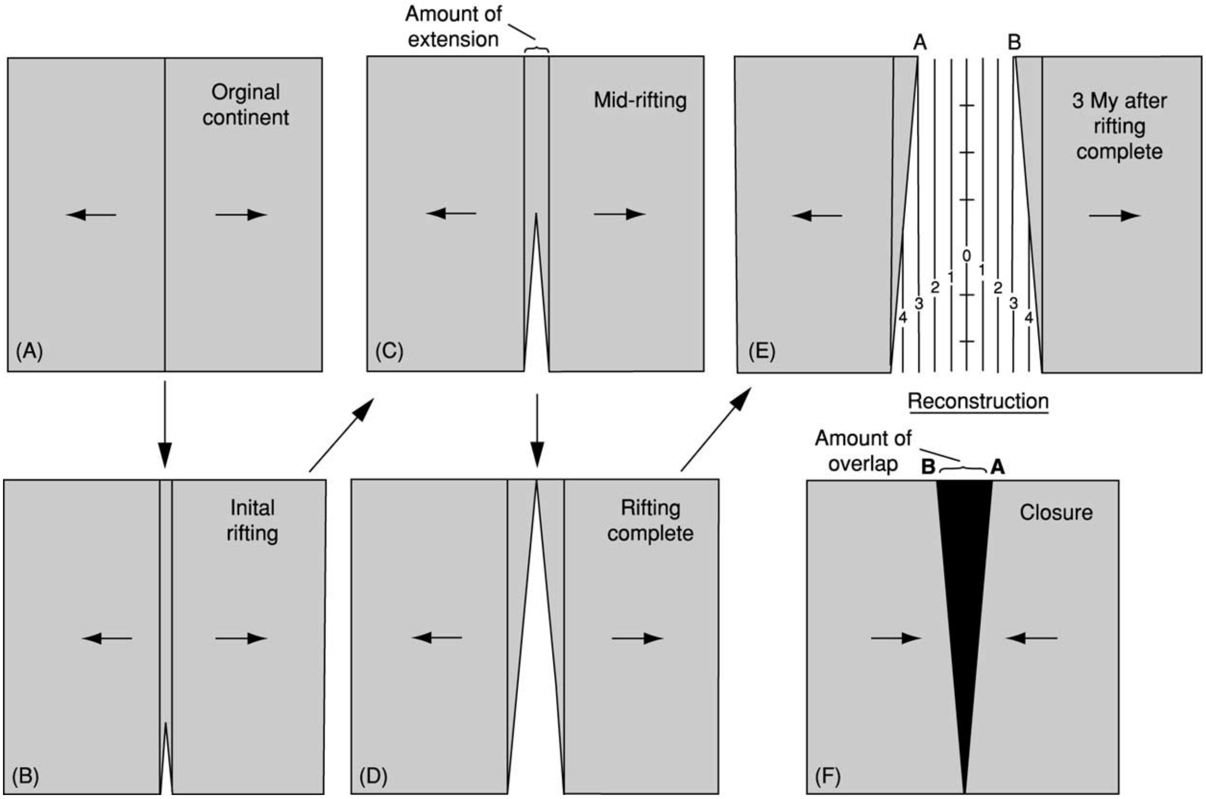


Figure 7 Propagating rift model for continental breakup (map view). (A) Original continent under tension; (B) initial rifting, with the amount of extension represented between parallel lines; (C) mid-rifting (seafloor spreading is occurring in the lower half of the continent while crustal thinning and extension occur in the upper half); (D) rifting complete (continental edges have undergone extension that increases in the direction of rifting); (E) 3 My after rifting is complete (oldest seafloor is found in the part of the ocean where rifting began; oldest isochrons converge with the ocean–continent boundary; continental edge is not an isochron); (F) reconstruction of the pre-rift configuration (extension due to rifting results in apparent overlap when the continents are returned to their pre-rift geometry). Reproduced with permission from Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677–10 688.

Continental Propagators

Continents break apart progressively as well, and thus continental margins are not isochrons but instead are a type of pseudofault boundary. During breakup, the margins bound V-shaped wedges of propagating rift lithosphere pointing in the direction of propagation. The relative azimuths of the margin and the first magnetic isochrons formed on the new seafloor-spreading centre give the propagation velocity. For example, the South Atlantic began forming at the southern tip of Africa–South America ~130 Ma and reached the area that became the great equatorial mid-Atlantic ridge (MAR) offset 30–40 My later, for an average northward propagation velocity of 10–15 km My^{-1}, similar to the spreading rate.

Predictable extensional rift deformation occurs ahead of the propagating spreading-centre tip (Figure 7). Plate reconstructions that take this new understanding of breakup deformation into account differ from classic reconstructions, such as the Bullard fits, which have equal amounts of continental overlap and underlap. Instead, increasing deformation ahead of the growing spreading centre means continental reconstruction fits should be exact where propagation begins, and should slow increasing overlap in the direction of propagation, with no areas of continental margin underlap (Figure 8).

Occasional rifts may propagate locally along globally unfavorable paths and eventually fail, resulting in failed rifts (aulocogenes) along margins. Systematic patterns of small failed rifts, or one huge one such as the Greenland–Canada failed rift in the Labrador Basin, would produce asymmetries in the conjugate margins. Occasional episodes of duelling propagation may also occur, forming ephemeral microplates eventually welded to one of the major plates, with corresponding margin asymmetries. One well-studied area of continental propagation is the Afar depression, where the Gulf of Aden propagator is presently breaking west into the African continent in Djibouti; a Red Sea propagator is simultaneously breaking south into the continent in Ethiopia (Figure 9). The interaction of these duelling propagators has produced changes in their propagation directions and the same kind of pervasive bookshelf faulting between the ridges as seen in oceanic rift propagation.

Implications

Rift propagation, which occurs on scales ranging from overlapping spreading centres with offsets of

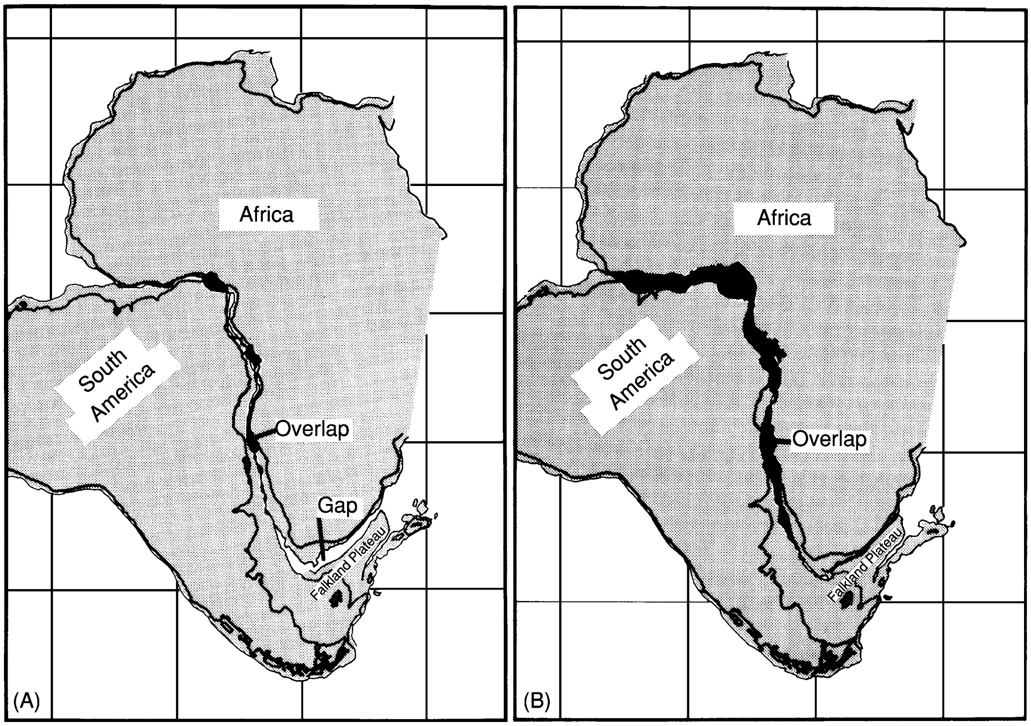


Figure 8 (A) The Bullard South America–Africa reconstruction, using the 1000-m isobath to represent the continental edge and the 3000-m isobath for the Falkland Plateau. (B) The Vink reconstruction, using the propagating rift model and the same isobaths, obtained by rotating South America 58.00° counterclockwise around a pole at 47.00° N, 33.80°. Dark regions represent areas of overlap. Reproduced with permission from Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677–10 688.

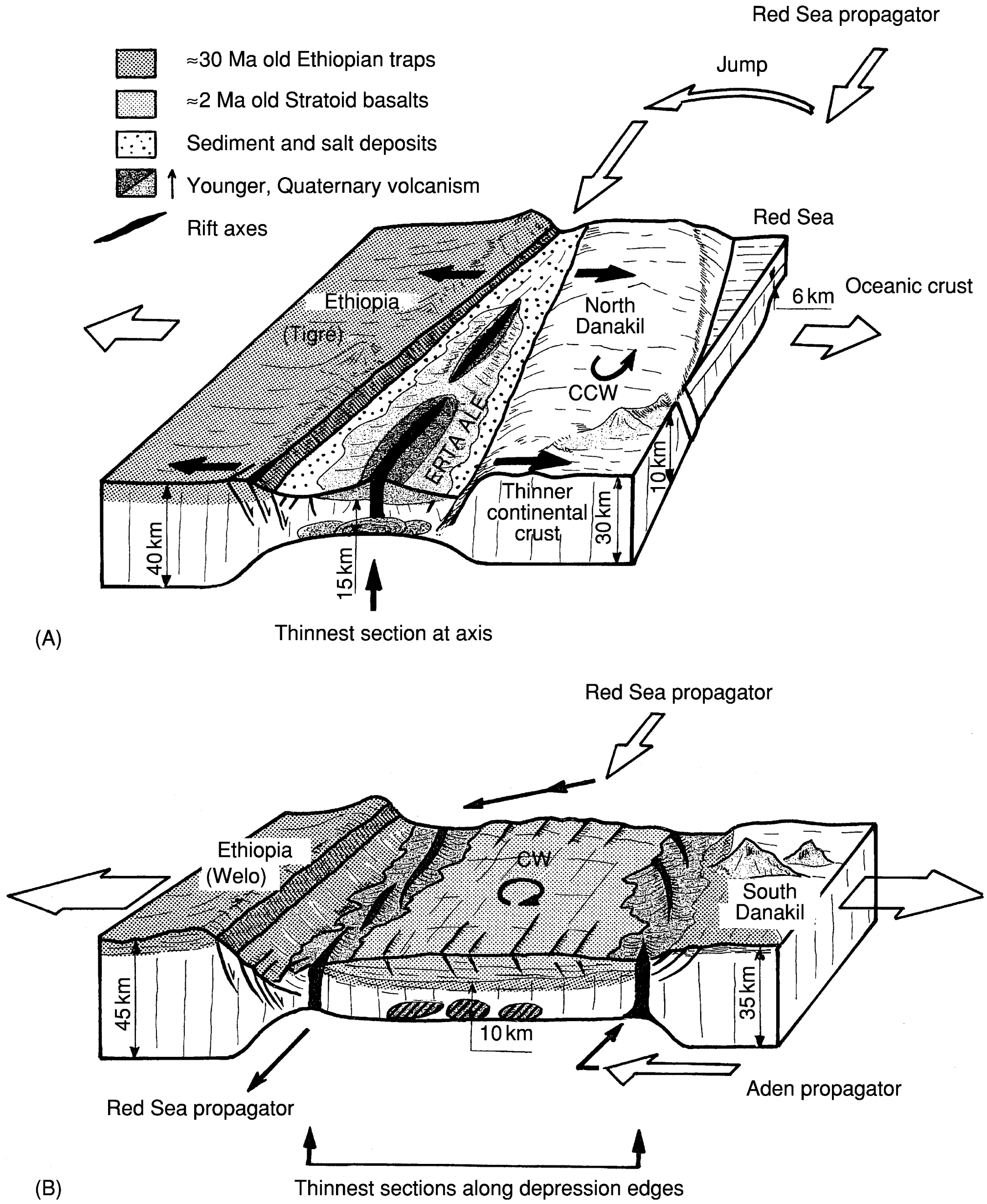


Figure 9 Schematic block diagrams of continental propagation in the (A) northern and (B) central Afar depression, Africa, viewed from the south-east. The zone between the duelling Aden and Red Sea propagators is rotating clockwise (CW) and deforming by bookshelf faulting. CCW, Counter-clockwise. Reproduced with permission from Manighetti I, Tapponnier P, Gillot PY, *et al.* (1998) Propagation of rifting along the Arabia–Somalia plate boundary: into Afar. *Journal of Geophysical Research* 103: 4947–4974.

only a few kilometres, through oceanic propagators with offsets on the order of 10–100 km, up to offsets of several hundred kilometres at microplate tectonic scales, and several thousand kilometres at continental rifting scales, appears to be the primary mechanism by which Earth's accretional plate boundary geometry is reorganized. Although conceptually simple, the propagating-rift hypothesis has important implications for plate tectonic evolution. It explains quantitatively the existence of several classes of structures, including pseudofaults, failed rifts, and zones of transferred lithosphere, that are oblique to ridges and transform faults and thus previously seemed incompatible with plate tectonic theory. It explains why passive continental margins are not parallel to the oldest seafloor isochrons, but are instead pseudofaults, bounding lithosphere created on propagating spreading centres and indicating the direction of the continental breakup propagators. It also explains the large-scale reorganization of many seafloor-spreading systems, including both the origination and the termination of many fracture zones, as well as the formation of some transient microplates that appear to be the modern analogues of large-scale spreading-centre jumps. This hypothesis provides a mechanistic explanation for the way in which many (if not all) spreading-centre jumps occur, why they occur in systematic patterns, and how spreading centres reorient when the direction of seafloor spreading changes. It also explains the origin of large areas of petrologically diverse seafloor, including the major abyssal ferrobasalt provinces. The common occurrence of rift propagation over a wide range of tectonic environments and spreading rates indicates that it represents an efficient mechanism of adjustment of extensional plate boundaries to the forces driving plate motions.

See Also

Geomorphology. **Plate Tectonics**. **Tectonics:** Seismic Structure At Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Bird RT and Naar DF (1994) Intratransform origins of mid-ocean ridge microplates. *Geology* 22: 987–990.

Hey RN (1977) A new class of pseudofaults and their bearing on plate tectonics: A propagating rift model. *Earth and Planetary Science Letters* 37: 321–325.

Hey RN, Duennebier FK, and Morgan WJ (1980) Propagating rifts on mid-ocean ridges. *Journal of Geophysical Research* 85: 3647–3658.

Hey RN, Sinton JM, and Duennebier FK (1989) Propagating rifts and spreading centers. In: Winters EL, Hussong DM, and Decker RW (eds.) *Decade of North American Geology: The Eastern Pacific Ocean and Hawaii,* pp. 161–176. Boulder, CO: Geological Society of America.

Hey RN, Johnson PD, Martinez F, *et al.* (1995) Plate boundary reorganization at a large-offset, rapidly propagating rift. *Nature* 378: 167–170.

Kleinrock MC and Hey RN (1989) Migrating transform zone and lithospheric transfer at the Galapagos 95.5° W propagator. *Journal of Geophysical Research* 94: 13 859–13 878.

Manighetti I, Tapponnier P, Gillot PY, *et al.* (1998) Propagation of rifting along the Arabia–Somalia plate boundary: into Afar. *Journal of Geophysical Research* 103: 4947–4974.

McKenzie D and Jackson J (1986) A block model of distributed deformation by faulting. *Journal of the Geological Society, London* 143: 349–353.

Naar DF and Hey RN (1991) Tectonic evolution of the Easter microplate. *Journal of Geophysical Research* 96: 7961–7993.

Schouten H, Klitgord KD, and Gallo DG (1993) Edge-driven microplate kinematics. *Journal of Geophysical Research* 98: 6689–6701.

Searle RC, Bird RT, Rusby RI, and Naar DF (1993) The development of two oceanic microplates: Easter and Juan Fernandez microplates, East Pacific Rise. *Journal of the Geological Society* 150: 965–976.

Tapponnier P, Armijo R, Manighetti I, and Courtillot V (1990) Bookshelf faulting and horizontal block rotations between overlapping rifts in southern Afar. *Geophysical Research Letters* 17: 1–4.

Vink GE (1982) Continental rifting and the implications for plate tectonic reconstructions. *Journal of Geophysical Research* 87: 10 677–10 688.

Seismic Structure At Mid-Ocean Ridges

S M Carbotte, Columbia University, New York, NY, USA

Introduction

Oceanic crust is created at mid-ocean ridges (*see* **Tectonics:** Mid-Ocean Ridges) as mantle material upwells and undergoes pressure-release melting in response to ongoing seafloor spreading. As mantle melts rise to the surface and freeze, they form an internally stratified crust of extrusive basalts and sheeted dykes underlain by layered and massive gabbros. Spreading rate has long been recognized as a fundamental variable governing crustal accretion at ridges, with first-order differences observed in a wide range of ridge properties. However, significant changes in ridge properties are also observed along the ridge axis at any given spreading rate, which suggests that factors other than the rate of plate separation contribute to the local supply and distribution of magma from the mantle. Seismic methods permit imaging of structures within the crust that result from magmatic processes at mid-ocean ridges and provide important insights into the role of spreading rate and magma supply in crustal creation.

Since the early days of seafloor exploration, seismic studies, which rely on the propagation of sound waves through rocks, have been the primary tool used to investigate the internal structure of the oceanic crust (*see* **Seismic Surveys**). These studies reveal two primary seismic layers, which are generally believed to correspond to lithological structures in the crust: seismic layer 2 corresponds to the dykes and basaltic lava flows that form the shallow crust, and layer 3 is associated with the massive and sheeted gabbros that form the lower crust. Seismic methods fall into two categories: reflection studies, which are based on the reflection of near-vertical seismic waves from interfaces where large contrasts in density and/or elastic properties are present, and refraction studies, which exploit the characteristics of seismic energy that travels horizontally as head waves through rock layers. Reflection methods provide continuous images of crustal horizons and permit efficient mapping of small-scale variations over large regions. Locating these horizons at their correct depths within the crust requires knowledge of the seismic velocity of crustal rocks, which is poorly constrained from reflection data. Refraction techniques provide detailed information on crustal velocity structure but typically result in relatively sparse measurements that represent large spatial averages. Hence, the types of information obtained from reflection and refraction methods are highly complementary, and these data are often collected and interpreted together.

At mid-ocean ridges, three crustal horizons are found where contrasts in elastic properties are sufficiently large that the horizons can be mapped with reflection techniques. These include seismic layer 2A (which is commonly assumed to correspond to the layer of lava flows (extrusives) that caps the oceanic crust), the shallow magma chamber from which the crust is formed, and the Moho (which marks the crust–mantle boundary). Each of these three structures and their main characteristics at mid-ocean ridges will be described here, and the implications of these observations for understanding how oceanic crust is created will be summarized. In the final section, changes in crustal structure at ridges spreading at different rates and the prevailing models to account for these variations will be described.

Seismic Layer 2A

Early Studies

Seismic layer 2A was first identified in the early 1970s from analysis of refraction data at the Reykjanes Ridge south of Iceland. This layer of low P-wave velocities (less than 3.5 km s^{-1}), which comprises the shallowest portion of the oceanic crust (Figure 1), was attributed to extrusive rocks with high porosities due to volcanically generated voids and extensive crustal fracturing. In the late 1980s a bright event corresponding to the base of seismic layer 2A was imaged for the first time using multichannel seismic-reflection data. This event is not a true reflection but rather a refracted arrival resulting from turning waves within a steep-velocity-gradient zone that marks the base of seismic layer 2A. Within this gradient zone, P-wave velocity rapidly increases to values typical of seismic layer 2B (more than 5.0 km s^{-1}) over a depth interval of about 100–300 m (Figure 1A). The 2A event is seen in the far offset traces of reflection data collected with long receiver arrays (more than

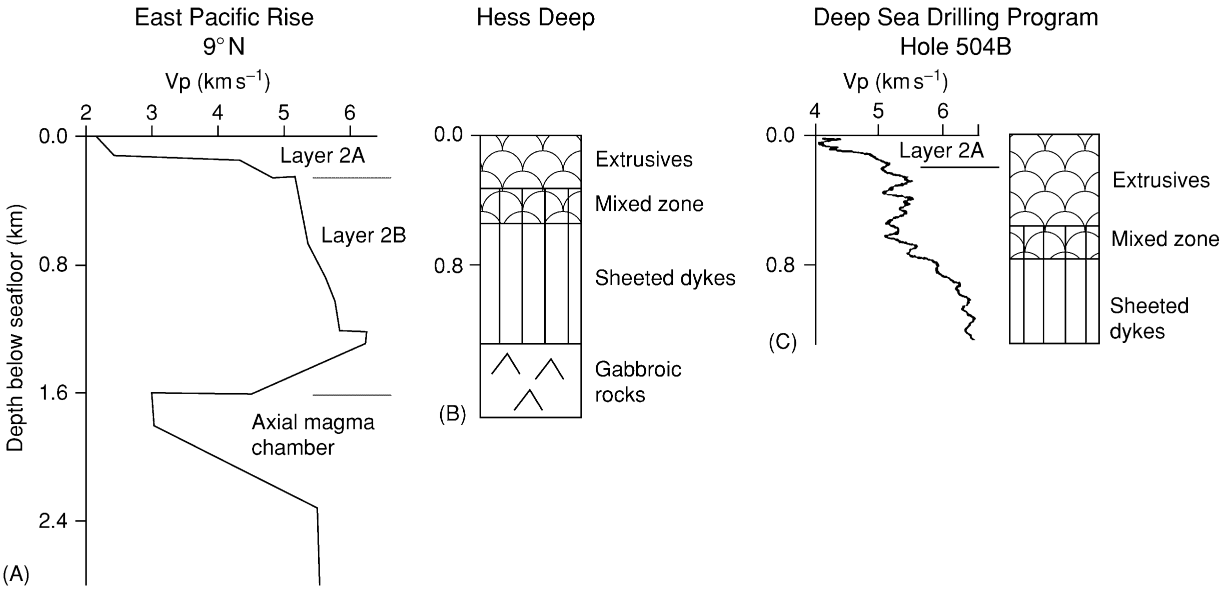


Figure 1 (A) Variations in seismic velocity with depth for newly formed crust at the East Pacific Rise. Layers 2A and 2B and the low velocities associated with the axial magma chamber are identified. (B) Lithological cross-section through the upper crust at Hess Deep, derived from submersible observations. (C) Comparison of P-wave velocities from *in situ* sonic logging within Deep Sea Drilling Program Hole 504B with the lithological units observed within the hole.

2 km) and can be successfully stacked, providing essentially continuous images of the base of layer 2A at mid-ocean ridges.

The Geological Significance of the Layer 2A–2B Transition: A Lithological Transition from Extrusives to Dykes or a Porosity Boundary within the Extrusives?

In most recent studies layer 2A near the ridge axis is assumed to correspond to extrusive rocks, and the base of layer 2A is assumed to correspond to a lithological transition to the sheeted-dyke section of the oceanic crust. The primary evidence for this lithological interpretation comes from studies at Hess Deep in the equatorial eastern Pacific. In this area, observations of fault exposures made from manned submersibles show that the extrusive rocks are on average 300–400 m thick, similar to the thickness of layer 2A measured near the crest of the East Pacific Rise (compare Figure 1A and B). However, the Hess Deep studies also revealed significant variability in the thickness of the extrusive layer (total range of 200–800 m) over horizontal distances of only a few hundred metres. Variability on this scale is well below the spatial resolution of seismic studies, which can provide only a smoothed view of the lithological heterogeneity that may be present.

Other researchers have suggested that the base of layer 2A may correspond to a porosity boundary within the extrusive section associated, perhaps, with a fracture front or hydrothermal alteration. This interpretation is based primarily on observations from a deep crustal hole located off the coast of Costa Rica, which was drilled as part of the Deep Sea Drilling Program (DSDP). Within this hole (504B) there is a velocity transition zone that is located entirely within the extrusive section (Figure 1C). Here, a thin high-porosity section of rubbly basalts and breccia with P-wave velocities of approximately 4.2 km s^{-1} overlies a thick lower-porosity section of extrusives with higher P-wave velocities (5.2 km s^{-1}). However, the relevance of these observations to the geological interpretation of the velocity structure of the ridge crest is questionable. Crust at DSDP hole 504B is 5.9 Ma old, and it is well established that the seismic velocity of the shallow crust increases with age owing to crustal alteration (see below). Indeed the velocities within the shallowest extrusives at DSDP 504B (*ca.* 4 km s^{-1}) are much higher than those observed at the ridge crest (2.5–3 km s^{-1}), indicating that significant crustal alteration has occurred (compare Figure 1A and C).

Conclusive evidence of the geological nature of seismic layer 2A will probably require drilling or observations of faulted exposures of the crust made where seismic observations are also available. At present, the bulk of the existing sparse information favours the lithological interpretation, and layer 2A is commonly used as a proxy for the extrusive crust. If this interpretation is correct, mapping the layer 2A–2B boundary provides direct constraints on the eruption and dyke-injection processes that form the uppermost part of the oceanic crust.

Characteristics of Layer 2A at Mid-Ocean Ridges

At the East Pacific Rise, layer 2A is typically 150–250 m thick within the region where crust is currently being formed (Figures 2 and 3). Only minor variations in the thickness of this layer are observed along the ridge crest, except near transform faults and other ridge offsets where the layer thickens.

Across the ridge axis, layer 2A approximately doubles or triples in thickness over a zone approximately 2–6 km wide, indicating extensive accumulation of extrusives within this wide region (Figures 4 and 5). This accumulation may occur through lava flows that travel up to several kilometres from their eruption sites at the axis, either over the seafloor or perhaps through subsurface lava tubes. Volcanic eruptions that occur off-axis may also contribute to building the extrusive pile.

Along the axis of the intermediate-spreading Galapagos spreading centre, located in the equatorial Pacific, layer 2A displays a wide range of thicknesses, which are closely linked with other characteristics of the ridge axis (Figure 6). Closest to the Galapagos Hotspot, a thin layer 2A is observed (150–350 m) beneath a shallow axial high, similar to that seen at fast-spreading ridges. Away from the hotspot-dominated portion of the ridge, the axial high disappears and layer 2A is thicker (*ca.* 300–600 m) and more variable along the axis. At the slow-spreading Mid-Atlantic Ridge the sparse existing data indicate that layer 2A does not systematically thicken away from the ridge axis, and it appears that the extrusive section is built largely within the floor of the median valley.

Axial Magma Chamber

Early Studies

Drawing on observations of the crustal structure of ophiolites (sections of oceanic or oceanic-like crust exposed on land) and the geochemistry of seafloor basalts, it was initially thought that mid-ocean ridges were underlain by large, essentially molten, magma reservoirs. However, prior to the 1980s, few actual constraints on the dimensions of magma chambers at ridges were available. Early seismic studies on the East Pacific Rise detected a zone of lower seismic

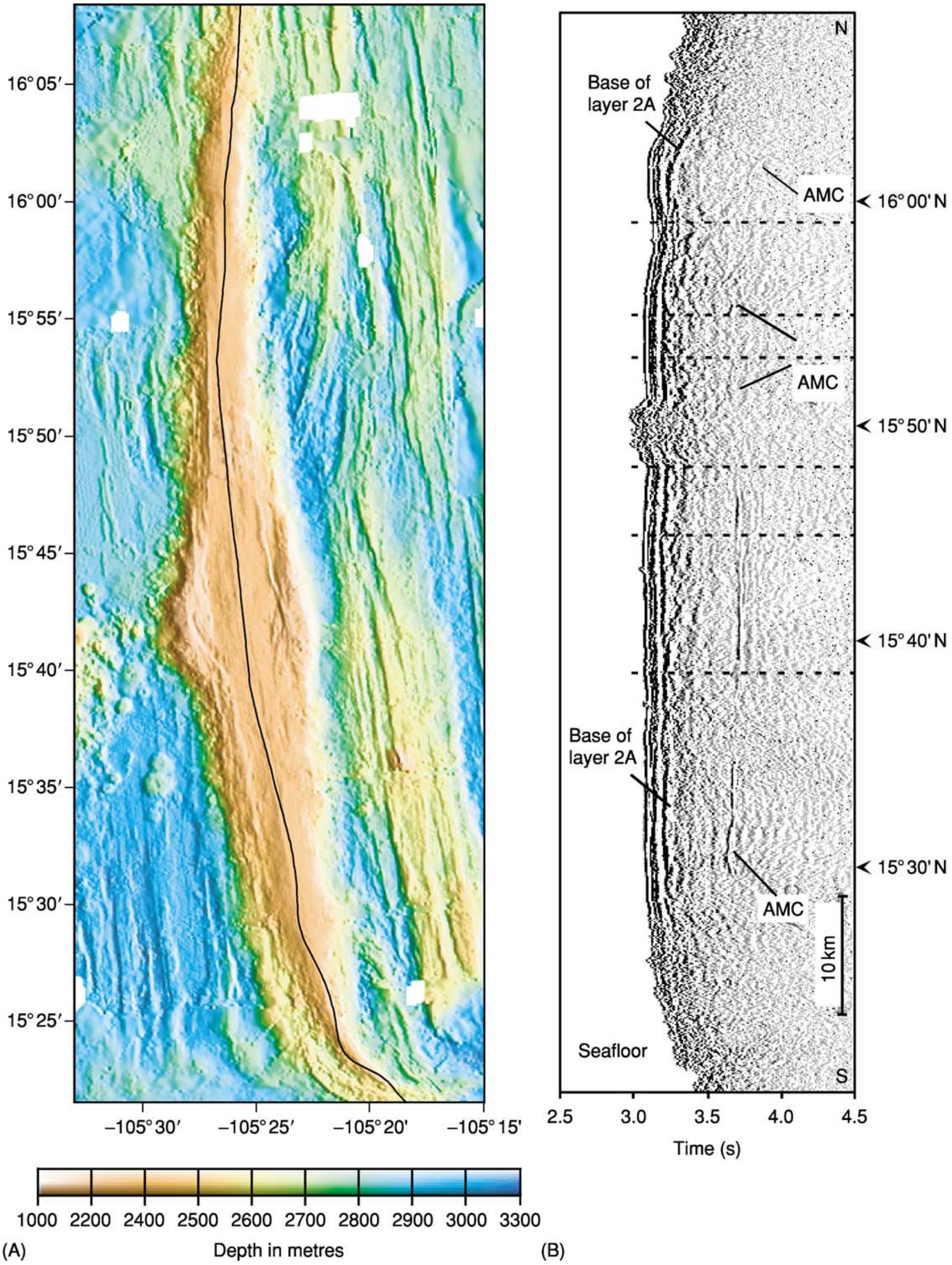


Figure 2 Example of a multichannel seismic line collected along the axis of the East Pacific Rise, showing the base of the extrusive crust (layer 2A) and the reflection from the top of the axial magma chamber (AMC). (A) The bathymetry of the ridge axis, with the location of the seismic profile indicated by the black line. (B) The dashed lines on the seismic section mark the locations of very small offsets that are observed in the narrow depression along the axis where most active volcanism is concentrated. From Carbotte SM, Ponce-Correa G, and Solomon A (2000) *Journal of Geophysical Research* 105: 2737–2759.

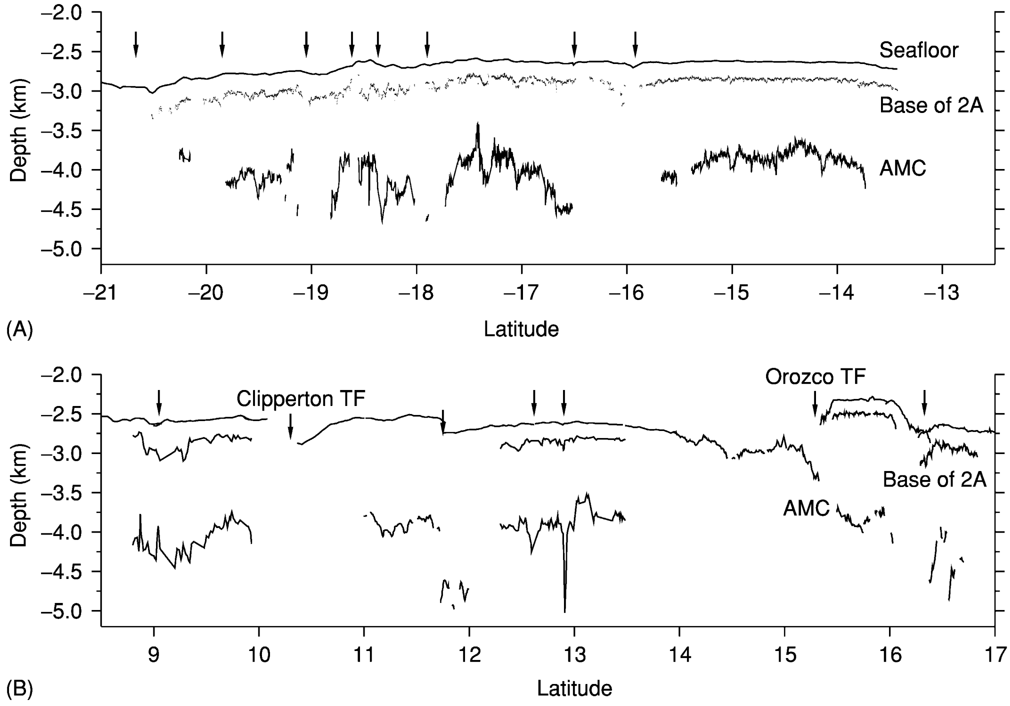


Figure 3 Cross-sections along the axis of (A) the southern and (B) the northern East Pacific Rise, showing depths to the seafloor, the base of the extrusive crust, and the axial magma chamber (AMC) reflection. This compilation includes all multichannel reflection data available along this ridge. Labelled arrows show the locations of transform faults (TF). Other arrows mark the locations of smaller discontinuities of the ridge axis known as overlapping spreading centres. (Top panel) Hooft EE, Detrick RS, and Kent GM (1997) *Journal of Geophysical Research* 102: 27319–27340; (bottom panel) Kent GM., Harding AJ, and Orcutt JA (1993) *Journal of Geophysical Research* 98: 13945–13696; Detrick RS, Buhl P, Vera E, Mutter JC, Orcutt JA, Madsen J, and Brocher T (1987) *Nature* 326: 35–41; Babcock JM, Harding AJ, Kent GM, and Orcutt JA (1998) *Journal of Geophysical Research* 103: 30451–30467; Carbotte SM, Ponce-Correa G, and Solomon A (2000) *Journal of Geophysical Research* 105: 2737–2759.

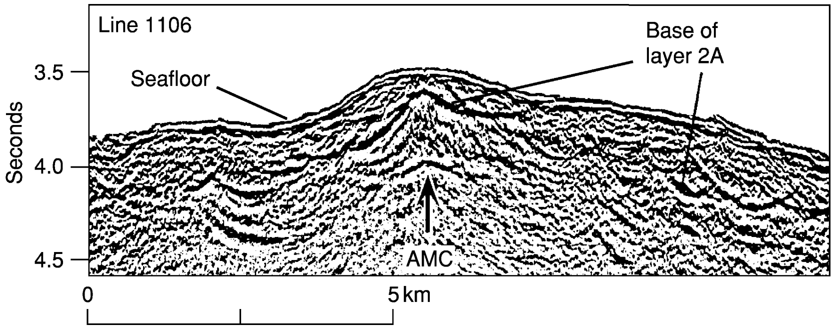


Figure 4 Example of a multichannel seismic profile shot across the ridge axis of the southern East Pacific Rise at 17°28′ S. The axial magma chamber (AMC) reflection and the event from the base of layer 2A can both be seen. From Carbotte SM, Mutter JC, and Wu L (1997) *Journal of Geophysical Research* 102: 10165–10184.

velocity beneath the ridge axis, as expected for a region containing melt. A bright reflector was also found, indicating the presence of a sharp interface with high acoustic impedance contrast within the upper crust. In the mid-1980s an extensive seismic reflection and refraction experiment was carried out on the northern East Pacific Rise by researchers from the University of Rhode Island, Lamont–Doherty Earth Observatory, and Scripps Institution of Oceanography. This study imaged a bright sub-horizontal reflector located 1–2 km below the seafloor along much of the ridge. In several locations this reflector was found to be phase reversed relative to the seafloor reflection, indicating that it resulted from an interface

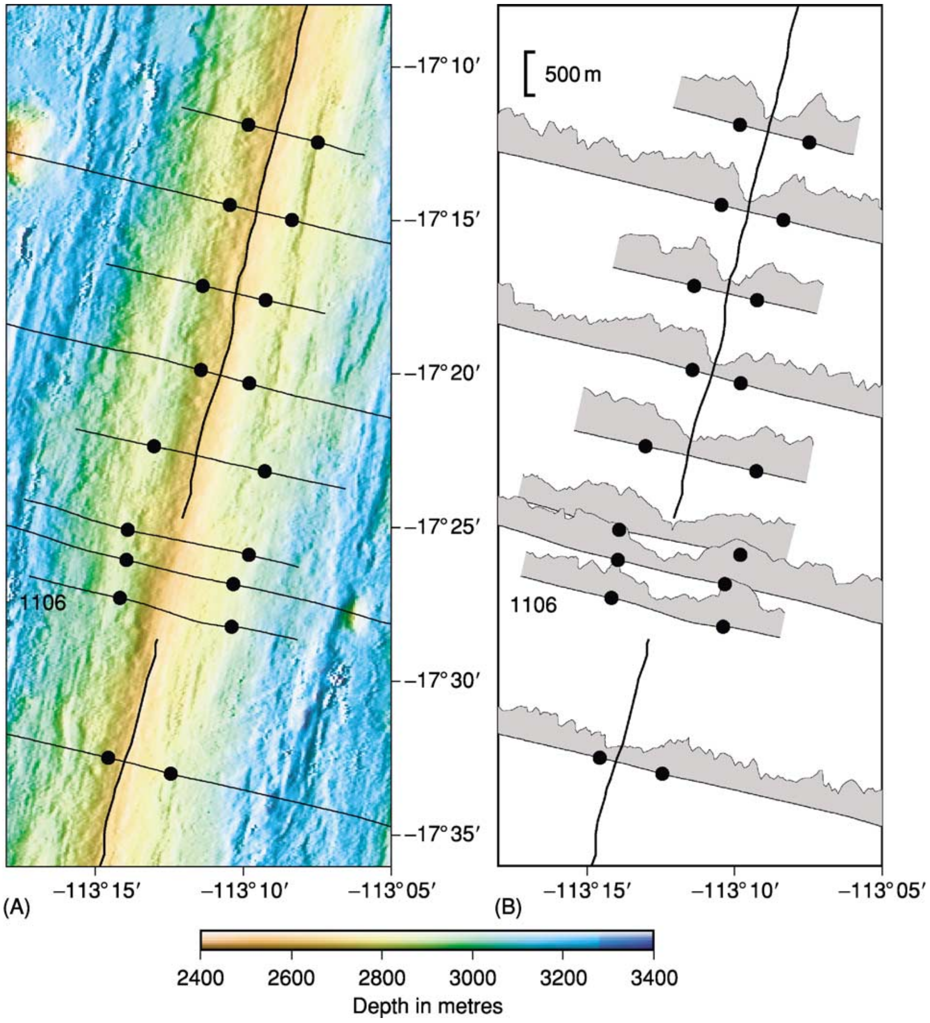


Figure 5 The thickening of the seismically inferred extrusive crust (layer 2A) across the axis of the southern East Pacific Rise. (A) Bathymetry map of the region with the locations of cross-axis seismic lines shown as light lines. The bold black line shows the location of the narrow depression along the ridge axis where most volcanic activity occurs. The black dots indicate the width of the region over which the seismically inferred extrusives accumulate, as interpreted from the data shown in Figure 5B. (B) Thickness of the extrusive crust inferred from the seismic data along each cross-axis line. Black dots mark where layer 2A reaches its maximum thickness away from the axis. Seismic line 1106 (shown in Figure 4) is labelled.

with an abrupt drop in seismic velocity. Based on its reversed phase and high amplitude, this event is now recognized as a reflection from a lens of magma located at the top of what is commonly referred to as an axial magma chamber.

Seismic refraction and tomography experiments show that this reflector overlies a broader region within which seismic velocities are reduced relative to normal crust (Figure 7). This low-velocity zone is approximately 5 km wide at shallow depths, possibly widening slightly at the base of the crust. Because of the relatively small velocity anomaly associated with much of this low-velocity zone (less than 1 km s^{-1}), this region is interpreted as hot largely solidified rock and crystal mush containing only a few percent partial melt.

Characteristics of the Axial Magma Chamber at Mid-Ocean Ridges

Several seismic-reflection studies have now been carried out along the fast-spreading East Pacific Rise, imaging over 1400 km of ridge crest (Figure 3). A reflection from the roof of the magma chamber is detected beneath about 60% of the surveyed region and can be traced continuously in places for tens of kilometres. This reflector is found at a depth of 1–2 km below the seafloor and deepens and disappears towards major offsets of the ridge axis, including transform faults and overlapping spreading centres (Figure 3). Most volcanic activity along the East Pacific Rise is concentrated within a narrow depression, less than 1 km wide, which is interrupted by

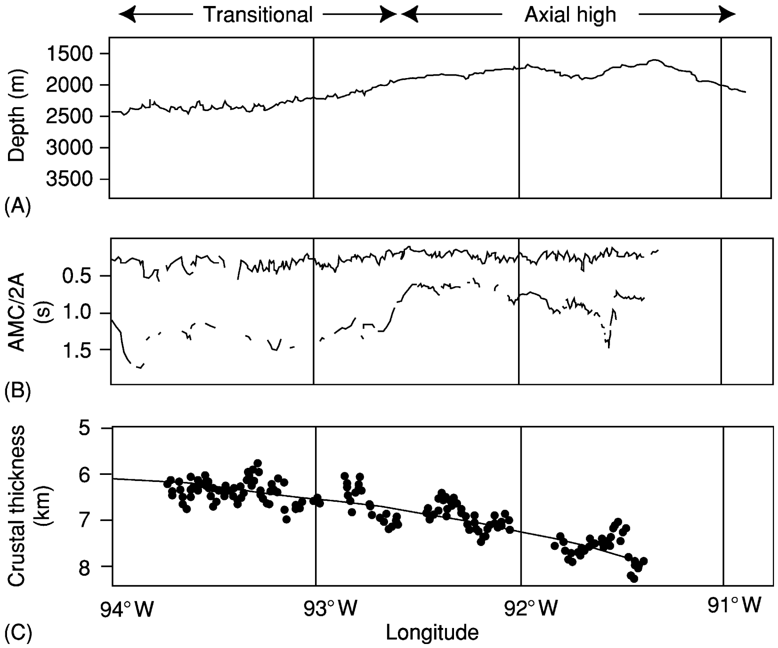


Figure 6 Crustal structure along the intermediate-spreading Galapagos Spreading Centre. (A) Seafloor depth along the ridge axis. (B) Two-way travel times to the base of layer 2A (top line) and the axial magma chamber (AMC) reflection (bottom line). (C) Crustal thickness derived from two-way travel time to Moho on the basis of velocities derived from refraction data. Black line shows best-fit polynomial to crustal-thickness data. (Reproduced from Detrick RS, Sinton JM, Ito G, *et al.* (2002) Correlated geophysical, geochemical, and volcanological manifestations of plume–ridge interaction along the Galápagos Spreading Centre. *Geochemistry Geophysics Geosystems* 3: 8501; DOI 10.1029/2002GC000350.)

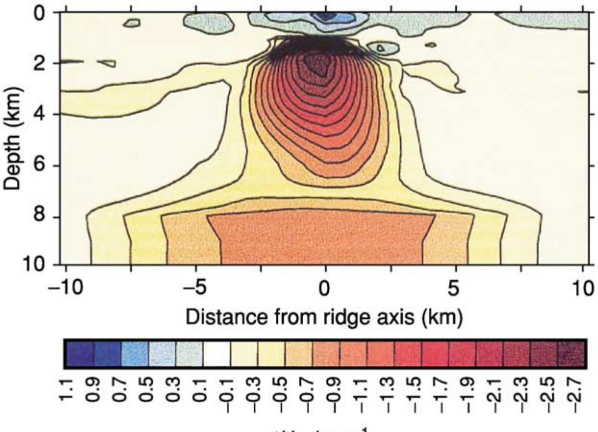


Figure 7 Seismic velocity structure of a fast-spreading ridge, showing the region of low velocities associated with melt and hot rock at the ridge axis. The velocity model is derived from a tomography experiment at 9°30′ N on the East Pacific Rise. (Reproduced from Dunn RA, Toomey DR, and Solomon SC (2000) Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise at 9°30′ N. *Journal of Geophysical Research* 105: 23 537–23 555.)

small steps or offsets, which may be the boundaries between individual dyke swarms. In many places, the magma-chamber reflector does not disappear beneath these offsets (Figure 3). However, changes in the depth and width of the reflector are often seen. Seismic-tomography studies show that the broader region of low velocities associated with the crustal magmatic system pinches and narrows beneath these small offsets. These results suggest that segmentation of the axial magma chamber is associated with the full range of offsets observed along the ridge axis.

Migration of seismic profiles shot perpendicular to the ridge axis reveals that the magma-chamber reflection arises from a narrow feature that is typically less than 1 km wide (total range from 200 m to 4 km; Figure 4). Refraction data and waveform studies of the magma-chamber reflection suggest that it arises from a thin body of magma a few hundred to perhaps a few tens of metres thick, leading to the notion of a magma lens or sill. Initial studies assumed that this lens contained pure melt. However, recent research suggests that much of the magma lens may have a significant crystal content (more than 25%), with regions of pure melt limited to pockets only a few kilometres or less in length along the axis.

Magma-lens reflections similar to those imaged beneath the East Pacific Rise have been imaged along intermediate-spreading ridges, including the Galapagos Spreading Centre (Figure 6), south-east Indian Ridge, and Juan de Fuca Ridge, and at the back-arc spreading centre in the Lau Basin. In these areas, magma-lens reflectors are typically found deeper in the crust, at depths of 2.5–3 km, although shallower reflections are observed in a few places.

Along the slow-spreading Mid-Atlantic Ridge, evidence for magma lenses has been found along the Reykjanes Ridge. Here, an intracrustal reflection at a depth of approximately 2.5 km is observed, which is similar to the depths of magma-lens events observed beneath the intermediate-spreading ridges. Seismic data have been collected elsewhere along the Mid-Atlantic Ridge with little evidence of magma-lens reflections, possibly owing to imaging problems associated with the rough seafloor topography typical of this ridge. However, there is evidence from refraction data and seismicity studies that large crustal magma bodies are not common beneath this ridge. Microearthquake data show that earthquakes can occur to depths of 8 km beneath the Mid-Atlantic Ridge, indicating that the entire crustal section is sufficiently cool for brittle failure. In other areas, slightly reduced velocities within the crust have been identified, indicating warmer temperatures and possibly the presence of small pockets of melt within the crust.

The prevailing model for magma chambers beneath ridges (Figure 8) incorporates both the geophysical constraints on chamber dimensions described above and geochemical constraints on magma-chamber processes. At fast-spreading ridges (Figure 8A), the magma chamber consists of a narrow and thin melt-rich magma lens, which overlies a broader crystal-mush zone and a surrounding region of hot but solidified rock. The dyke-injection events and volcanic eruptions that build the upper crust are assumed to tap the magma lens. The lower crust is formed from the crystal residuum within the magma lens and from the broader crystal-mush zone. Observations of ophiolites suggest that the injection of sills that tap magma directly from the upper mantle also contributes to the lower crustal section. Both seismic and seafloor compliance studies from the 9°–10° N region of the East Pacific Rise indicate melt accumulation at the base of the crust, within either melt-rich sills or a broader partially molten zone.

At slow-spreading ridges (Figure 8B) a short-lived dyke-like crystal-mush zone without a steady-state magma lens is envisioned. At these ridges volcanic eruptions occur and the crystal-mush zone is replenished during periodic magma-injection events from the mantle. Observations of seafloor fault exposures of crust created at slow-spreading ridges reveal a heterogeneous crustal section, where altered and deformed lower-crustal and upper-mantle rocks are unconformably overlain by lavas in some locations. Crustal accretion at these ridges is inferred to be a highly episodic process, with the internal structure of the crust being strongly disrupted by faulting.

Moho

The base of the crust is marked by the Mohorovicic Discontinuity, or ‘Moho’, where P-wave velocities increase from values typical of lower-crustal rocks (7–7.5 km s^{-1}) to mantle velocities (more than 8.0 km s^{-1}). The change in P-wave velocity is often sufficiently abrupt that a sub-horizontal Moho reflection is observed in reflection data, from which the base of the crust can be mapped. Depths to the Moho derived from seismic-refraction studies provide our best estimates of crustal thickness and are used to study how total crustal production varies in different ridge settings.

Characteristics of the Moho at Mid-Ocean Ridges

Reflection Moho is imaged in much of the data collected at the East Pacific Rise (Figure 9). It can often be traced beneath the region of lower-crustal velocities found at the ridge and occasionally beneath the magma-lens reflection itself. Where information on crustal velocities is available, average crustal thicknesses of 6–7 km are measured. There is no evidence for thickening away from the ridge crest, indicating that the crust acquires its full thickness within a narrow zone at the axis.

Unlike at the fast- and intermediate-spreading ridges, at the slow-spreading Mid-Atlantic Ridge the Moho is difficult to identify in seismic-reflection data, possibly owing to poor imaging conditions or a difference in the geological nature of this boundary. At this ridge, information on crustal thickness and variations in crustal structure is derived primarily from seismic-refraction studies. Average crustal thicknesses are similar to those observed at fast-spreading ridges (6–7 km). However, significant crustal thinning is observed (by 1–4 km) towards transform faults and smaller ridge offsets (Figure 10B). These results are interpreted as reflecting a three-dimensional pattern of mantle upwelling or melt migration to the ridge, resulting in greater crustal production within the central regions of ridge segments away from ridge offsets.

Variations in crustal thickness within ridge segments are also observed along the East Pacific Rise (Figure 10A). However, in the region with the best data constraints (9°–10° N), the spatial relationships are opposite to those observed at the Mid-Atlantic Ridge. Here, the crust is approximately 1 km thinner, not thicker, in the portion of the segment where a range of ridge-crest observations indicate that active

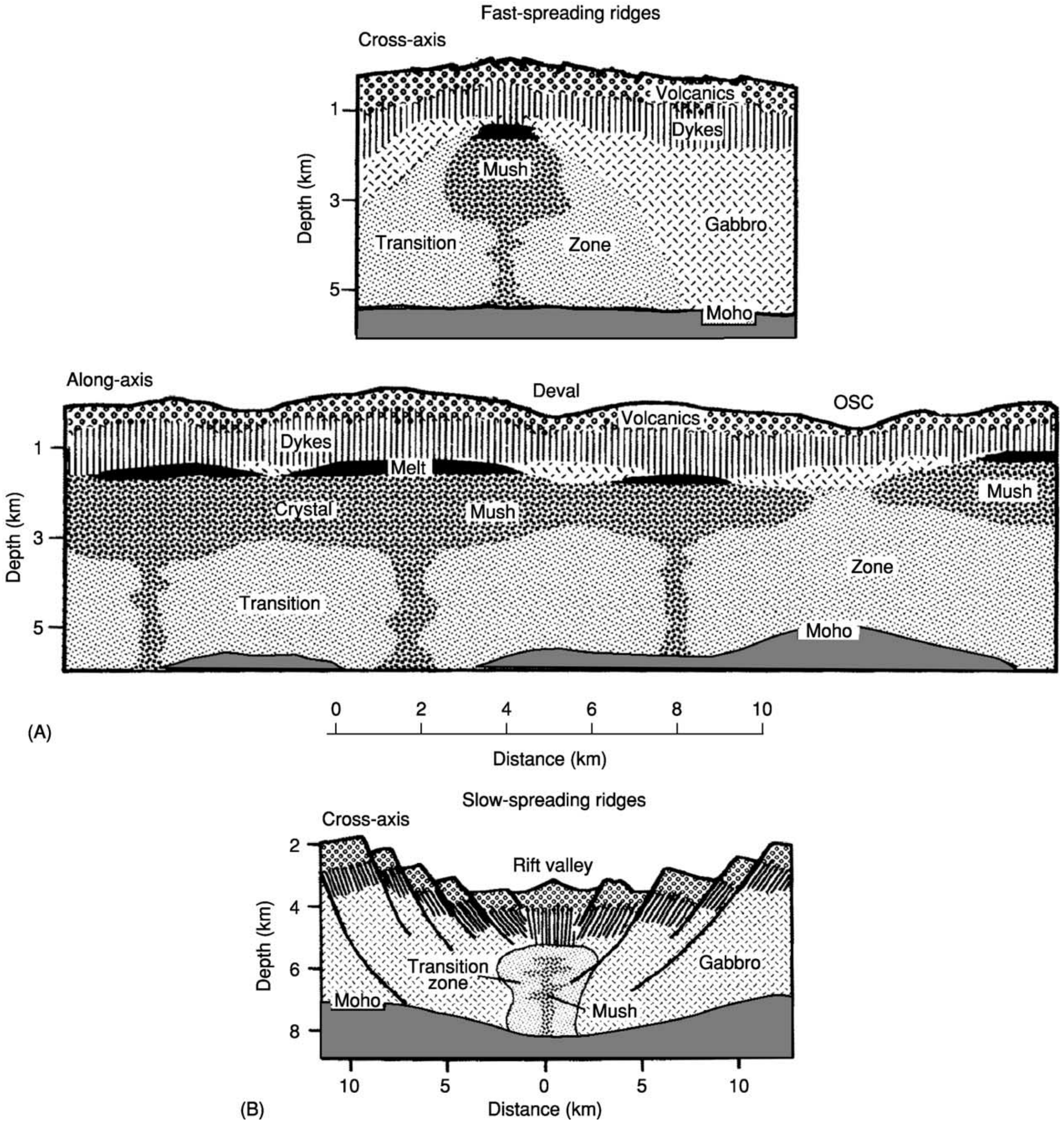


Figure 8 Schematic representation of the axial magma chamber beneath (A) fast- and (B) slow-spreading ridges. (A) At a fast-spreading ridge a thin zone of predominantly melt (black region) is located 1–2 km below the seafloor, grading downwards into a partially solidified crystal-mush zone. This region is in turn surrounded by a transition zone of solidified but hot rock. Along the ridge axis, the 'melt' sill and crystal-mush zone narrows and may disappear at the locations of ridge discontinuities (labeled Deval and OSC in the along-axis profile). (B) At a slow-spreading ridge a steady-state melt region is not present. Magma is periodically injected into the crust from the mantle with volcanic eruptions and the emplacement of small intrusive bodies, which crystallize to form the oceanic crust. (Reproduced from Sinton JA and Detrick RS (1992) Mid-ocean ridge magma chambers. *Journal of Geophysical Research* 97: 197–216.)

crustal accretion is focused. At this fast-spreading ridge the presence of a steady-state magma chamber and a broad region of hot rock (Figure 8) may permit efficient redistribution of magma away from regions of focused delivery from the mantle. The absence of a steady-state magma chamber beneath the slow-spreading Mid-Atlantic Ridge may prohibit significant along-axis transport of magma, such that, at this ridge, thicker crust accumulates at the site of focused melt delivery.

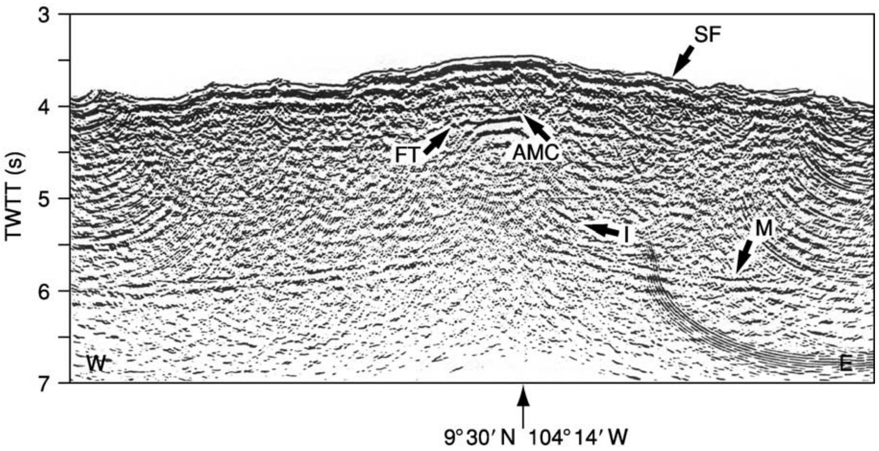


Figure 9 Multichannel seismic line crossing the East Pacific Rise at 9°30′ N, showing the Moho reflection (M), the seafloor (SF), the axial magma chamber (AMC) reflection, and other intracrustal reflections (FT, I). (Reproduced from Barth GA and Mutter JC (1996) Variability in oceanic crustal thickness and structure: multichannel seismic reflection results from the northern East Pacific Rise. *Journal of Geophysical Research* 101: 17 951–17 975.)

Variations in Mid-Ocean Ridge Structure

The seismic observations described in the previous sections reveal significant differences in the internal structure of the crust at fast- and slow-spreading ridges as well as within individual spreading segments, with important implications for crustal creation at mid-ocean ridges. Large gradients in crustal structure are observed at slow-spreading ridges, with crustal thickness often varying by a factor of two within individual ridge segments (Figure 10B). In comparison, at fast-spreading ridges only minor variations in crustal structure are observed within ridge segments, and crustal accretion appears to be a more uniform process at these rates. Indeed, throughout the fast-spreading-rate range (85–150 mm year^{-1}), average crustal structure is remarkably constant, with magma lenses imaged beneath much of the axis at similar widths and depths (Figure 11). In contrast, at slow-spreading ridges, steady-state magma bodies within the crust are rarely detected. Intermediate-spreading ridges (less than 85 mm year^{-1}) are of particular interest because they display characteristics from across the spreading-rate spectrum, the distribution of which appears to be closely linked with spatial variations in the supply of magma to the ridge. Where the ridge forms a shallow axial high, magma bodies have been observed at the shallow depths (less than 2 km) characteristic of fast-spreading ridges. Beneath the shallowly rifted sections that are more typical of intermediate-spreading ridges, magma bodies lie at a deeper level within the crust of 2.5–3 km (Figure 11).

At fast-spreading ridges, the shallowest crust, defined by seismic layer 2A, is uniformly thin along the ridge axis (*ca.* 200 m; Figure 12) and commonly thickens over a region several kilometres wide about the axis. In comparison, at slow-spreading ridges, the sparse available data suggest that a thicker layer 2A is developed along the axis and that full accretion of this layer occurs within a narrow region confined to the axial valley. Along some sections of intermediate-spreading ridges, layer 2A thickens away from the axis, as observed at the fast-spreading ridges, whereas in other regions, this layer appears to acquire its full thickness within the innermost axial zone. Assuming that layer 2A corresponds to the extrusive section, these differences in the accumulation of this layer could reflect differences in eruption parameters such as eruptive volumes, lava-flow viscosity and morphology, and the dominance of fissure versus point-source eruptions. Where a wide zone of extrusive-layer thickening is observed along fast- and portions of intermediate-spreading ridges, low-viscosity lobate and sheet flows may predominate, forming thin flows that travel for significant distances from eruptive fissures at the axis. Large-volume pillow-flow eruptions and eruptions that quickly localize at point sources forming local volcanic constructions may be more common at the intermediate- and slow-spreading ridges, where little thickening of the extrusive layer away from the axis is inferred from the seismic data. The bounding faults of the axial valleys typically present at these ridges may serve to dam any far-travelling lobate and sheet flows, giving rise to full accumulation of the extrusives at the axis.

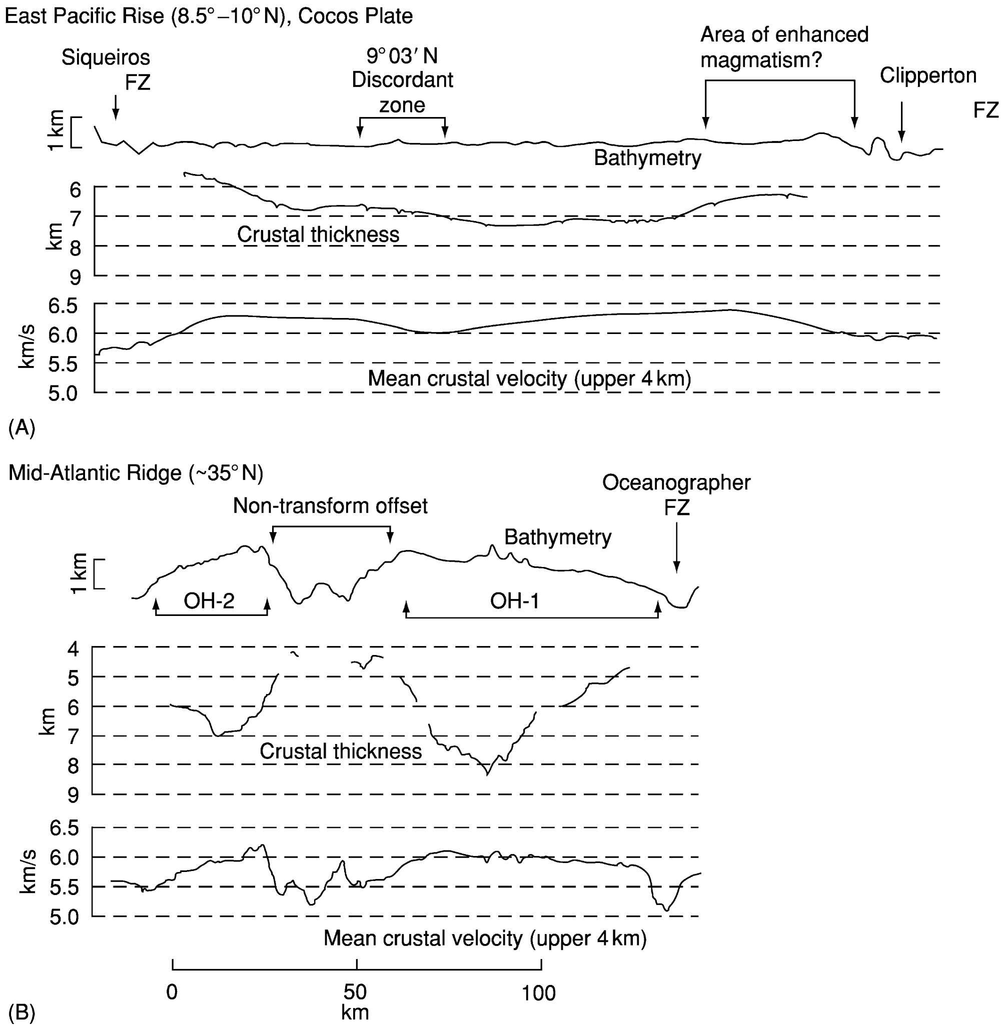


Figure 10 Comparison of crustal structures at (A) fast- and (B) slow-spreading ridges, showing along-axis variations in crustal thickness and the mean seismic velocity of the upper crust. (B) At slow-spreading ridges, the crust thins towards fracture zones and non-transform offsets. (A) Changes in crustal thickness towards discontinuities are more modest at fast spreading rates. (Reproduced from Canales JP, Detrick RS, Toomey DR, and Wilcock WSD (2002) Segment-scale variations in crustal structure of 150- to 300-ky-old fast spreading oceanic crust (East Pacific Rise, 8°15′ N–10°15′ N) from wide-angle seismic refraction profiles. *Geophysical Journal International* 152: 766–794.)

Although first-order differences are observed in a wide range of ridge properties with differences in spreading rate, several aspects of the seismic structure of ridges are surprisingly similar at all rates. The average thickness of the extrusive layer away from the ridge axis is comparable (*ca.* 350–650 m), and the total volume of extrusives produced by seafloor spreading may be largely independent of spreading rate. Average crustal thickness is also similar (6–7 km) across almost the entire spreading-rate range, and total crustal production does not depend on spreading rate except at the slowest rates (less than 15 mm year^{-1}; Figure 13). Below rates of 15 mm year^{-1}, the crust is thinner (2–4 km) and more variable in thickness, possibly because enhanced conductive heat loss in the uppermost mantle results in reduced melting.

What Controls the Depth at Which Magma Chambers Reside at Ridges?

Two main hypotheses have been put forward to explain the depths at which magma chambers are found at ridges. One hypothesis is based on the concept of a level of neutral buoyancy for magma within the oceanic crust. This model predicts that magma will rise until it reaches a level where the density of the surrounding country rock equals that of the magma. However, magma lenses at ridges lie at considerably greater depths than the neutral-buoyancy level

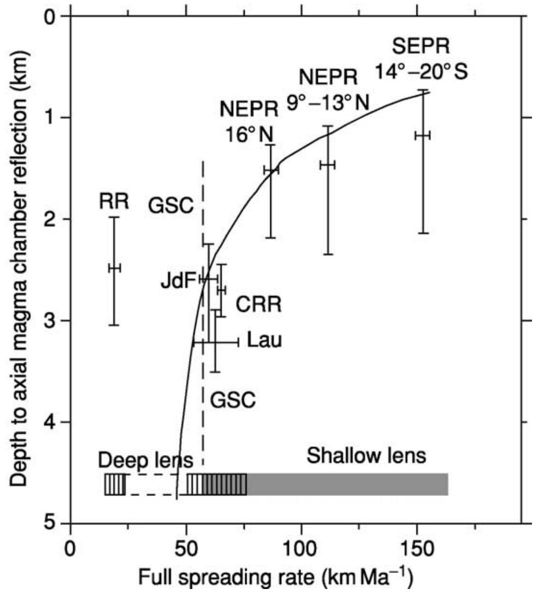


Figure 11 Average depth of magma-lens reflections beneath ridges versus spreading rate. Magma lenses lie within two distinct depth ranges of 1–2 km for fast-spreading ridges and 2.5–4 km for intermediate- and slow-spreading ridges. Both shallow and deep lenses are observed at some intermediate-spreading ridges. The curved line shows the depth to the 1200°C isotherm calculated from the ridge thermal model of Phipps Morgan J and Chen YJ. Data from different ridges are labelled: RR, Reykjanes Ridge; JdF, Juan de Fuca Ridge; GSC, Galápagos Spreading Centre; CRR, Costa Rica Rift; Lau, Lau Basin; NEPR; northern East Pacific Rise; SEPR, southern East Pacific Rise.

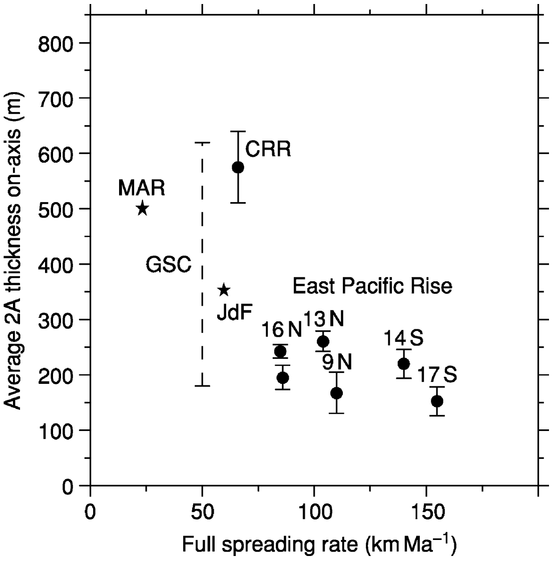


Figure 12 Thickness of the extrusive crust at the ridge axis versus spreading rate. For data obtained from detailed reflection surveys, average thicknesses are shown by black dots with standard deviations where available (solid lines) or thickness ranges (dotted lines). Data derived from other seismic methods are shown by stars. Data for the East Pacific Rise are labelled by survey location. CRR, Costa Rica Rift; MAR, Mid-Atlantic Ridge; JdF, Juan de Fuca Ridge; GSC, Galápagos Spreading Centre.

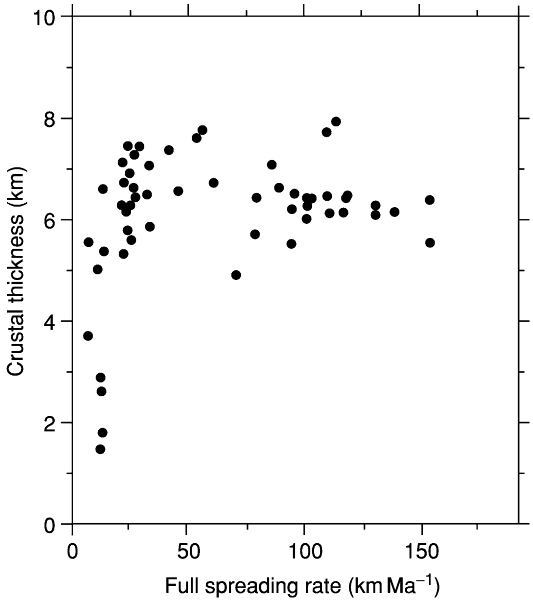


Figure 13 Crustal thickness versus spreading rate. Crustal thicknesses are determined from seismic data obtained away from fracture zones. (Reproduced from Bown JW and White RS (1994) Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth and Planetary Science Letters* 121: 435–449.)

predicted if the density of the magma is equivalent to that of lavas erupted onto the seafloor (2700 kg m^{-3}). Either the average density of magma is greater or mechanisms other than neutral buoyancy control magma-lens depth.

The alternative model hypothesizes that magma-chamber depth is controlled by the thermal structure of the ridge axis. In this model, a mechanical boundary, such as a freezing horizon or the brittle–ductile transition, prevents magma from rising to its level of neutral buoyancy. The depth of this boundary within the crust will be primarily controlled by the thermal structure of the ridge axis, which is expected to vary with spreading rate. The inverse relation between spreading rate and depth to low-velocity zones at ridges apparent in early seismic datasets provided compelling support for this hypothesis. Numerical models of ridge thermal structure have been developed that predict systematic changes in the depth to the 1200° C isotherm (a proxy for basaltic melts) with spreading rate that match the first-order depth trends for magma lenses (Figure 11). This model predicts a minor increase in lens depth within the fast-spreading-rate range and an abrupt transition to deeper lenses at intermediate spreading rates, consistent with the present dataset. The numerical models also predict that, at intermediate spreading rates, small variations in magma supply to the ridge can give rise to large changes in axial thermal structure. These models are supported by recent observations of the Galapagos spreading centre and the Southeast Indian Ridge. At these ridges, abrupt steps in the depths of crustal magma bodies occur where the ridge axis changes from an axial high to a shallowly rifted valley, although differences in crustal thickness (a proxy for magma supply) are modest (e.g. Figure 6). These recent investigations of intermediate-spreading ridges highlight the important role of spatial variations in magma supply independent of spreading rate in the process of crustal accretion at mid-ocean ridges.

See Also

Earth: Mantle; Crust. **Igneous Processes**. **Plate Tectonics**. **Seismic Surveys**. **Tectonics:** Mid-Ocean Ridges; Propagating Rifts and Microplates At Mid-Ocean Ridges. **Volcanoes**.

Further Reading

Bown JW and White RS (1994) Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth and Planetary Science Letters* 121: 435–449.

Buck WR, Delaney PT, Karson JA, and Lagabrielle Y (eds.) (1998) *Faulting and Magmatism at Mid-Ocean Ridges*. Geophysical Monograph 106. Washington, DC: American Geophysical Union.

Hooft EE and Detrick RS (1993) The role of density in the accumulation of basaltic melts at mid-ocean ridges. *Geophysical Research Letters* 20: 423–426.

Jacobson RS (1992) Impact of crustal evolution on changes of the seismic properties of the uppermost oceanic crust. *Reviews of Geophysics* 30: 23–42.

Karson JA and Christeson G (2003) Comparison of geological and seismic structure of uppermost fast-spread oceanic crust: insights from a crustal cross-section at the Hess Deep Rift. In: Goff J and Holliger K (eds.) *Heterogeneity in the Crust and Upper Mantle: Nature, Scaling and Seismic Properties*, pp. 99–129. New York: Kluwer Academic.

Kent GM, Singh SC, Harding AJ, *et al.* (2000) Evidence from three-dimensional seismic reflectivity images for enhanced melt supply beneath mid-ocean ridge discontinuities. *Nature* 406: 614–618.

Phipps Morgan J and Chen YJ (1993) The genesis of oceanic crust: magma injection, hydrothermal circulation, and crustal flow. *Journal of Geophysical Research* 98: 6283–6297.

Purdy GM, Kong LSL, Christeson GL, and Solomon SC (1992) Relationship between spreading rate and the seismic structure of mid-ocean ridges. *Nature* 355: 815–817.

Sinton JA and Detrick RS (1992) Mid-ocean ridge magma chambers. *Journal of Geophysical Research* 97: 197–216.

Solomon SC and Toomey DR (1992) The structure of mid-ocean ridges. *Annual Review of Earth and Planetary Science* 20: 329–364.

Mountain Building and Orogeny

M Searle, Oxford University, Oxford, UK

Introduction

The term orogeny, derived from the Greek word ‘oros’, meaning mountain, and ‘genesis’, meaning birth or origin, encompasses all the processes of mountain building. Orogenic belts generally occur along plate margins and are characterized by thickened crust, metamorphism, magmatism, flexure of the lithosphere, and large-scale crustal deformation. The average thickness of the oceanic crust is about 5 km, and that of continental crust around 35 km. In the Himalaya–Tibet region the crust has reached 75–80 km thick, double the normal thickness. High topography along mountain belts usually accords with crustal and lithospheric thickening as a result of plate collision.

Most mountain ranges are the result of plate collision processes at convergent plate margins. Oceanic crust is dominantly composed of basaltic or gabbroic rocks, and the oceanic lithosphere is relatively strong and dense. Oceanic crust, composed mainly of olivine and pyroxene, can be subducted along Benioff zones and recycled back into the mantle. Continental crust is composed dominantly of granites, gneisses, and upper crustal sedimentary rocks. The quartz and feldspar-rich continental crust is weaker and more buoyant, and cannot easily subduct into the denser mantle.

Following the advent of plate tectonic theory in the 1960s it was proposed by J T Wilson that the process of orogeny was a ‘cycle’ beginning with rifting of continents and development of passive ‘Atlantic-type’ continental margins, followed by seafloor spreading and ocean basin formation, and ending with subduction, ocean closure, and finally, continental collision. This process became known as the Wilson cycle.

Mountain belts can be broadly categorized into three types of plate collision zones. These occur as the result of the collision between two oceanic plates (e.g., Mariana–Philippine arc or the Caribbean island arc), secondly between a continental plate and an oceanic plate (e.g., Andes, or North American Cordillera), and thirdly between two continental plates (e.g., Alpine–Himalayan belt). Mountain belts can also form under the oceans, for example along the Mid-Atlantic ridge, or above hot spots, such as Hawaii or the Canary Islands, or in oceanic plateaux. Other mountain belts can form in the middle of continents, for example along the East African rift, where the mantle has domed up beneath linear rift valleys, as the continental plate begins to split apart.

Oceanic Island Arc Belts

Oceanic crust is generated along mid-ocean ridges, and becomes progressively older, colder, and denser with increasing distance from the ridge axis. When two oceanic plates converge, one plate flexes and bends beneath the other and begins to subduct. The subduction zone is marked by a deep ocean trench, an inclined zone of deep earthquakes and low heat flow (Wadati-Benioff zone), along which isotherms are buckled down. Subduction of slivers of sedimentary and basaltic rocks along the subduction zone can result in high-pressure metamorphism typical of the blueschist and eclogite facies. Typical examples of subduction zone plate boundaries are the Mariana trench between the Pacific and Philippine plates, or the Tonga–Kermadec trench between the Pacific and Indo-Australian plates. In both cases the mountain belt, comprising the fore arc and island arc complex is largely beneath sea-level. Andesitic island arcs are generated above subduction zones and are composed of explosive calc-alkaline volcanoes. Island arcs can produce small mountain ranges such as those along several West Pacific plate margins, or larger mountain ranges such as the Java–Banda arc in Indonesia. Marginal oceanic basins may open up behind the arc, forming back-arc spreading centres. Along the fore-arc region accretionary prisms of stacked thrust sheets can form as a result of scraping off sediments from the down-going plate. Subduction-related mélanges, including sepentinite mélanges with high-pressure rock clasts, are also typical of fore-arc regions. Ancient examples include the Franciscan complex in California.

Mature island arcs can produce significant mountain ranges, such as in Japan, where andesitic volcanoes are formed above a deep subduction zone. The Japan crust also consists of continental rocks, paired metamorphic belts, typical of fore-arc regions, and intrusive granitic magmas derived from partial melting of the down-going slab. The Japan Sea is a narrow ocean basin between the active margin in Japan and the passive continental margin along the east coast of mainland Asia.

Ophiolites and Mountain Building

Ophiolites are allochthonous sequences of oceanic crust and upper mantle emplaced onto continental margins. Some ophiolites are preserved as highly deformed slivers of serpentinized peridotites along ancient suture zones, or sites of plate collision. Some of the best preserved ophiolites occur as obducted thrust sheets emplaced over passive continental margins, typified by the Semail ophiolite in Oman, the Troodos ophiolite in Cyprus, and the Bay of Islands ophiolite in Newfoundland. In each of these cases a passive continental margin converged with an oceanic plate, which, instead of subducting downwards into the mantle, was detached and obducted on top of the continent. The subduction zone dipped away from the continental margin and the ophiolite complex was one of a series of thin-skinned thrust sheets emplaced onto the depressed continental margin.

In the Oman mountains of eastern Arabia, the ophiolite forms a thrust sheet over 15 km thick, approximately 700 km long, and over 100 km wide . It was formed at a spreading centre above an active subduction zone, along which old, cold oceanic basalts were subducted to depths of about 40 km, metamorphosed to garnet + clinopyroxene amphibolites, and accreted to the base of the mantle sequence peridotite. During the latest stages of ophiolite emplacement and mountain building the thinned leading edge of the passive continental margin was dragged down the subduction zone to depths of around 80 km or more and metamorphosed to eclogite facies. When the continental margin finally choked the subduction zone due to the buoyancy of the continental rocks, the deeply buried eclogites were exhumed rapidly back to the surface along the same subduction zone. The mountain building phase in Oman lasted approximately 25 My during the Late Cretaceous, when approximately 300–400 km width of the Tethyan ocean was closed.

Ophiolite obduction was the earliest phase of mountain building in the Alps and the Himalayas. Later continent–continent collision frequently overprints the earlier phases of orogeny, so mountain belts such as Oman are extremely important for deciphering subduction–obduction processes in the mountain-building cycle. A modern day example of a similar tectonic setting can be seen along the North Australian continental margin, which is currently subducting northwards beneath the accretionary prism of Timor island and the Banda volcanic arc of eastern Indonesia.

Andean-Type Mountain Building

The Andes mountains of South America extend for over 5000 km from Venezuela to southern Chile, and have formed as a result of the collision between the South American continental plate and the oceanic lithosphere of the Nazca and Pacific plates (Figure 1). The oceanic trench, the surficial trace of the subduction zone, lies about 70 km offshore the South American continental margin and reaches a depth of about 7 km off northern Chile. Earthquakes define a slab-like subduction zone inclined at 30°–40° eastwards beneath the Bolivian Andes, although to the south, the subducted lithosphere is almost horizontal beneath southern Peru and northern Chile before descending steeply into the mantle.

The geology of the Andes is dominated by linear belts of subduction-related granite–granodiorite batholiths and calc-alkaline volcanoes with shallow basins filled with continental red-beds. It has been estimated that around 300 km of crustal shortening has occurred in the central Andes as a result of east–west compression. Crustal shortening has resulted in increased crustal

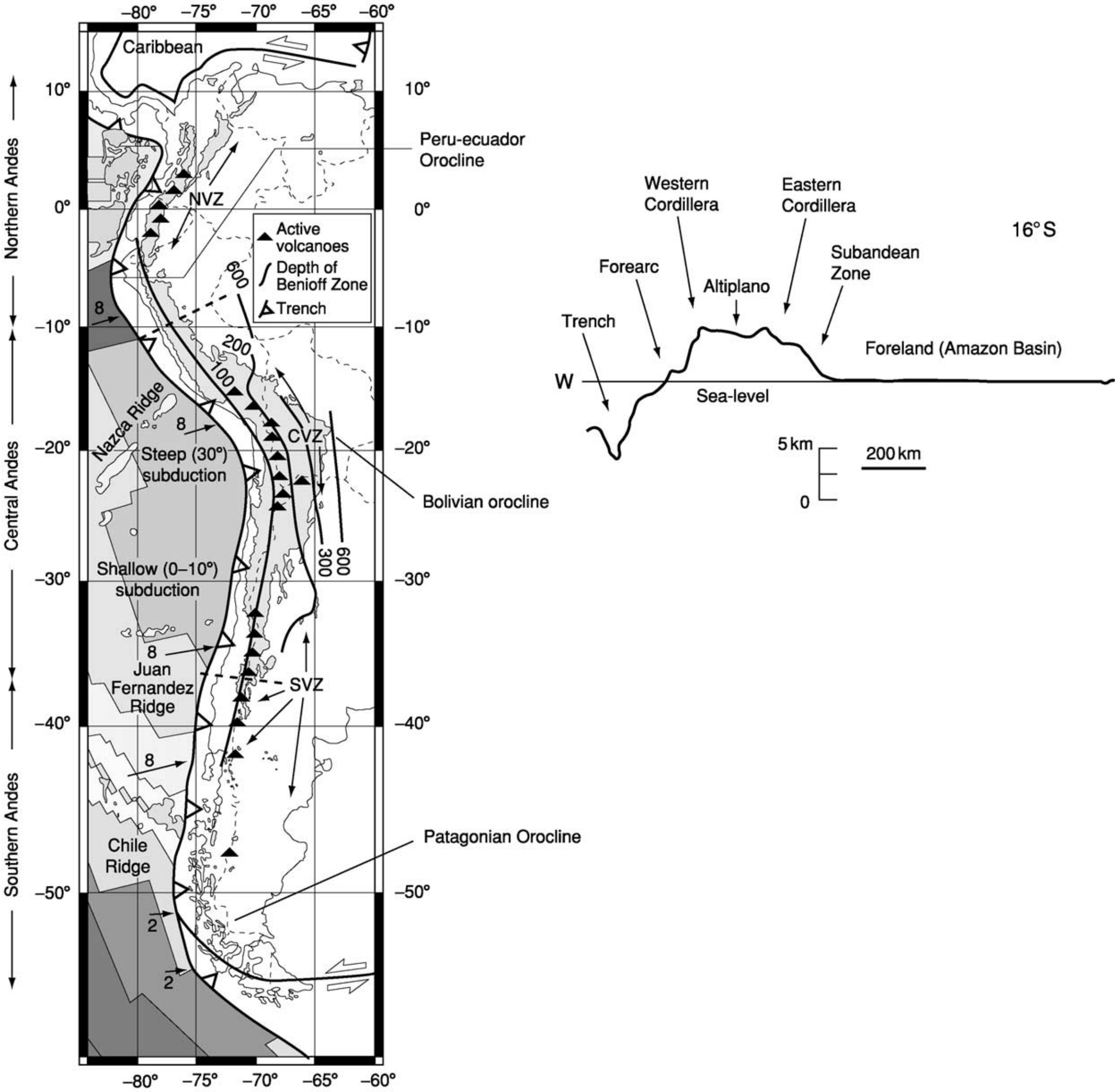


Figure 1 Map of the Andes in South America and topographic profile across the Central Andes showing the main tectonic features. Oceanic lithosphere is being subducted beneath the trench, and the oceanic crust is shaded according to age. Location of active volcanoes and the depth of the Benioff zone beneath the Andes is also shown. After Hancock PL and Skinner BJ (2000) *Oxford Companion to the Earth*, p. 21. © Oxford University Press.

thickening, ranging from ~35 km thickness east of the Andes, up to 80 km beneath the Bolivian Altiplano. The Altiplano has an average elevation of 4000 m above sea-level, and both the elevation and width of the Andes decrease both to the south and north. Along the eastern margin of the Bolivian Altiplano, a fold and thrust belt is present where relatively thin-skinned thrust sheets verge eastwards onto the Amazon foreland.

The North American Cordillera and Rocky Mountains have many geological features similar to those of the Andes, but also show important strike-slip, or transcurrent, faults like the San Andreas fault, movement along which has resulted in large-scale horizontal motion of crustal plates. Oceanic subduction beneath the continental margin has resulted in accretion of subduction-related deformed ophiolites, high-pressure metamorphic rocks, and mélanges. Granite batholiths, similar to the Andean batholiths, are aligned along mountain ranges like the Sierra Nevada, parallel to the continental margin. Andesitic volcanoes like the recently active Mount St Helens are regularly spaced along the range, along a linear belt roughly 200 km east of the trench.

Alpine-Type Mountain Building

The Alpine–Himalayan mountain belt formed as a result of the collision of the Africa–Arabia and Indian plates in the south with the Eurasian plate in the north. The Tethyan ocean, which separated these Gondwana continental fragments to the south from the Laurasian continent to the north, was mainly an east–west-aligned Permian and Mesozoic ocean that closed during the Late Cretaceous. The zone of collision stretches from the European Alps east along the Zagros mountains of Iran to the Himalaya. There are significant differences between the Alps, the Zagros, and the Himalayan mountain ranges, indicating that continental collision processes vary significantly along strike. The Oman mountains in Arabia show the early ophiolite obduction stage of the collision process, prior to continental collision. The Zagros mountains of Iran represent the initial continent–continent collision stage, and the Himalaya–Tibet orogeny represents the final stage of the continental collision process.

Seismic reflection and refraction profiles have revealed the deep crustal structure of the Alps. Together with surface geology, it has become possible to construct a geological section across the Alps (Figure 2). The Alpine orogeny includes a Cretaceous event and a Tertiary event, both of which involved subduction of crustal rocks down to eclogite facies depths and subsequent exhumation. Remnant ophiolites, marking the Tethyan suture zone, are preserved in the Valais region, and along the Ivrea zone. The Zermatt-Saas ophiolites have been subducted to over 80 km depth, metamorphosed to eclogite facies, and then exhumed rapidly during the early stage of Alpine orogeny. They were subsequently deformed by south-vergent backfolding and backthrusting along the Insubric line, a large-scale strike-slip and thrust fault that separates the European crust from the Adriatic crust.

Approximately 500 km of north–south convergence between Europe and Africa has been estimated from the restoration of cross-sections across the Alpine orogeny. Part of this shortening has been taken up by folding and thrusting of upper crustal, mainly sedimentary rocks of the Austro-Alpine, Penninic, and Helvetic nappes (large-scale recumbent folds overlying a thrust plane). The lower crust shows wedging of basement massifs including the Aar, Gotthard, and Mont Blanc thrust sheets in the northwest and the Monte Rosa, Gran Paradiso, and Dora Maira massifs in the south-east. The latter shows coesite-bearing eclogites, indicative of ultra-high-pressure metamorphism and deep subduction (around 100 km depth) of continental crustal material.

The later stage of the Alpine orogeny was characterized along the north and northwest by foreland-propagating thrusts extending from the Helvetic nappes into the Swiss molasse basin. The molasse basin developed in front of the rising Alps by flexural buckling of the European lithosphere caused by the excess load of the Alps. Along the south and south-east parts of the orogeny, the South Alpine nappes and Insubric Line structures show south-vergent backfolds and backthrusts, producing a retroshear, or bivergent orogeny with folds and thrusts verging in both directions.

Himalayan-Type Mountain Building

The Himalayan range stretches for over 2500 km from north-west Pakistan, eastwards across northern India, Nepal, Bhutan, and southern Tibet to south-west China, and is the type-example of a mountain range formed as a result of the collision of two continents. The Indian plate was a part of the southern supercontinent Gondwana until it rifted away from southern Africa, Madagascar, and Antarctica approximately 150 My ago. As the Indian Ocean spreading ridges formed in between the continents, India was pushed northwards at rapid plate tectonic rates of around 20 cm year^{-1}, until it collided with the southern margin of Asia approximately 50 My ago at equatorial latitudes. Palaeomagnetic data suggest that India underwent a marked reduction in the rate of northward motion since the time of collision to around 4–5 cm year^{-1}. The Indian plate has also undergone

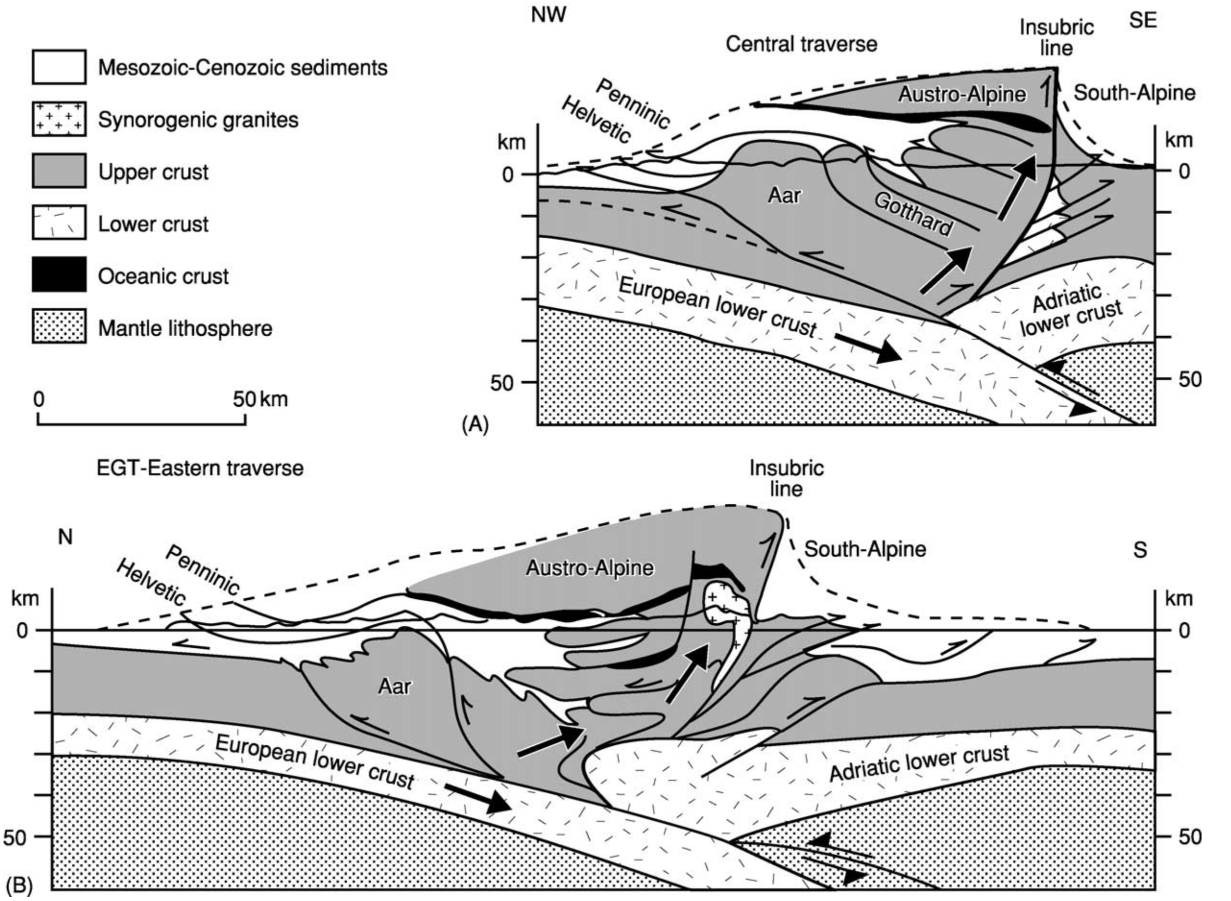


Figure 2 Two sections across the European Alps, showing the overall structure with the stacking of upper crustal sheets above the subducting lower crust. After Schmid *et al.* (1996) Geophysical–geological transect and tectonic evolution of the Swiss–Italian Alps. *Tectonics* 15: 1036–1064.

20°–30° of counterclockwise rotation. GPS measurements suggest that the present rate of contraction across the Himalaya is about $17.5 \pm 2\,\mathrm{mm\,year^{-1}}$. Earthquake focal mechanisms show that the majority of earthquakes are a result of north–south compression and occur along the main Himalayan fault, the active plate boundary along which the Indian Shield underthrusts the Lesser Himalaya.

The collision of India with Asia was the most recent of a series of continental plate collisions, which successively accreted smaller continental blocks onto the stable Siberian shield. Since the initial continental collision, India has moved northwards, indenting into Asia by over 2500 km and creating renewed uplift of all the mountain ranges along the southern margin of Asia, including the Pamir, Hindu Kush, the Karakoram, and the Tibetan plateau (Figure 3). The effects of the Indian plate collision extend northwards as far as the Tien Shan and Altay ranges along the border of inner Mongolia.

The earliest effects of the India–Asia collision occurred along the Indus suture zone where the two plates first met. This zone is marked by oceanic rocks including ophiolites, slabs of oceanic crust, and upper mantle emplaced onto continental margins, deep-sea sediments, and occasionally high-pressure metamorphic rocks, indicative of subduction zones. The youngest marine sediments along the suture zone are Early Eocene *Nummulites*-bearing limestones deposited between 52 and 49 My ago. These rocks are frequently used as a proxy for dating the collision. As the collision progressed, the sedimentary rocks along the northern continental margin began to shorten by folding and thrusting processes. The northern mountain ranges of the Himalaya in Zanskar, Spiti, northern Nepal, and southern Tibet show spectacular folding and thrusting of these Tethyan sedimentary rocks. Approximately 150 km of shortening of upper crustal rocks has been estimated across the Tethyan zone. The lower crust that originally underlay these upper crust sediments must have been detached and underthrust northwards beneath the southern margin of the Asian plate (Lhasa block).

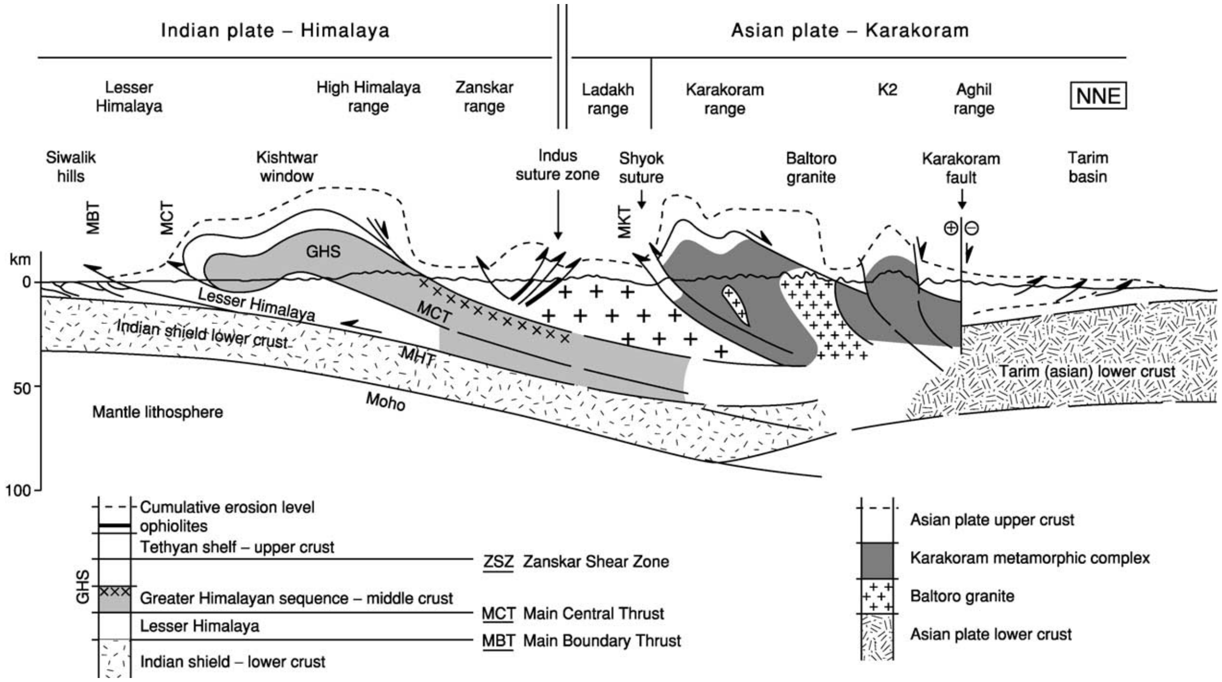


Figure 3 Geological profile across the western Himalaya and central Karakoram mountains. The middle crust metamorphic rocks in the Greater Himalaya and southern Karakoram are shaded and also contain crustal melt granites. The dashed line above shows the approximate cumulative erosion level, the amount of rock material eroded off the mountain range during the Tertiary orogeny. After Searle (1991) *Geology and Tectonics of the Karakoram Mountains*. Chichester, England: Wiley.

Crustal shortening and thickening resulted in increased temperatures and pressures in the deeper crust, causing regional metamorphism along the Greater Himalaya. Many of the highest peaks along the Himalaya are composed of the exhumed deep crustal metamorphic rocks. Dating of monazites by U–Th–Pb methods, and garnets using Sm–Nd isotopes from kyanite- and sillimanite-bearing gneisses in the Greater Himalaya shows that peak *P*–*T* conditions were initially reached between 35 and 30 My ago, and that temperatures remained high (above 600°C, in the kyanite and sillimanite stability fields) for at least 15 My after that. Around 20 My ago during the Early Miocene, temperatures peaked during a widespread sillimanite-grade metamorphic event that culminated in partial melting of the crust. This melting event resulted in migmatisation of the gneisses and generation of crustal melt leucogranites. The Himalayan leucogranites contain tourmaline, garnet, muscovite, and biotite, and they have very distinctive isotope chemistry (extremely high ^{87}Sr/^{86}Sr ratios) characteristic of granites produced from the melting of continental crust. The leucogranites were formed at relatively shallow depths (15–25 km) and were extruded southwards as giant sill complexes.

During the southward extrusion of the Greater Himalayan slab of metamorphic and granitic rocks, the metamorphic isograds were folded and inverted along a giant thrust fault shear zone termed the 'Main Central Thrust'. This thrust fault was active during the Early Miocene (around 20–15 My ago) and carried the entire Greater Himalaya southwards over the relatively unmetamorphosed rocks of the Lesser Himalaya. During this event the metamorphic isograds were recumbently folded and sheared by crustal scale thrusting. As a consequence, an inverted metamorphic isograd sequence characterizes the main central thrust high strain shear zone, with sillimanite- and kyanite-grade rocks structurally above low-grade biotite and garnet-grade rocks. The extrusion of this ductile mid-crustal layer, or channel, of partially molten high-grade gneisses, migmatites, and leucogranites was active during the Miocene (20–17 My ago) with major thrust-related shear zones along the base (Main Central Thrust) and a major normal fault shear zone along the top (South Tibetan detachment).

Around 10 My ago thrusting propagated southwards from the Main Central Thrust into the Lesser Himalaya and eventually to the 'Main Boundary Thrust', the southern boundary of the Himalaya. Active thrust faults have developed during Quaternary and recent times, extending south into the Siwalik hills where the active Himalayan front occurs today. The loading of the Himalaya caused the Indian plate to flex down and create the Siwalik molasse basin. This basin accumulated all the debris eroded from the rising Himalaya, transported south by rivers. The rivers converge into the Indus River in the west and the Ganges in the east. Sediments eroded from the Himalaya have been transported by these rivers to the Indus Fan in the Arabian Sea and the Bengal Fan in the Bay of Bengal.

Tibetan Plateau

The Tibetan plateau is the largest area of uplifted crust on Earth. The plateau extends for over 3000 km east–west and 1500 km north–south. It lies at an average elevation of just over 5 km above sea-level. The interior of the plateau is very flat with internal drainage, low precipitation, and low erosion rates. The margins of the plateau are ringed by mountain ranges including the Himalaya along the south, the Karakoram and Pamirs to the south-west, the Tien Shan and Kun Lun to the north, and the Long Men Shan along the east. Earthquakes reveal that the high plateau is currently undergoing east–west extension, whilst the margins of the plateau show compression or strike-slip faulting.

The geology of Tibet shows that the plateau region includes several different continental plates that were progressively accreted to the southern continental margin of Asia throughout the Phanerozoic (Figure 4). The most recent and probably the largest of these was the final plate collision, that of India with Asia. The crust beneath the Tibetan plateau is between 65 and 70 km thick, double that of normal continental crust. Several different models have been proposed to account for the thick crust and approximately 1000 km of crustal shortening required. The extreme end-member models include underthrusting of India beneath the entire Tibetan plateau, a model proposed initially by Emile Argand in 1924, and homogeneous thickening of the plateau with very little underthrusting of Indian material at depth.

Recent deep crustal seismic reflection and refraction profiling of Tibet has revealed that the Indian plate lower crust probably underthrusts southern Tibet only as far as the Bangong suture zone approximately 450 km north of the Indus–Yarlung Tsangpo suture zone. The equivalent amount of shortening in the Indian plate upper crust has been taken up by intense folding, thrusting, and crustal thickening in the Tethyan zone and Greater Himalaya. The seismic profiling has also managed to trace the prolongation of the surface faults in the Himalaya, northwards beneath the southern part of the plateau. Seismic and structural data has revealed that the Main Central Thrust and the South Tibetan detachment normal fault bound a mid-crustal layer of hot, partially molten rock that extends southwards to the high Himalaya. Magnetotelluric studies have revealed the presence of ‘bright spots’ indicative of pockets of fluid or magma at relatively shallow depths beneath southern Tibet today. These have been interpreted as pockets of granitic magma forming today in a structural position similar to those of the 20- to 17-My-old leucogranites cropping out in the high Himalaya. This layer of partial-melt migmatites, high-grade gneisses, and leucogranites was extruded out from beneath the southern part of the Tibetan plateau as a channel of ductile-deforming rock bounded by a rheologically stronger upper crust (Tethyan zone) and lower crust (underplated Indian shield Precambrian and Early Palaeozoic rocks).

Whereas the Tibetan plateau shows little relief despite being 5 km above mean sea-level, the neighbouring Karakoram range of northern Pakistan and Ladakh shows the highest relief of all, with many 7000- to 8000-m-high mountains and deeply incised river valleys. The Karakoram crust has also been tilted, revealing the deep crustal geology not exposed in Tibet (Figure 3). The Karakoram shows multiple episodes of crustal thickening and regional metamorphism spanning the past 65 My and multiple episodes of crustal melting, resulting in granite magmatism. A series of pre-50-My-old granite–granodiorite intrusions indicate that the southern margin of Asia was probably a subduction-related Andean-type continental margin prior to the Indian plate collision 50 My ago. The climax of mountain building in the Karakoram occurred between 24 an – 15 My ago with the emplacement of the huge Baltoro granite batholith, a series of intrusions of biotite monzogranite to biotite–muscovite–garnet leucogranite. A suite of 24- to 22-My-old lamprophyre dykes intruding around the Baltoro granites indicates that parts of the upper mantle were melting at the same time as the lower crust. The ages of the Baltoro granites are similar to the age of the Greater Himalayan leucogranites to the south, which span 24–12 My ago, with the majority between 21 and 17 My ago. This suggests that following the India–Asia collision, crustal thickening, metamorphism, and magmatism spread both across the south Asian margin in the Karakoram and across the north Indian plate margin along the Himalaya.

Earthquake distribution across Tibet shows that the entire plateau region is deforming internally today, and not acting as a rigid plate. Earthquakes in the high plateau show that the crust is undergoing east–west extension, whereas earthquakes in the mountain ranges bordering the high plateau are

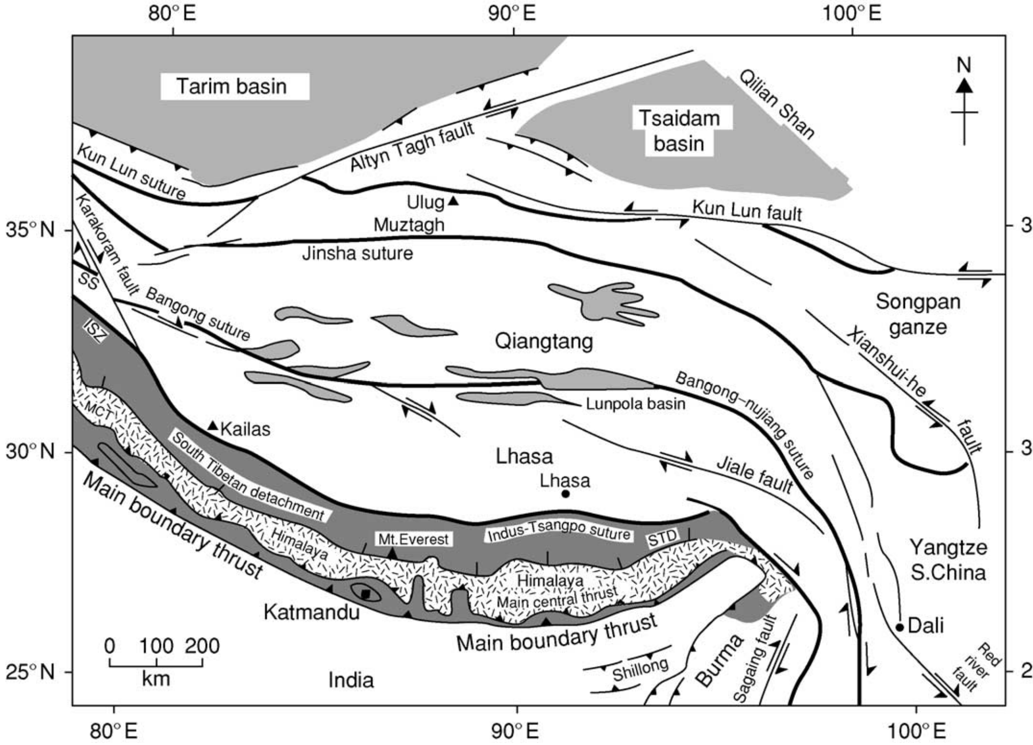


Figure 4 Map of the Tibet region showing the major plate boundaries. The Himalayas are shaded along the southern margin of the plateau. The stable continental blocks of the Tarim basin and Tsaidam basin in the north are shown, together with the major strike-slip faults bounding the plateau. Suture zones are progressively younger towards the south from the Kun Lun to the Jinsha, Bangong-Nujiang, and Indus–Tsangpo suture zones.

mostly a result of either compression or strike-slip deformation. Southern Tibet shows about eight large graben systems, bounded by north–south-aligned steep normal faults. These extensional rifts cut across the northernmost Himalayan range, the Tethyan zone, but die out to the south and do not extend into the Greater Himalayan zone. The rifts are associated with hot springs and active geothermal systems.

The high plateau is bounded by a series of large-scale strike-slip faults, notably the sinistal Altyn Tagh and Kun Lun faults along the north, the dextral Karakoram fault along the south-west, and a series of arcuate, dextral strike-slip faults swinging around the eastern syntaxis (Jiale, Xianshui-he faults). The distribution and slip motion on these faults suggested that the thickened crust of the Tibetan plateau was being extruded horizontally, eastwards out of the way of the Indian plate indentor, a process known as continental extrusion. Despite being impressive faults, many of which show active offsets of Quaternary glacial features, there is relatively limited total geological offset along them, and their timing is generally not concomitant with collision.

The Tibetan plateau region is also remarkable for having a series of shoshonitic volcanic rocks that erupted intermittently over the past 45 My. These volcanic rocks were derived from partial melting of hot asthenospheric mantle. The ages of these volcanics reveal that by 13 My ago, southern Tibet no longer had a hot upper mantle, as relatively cold the Indian plate lithosphere was underplating from the south; however, central and northern Tibet did keep a hot mantle, indicating that any cold lithospheric root to the thickened plateau must have dropped off intermittently and diachronously, rather than as a single catastrophic event.

Conclusions

Although mountain belts can occur along constructive plate margins or even in some intraplate regions, most orogenic belts are the result of plate collision processes along destructive plate margins. Ocean–ocean plate collisions result in island arcs, accretionary prisms, and deep subduction along oceanic trenches. Ocean–continent plate collisions result in Andean-type mountain belts, characterized by calc-alkaline volcanic arcs and granite–granodiorite batholiths. Continent–continent collision can result in a wide array of mountain belts characterized by the relatively simple folding and thrusting of the Zagros mountains, to the more complicated structure and metamorphism seen in the Alps and Himalaya. The largest and most extensive such collision, that between India and Asia, resulted in thousands of kilometres of crustal shortening, crustal thickening, metamorphism, melting, and exhumation as seen in the Himalaya, Karakoram, and Tibetan plateau.

See Also

Europe: The Alps. **Plate Tectonics**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges.

Further Reading

Harrison TM, Copeland P, Kidd WSF, and Yin An (1992) Raising Tibet. *Science* 255: 1663–1670.

Hodges KV (2000) Tectonics of the Himalaya and southern Tibet from two perspectives. *Geological Society of America Bulletin* 112: 324–350.

Keary P and Vine FJ (1996) *Global Tectonics*. Oxford: Blackwell Science.

Schmid SM, Pfiffner OA, Froitzheim N, Schönborn G, and Kissling E (1996) Geophysical–geological transect and tectonic evolution of the Swiss–Italian Alps. *Tectonics* 15: 1036–1064.

Searle MP (1991) *Geology and Tectonics of the Karakoram Mountains*. Chichester, England: Wiley.

Neotectonics

I Stewart, University of Plymouth, Plymouth, UK

Introduction

Neotectonics concerns the study of horizontal and vertical crustal movements that have occurred in the geologically recent past and which may be ongoing today. Though most crustal movements arise directly or indirectly from global plate motions (i.e., tectonic deformation), neotectonic studies make no presumption about the mechanisms driving deformation. Consequently, 'movements' is a vague catch-all term that encompasses a myriad of competing deformation processes, such as the gradual pervasive creep of tectonic plates, discrete (seismic) displacements on individual faults and folds, and distributed tilting and warping through isostatic readjustment or volcanic upheaval. The phrase 'geologically recent past' is also appropriately vague. Early attempts to define the discipline by arbitrary time windows (e.g., Late Cenozoic, Neogene, or Quaternary) have given ground to a more flexible notion that envisages neotectonism starting at different times in different regions. The onset of the neotectonic period, or the 'current tectonic regime', depends on when the contemporary stress field of a region was first imposed. For instance, the current tectonic regime began in the Middle Quaternary ($\sim$700 000 years ago) in the Apennines of central Italy, and even more recently (<500 000 years ago) in California; in contrast, in eastern North America, the present-day stress regime has been in existence for at least the past 15 million years.

Typically then, neotectonic movements have been in operation in most regions for the past few million years or so. Over such prolonged intervals, neotectonic actions are revealed by the stratigraphic build-up of sediments in inland and marine basins, the burial or exhumation histories of rocks, and the geomorphological development of landscapes. Geological studies of palaeobotany and palaeoclimate, numerical models of landscape evolution, and techniques such as fission-track analysis and cosmogenic dating are among the disparate tools unravelling this long-term tectonic activity. Over periods of many tens of to several hundreds of thousands of years, the actions of individual tectonic structures (faults and folds) can be determined, unmasked by their deformation of geomorphic markers, such as marine and fluvial terraces, and tracked with reference to the Late Pleistocene glacial–eustatic time-frame. The apparently smooth deformation rates discerned over intermediate time-scales are revealed to be episodic and irregular when faults and folds are examined over Holocene (10 000 years) time-scales. Over millennial time-scales, secular variations in the activity of tectonic structures can be gleaned from a diverse set of palaeoseismological approaches, from interpreting the stratigraphy of beds that have been affected by faulting, to detecting disturbances in the growth record of trees or coral atolls.

Active Tectonics

Although neotectonic movements continue up to the present day, the term 'active tectonics' is typically used to describe those movements that have occurred over the time-span of human history. Active tectonics deals with the societal implications of neotectonic deformation (such as seismic-hazard assessment, future sea-level rise, etc.), because it focuses on crustal movements that can be expected to recur within a future interval of concern to society. Contemporary crustal movements may be discerned in Earth surface processes and landforms, such as in the sensitivity of alluvial rivers to crustal tilting. In addition, geomorphological and geological studies are important in recording the surface expression of Earth movements such as earthquake ground ruptures, which, due to their subtle, ephemeral, or reversible nature, are unlikely to have been preserved in the geological record. However, active tectonics also employs an array of high-technology investigative practices; prominent among these are the monitoring of ongoing Earth surface deformation using space-based or terrestrial geodetic methods (tectonic geodesy), radar imaging (interferometry) of ground deformation patterns produced by individual earthquakes and volcanic unrest, and the seismological detection and measurement of earthquakes (seismotectonics). These techniques are applied globally via the World-Wide Standardized Seismograph Network and regionally via local seismographic coverage.

The modern snapshots of tectonism can be pushed back beyond the twentieth century through the analysis of historical accounts and maps to infer past land surface changes or to deduce the parameters of past seismic events (historical seismology). In addition, earthquakes can leave their mark in the mythical practices and literary accounts of ancient peoples, recorded in the stratigraphy of their site histories and in the damage to their buildings (archaeoseismology). The time covered by such human records varies markedly, ranging from many thousands of years in the Mediterranean, Near East, and Asia to a few centuries across much of North America. Generally records confirm that regions that are active today have been consistently active for millennia, thereby demonstrating the long-term nature of crustal deformation, but occasionally records reveal that some regions that appear remarkably quiet from the viewpoint of modern seismicity (such as the Jordan rift valley) are capable of generating large earthquakes. In reality, the distinction between neotectonics and active tectonics is artificial; the terms simply describe different time slices of a continuum of crustal movement. This continuum is maintained by the persistence of the contemporary stress field, which means that inferences of past rates and directions of crustal movement from geological observations can be compared directly with those measured by modern geodetic and geophysical methods.

Although the terms 'neotectonic' and 'active' are somewhat blurred and are often used interchangeably, societal demands (for instance, regulatory authorities for seismic hazard, nuclear safety, etc.) often require the incidence of tectonic movements to be defined strictly. For instance, in the United States, under California law, an 'active fault' is presently defined as one that has generated surface-rupturing earthquakes in the past 11 000 years (i.e., the time period was established to relate to the time when the Holocene was considered to have begun). Other regulatory bodies recognize a sliding scale of fault activity: Holocene (activity in the past 10 000 years), Late Quaternary (activity in the past 130 000 years), and Quaternary (activity in the past 1.6 million years). Neotectonic faults, by comparison, are simply those that formed during the imposition of the current tectonic regime. 'Real' structures, of course, are unconstrained by such legislative concerns. Many modern earthquakes rupture along older (i.e., palaeotectonic) basement faults. Indeed, it is important to recognize that any fault that is favourably oriented with respect to the stress currently being imposed on it has the potential to be activated in the future, regardless of whether it has moved in the geologically recent past.

Global Tectonics

A useful way to differentiate styles and degrees of neotectonic activity is in terms of tectonic strain rate, which is a measure of the velocity of regional crustal motions and, in turn, of the consequent tectonic strain build-up. Crustal movements are most vigorous, and therefore most readily discernible, where plate boundaries are narrow and discrete. In these domains of high tectonic strain, frequent earthquakes on fast-moving ($>10\,\text{mm year}^{-1}$) faults ensure that a century or two of historical earthquakes and a few years of precise geodetic measurements are sufficient to capture a consistent picture of the active tectonic behaviour. Intermediate tectonic strain rates characterize those regions where plate–boundary motion is distributed across a network of slower moving (0.1–$10\,\text{mm year}^{-1}$). Examples of such broad deforming belts are the Basin and Range Province of the western United States or the Himalayan collision zone, where earthquake faults rupture every few hundred or thousand years, ensuring that the Holocene period is a reasonable time window over which to witness the typical crustal deformation cycle. In contrast, areas with low strain rates ensure that intraplate regions, often referred to as 'stable continental interiors', are low-seismicity areas with slow-moving ($<0.1\,\text{mm year}^{-1}$) faults that rupture every few tens (or even hundreds) of thousands of years, making the snapshot of human history an unreliable guide to the future incidence of tectonic activity.

The global pattern of present-day crustal motions can be accounted for by 'plate tectonic' theory, that elegant kinematic framework in which rigid plates variously collide, split apart, and slide along their actively deforming boundaries (*see* **Plate Tectonics**). Closer inspection, however, reveals that the basic rules that govern global plate motions (i.e., rigid blocks separated by narrow deforming boundaries) break down at the regional and local scales. This is particularly so on the continents, where a patchwork of preexisting geology and structure ensures that tectonic stresses are not applied in a uniform, straightforward fashion. Studies of how the contemporary stress field varies across Earth's surface distinguish between first- and second-order stress provinces. First-order provinces have stresses generally uniformly oriented across several thousands of kilometres. The largest of these are the midplate regions of North America and western Europe, where the stress fields are largely the far-field product of ridge push and continental collision. In contrast, first-order stress provinces in tectonically active areas are dominated by the downgoing pull of subducting slabs and the resistance to subduction. Second-order stress provinces are smaller, typically less than 1000 km across, and are related to crustal flexure induced by thick sequences of sediments and postglacial rebound, and to deep-seated rheological contrasts. Although the bulk of Earth's crust is in compression, significant regions of extension occur. In both the continents and the oceans, these extensional domains are long and narrow and correspond to topographically high areas, though notable exceptions are the Basin and Range provinces and the Aegean region of the eastern Mediterranean. Most first-order stress provinces, and many second-order stress provinces, coincide with distinct physiographic provinces.

Glacial Isostatic Adjustment

Plate-driving forces may exert the dominant control on the contemporary stress field, but another process contributes to crustal deformation at a global scale. That process is glacial isostatic adjustment (GIA), the physical response of Earth's viscoelastic mantle to surface loads imposed and removed by the cycles of glaciation and deglaciation, to which the planet has been subjected for the past 900 000 years (*see* **Tertiary To Present:** Pleistocene and The Ice Age). Because large ice-mass fluctuations induce the subcrustal flow of material, measurable crustal deformation extends for thousands of kilometres beyond the limits of the former ice margins; consequently, the effects of GIA are felt globally. In addition, though the crust's elastic response to ice-sheet decay is geologically immediate, the delayed viscoelastic response of the mantle ensures that GIA can persist long after the ice has gone. Although the effects of GIA can now be detected from space geodesy, its legacy is most clearly visible in the worldwide pattern of postglacial sea-level changes. Regions that were ice covered at the Last Glacial Maximum are uplifting (i.e., relative sea-level is currently falling) as a consequence of postglacial rebound of the crust. Likewise, the regions peripheral to the former ice sheets are subsiding (i.e., relative sea-levels are rising) due to collapse of the 'glacial forebulge'. The effect of this subsidence outside the area of forebulge collapse is to draw in water from the central ocean basins, which is compensated by uplift in the ocean basin interiors in the far-field of the ice sheets. The final GIA component is the hydro-isostatic tilting of continental coastlines due to the weight applied to Earth's surface by the returning meltwater load, which produces a 'halo' of weak crustal subsidence around the world's major landmasses. For the most part, geological studies of Holocene relative sea-level changes are consistent with the uplift/subsidence pattern predicted by global viscoelastic theory. The key areas of misfit are along plate boundary seaboards (especially subduction zones), where tectonic deformation dominates, and those areas 'contaminated' by local anthropogenic effects (groundwater extraction, etc.).

The neotectonic implications of GIA are not confined to the coastline. Glacial rebound is now widely considered as an effective mechanism for exerting both vertical and horizontal stresses, not only within the limits of the former ice sheets, but for several hundred kilometres outside. Within the former glaciated parts of eastern North America and northern Europe, both tectonic and rebound stresses are required to explain the distribution and style of both postglacial and contemporary seismotectonics. Outside in the ice-free forelands, predicted glacial strain rates are still likely to be one to three orders of magnitude higher compared to tectonic strain rates typical of continental interiors. Consequently, some workers argue that an apparent 'switching on' of Holocene earthquake activity in the eastern United States and the occurrence of atypically large seismic events, such as the great (magnitude >8) earthquakes that struck the Mississippi valley area of New Madrid in 1811–1812, may be associated with areas where glacial strains are particularly high. Glacial loading and unloading may also disturb the build-up of tectonic strain at glaciated plate boundaries, such as today in Alaska or previously when the Cordilleran ice sheet capped part of the Cascadia subduction zone. More recently, the isostatic component of glacier erosion in the mountain-building process is becoming appreciated.

Global Perspective

The worldwide patterns of vertical and horizontal crustal movements arise from the global effects of plate motions and glacial isostatic adjustment. Regionally and locally, this is augmented by flexure from eustatic or sediment loading, volcanic deformation, or anthropogenic change (e.g., dam impoundment). Though many neotectonic investigations seek to disentangle movements arising from the imposition of tectonic strains from those augmented by non-tectonic processes, this is often a fruitless holy grail; because deformation of Earth's crust typically induces compensatory flow underlying the mantle, neotectonic movements are applied globally. Nevertheless, these disparate contributory mechanisms, coupled with the varying time-scales over which their actions can de discerned, ensure that neotectonics encompasses a remarkable breadth of research disciplines. Few other fields easily blend topics as disparate as space science, seismology, Quaternary science, geochronology, structural geology, geomorphology, geodesy, archaeology, and history. It is this interdisciplinary marriage that makes neotectonics particularly exciting and especially challenging.

See Also

Plate Tectonics. **Tertiary To Present:** Pleistocene and The Ice Age.

Further Reading

Burbank DW and Anderson RS (2001) *Tectonic Geomorphology.* Oxford: Blackwell.

Peltier WR (1999) Global sea-level rise and glacial isostatic adjustment. *Global and Planetary Change* 20: 93–123.

Stewart IS and Hancock PL (1994) Neotectonics. In: Hancock PL (ed.) *Continental Deformation,* pp. 370–409. London: Pergamon Press.

Stewart IS, Sauber J, and Rose J (2000) Glacio-seismotectonics: ice sheets, crustal deformation and seismicity. *Quaternary Science Reviews* 14/15: 1367–1390.

Vita-Finzi C (2002) *Monitoring the Earth.* Harpenden: Terra Publishing.

Ocean Trenches

R J Stern, The University of Texas at Dallas, Richardson, TX, USA

Introduction

An oceanic trench is a long, narrow, and generally very deep depression of the seafloor. Oceanic trenches are the deepest places on the Earth's solid surface and range down to 11 km below sea-level. These tremendous depths mark fundamental breaks in the Earth's lithosphere, the great plates that we all ride on (*see* **Plate Tectonics**). If mid-ocean ridges are where the Earth turns itself inside out, trenches are where the Earth swallows its skin. The asymmetry of trenches reflects a deeper phenomenon: as one plate bends down to return to the mantle, the other plate strains to fill the growing void. The depths of trenches are governed by many things, most importantly sediment flux but also the age of the downgoing lithosphere, the convergence rate, intermediate slab dip, and even the width of the sinking plate. Trenches are sites where fluids are 'squeezed' out of the subducted sediments and a newly recognized biosphere thrives.

Early Years of Study

Trenches are the most spectacular morphological features on the Earth's solid surface, but they were not clearly defined until the late 1940s and 1950s. The depths of the oceans were scarcely imagined until we began to lay telegraph cables between the continents in the late nineteenth and early twentieth centuries. The elongated bathymetric expression of trenches was not recognized early, and the term 'trench' does not appear in Murray and Hjort's classic oceanography overview. Instead they used the term 'deep' to describe the deepest parts of the ocean, such as the Challenger Deep, which is now recognized as the greatest gash on the solid surface of the Earth. Experiences in the World War I battlefields emblazoned the concept of a trench as an elongate depression defining an important boundary, so it is no surprise that the term 'trench' was used to describe natural features in the early 1920s. The term was first used in a geological context by SJ Scofield two years after the war ended to describe a structurally controlled depression in the Rocky Mountains. James Johnstone, in his 1923 textbook *An Introduction to Oceanography*, first used the term in its modern sense to describe a marked elongate depression of the seafloor.

During the 1920s and 1930s, Vening Meinesz developed a unique gravimeter that could measure gravity in the stable environment of a submarine and used it to measure gravity over ocean trenches. His gravity measurements revealed that trenches are sites of downwelling in the solid Earth. The concept of downwelling at trenches was characterized by DT Griggs in 1939 as the tectogene hypothesis, for which he developed an analogous model using a pair of rotating drums. The war in the Pacific led to great improvements in bathymetry, especially in the western and northern Pacific, and the linear nature of the trenches became clear. The rapid growth of deep-sea research efforts, especially the widespread use of echo-sounders in the 1950s and 1960s, confirmed the morphological utility of the term. The important trenches were identified and sampled, and their greatest depths were sonically plumbed. The heroic phase of trench exploration culminated in the 1960 descent of the bathyscaphe *Trieste*, which set an unbeatable world record by diving to the bottom of the Challenger Deep. Following Dietz' and Hess' articulation of the seafloor-spreading hypothesis in the early 1960s and the plate-tectonic revolution in the late 1960s, the term 'trench' has been redefined so that it now has tectonic as well as morphological connotations.

Trenches mark one of the most important types of natural boundary on the Earth's solid surface, that between two lithospheric plates. There are three types of lithospheric-plate boundary: divergent (where lithosphere and oceanic crust are created at mid-ocean ridges (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mid-Ocean Ridges), convergent (where one lithospheric plate sinks beneath another and returns to the mantle (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges), and transform (where two lithospheric plates slide past each other). Trenches are the spectacular and distinctive morphological features of convergent plate boundaries (Figure 1). Plates move together along convergent plate boundaries at rates that vary from a few millimetres to ten or more centimetres per year. Trenches form where oceanic lithosphere is subducted at a convergent plate margin,

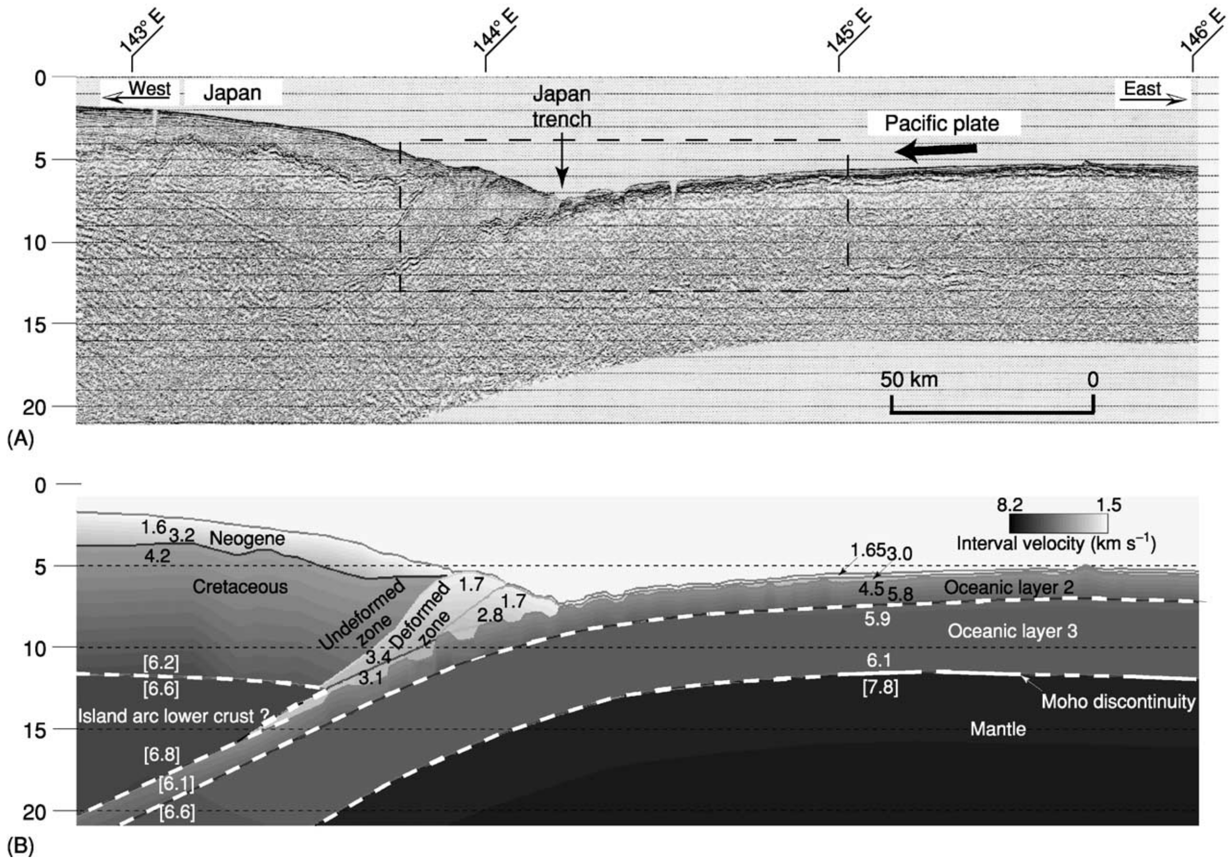


Figure 1 Profile across a typical trench (the Japan Trench near 39° N). (A) Seismic-reflection image (pre-stack depth migration) and (B) interpretation, including crustal units and seismic velocities (in km s^{-1}). Dashed box in (A) indicates the region shown in detail in Figure 3A. Vertical exaggeration in (B) is four times. The model is shaded according to the seismic velocities, and selected digital values are also shown. Modified from Tsuru T, Park J-O, Takahashi N, *et al.* Tectonic features of the Japan Trench convergent margin off Sanriku, northeastern Japan, revealed by multichannel seismic reflection data. *Journal of Geophysical Research* 105: 16 403–16 413.

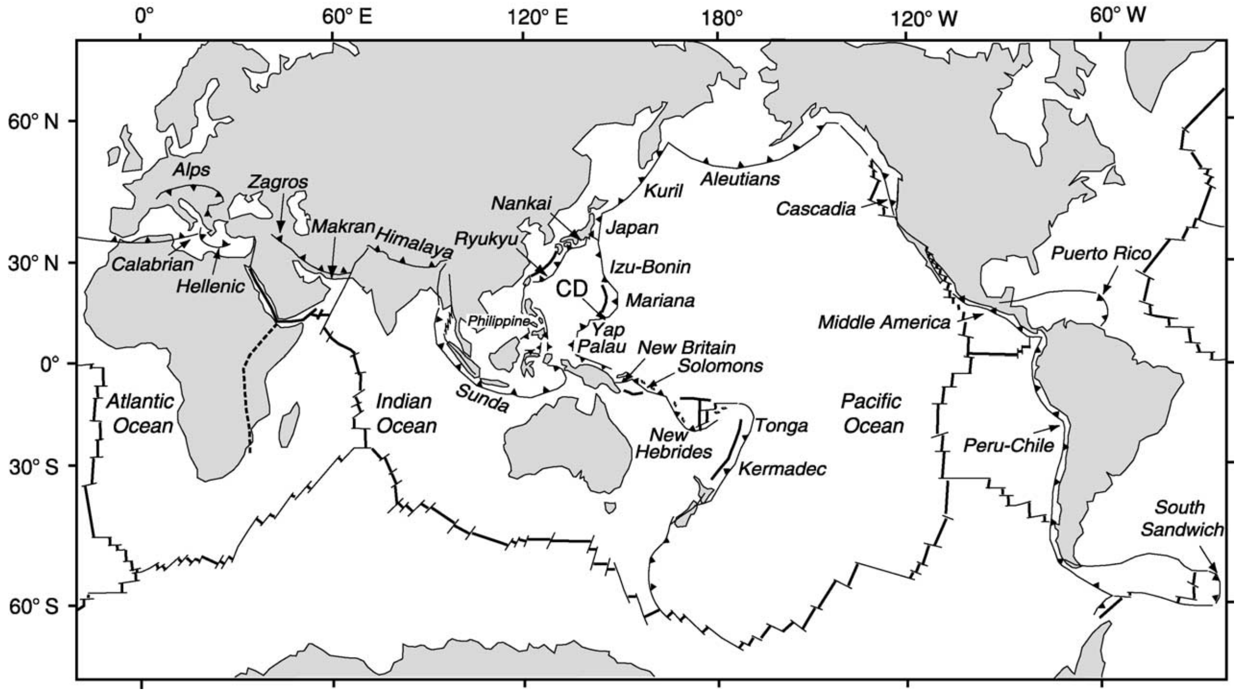


Figure 2 Locations of the world's major trenches and collision zones. CD is the Challenger Deep.

presently at a global rate of about a tenth of a square metre per second.

Geographical Distribution

There are about 50 000 km of convergent plate margins in the world, mostly around the Pacific Ocean – the reason that they are sometimes called 'Pacific-type' margins – but they also occur in the eastern Indian Ocean, and there are relatively short convergent-margin segments in the Atlantic and Indian Oceans and in the Mediterranean Sea (Figure 2). Trenches are sometimes buried and lack bathymetric expression, but the fundamental structures that they represent mean that the name should still be applied in these cases. This applies to the Cascadia, Makran, southern Lesser Antilles, and Calabrian trenches (Table 1). Trenches, magmatic (island) arcs, and zones of earthquakes that dip under the magmatic arc as deeply as 700 km are diagnostic of convergent plate boundaries and their deeper manifestations, subduction zones. Trenches are related to but distinguished from zones of continental collision, where continental lithosphere enters the subduction zone. When buoyant continental crust enters a trench, subduction eventually stops and the convergent plate margin becomes a collision zone. Features analogous to trenches are associated with collision zones; these include sediment-filled

Table 1 Maximum trench depth

Trench	*Maximum depth (m)*	*Accretionary prism?*
Challenger, Mariana	10 920	No
Tonga	10 800	No
Philippine (East Mindanao)	10 057	No
Kermadec	10 047	No
Izu-Bonin	9 780	No
Kuril	9 550	No
North New Hebrides	9 175	No
New Britain	8 940	No
Yap	8 650	No
Puerto Rico	8 605	No
South Sandwich	8 325	No
South Solomons	8 322	No
Peru–Chile	8 170	No
Japan	8 130	No
Palau	8 055	No
Aleutians	7 680	Yes
Ryukyu	7 460	Yes
Sunda	7 125	Yes
Middle America	6 660	No
Hellenic	5 092	Yes
Nankai	4 900	Yes
Calabrian	4 200	Yes, no morphological trench
Makran	3 200	Yes, no morphological trench
Cascadia	3 136	Yes, no morphological trench

foredeeps, referred to as peripheral foreland basins, such as that which the Ganges and Tigris–Euphrates rivers flow along.

Morphological Expression

Trenches are the centrepieces of the distinctive physiography of a convergent plate margin. Transects across trenches yield asymmetric profiles, with relatively gentle (*ca.* 5°) outer (seaward) slopes and steeper (*ca.* 10–16°) inner (landward) slopes. This asymmetry is due to the fact that the outer slope is defined by the top of the downgoing plate, which must bend as it begins its descent. The great thickness of the lithosphere requires that this bending be gentle. As the subducting plate approaches the plate boundary, it is first bent upwards to form the outer swell and then descends to form the outer trench slope. The outer trench slope is disrupted by a set of subparallel normal faults, which staircase the seafloor down into the trench (Figure 3). The plate boundary is defined by the trench axis itself. Beneath the inner trench wall, the two plates slide past each other along the subduction dêcollement, which intersects the seafloor along the base of the trench. The overriding plate generally contains a magmatic arc and a fore arc. The magmatic arc is created as a result of physical and chemical interactions between the subducted plate at depth and the asthenospheric mantle associated with the overriding plate. The fore arc lies between the trench and the magmatic arc. Fore arcs have the lowest heat flow of any place on Earth because there is no asthenosphere (convecting mantle) between the forearc lithosphere and the cold subducting plate.

The inner trench wall marks the edge of the overriding plate and the outermost fore arc. The fore arc consists of igneous and metamorphic basement, and this basement may act as a buttress to a growing accretionary prism, depending on how much sediment is supplied to the trench. If the sediment flux is high, material will be transferred from the subducting plate to the overriding plate. In this case an accretionary prism grows, and the location of the trench migrates away from the magmatic arc over the life of the convergent margin. Convergent margins with growing accretionary prisms are called accretionary convergent margins and make up nearly half of all convergent margins. If the sediment flux is low, material will be transferred from the overriding plate to the subducting plate by a process of tectonic ablation known as subduction erosion and carried down the subduction zone. Fore arcs undergoing subduction erosion typically expose igneous rocks. In this case,

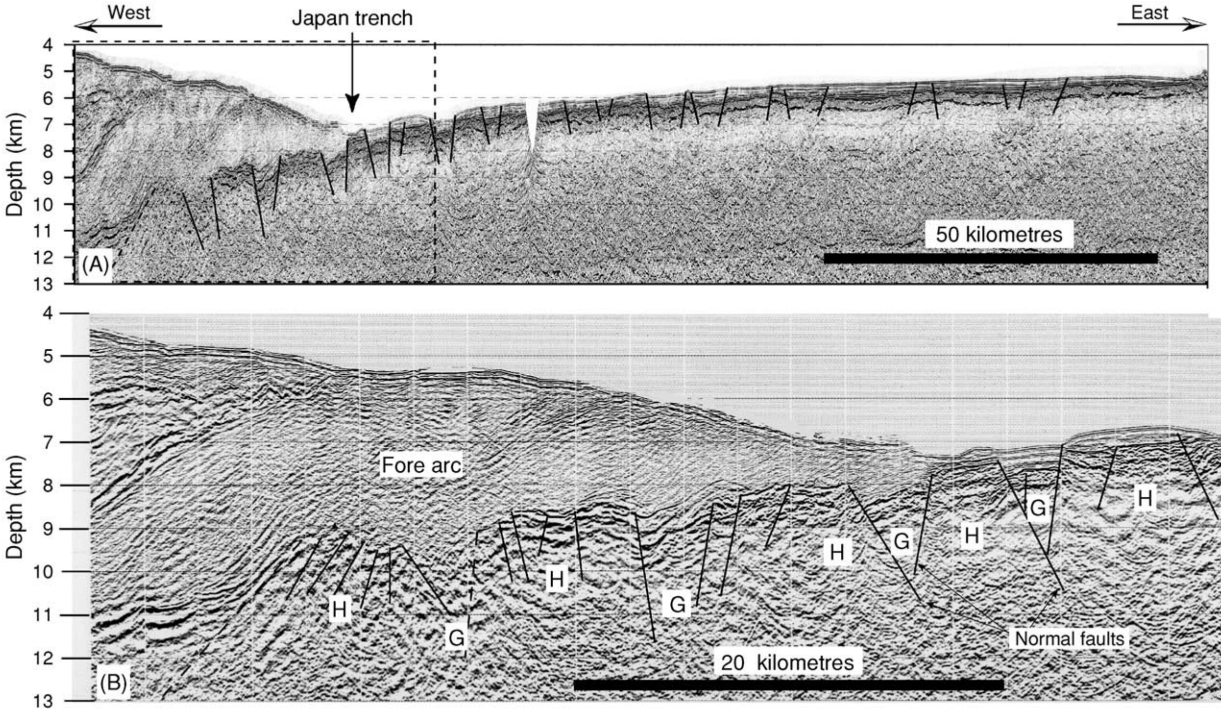


Figure 3 Horst (H) and graben (G) structures with normal faults associated with the subduction of the Pacific plate in the Japan Trench. Vertical exaggeration is four times in (A) and twice in (B). Dashed box in (A) indicates the region shown in detail in (B). Modified from Tsuru T, Park J-O, Takahashi N, *et al.* Tectonic features of the Japan Trench convergent margin off Sanriku, northeastern Japan, revealed by multichannel seismic reflection data. *Journal of Geophysical Research* 105: 16 403–16 413.

the location of the trench will migrate towards the magmatic arc over the life of the convergent margin. Convergent margins experiencing subduction erosion are called nonaccretionary convergent margins and comprise more than half of convergent plate boundaries. This is an oversimplification, because a convergent margin can simultaneously experience sediment accretion and subduction erosion.

The asymmetric profile across a trench reflects fundamental differences in materials and tectonic evolution. The outer trench wall and outer swell are composed of seafloor, which takes a few million years to move from where subduction-related deformation begins near the outer trench swell to where the plate sinks beneath the trench. In contrast, the inner trench wall is deformed by plate interactions throughout the life of the convergent margin. The fore arc is continuously subjected to subduction-related earthquakes. This protracted deformation and shaking ensures that the slope of the inner trench wall is controlled by the angle of repose of whatever material it is composed of. Because they are composed of igneous rocks instead of deformed sediments, nonaccretionary trenches have steeper inner walls than accretionary trenches.

Filled Trenches

The composition of the inner trench slope is determined by sediment supply, which also exerts a first-order control on trench morphology. Active accretionary prisms are especially common in trenches that are located near continents where large rivers or glaciers reach the sea. These filled trenches cause some confusion because they are tectonically indistinguishable from other convergent margins but lack the bathymetric expression of a trench. The Cascadia margin off the north-west coast of the USA is a filled trench, the result of sediments delivered by the rivers of the north-western USA and south-west Canada. The Lesser Antilles convergent margin shows the importance of proximity to sediment sources for trench morphology. In the south, near the mouth of the Orinoco River, there is no morphological trench, and the fore arc and the accretionary prism have a total width of almost 500 km. The accretionary prism is so large that it forms the islands of Barbados and Trinidad. Northwards the fore arc narrows, the accretionary prism disappears, and north of 17°N only the morphology of a trench is seen. In the extreme north, far away from sediment sources, the Puerto Rico Trench is over 8600 m deep, and there is no active accretionary prism. A similar relationship between proximity to rivers, fore arc width, and trench morphology can be observed from east to west along the Alaskan–Aleutian convergent margin. The convergent plate boundary off the coast of Alaska changes along its strike from a filled trench with a broad fore arc in the east (near the coastal rivers of Alaska) to a deep trench with a narrow fore arc in the west (near the Aleutian islands). Another example is the Makran convergent margin off the coasts of Pakistan and Iran, where the trench is filled by sediments from the Tigris–Euphrates and Indus rivers. Thick accumulations of turbidite deposits along a trench can be supplied by down-axis transport of sediments that enter the trench 1000–2000 km away, as is found in the Peru–Chile Trench south of Valparaiso and in the Aleutian Trench. Convergence rate can also be important in controlling trench depth, especially for trenches near continents, because slow convergence may mean that the capacity of the convergent margin to dispose of sediment is exceeded.

Thus, trench morphology can be expected to evolve continuously as oceans close and continents converge. While the ocean is wide, the trench may be far away from continental sources of sediment and so may be deep. As the continents converge, the trench may increasingly be filled with continental sediments and shoals. A simple definition of the transition from subduction to collision is when a plate boundary previously marked by a trench has filled sufficiently to rise above sea-level.

Accretionary Prisms and Sediment Transport

Accretionary prisms grow by frontal accretion – where sediments are scraped off, bulldozer-fashion, near the trench – or by underplating of subducted sediments and perhaps oceanic crust along the shallow parts of the subduction décollement. Frontal accretion over the life of a convergent margin results in the youngest sediments being found in the outermost part of the accretionary prism and the oldest sediments being found in the innermost portion. Older (inner) parts of the accretionary prism are much more lithified and have steeper structures than the younger (outer) parts. Underplating is difficult to detect in modern subduction zones but may be recorded in ancient accretionary prisms, such as the Franciscan Group of California, in the form of tectonic mélanges and duplex structures. Different modes of accretion are reflected in the morphology of the inner slope of the trench, which generally shows three morphological provinces. The lower slope comprises imbricate thrust slices, which form ridges. The middle part of the slope may comprise a bench or terraces. The upper slope is smoother but may be cut by submarine canyons.

Because accretionary convergent margins have high relief, are continuously deformed, and accommodate a large flux of sediments, they are sites of vigorous sediment dispersal and accumulation. Sediment transport is controlled by submarine landslides, debris flows, turbidity currents, and contourites. Submarine canyons transport sediment from beaches and rivers down the upper slope. These canyons are formed by channelized turbidites and generally lose definition with depth because continuous tectonic readjustments disrupt the channels. Sediments move down the inner trench wall via channels and a series of fault-controlled basins. The trench itself serves as an axis of sediment transport. If enough sediment moves into the trench, it may be completely filled, and turbidity currents will then be able to carry sediment well beyond the trench and may even surmount the outer swell. Sediments from the rivers of south-west Canada and the north-western USA spill over where the Cascadia trench would be and reach the Juan de Fuca spreading ridge several hundred kilometres to the west.

The slope of the inner trench wall of an accretionary convergent margin continuously adjusts to the thickness and width of the accretionary prism. The prism maintains a ‘critical taper’, established by the Mohr–Coulomb failure criterion for the pertinent materials. A package of sediments scraped off the downgoing lithospheric plate will deform until it and the accretionary prism that it has been added to attain a critical-taper (constant slope) geometry. Once critical taper is attained, the wedge slides stably along its basal décollement. Strain rate and hydrological properties strongly influence the strength of the accretionary prism and thus the angle of critical taper. Fluid pore pressure can modify rock strength and is an important determinant of critical taper angle. Low permeability and rapid convergence may lead to pore pressures that exceed lithostatic pressure and result in a relatively weak accretionary prism with a shallowly tapered geometry, whereas high permeability and slow convergence lead to lower pore pressures, stronger prisms, and steeper geometry.

The Hellenic Trench system is unusual because its convergent margin subducts evaporites. The slope of the southern flank of the Mediterranean Ridge (its accretionary prism) is low, about 1°, which indicates very low shear stress on the décollement at the base of the wedge. Evaporites influence the critical taper of the accretionary complex, because their mechanical properties differ from those of siliciclastic sediments and because of their effect upon fluid flow and fluid pressure, which control effective stress. In the 1970s, the linear deeps of the Hellenic Trench south of Crete were interpreted as being similar to trenches in other subduction zones, but, with the realization that the Mediterranean Ridge is an accretionary complex, it became apparent that the Hellenic Trench is actually a starved fore arc basin, and that the plate boundary lies south of the Mediterranean Ridge.

Water and Biosphere

The volume of water escaping from within and beneath the fore arc results in some of the Earth’s most dynamic and complex interactions between aqueous fluids and rocks. Most of this water is trapped in pores and fractures in the upper lithosphere and the sediments of the subducting plate. The average fore arc is underlain by a solid volume of oceanic sediment that is 400 m thick. This sediment enters the trench with 50–60% porosity. The sediment is progressively squeezed as it is subducted, reducing void space and forcing fluids out along the décollement and up into the overlying fore arc, which may or may not have an accretionary prism. Sediments accreted to the fore arc are another source of fluids. Water is also bound in hydrous minerals, especially clays and opal. The increasing pressure and temperature experienced by the subducted materials convert the hydrous minerals to denser phases that contain progressively less structurally bound water. Water released by dehydration accompanying phase transitions is another source of fluid introduced to the base of the overriding plate. These fluids may travel diffusely through the accretionary prism, via interconnected pore spaces in sediments, or may follow discrete channels along faults. Sites of venting may take the form of mud volcanoes or seeps and are often associated with chemosynthetic communities. Fluids liberated in the shallowest parts of the subduction zone may also escape along the plate boundary but have rarely been observed to drain along the trench axis. All of these fluids are dominated by water but also contain dissolved ions and organic molecules, especially methane. Methane is often sequestered in an ice-like form (clathrate) in the fore arc. Gas hydrates are a potential energy source and can rapidly break down. The destabilization of gas hydrates has contributed to global warming in the past and will probably do so in the future (*see* **Petroleum Geology:** Gas Hydrates).

Chemosynthetic communities thrive where fluids seep out of the fore arc. Cold seep communities have been discovered on inner trench slopes in the western Pacific, especially around Japan, in the Eastern Pacific, along the North, Central, and South American coasts from the Aleutian to the Peru–Chile Trenches, on the Barbados prism, in the Mediterranean, and in the Indian Ocean, along the Makran and Sunda convergent margins. These communities have been found down to depths of 6000 m. They have received much less attention than the chemosynthetic communities associated with hydrothermal vents. Chemosynthetic communities are located in a variety of geological settings: above over-pressured sediments in accretionary prisms, where fluids are expelled through mud volcanoes or ridges (Barbados, Nankai, and Cascadia); along active erosive margins with faults; and along escarpments caused by debris slides (Japan Trench, Peruvian margin). Surface seeps may be linked to massive hydrate deposits and destabilization (e.g. Cascadia margin). High concentrations of methane and sulphide in the fluids escaping from the seafloor are the principal energy sources for chemosynthesis (*see* **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges).

Empty Trenches and Subduction Erosion

Trenches distant from an influx of continental sediments lack an accretionary prism, and the inner slope of such trenches is commonly composed of igneous or metamorphic rocks. Nonaccretionary convergent margins are characteristic of (but not limited to) primitive arc systems. Primitive arc systems are those that are built on oceanic lithosphere, such as the Izu–Bonin–Mariana, Tonga–Kermadec, and Scotia (South Sandwich) arc systems. The inner trench slopes of these convergent margins expose the crust of the fore arcs, including basalt, gabbro, and serpentinized mantle peridotite. These exposures allow easy access to materials from the lower oceanic crust and upper mantle, and provide a unique opportunity to study the magmatic products associated with the initiation of subduction zones. Most ophiolites are probably formed in a fore-arc environment during the initiation of subduction, and this setting favours ophiolite emplacement during collision with blocks of thickened crust. Not all nonaccretionary convergent margins are associated with primitive arcs. Trenches adjacent to continents where there is a low influx of sediments from rivers, such as the central part of the Peru–Chile Trench, may also lack an accretionary prism.

The igneous basement of a nonaccretionary fore arc may be continuously exposed by subduction erosion. This transfers material from the fore arc to the subducting plate and can take the form of frontal erosion or basal erosion. Frontal erosion is most active in the wake of seamounts being subducted beneath the fore arc. Subduction of large edifices (seamount tunnelling) oversteepens the fore arc, causing mass failures that carry debris towards and ultimately into the trench (Figure 4). This debris may be deposited in graben of the downgoing plate and subducted with it. In contrast, structures resulting from basal erosion of the fore arc are difficult to recognize on seismic-reflection profiles, so the occurrence of basal erosion is difficult to confirm. Subduction erosion may also diminish a once-robust accretionary prism if the flux of sediments into the trench diminishes.

Nonaccretionary fore arcs may also be sites of serpentinite mud volcanism. Serpentinite mud volcanoes form where fluids released from the downgoing plate percolate upwards and interact with the cold mantle lithosphere of the fore arc. Peridotite is hydrated into serpentinite, which is much less dense than peridotite and so will rise diapirically when there is an opportunity to do so. Some nonaccretionary fore arcs, for example the Marianas, are subjected to strong extensional stresses, which allows buoyant serpentinite to rise to the seafloor and form serpentinite mud volcanoes. Chemosynthetic communities are also found on non-accretionary margins such as the Marianas, where they thrive on vents associated with serpentinite mud volcanoes.

Outer Trench Swell

The outer rise is where the descending plate begins to flex and fault as it approaches the subduction zone. Here, the lithosphere is bent upwards by plate stresses, just as the plate is bent downwards in the trench – in neither case is the plate in isostatic equilibrium. Typically, the gravity over the outer swell is about 50 mGals higher than expected from isostasy, while gravity over the trench is about 200 mGals less than that expected from isostatic considerations. The bending of the plate is associated with tension in the upper 20 km, and shallow earthquakes, caused by tensional failure induced by the downward bending of the oceanic plate, are common: about 20 extensional outer-rise earthquakes of magnitude 5 or greater occur annually. Most axes of tension are perpendicular to the trench, regardless of the direction of relative motion between the two plates, indicating that failure is controlled by bending stresses in the plate. Plate bending also causes deeper (down to 50 km) earthquakes due to compression. The width of the outer rise is directly related to the flexural rigidity of the lithosphere. The thickness of the elastic lithosphere varies between 20 km and 30 km for most trench profiles. Faulting related to plate bending and stair-stepping of the descending slab into the trench may allow seawater to infiltrate deep into the crust and perhaps into the upper mantle. Faulting of the downgoing plate results in a horst-and-graben structure, which allows sediment that reaches the trench to be deposited in graben and carried downward. This faulting also breaks up seamounts as they approach the trench (Figure 5). The mechanism of frontal erosion may operate through the combined effects of seamount tunnelling, mass wasting and

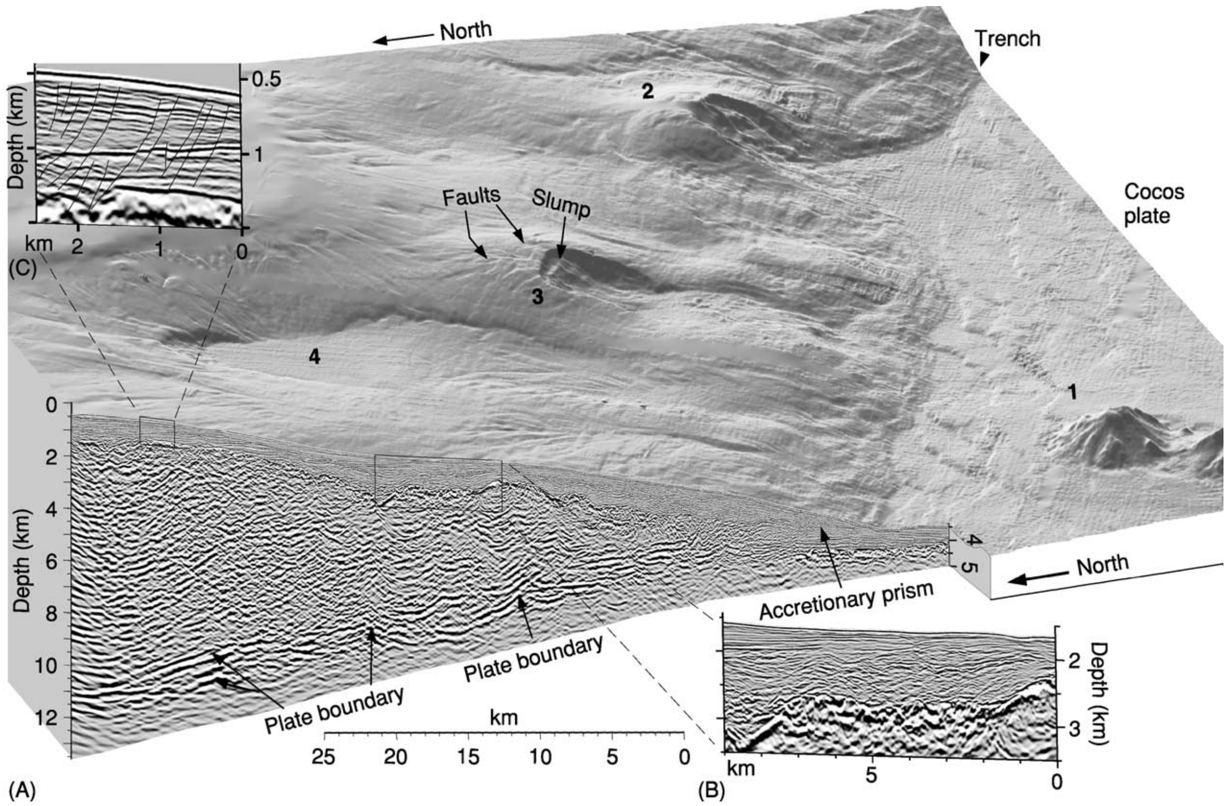


Figure 4 Downgoing seamounts and subduction erosion in the Middle America Trench off the coast of Costa Rica, where the Cocos Plate is being rapidly subducted (80 mm yr^{-1}). (A) Four seamounts in various stages of subduction (1–4) are particularly well manifested in the bathymetry of the inner trench wall. Seamount 1 (about 1 km tall) is approaching the trench and will enter it in about 200 000 years. Seamount 2 entered the trench about 200 000 years ago and is destroying the inner trench wall. The collision has caused oversteepening, with relief locally exceeding 0.5 km, leading to collapse at the sides and especially in the wake of the seamount. Note the slump and fractures. Oversteepening causes submarine landslides and flows of debris towards the trench, rebuilding the angle of repose and flooding the trench floor. Seamount 3 entered the trench about 400 000 years ago and has been swallowed beneath the accretionary prism. Debris continues to be shed from the impact zone and flows into the trench. Seamount 4 entered the trench about 600 000 years ago, and the region above it has almost completed its collapse above the sunken seamount. Subduction erosion usually occurs when debris flows fill graben (Figure 3B) in the downgoing plate and are carried down. Note also the seismic-reflection profile across the Central American fore arc. (B) Detail of part of the seismic-reflection profile. The frontal *ca.* 40 km of the margin has a rough margin-wedge top produced by seamount subduction. (C) Where the margin wedge is more than 6–8 km thick its top is smooth, cut only by normal faulting. Modified from Ranero C and von Huene R (2000) Subduction erosion along the Middle America convergent margin. *Nature* 404: 748–752.

transport to the trench, deposition in a graben on the downgoing plate, and descent into the mantle.

Controls on Trench Depth

There are several factors that control the depths of trenches. The most important is the supply of sediment, which may fill the trench so that there is no bathymetric expression. It is therefore not surprising that the deepest trenches are all nonaccretionary. Table 1 shows that all trenches deeper than 8000 m are nonaccretionary. In contrast, all trenches with growing accretionary prisms are shallower than 8000 m. A second factor controlling trench depth is the age of the lithosphere at the time of subduction. Because oceanic lithosphere cools and thickens as it ages, it subsides. The older the seafloor, the deeper it lies, and this controls the minimum depth from which the seafloor begins its descent. This obvious correlation can be removed by looking at the relative depth (Δd), which is the difference between the regional seafloor depth and the maximum trench depth. The relative depth is affected by the age of the lithosphere at the trench, the convergence rate, and the dip of the subducted slab at intermediate depths. Finally, narrow slabs can sink and roll back more rapidly than broad plates, because it is easier for the underlying asthenosphere to flow around the edges of the sinking plate. Such slabs may have steep dips at relatively shallow depths and so may be

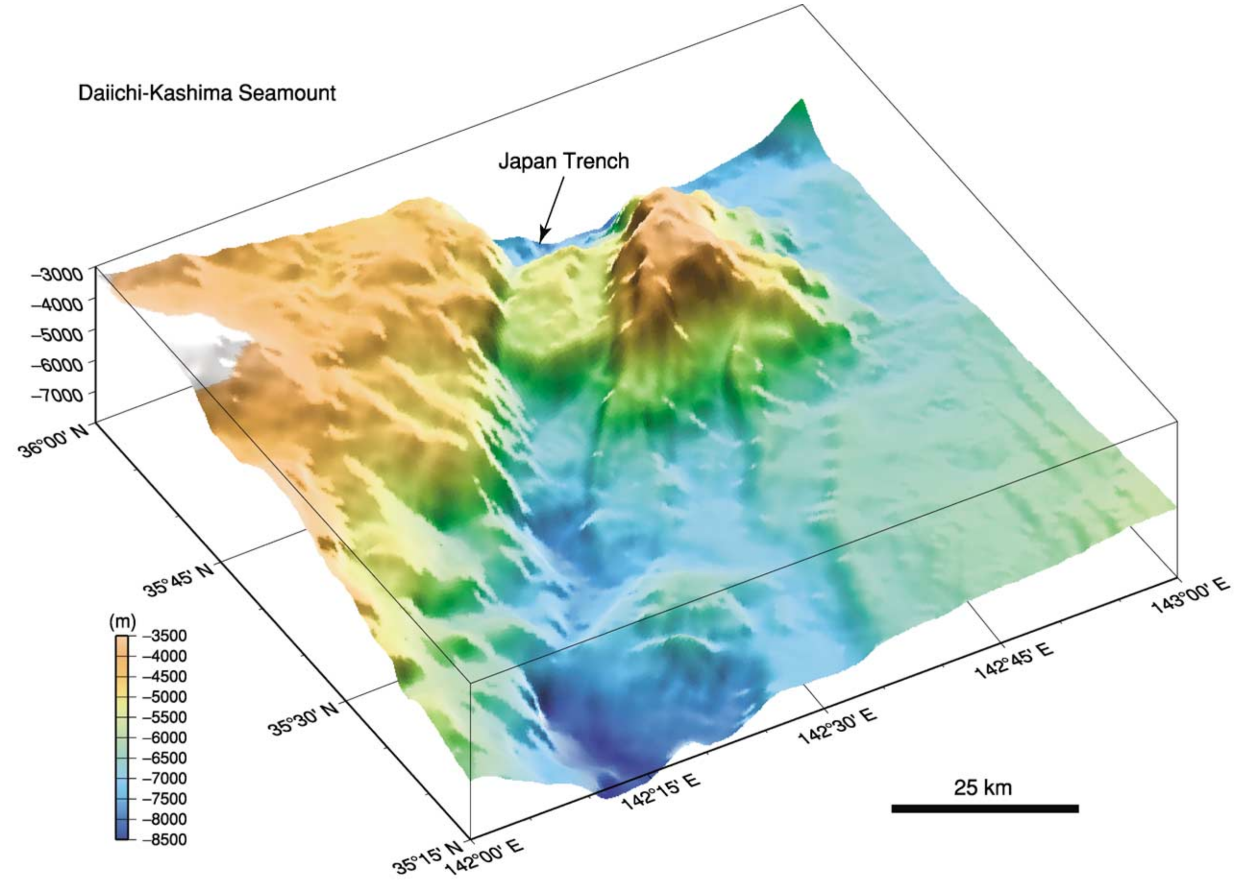


Figure 5 Bathymetric profile showing normal faulting affecting the Daiichi-Kahima Seamount as it enters the Japan Trench, east of Tokyo. The seamount, which was originally a guyot (conical sides and a flat top), has been cut by a normal fault that drops its western third by about 1 km. Smaller normal faults related to bending of the plate as it approaches the trench affect the eastern part of the seamount and the seafloor around it. Data from JODC-Expert Grid Data for Geography – 500 m (J-EGG500) Japan Oceanographic Data Center. View is from 225° (azimuth) and 30° (elevation), illumination is from the east. Vertical exaggeration is 5.5 times. Figure generated by Tomoyuki Sasaki of the Ocean Research Institute, University of Tokyo.

associated with unusually deep trenches, such as the Challenger Deep.

See Also

Analytical Methods: Gravity. **Earth:** Mantle; Crust. **Plate Tectonics**. **Seismic Surveys**. **Tectonics:** Convergent Plate Boundaries and Accretionary Wedges; Mid-Ocean Ridges.

Further Reading

Clift P and Vanucchi P (2004) Controls on Tectonic accretion versus erosion in Subduction Zones: Implications for the Origin and Recycling of the Continental Crust. *Reviews of Geophysics* v. 42.

Fisher RL (1997) Deep-sea trench. In: *Encyclopedia of Science and Technology,* 8th edn. McGraw-Hill.

Fisher RL and Hess HH (1963) Trenches. In: Hill MN (ed.) *The Sea: volume 3. The Earth Beneath the Sea,* pp. 411–436. New York: Wiley-Interscience.

Gvirtzman Z and Stern RJ (2004) Bathymetry of Mariana trench-arc system and formation of the Challenger Deep as a consequence of weak plate coupling. *Tectonics* v. 23.

Hamilton WB (1988) Plate tectonics and island arcs. *Geological Society of America Bulletin* 100: 1503–1527.

Hawkins JW, Bloomer SH, Evans CA, and Melchior JT (1984) Evolution of intra-oceanic arc–trench systems. *Tectonophysics* 102: 175–205.

Jarrard RD (1986) Relations among subduction parameters. *Reviews of Geophysics* 24: 217–284.

Ladd JW, Holcombe TL, Westbrook GK, and Edgar NT (1990) Caribbean marine geology: active margins of the plate boundary. In: Dengo G and Case J (eds.) *The Geology of North America: volume H. The Caribbean Region,* pp. 261–290. Boulder: Geological Society of America.

Murray J, Sir, Hjort J, Johan J, Appellöf A, Gran HH, and Helland-Hansen B (1912) *The depths of the ocean, a general account of the modern science of oceanography based largely on the scientific researches of the*

Norwegian steamer Michael Sars in the North Atlantic. xx, p. 821. London: Macmillan.
Sibuet M and Olu K (1998) Biogeography, biodiversity and fluid dependence of deep-sea cold-seep communities at active and passive margins. *Deep Sea Research Part II: Topical Studies in Oceanography* 45: 517–567.
Smith WHF and Sandwell DT (1997) Global sea floor topography from satellite altimetry and ship depth soundings. *Science* 277: 1956–1962.
Stern RJ (2002) Subduction zones. *Reviews of Geophysics* 10.1029/2001RG0001.
von Huene R and Scholl DW (1993) The return of sialic material to the mantle indicated by terrigeneous material subducted at convergent margins. *Tectonophysics* 219: 163–175.
Watts AB (2001) *Isostasy and Flexure of the Lithosphere.* Cambridge: Cambridge University Press.
Wright DJ, Bloomer SH, MacLeod CJ, Taylor B, and Goodlife AM (2000) Bathymetry of the Tonga Trench and forearc: a map series. *Marine Geophysical Researches* 21: 489–511.

Rift Valleys

L Frostick, University of Hull, Hull, UK

Introduction

Rift valleys are generally long narrow depressions in the Earths' crust that often contain major lakes and rivers (e.g. Lake Baikal and the Rhine and Rio Grande rivers; see Figure 1). They form as a result of the crust being stretched by plate-tectonic movements, for example during the extension that leads to the opening of new oceans. Rifts have been a feature of geological processes since the solid crust first formed, and early examples of rift valleys have been recognized in rocks as old as the Precambrian, more than 700 Ma ago. Those that have developed during more recent geological time form impressive features of the landscape, with the sides of the rift rising several kilometres above the rift floor. One good example is the East African Rift, which dominates the landscape from Ethiopia in the north to Malawi in the south (*see* **Africa:** Rift Valley).

Rift valleys have been the subject of considerable research and interest for more than a century, partly because of the inherent beauty of deep valleys but also because they are sites where both fossiliferous and economically important deposits are found (e.g. fossil hominids in East Africa, and the oil and gas found in ancient rift deposits of the North Sea; Figure 2). The reason rifts are so productive is linked directly to their mechanism of formation and to the geological structure that sits beneath the very visible surface valley.

Morphology and Structure

The surface expressions of all rifts are similar; there is a central depression or valley flanked by two uplifted shoulders, each of which is cut by a series of faults that step down towards the central lowland. The original interpretation of the geological structure that underlies this morphology was that it is like the keystone of an arch: a central piece of rock drops down between two faults (Figure 3A) to form a structure known by the German word ‘graben’. This interpretation was schooled by the surface appearance of rifts and generally ignored the fact that the flat bottoms of rift valleys are almost always clothed in sediments, which mask the underlying geology. With the development of geophysical techniques, especially seismology, that have allowed geologists to glimpse the subsurface, it has become evident that the structure of many rifts is asymmetric, with one margin being higher than the other and the floor being tilted (Figure 3B). The higher margin is cut by the largest faults, which are referred to as the main border faults. The faults are generally normal ones, with one side of a slightly inclined fracture dropping downwards relative to the other and resulting in a near-vertical step in the geology. The lower of the two margins is generally associated with a series of smaller faults inclined both towards and away from the rift axis, which fragment the geology into a series of small horst blocks with intervening valleys. These are termed synthetic faults when the fault plane is inclined in the same direction as that of the border fault and antithetic faults when the inclinations are opposed.

The precise location and magnitude of the faulting is controlled by the thickness and character of the crust undergoing rifting and by the presence of pre-existing lines of weakness, which can be exploited by new faults if they have the right orientation. In some parts of rifts there are a few very large faults, which can have total vertical displacements in excess of 2 km; in other parts there are much smaller faults and the overall displacement is much less. Often the rocks between subparallel marginal faults are

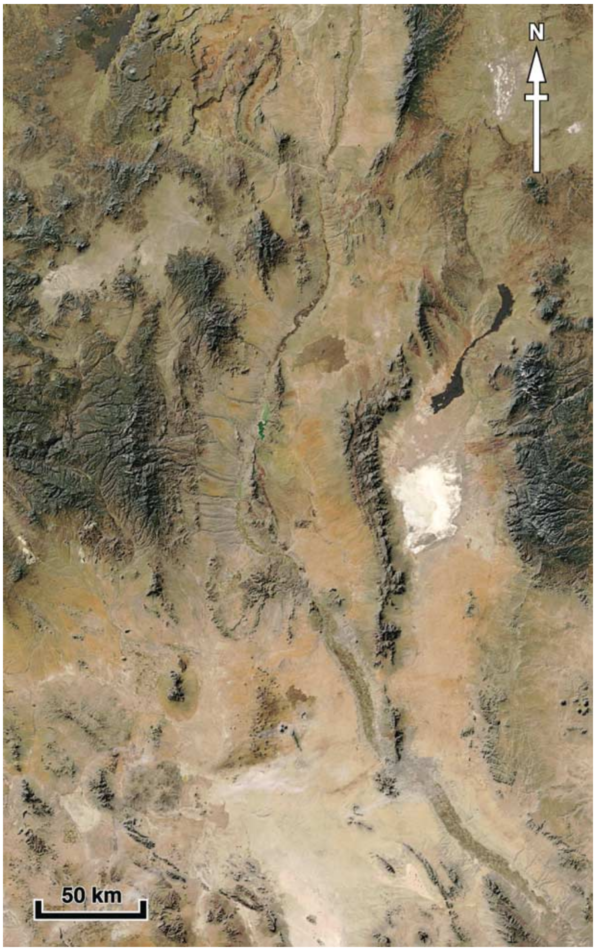


Figure 1 Satellite remote-sensing image of the Rio Grande Rift and River, USA. Image taken by the TERRA satellite using the MODIS instrument (moderate-resolution imaging spectroradiometer) and enhanced with SRTM30 shaded relief.

tilted away from the rift axis, forming small valleys between the fault blocks.

The margins of rift valleys cut by the main border fault are almost always subject to uplift, which accentuates the topographical step down to the central floor. The distance from margin to margin varies considerably, but continental examples are normally between 30 km and 200 km wide. However, the largest rift system in the world is invisible beneath the surface of our major oceans. The mid-ocean ridges are the centres at which ocean floor is created, and they dominate the seafloor topography (*see* **Tectonics:** Mid-Ocean Ridges). Typically they are between 1000 km and 2000 km wide and 2–3 km high, and they are the foci of considerable volcanic and earthquake activity.

The length of continental rifts varies considerably. The East African Rift is perhaps the best-documented example, and it runs from Afar in Ethiopia in the north to the Zambesi river in the south, a total distance of over 3000 km. However, the rift is not equally deep all along its length; there are deeper areas, often occupied by lakes, and shallower areas, often flanking lake basins. The alternation of shallow and deep areas along the rift reflects segmentation in the underlying structure. Rifts are cut into segments by cross-rift structures that delimit sections dominated by faulting on a particular margin and in a particular orientation. In adjacent segments the main border fault can be on opposite margins, have different orientations, or move laterally with regard to the rift axis. Where such changes occur, the basement rocks come closer to the surface and effectively isolate parts of the rift, ponding water to form lakes. The cross-rift structures in recent rifts may be sufficiently prominent to be seen from space (Figure 4), and the string of lakes that results is clearly seen in the African rift (Figure 5).

Origin

Although it is generally accepted that rifts form where the crust is stretched as a result of tension, the plate-tectonic setting of rifts is very variable. Rifts can be associated with all three plate-margin types (constructive, destructive, and conservative) and are also found within otherwise-stable plates. There is therefore no single underlying mechanism of formation that can explain all rifts, and hence, the ways in which a rift develops and evolves are very variable. Whatever the mechanism, all rifts are situated in areas where the crust has been stretched and has thinned, much like the thinning that occurs in warm wax when it is pulled apart. As the crust thins, hot low-density mantle material is pulled upwards towards the surface, resulting in relatively high heat flows in the rocks around most rifts. In some rifts the heating of the crust is linked with the development of major domes, which cover vast areas, and with major volcanic outpourings. A good example of areas of doming linked with volcanism can be seen in the East African Rift, where there are two domes, each over 1000 km in diameter.

Volcanic activity can commence at an early stage of rift development and can be extensive; for example, in East Africa an area of over 500 000 km^2 is covered with rift-based volcanic rocks, and many of the most notable peaks of the area are, in fact, active or dormant volcanoes, e.g. Kilimanjaro. The rock types that spill out of the volcanoes are very varied and range from basic to acidic in nature. They also include some rocks that are rarely found outside rifts, such as carbonatites (*see* **Igneous Rocks:** Carbonatites), which

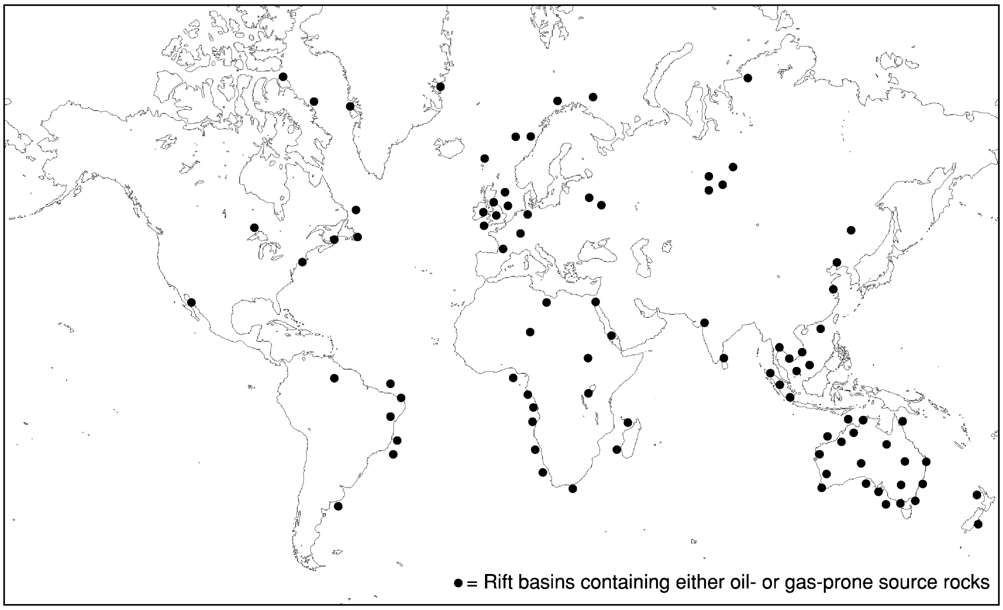

● = Rift basins containing either oil- or gas-prone source rocks

Figure 2 The locations of rift-valley deposits that contain oil and gas (black spots). Adapted from Katz BJ (1995) A survey of rift basin source rocks. In: Lambiase JJ (ed.) *Hydrocarbon Habitat in Rift Basins*. Geol. Soc. London Spec. Publ. 80: 213–240.

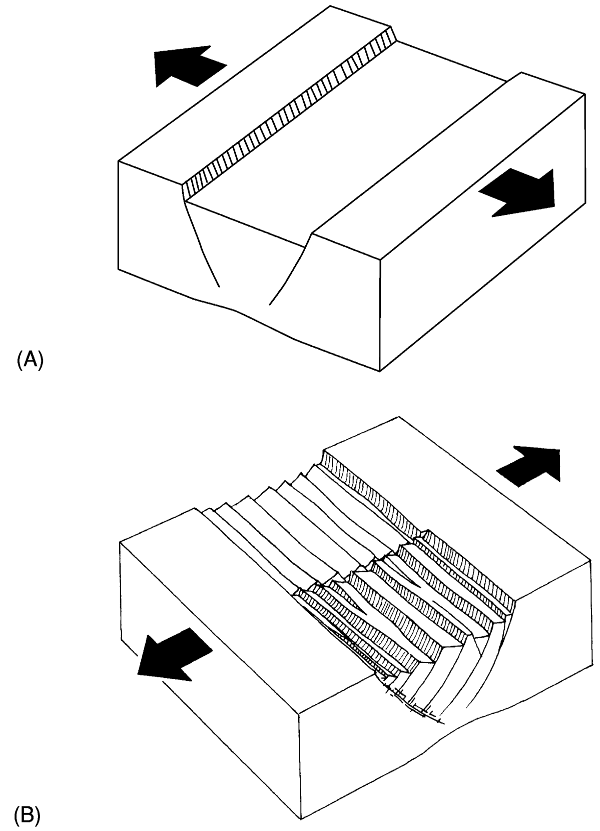


Figure 3 (A) Traditional interpretation of the structure of rifts – a full graben with symmetrical fault scarps. (B) The half-graben model of rift structure, with strongly asymmetrical fault scarps. Adapted from Perrodon A (1982) Rifts and Fossil Energy Rources, *Ancient Rifts and Troughs - Symposium of the French National Centre of Scientific Research (CNRS) Marseilles, Nov–Dec 1982.*

contain high levels of various salts that can be dissolved and swept into lakes, making them saline.

The Impact of Rifting on Hydrology, Climate, and Ecology

The development of a rift has a major impact on the overall environment. Continental rifts are often formed in major landmasses that are crossed by large rivers that have spent millions of years wearing down the surface. Rifting produces new topography to which the surface processes adapt, and the consequences of rifting can include the disruption of weather patterns, the diversion of rivers, and the formation of new lakes.

The uplift that occurs along the rift margins will decrease the ambient temperature and can increase rainfall. In contrast, the rift-valley floor remains warmer and can experience a rain-shadow effect since rain forms where clouds are forced to rise, and this occurs preferentially on the higher topographical areas of the flanks. As a result there are often different types of vegetation on the flanks and in the valley bottom. This can include the development of rainforest if the rift is in the tropics. Both the topography and the contrast in habitats across the rift valley can act as barriers to the migration of animals and, to a lesser extent, the spread of plant species. Where a rift forms in a previously undivided continent it produces new diversity of habitat and isolates individuals, both of which can boost evolution. This is one of the reasons

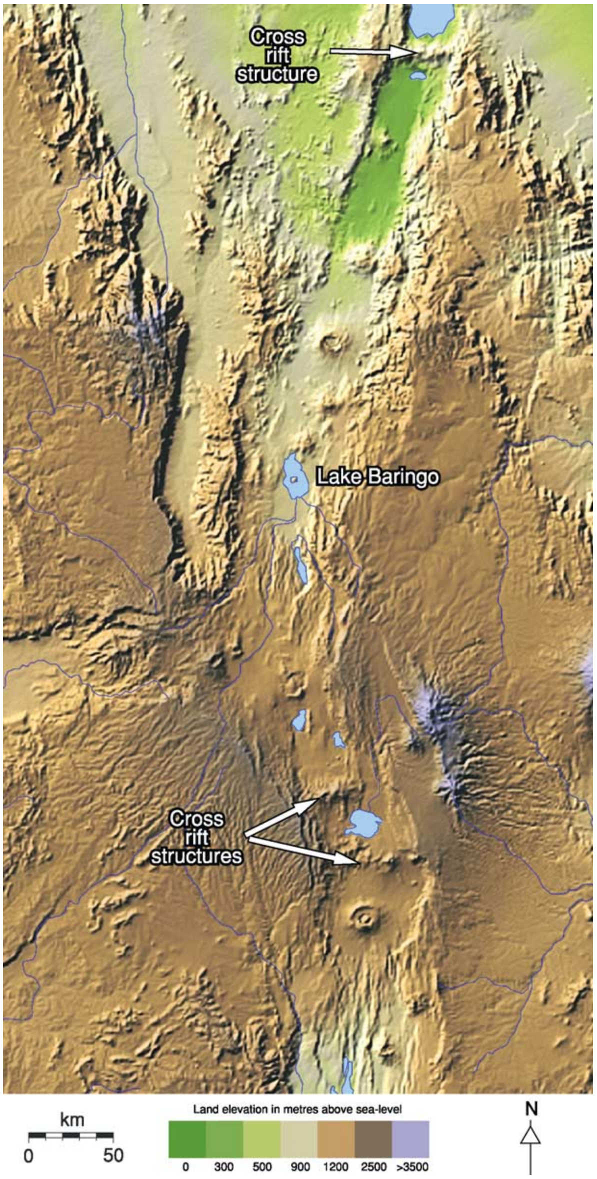


Figure 4 Cross-rift structures seen clearly in this satellite image of the East African Rift. This is a shaded relief map produced from SRTM30 data with colour added to indicate land elevations.

why many early hominid remains have been located in rift sediments. Of particular importance in this regard are the deposits in the East African Rift, which have been excavated and studied by many palaeoanthropologists, including successive members of the Kenyan Leakey family.

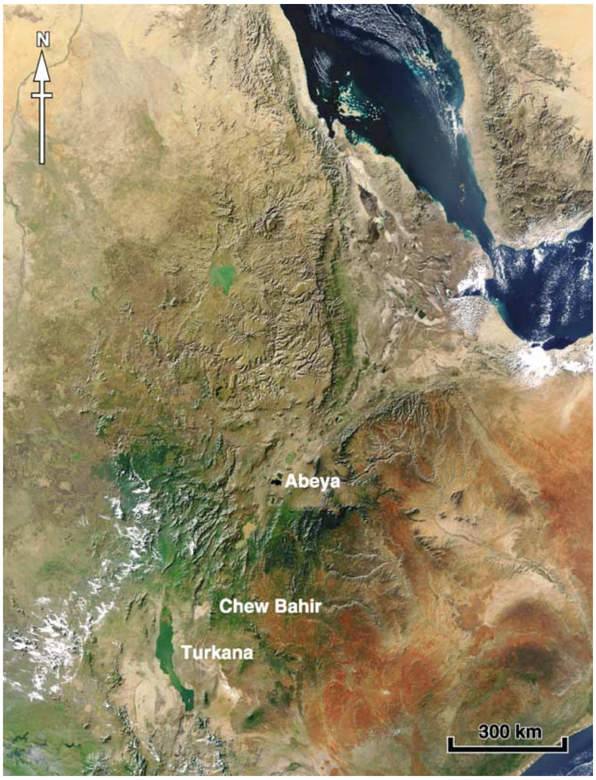


Figure 5 Satellite remote-sensing image of the northern Kenyan–Ethiopian section of the East African Rift showing a string of separate lakes. Image taken by the TERRA satellite using the MODIS instrument (moderate-resolution imaging spectroradiometer) and enhanced with SRTM30 shaded relief.

The development of the new topographical features of a rift will disrupt pre-existing continental drainage patterns and result in a very different landscape. Before rifting the topography is often subdued, with a very small number of large and ancient rivers draining towards the continental margin. The impact of rifting on the rivers will depend on their orientation. If the new rift parallels existing rivers, it can capture most or all of them, but if it cuts across the general trend of the drainage, streams can be beheaded, diverted, or even reversed. Domed sections of rifts are particularly effective at diverting rivers. The effect is similar to that of piling up soil into a heap so that water falling onto it is shed in all directions, giving rise to a radial pattern of channels.

Active faults form steps in the landscape, which can also affect rivers. Where there is more than one parallel fault, rivers can be caught between the two steps and flow parallel to the rift axis for many hundreds of miles. Margin uplift and tilting of fault blocks both tend to divert rivers away from rift basins. Drainage that has been affected in this way is often captured by continental river systems; for example, the Nile was augmented by the waters of rivers that were diverted away from the Red Sea and Gulf of Aden rifts. In some rifts cross-rift structures form topographical barriers that pond up drainage systems to form lakes. Good examples of this aspect of rifts are Lakes Baikal and Magadi in the Baikal and East African Rifts, respectively (Figure 5). In some rifts

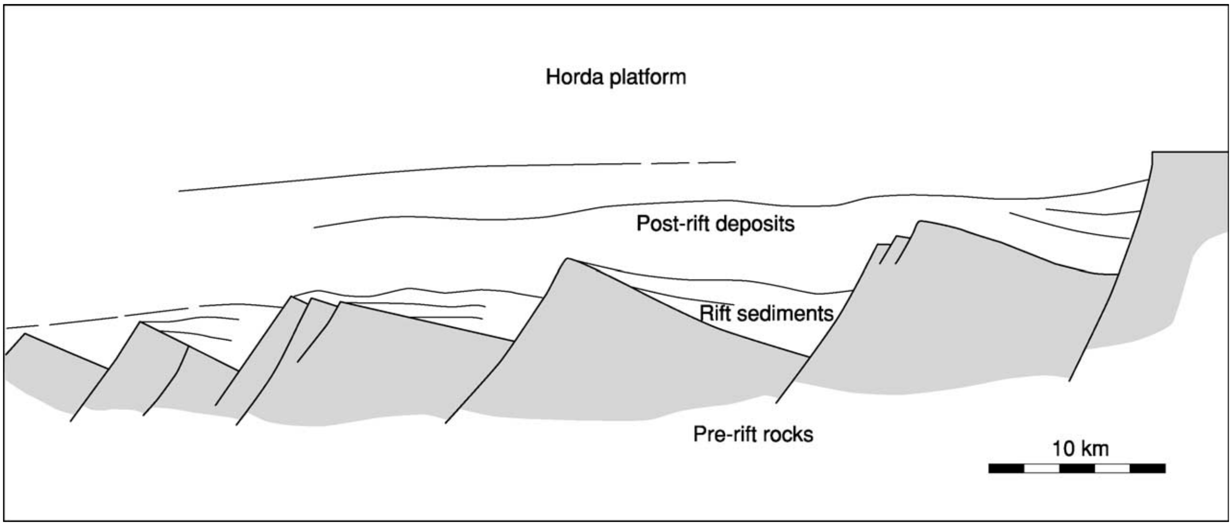


Figure 6 Diagrammatic cross section of the rift and post-rift sedimentary fills of the Horda platform in the Norwegian North Sea. Note the tilting of the lower deposits, labelled 'rift sediments', which were emplaced during the period of active rift faulting. Adapted from Steel R and Ryseth A (1990) The Triassic-Early Jurassic succession in the northern North Sea: megasequence stratigraphy and intra-Triassic tectonics. In: Hardman RFP and Brooks J (eds.) *Tectonic Events Responsible for Britain's Oil and Gas Resources*. Geol. Soc. London Spec. Publ. 55: 139–168.

there are no large lakes, and in these there is often a large river that drains the length of the valley, for example in the Rhine and Rio Grande Rifts. Fault scarps are rarely crossed by large rivers, but small streams flow down the step created by the fault, cutting small steep valleys and depositing sediment in alluvial fans at the base of the slope.

Rifts as Sedimentary Basins

In an active rift the floor of the basin can continue to subside for millions of years. Throughout that time, sediments will be swept into the basin to fill the hole being created and will build up into a sequence of deposits with common characteristics that make it possible to identify a rift basin long after the surface expression has disappeared. The main characteristics are wedge-shaped deposits thickening towards the main border fault and river or lake deposits in the deepest part of the basin flanked by alluvial-fan and river deposits on either side. As subsidence is often asymmetrical, lower layers become progressively tilted relative to the surface (Figure 6) as whole sequences of sediment wedges are superimposed on each other. If subsidence ceases, the structure will eventually lose its topographical expression and the basin will stop accumulating sediments. Rifts can remain active for very long periods of time and will therefore accumulate many

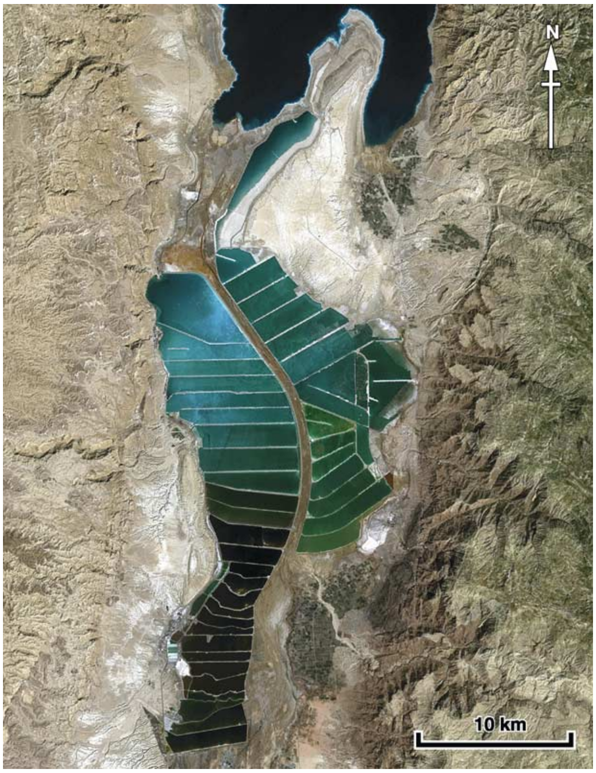


Figure 7 Satellite image of the salt pans at the southern end of the Dead Sea. The very saline waters of this lake are directed into shallow pans and evaporated to extract salts rich in bromine. This is a LANDSAT-7 ETM image using bands 3, 2, and 1 for red, green, and blue colour channels.

kilometres of sediment. Good examples of ancient buried rifts are the Reconcavo basin in Brazil, the Jeanne d'Arc basin off the coast of Newfoundland, and the Triassic Richmond and Taylorsville basins of eastern Virginia, USA.

Economic Deposits in Rifts

The sedimentary sequences that fill rifts can host a range of economic deposits, which are exploited around the world. Of these, the most significant are oil and gas; for example the Jurassic basins of the North Sea between the UK and Norway and the Tertiary basins of the Red Sea are rifts that contain appreciable reserves of oil and gas (Figure 2). The oil is derived from organic-rich fine sediments that accumulated in the deeper quieter part of the basin and which, when buried and heated, migrated into adjacent deposits and filled the pore spaces. Salts that accumulate in both marine and nonmarine rifts have a long history of exploitation; for example, the bromine-rich salts of the Dead Sea Rift (Figure 7) have been extracted since the Bronze Age – a period of more than 4000 years. The processes of rock erosion and sediment deposition can lead to the enrichment of parts of the new rift deposit with heavy minerals such as metals and gemstones, and the sediments themselves can be an extractable resource, particularly if the rift is close to areas of human population where there is a demand for sands and gravels as building materials.

Perhaps the most significant property of rifts is the spectacular scenery associated with them, which attracts tourists and is economically significant. One good example of this is Death Valley in the USA (Figure 8), which is in an area of low rainfall and as such has sparse vegetation, allowing the structure to be seen very clearly even by non-geologists.

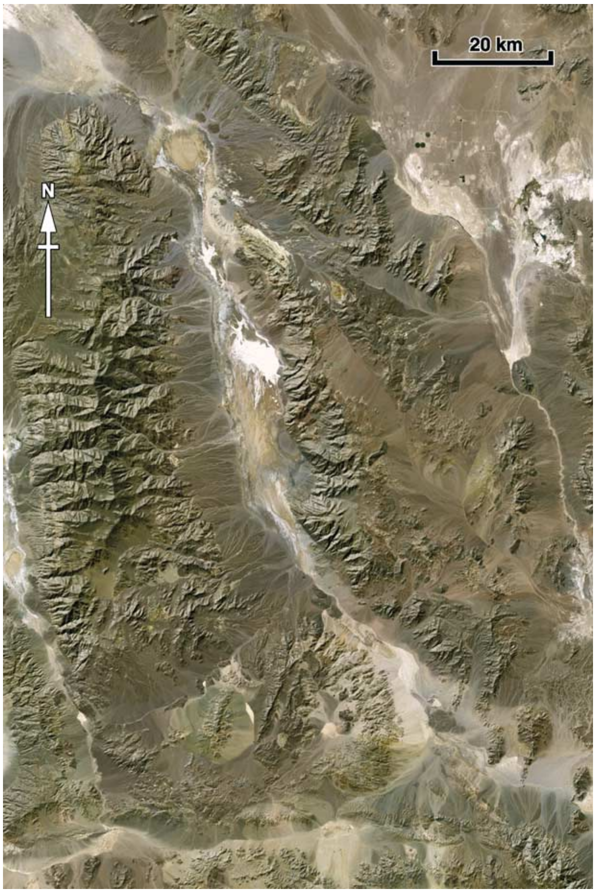


Figure 8 Satellite image of Death Valley, USA, showing clearly the fault scarps and surface deposits. This is a LANDSAT-5 TM image using bands 3, 2, and 1 for red, green, and blue colour channels.

See Also

Africa: Rift Valley. **Geomorphology**. **Igneous Rocks:** Carbonatites. **Petroleum Geology:** Reserves. **Sedimentary Environments:** Alluvial Fans, Alluvial Sediments and Settings; Lake Processes and Deposits. **Tectonics:** Faults; Mid-Ocean Ridges.

Further Reading

Allen PA and Allen JR (1990) *Basin Analysis Principles and Applications*. Oxford: Blackwells.

Frostick LE and Reid I (1989) Is structure the main control on river drainage and sedimentation in rifts? *Journal of African Earth Sciences* 8: 165–182.

Frostick LE and Steel RJ (eds.) (1993) *Tectonic Controls and Signatures in Geological Successions* Special Publication 20. International Association of Sedimentologist: Blackwell Scientific Publications.

Gawthorpe RL and Hurst JM (1993) Transfer zones in extensional basins their structural style and influence on drainage development and stratigraphy. *Journal of the Geological Society* 150: 1137–1152.

Gupta S and Cowie P (2000) Processes and controls on the stratigraphic development of extensional basins. *Basin Research* 12: 185–194.

Lambiase JJ (ed.) (1995) Hydrocarbon habitat in rift basins. Special Publication 90. London: Geological Society.

Leeder MR, Mack GH, and Salyards SL (1996) Axial transverse fluvial interactions in half graben: Plio-Pleistocene Rio Grande rift, New Mexico, USA. *Basin Research* 8: 225–242.

Palmason G (ed.) (1982) *Continental and Oceanic Rifts*. Geodynamics Series A, volume 8. Washington: American Geophysical Union.

Selley RC (ed.) (1997) *African Basins. Sedimentary Basins of the World 3*. Amsterdam: Elsevier.

Summerfield MA (1991) *Global Geomorphology: An Introduction to the Study of Landforms*. Harlow: Longman.

TEKTITES

G J H McCall, Cirencester, Gloucester, UK

Introduction

Tektites, natural glass objects of unknown origin, were described in the tenth century in China and in the eighteenth century in Europe; Charles Darwin described one of these objects from Western Australia while on the Beagle, and four strewn fields were recognized by the twentieth century. An immense volume of work in the past 50 years, involving many state-of-the-art laboratory techniques, has conclusively shown that tektites are the product of a handful of the ~170 terrestrial large-scale events in the geological record. A number of similar events that produced analogous, but somewhat different, glass bodies are also recognized in the geological record, ranging back through the Pleistocene, Pliocene, Oligocene, Eocene, Cretaceous–Tertiary (K–T) boundary, Late Devonian, and possibly even the Archaean. The primary processes of formation of the splash forms of tektites (projection from the target site) and the secondary ablation process that produces flanged, glassy objects (on descent to the strewn field) are now well understood, but much is still unresolved concerning the processes at the target site and the manner of dispersion; the largest of the strewn fields represents a dispersion that covered one-tenth of Earth's surface with glassy objects.

Historical

For centuries, tektites were one of the mysteries of geology. Found in certain sediments of Cenozoic or Quaternary age or strewn on the surface of present-day salt lakes or sand dunes, restricted to four well-defined strewn fields of different geological ages, tektites were once the subject innumerable hypotheses, invoking such visitations as atmosphere-grazing or-skipping comets and lunar ejections. The late twentieth century saw the advent of the Space Age and intensified research showed conclusively that tektites are terrestrially sourced and are the product of immense-scale impact events. Further occurrences of related objects in rocks of Pleistocene, Pliocene, and Eocene ages, at the Cretaceous–Tertiary (K–T) time boundary, in the Late Devonian and possibly in the Precambrian, were recognized at the same time, and research on seafloor sediments in drill cores revealed microscopic tektites (microtektites) associated with three of the four originally defined strewn fields.

In Europe, Cro-Magnon man valued tektites 30 000 years ago, using the glass for ornaments. The Chinese recorded the occurrence of these objects in the *Record of Heterodoxy outside Nanling*, compiled in the tenth century. The objects found strewing fields in the Leizhou Peninsula were attributed to thunderstorms, being termed *Lei-gong-mo* or *Lei-gong-shih* (the 'ink stocks' or 'stools' of the 'thunder god') (Figure 1A). Tektites from the Austrian empire (the present-day Czech Republic) were described by Josef Mayer in 1788. Because of the glass industries there, a connection with prehistoric glass making was suspected. Charles Darwin was shown a tektite at Albany, Western Australia, in the 1830s and deemed it to be a volcanic 'bomb' (Figure 1B). In the 1930s, Alfred Lacroix described occurrences of these objects in Indochina and the Ivory Coast, and Henryk Stenzel described tektites from Texas.

Explanations advanced for the formation of tektites have included lightning strikes, volcanic eruption, prehistoric and historic glass manufacture, burning coal seams, and desiccation of silica gel masses. Extra-terrestrial origin related to meteorites was advanced first by Charles Walcott in 1898 and Eduard Suess in 1900, and this was followed by a rash of extraterrestrial hypotheses (for example, oxidation of the tail of comets; plastic seepings from meteorites passing through the atmosphere; debris from an Earth-like celestial planet blanketed by sedimentary rocks; light-metal meteorites plunging into the atmosphere and producing glass; cometary flight skipping through Earth's atmosphere). Ejected material following lunar impacts became popular as a proposed source of tektites from 1940 up to the time of the Apollo XI landing and the Luna unmanned sample recovery mission, which together dealt a death blow to such provenance because of the data on the silica-poor nature of the lunar surface rocks. Spencer, in 1933, appears to have been the first scientist to come to the right answer – "some form of spallation of melts from impacts on the Earth". In 1962, Ross Taylor, in Australia, concluded from the isotopic evidence that the material of tektites came from Earth, and the immense volume of research carried out on tektites since the Apollo XI landing has confirmed this relationship; besides geochemical, mineralogical, and isotopic evidence, the Ivory Coast strewn field has been conclusively related to the Bosumtwi, Ghana, impact crater, and the Central European strewn field has been related to the Ries

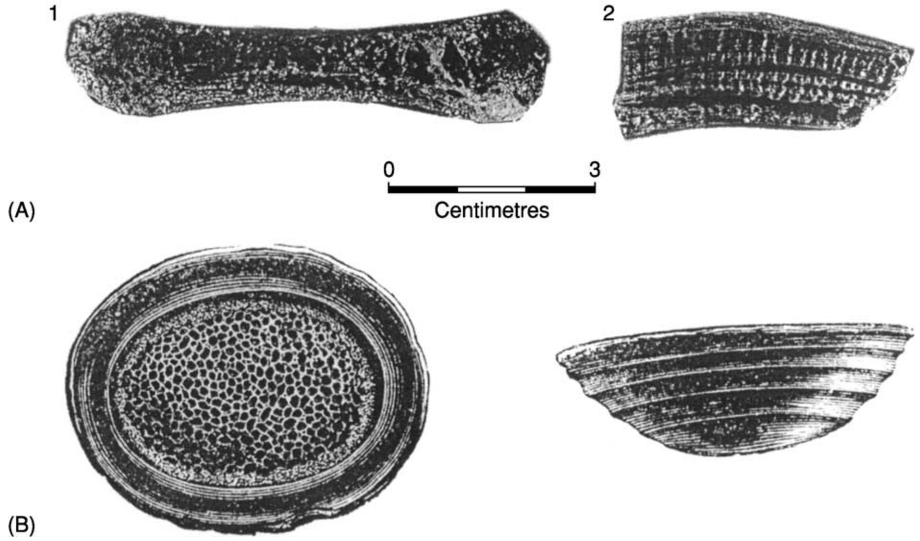


Figure 1 (A) Tektites from China, showing the (1) dumbbell shape (2) layered, Muong Nong type. (B) Charles Darwin's drawing of a flanged button australite: note the ring waves on the anterior surface. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

impact structure in south-east Germany (though there remain some aspects of this relationship still difficult to explain).

Strewn Fields

There are four strewn fields, all of which are of Late Eocene age or younger. However, there are other occurrences of related glassy objects and two of anomalous natural glasses, both of which are not tektites by definition, but are almost certainly closely related to them. The four strewn fields are summarized in Table 1; their distribution is shown together with other related occurrences in Figure 2.

Tektites commonly display primary regularly shaped 'splash' forms characteristic of spinning masses of melt solidifying while travelling through the atmosphere following ejection (though some are irregularly shaped); they may also have superimposed on these forms secondary shapes that were the product of ablation during their descent through the atmosphere hundreds or thousands of kilometres away from the target area. They may then be further modified by terrestrial agencies after falling to Earth.

North American Strewn Field

In North America, tektites were first discovered in Grimes County, Texas, USA, where they occur in the Eocene Jackson Formation and overlying Pleistocene gravels: they were called 'bediasites', a name derived from the Bidai Indians of south-eastern Texas. Bediasites consist of black glass, commonly deeply etched (Figure 3A), and shapes include splash-form spheres, discs, teardrops, and peardrops. Secondary ablation shapes are extremely rare, but at least one example has been reported (Figure 3B). Tektites subsequently recovered from Dodge County, Georgia, are similar, and the single recovery from Martha's Vineyard is like these, rather than those from Texas. All specimens yield the same radiometric age (K/Ar, Ar/Ar, or fission track), dating to 34.9 Ma, consistent with their presence in Eocene sediments. Microtektites of the same radiometric and microfossil controlled age are known from a number of drill cores in the Caribbean, and both microtektites and tektite fragments have been found in Uppermost Eocene sediments on land in Barbados; microtektites of this age have also been found in a drill core from as distant a site as the Weddell Sea, Antarctica. The source of this strewn field is believed to be the 85-km-diameter Chesapeake Bay impact structure in Delaware, which is of the right age.

Central European Strewn Field

Tektites have long been known to occur in the Czech Republic, Austria, and near Dresden and Kottbus in Germany, where they are called 'moldavites'. Many are of a greenish, translucent tint, and these have been used as gemstones. They occur in Miocene sediments, consistent with their radiometric age dating to 15.1 Ma, but also are found in reworked Pleistocene gravels. They exist in a variety of splash forms (ovoids, discs, teardrops, and rods) and their surface is commonly rough (Figure 4). Again, secondary shapes due to ablation are extremely rare.

Table 1 Tektite distribution

Strewn field	*Area covered*	*Age (isotopically determined)*
North American (~4000 tektites recovered)	Texas, Georgia, Martha's Vineyard, Barbados (microtektites in the Caribbean and Weddell Sea)	34.9 ± 2.5 Ma (Eocene)
Central European (55 000 tektites recovered)	Czech Republic, Austria, Germany	15.1 ± 0.1 Ma (Miocene)
Ivory Coast (~200 tektites recovered)	Ivory Coast (microtektites offshore West Africa)	~1 Ma (Pleistocene)
Australasian (South-east Asia, >600 000 tektites recovered; Australia, ~100 000 tektites recovered)	China, Indochina, Thailand, Malaysia, Philippines, Indonesia, Australia, Central Indian Ocean (microtektites over wide area of the Indian Ocean, China Seas, and around Australia)	0.77–0.78 Ma (Pleistocene)

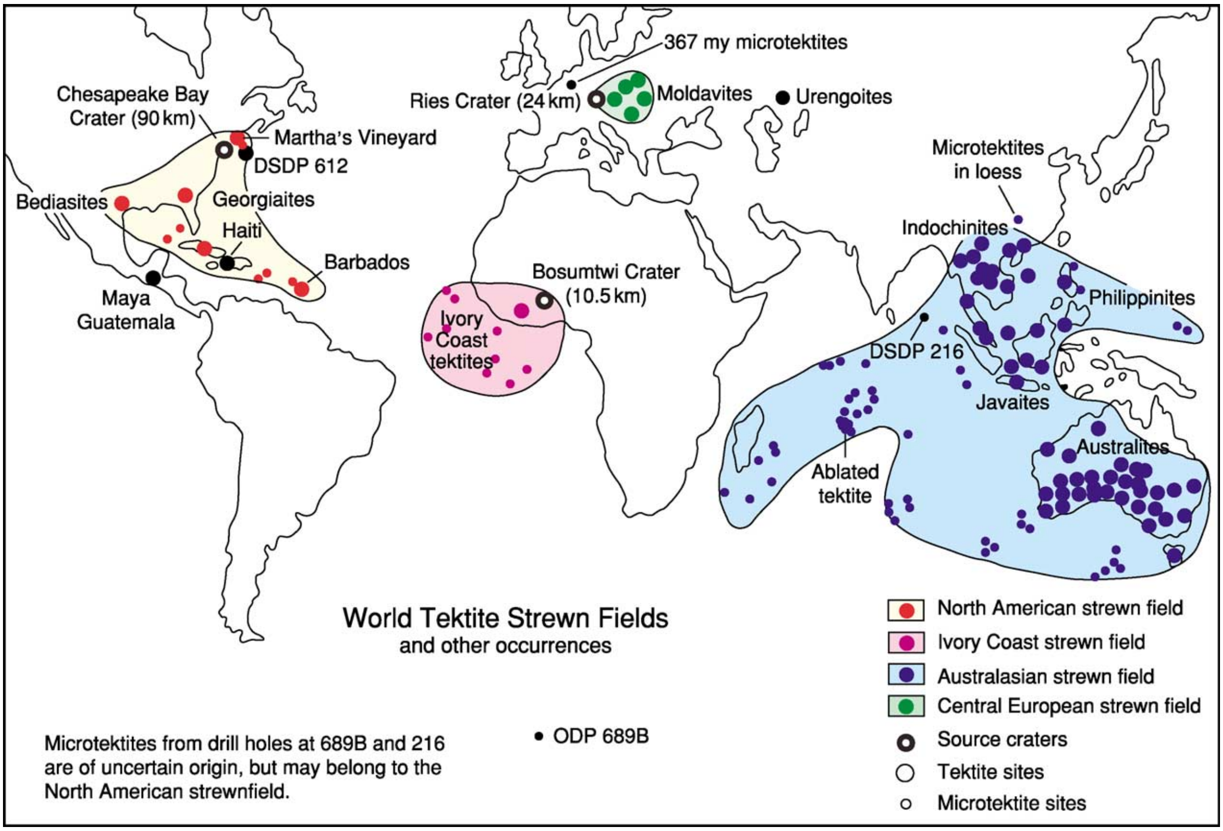


Figure 2 Strewn field distribution of tektites and sites of related occurrences. Deep-Sea Drilling Project (DSDP) and Ocean Drilling Program (ODP) sites are also indicated. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House, and McNamara K and Bevan A (2001) *Tektites*. Perth: Western Australian Museum.

Microtektites are unknown in this strewn field. The source is believed to be the 24-km-diameter Ries impact structure in south-east Germany, which is of the correct radiometric and stratigraphic age, but there are some unsolved questions relating to how exactly they were expelled and what rock formation supplied the material to form the melt and subsequent glass.

Ivory Coast Strewn Field

A very small number of tektites have been recovered from an area to the west of the Camoe River, near Ouelle, in the Ivory Coast, West Africa. They occur in surficial alluvial deposits over Precambrian (Birrimian) rocks in a gold-mining area. They are of black pitted glass and several splash forms are

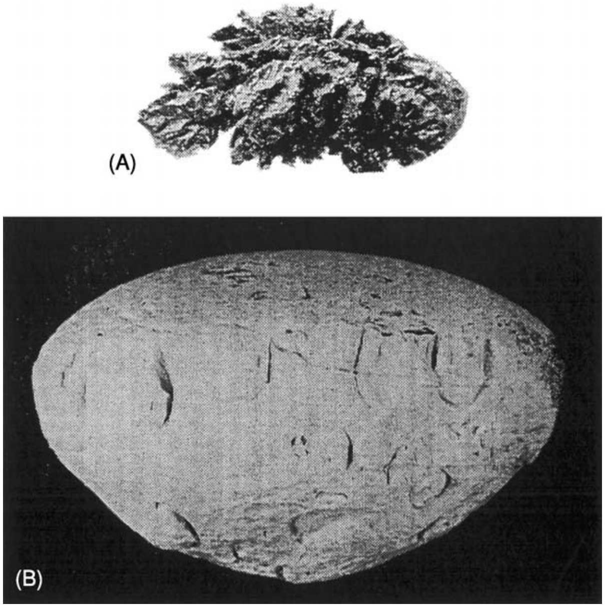


Figure 3 Two bediasites from Texas, USA. (A) Specimen deeply etched by terrestrial agencies. (B) Specimen showing secondary ablation (maximum dimension, 25 mm). Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

recorded (discs, spheroids, teardrops, and peardrops) (Figure 5). Radiometric ages are ~1 million years. An extensive area of the Atlantic Ocean off the coast of West Africa has yielded numerous microtektites in drill cores, and these occur close to the Jaramillo Magnetic Reversal dated at 0.97 Ma, consistent with the radiometric age of the tektites. This strewn field is attributed to the Bosumtwi Crater to the east, in Ghana, as the source impact structure, this being confirmed by radiometric dating, geochemistry, and isotopic methods. Very sophisticated methods related to Os/Re isotopes have shown that the bulk of the osmium in the tektites is extraterrestrial.

Australasian Strewn Field

The Australasian strewn field covers one-tenth of Earth's surface. The northern part of the field includes China, Indochina, Thailand, Malaysia, the Philippines, Borneo, and Indonesia (especially Java, Belitung, and Flores), and the microtektites from marine cores adjacent to these territories are also included. Though several names exist for these tektites (e.g., indochinites, thailandites, javanites, billitonites, rizalites, and philippinites), radiometric dating yields the same age of formation (0.77–0.78 Ma) for all specimens; thus they all represent the same event and are best referred to as 'South-east Asian tektites'. They differ greatly in physical form over the geographical range in which they are found; these differences are explained by climatic influences and the projected swathe of their descent to Earth. For example, the moist climate in the north part of the strewn field can be correlated with much more pitted, etched, and grooved surfaces, and the comparative lack of secondary ablation forms of northern specimens can be contrasted with specimens from the (distal) south.

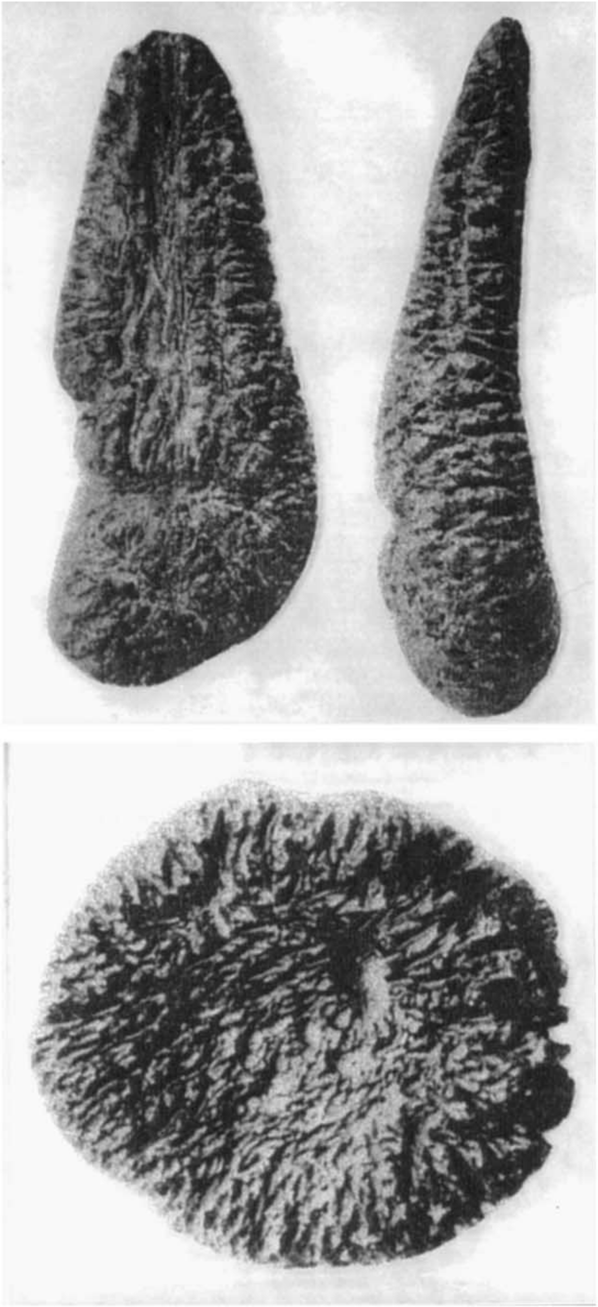


Figure 4 Moldavites from the Czech Republic, showing two teardrops and a discoidal form. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

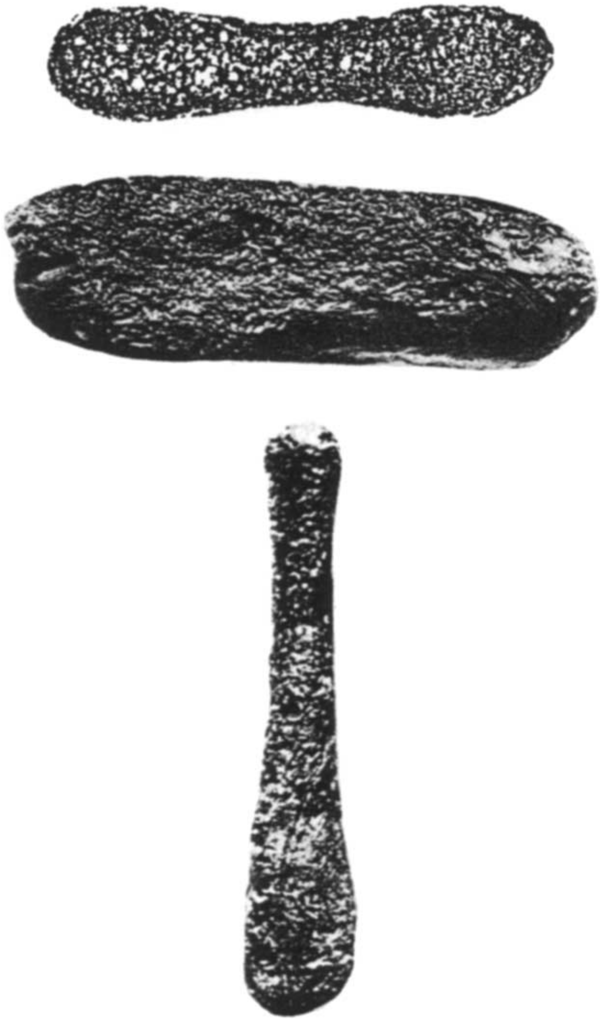


Figure 5 (Top) Dumbbell-shaped tektite, 9 cm long, from the Ivory Coast; (middle and bottom) tektites from Indochina, showing the boat shape and teardrop form. Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

Most specimens exhibit only splash-form shapes (spheres, ovoids. discs, dumbbells, boat shapes, teardrops, and peardrops). They typically occur in superficial deposit profiles and may be above or below laterite horizons.

Australasian strewn field tektites include the layered and irregularly shaped Muong Nong-type tektites, which were first identified in Laos. These may weigh up to 24 kg and they occur over hundreds, if not thousands, of kilometres, across the strewn field in Indonesia and Thailand (and possibly the Philippines), and the manner of their transport such distances from their source remains unexplained. These tektites also include suites of relict heavy and refractory minerals familiar in terrestrial sediments (quartz, corundum, monazite, rutile, zircon, chromite, andalusite, sillimanite, and kyanite). Lechatelierite and coesite are reported; the former is never found in volcanic glasses and the latter is a polymorph of quartz associated with high pressures in impact explosion processes, though it can occur in tectonic extreme pressure situations (*see* **Ultra High Pressure Metamorphism**). There are intermediate layered tektites in the Indochina collections, but these are splash-form shaped and are not irregular. Muong Nong-type tektites are of rare occurrence in the North American and Central European strewn fields; are not represented at all in the Ivory Coast strewn field or in the southern Australian part of the Australasian strewn field. Immense numbers of tektites may be recovered from a single site in Indochina, and several hundred thousand are reported from Da Lat alone.

The southern part of the Australasian strewn field covers the Australian continent. The tektites here show all the well-known splash forms, but ablated forms of these, particularly perfectly flanged buttons of relatively smooth black glass, are not uncommon (Figure 6A), though the flanges break off easily, leaving chatter-marked collars where the flanges have separated (Figures 6B and 12), thus producing the most common form ('cores'). Dumbbells, boat shapes, and teardrops may also show ablation flanges (Figure 7). Some australites have ablated away so much of the body of the splash form that they are preserved as flat discs (Figure 8). Radiometric dating yields the same age as for South-east Asian tektites (0.77–0.78 Ma), and the specimens represent the same event. It has been claimed, based on stratigraphic evidence, that many Australian tektites fell to Earth very much later than this, and that an 'age paradox' is at work, but this is now refuted (though radiometric dating does indicate that there is a small cluster of older ~10-million-year-old Na-rich tektites near the Western Australia/South Australia border). A feature of the Australasian subfield is that large splash forms cluster around certain localities, an unexplained development. Australites are mainly found on the surface of salt-pan lakes in Western Australia, where they are washed in; they are also found in diamondiferous deposits in the north of Western Australia, on and between sand dunes in South Australia, and in Quaternary sediments in Victoria.

Microtektites were first discovered in the seas off the coasts of South-east Asia and Australia. Their stratigraphic and radiometric ages are consistent with being part of the strewn field. Some contain coesite and stishovite, both of which are high-pressure polymorphs of quartz associated with impact explosion processes; stishovite is found only associated with these types of processes. A single tektite has been

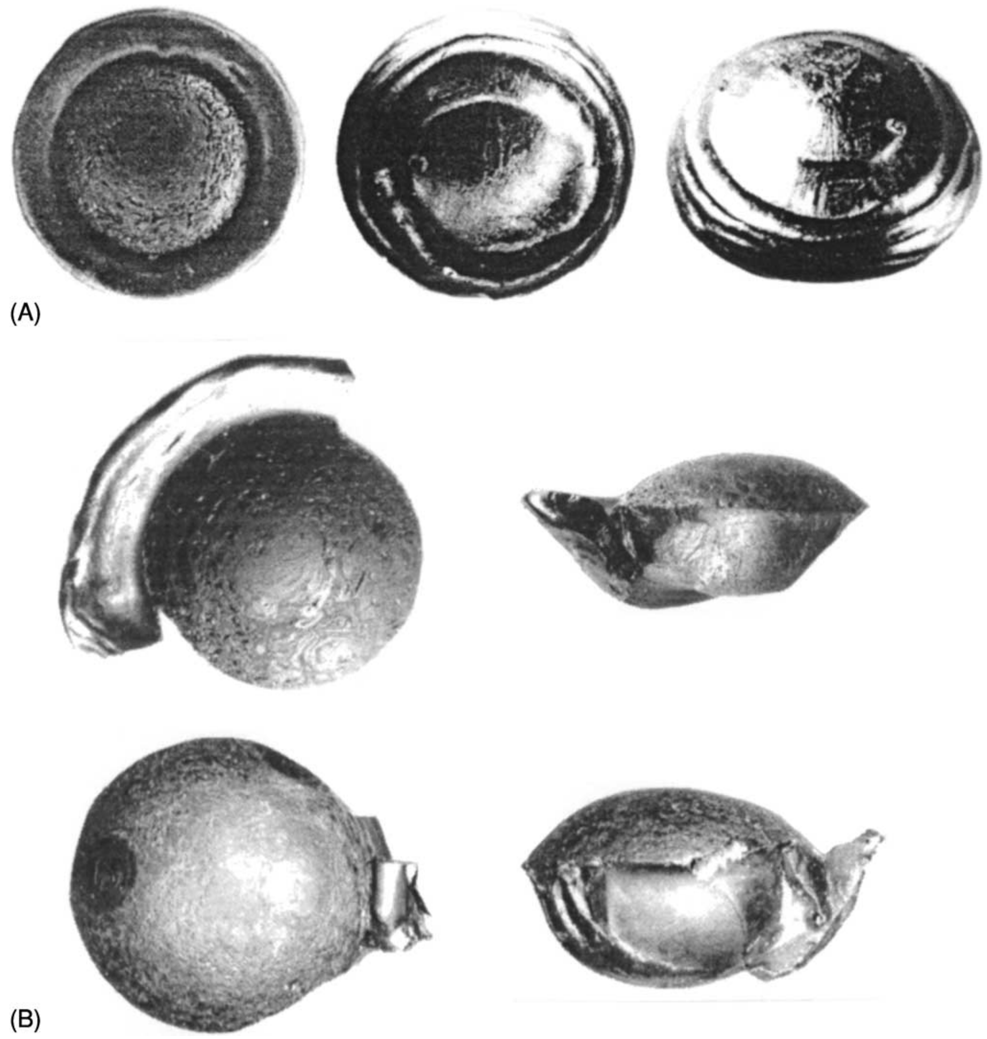


Figure 6 Australites from the Finke River, Central Australia. (A) Three views of a perfect flanged button, showing a posterior surface in flight and two anterior surfaces with ring waves produced by ablation (maximum dimension, 20 mm). (B) Four views of australites with flanges partly preserved as they separated from the remnant 'core' (maximum dimension, 17.5 mm). Photographs by WH Cleverly; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

recovered from a grab sample in the middle of the Indian Ocean (Figure 9) and is attributed to the same event leading to the other Australasian strewn field tektites. Microtektites have been recovered from cores all the way across the Indian Ocean, to sites not distant from Madagascar.

The source of the Australasian tektites remains a complete mystery. Tonle Sap Lake in Cambodia was investigated as a possible impact site, but no evidence was found of impact there. Studies of progressive changes in populations of microtektites and delineation of the restricted area of occurrence of coesite and stishovite content suggest a source in Indochina, in Cambodia, not far from Tonle Sap. The whole character of the strewn field suggests that China and Indochina are proximal to the source, and Australia is distal. It is inexplicable that the source structure, of such a geologically young age, and presumably larger than the 85-km-diameter Chesapeake Bay structure, in view of the strewn field dimensions, is not preserved to some extent. This has led to suggestions that the Australasian tektites derived from an explosive event in the atmosphere, as has been widely accepted for the Tunguska event in Siberia in 1908.

Microtektites

These microscopic glass bodies are seldom larger than 1 mm in diameter. They display the familiar splash forms in miniature (mostly spheres, but also

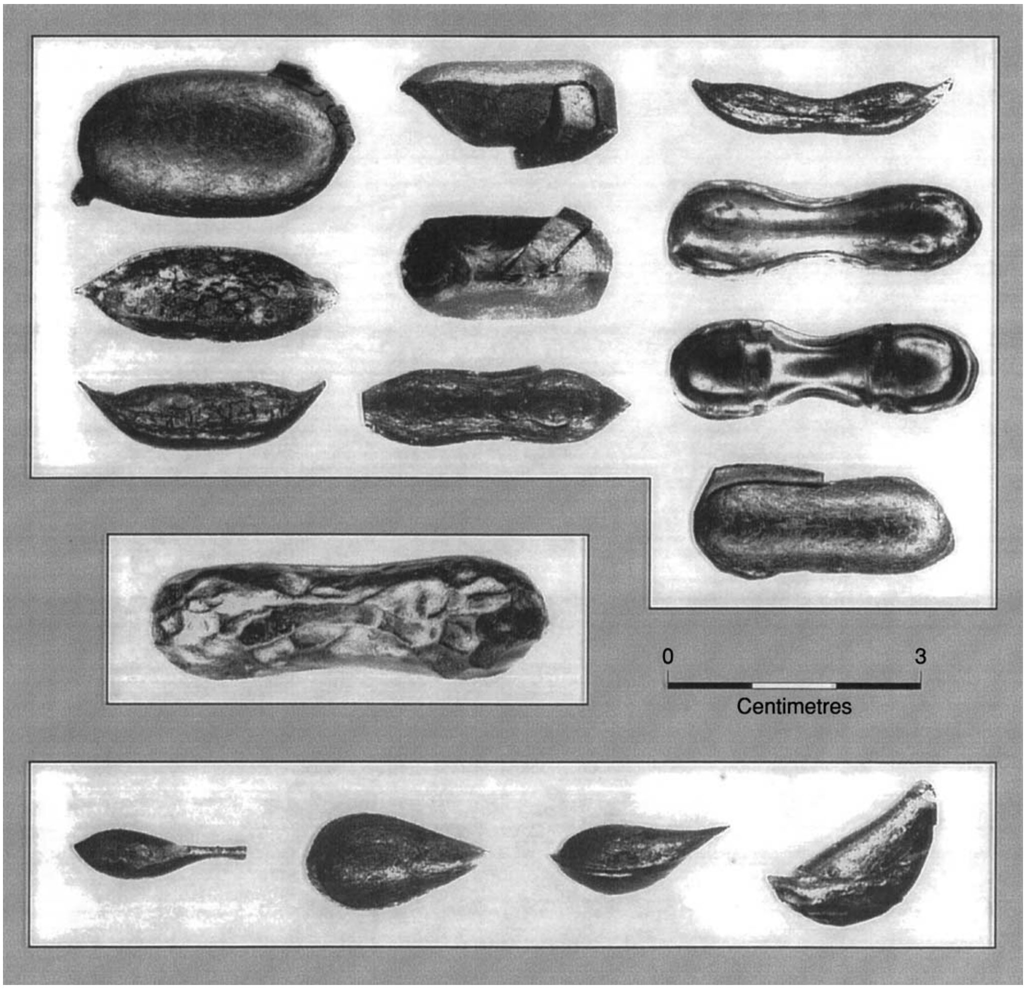


Figure 7 Boat-shaped, dumbbell, and teardrop australites from Menangina and Gindalbie Sheep Stations, Western Australia, showing flanges in various stages of detachment from the 'cores'. Photographs by WH Cleverly; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

dumbbells, teardrops, peardrops, and irregular forms) (Figure 10A–C). They occur in deep-sea cores, but the horizons are commonly not entirely sharply defined because of the action of seafloor scavengers. Microtektites do not normally occur in soils on land, because, even in the time-span in which the australites have littered Australia (from 0.77 to 0.78 Ma) on land, the action of groundwater would have dissolved them. The one known occurrence on land is in Eocene sediments in Barbados.

Tektite Composition, Experimental Data, and Theoretical Considerations

Tektites are silica-rich glasses, ranging from an average of 68% (Ivory Coast) to 80% (North American, georgiaites) silica. Refractive indices range from 1.48 to 1.51 and specific gravities from 2.3 to 2.5. Each strewn field has its peculiar range of values. The whole-rock analyses have a different character, as compared to volcanic glasses, and the glass is almost anhydrous, in contrast to the values for volcanic glasses. Microtektites show greater compositional variation than do the tektites found on land because they have a much smaller volume and can be formed of anomalous fractions of the glass, whereas the larger tektites average out. Lechatelierite, coesite, and stishovite are commonly present in tektites and microtektites, whereas they are unknown in volcanic glasses.

Wind-tunnel experiments in 1963 on tektite glass and gelatine by Chapman and Larsen in the United States reproduced exactly the flanged button form of ablated australite tektite found in Australia, and left no doubt concerning the mode of origin of the australites (Figures 11 and 12). Theorizing on the origin of tektites in 1998, H Jay Melosh concluded that

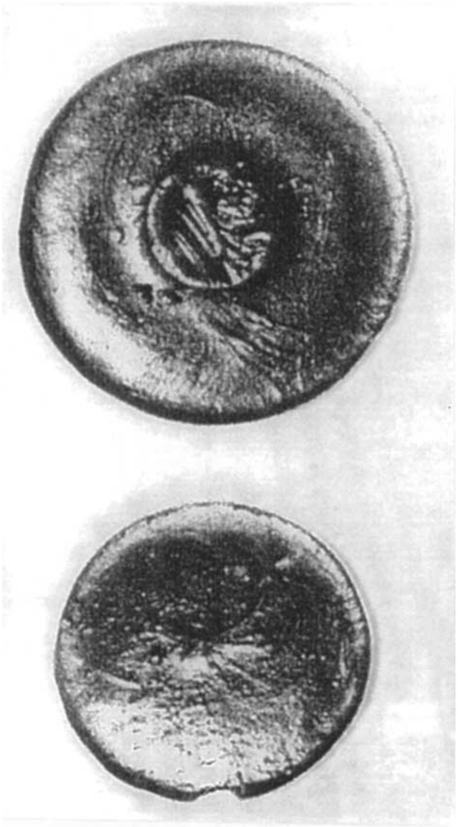


Figure 8 Ablated spherical australites that have lost all but a small relic of the splash-form sphere, due to flowage of melt on ablation to the enlarged flange. This type of tektite is found only at Port Campbell, Victoria at the extreme distal end of the strewn field. Photograph by G Baker; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

immense shock pressures of 100 GPa occurred during impact explosion processes accompanied by temperatures of up to 50 000 K. Thus arguments related to requirements for high temperatures in glass technology, which have been advanced to refute terrestrial impact generation of tektites, are invalid, because the temperatures involved in impact processes are many orders greater than had been thought.

Related Occurrences

Libyan Desert Glass and Mount Darwin Glass

Anomalous natural glass objects have been found in an area in the extreme south-west of Egypt. Irregular masses weighing up to 800 kg strew the desert in inter-dune corridors. These objects are highly siliceous, lechatelierite-bearing glass, quite unlike tektites in that they contain 98.2% silica. They seem to be ejected impactites formed of Nubian Sandstone, thrown out

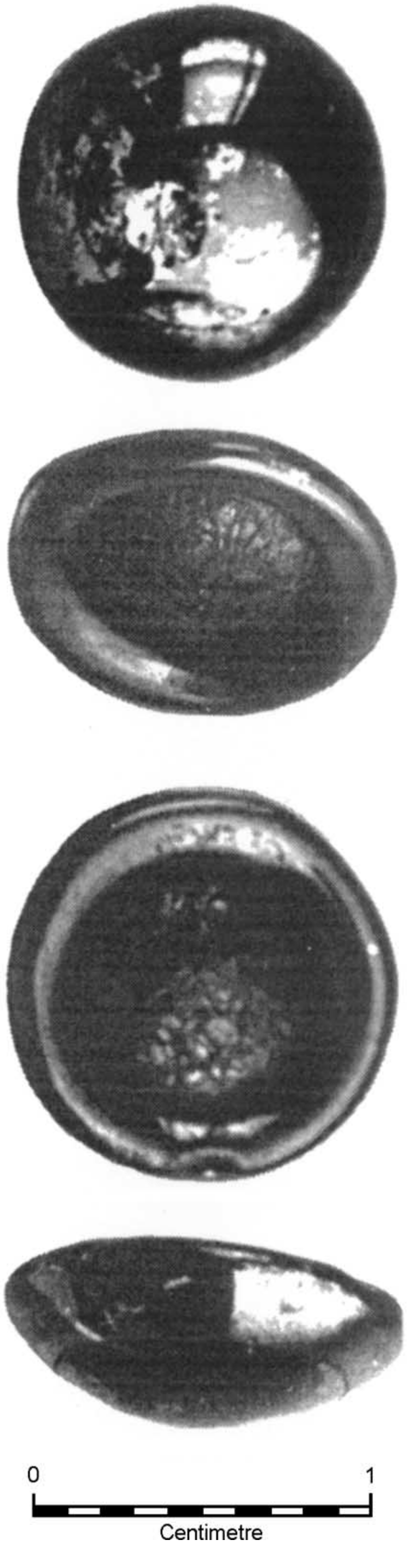


Figure 9 An ablated discoidal tektite with a weakly developed flange from a grab sample in the Central Indian Ocean. Photograph by BP Glass; reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky.* Bath: Geological Society Publishing House.

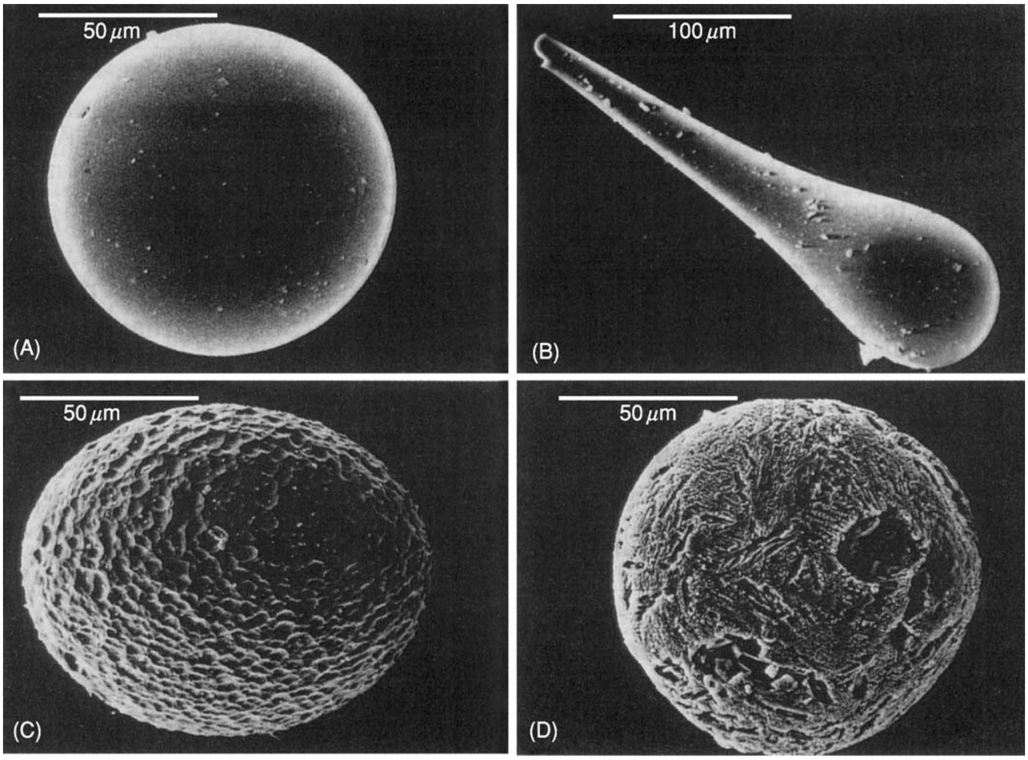


Figure 10 (A–C) Microtektites and (D) microcrystite (clinopyroxene microspherule), – all from Deep-Sea Drilling Project 689, Maud Rise, Weddell Sea. The microtektites are equated with the North American strewn field and the microcrystite is equated with the Popigai impact structure, northern Siberia. Reproduced with permission from Glass BP and Koeberl C (1999) Ocean Drilling Project Hole 689B spherules and Upper Eocene microtektite and clinopyroxene-bearing spherule strewn fields. *Science* 34: 197–208.

during the Oligocene 29 million years ago from two small craters (named BP and Oasis) that are situated about 100 km to the west. A minute fragment of an iron meteorite has been found with these glass objects. The glass was used to make Acheulian scrapers.

Mount Darwin Glass

Irregular masses of anomalous glass have found on the surface in western Tasmania and near Mt Macedon in Victoria. The glass mass is layered, not unlike Muong Nong-type tektites, but the silica content is much higher (88%, compared to 73%). Radiometric dating indicates formation 0.73 Ma, but the occurrence of these masses clearly has nothing to do with australites. A 1000-m-diameter crater has been recognized close to the occurrences of these glass objects in Tasmania, but there is no evidence of impact on excavation, though this site may be the source of the glass. The similar glass found at Mt Macedon, across the Bass Strait, 560 km to the north, is unexplained.

Zhamanshinites and Irghizites

Slags and glasses have been reported from the 13.5-km Zhamanshin impact structure in Kazakhstan, north of the Aral Sea (see Figure 2), overlying Palaeogene country rock. The best radiometric age derived dates these objects to 1.09 Ma. Zhamanshinites contain rock fragments and are impactite glasses. Irghizites occur within the bounds of the crater structure, not in an external strewn field, and are composed of small ‘micro-irghizite’ particles, welded together. The silica content of the irghizites is 72–79%, not unlike that of tektites, but the water content is slightly higher. The interest in these objects is that they may represent the separation of microtektites at source.

Urengoites

Three fragments of tektite-like glass have been found buried in Siberia at two sites 40 km apart (see Figure 2); based on radiometric dating, these urengoites were formed at 22–24 Ma.

The Eltanin Glasses

The 25-km-wide Pliocene Eltanin structure on the floor of the southern Pacific Ocean (see Figure 2) has both associated minute fragments of a mesosiderite (stony iron) meteorite and microscopic glass spherules; these objects have been recovered in piston

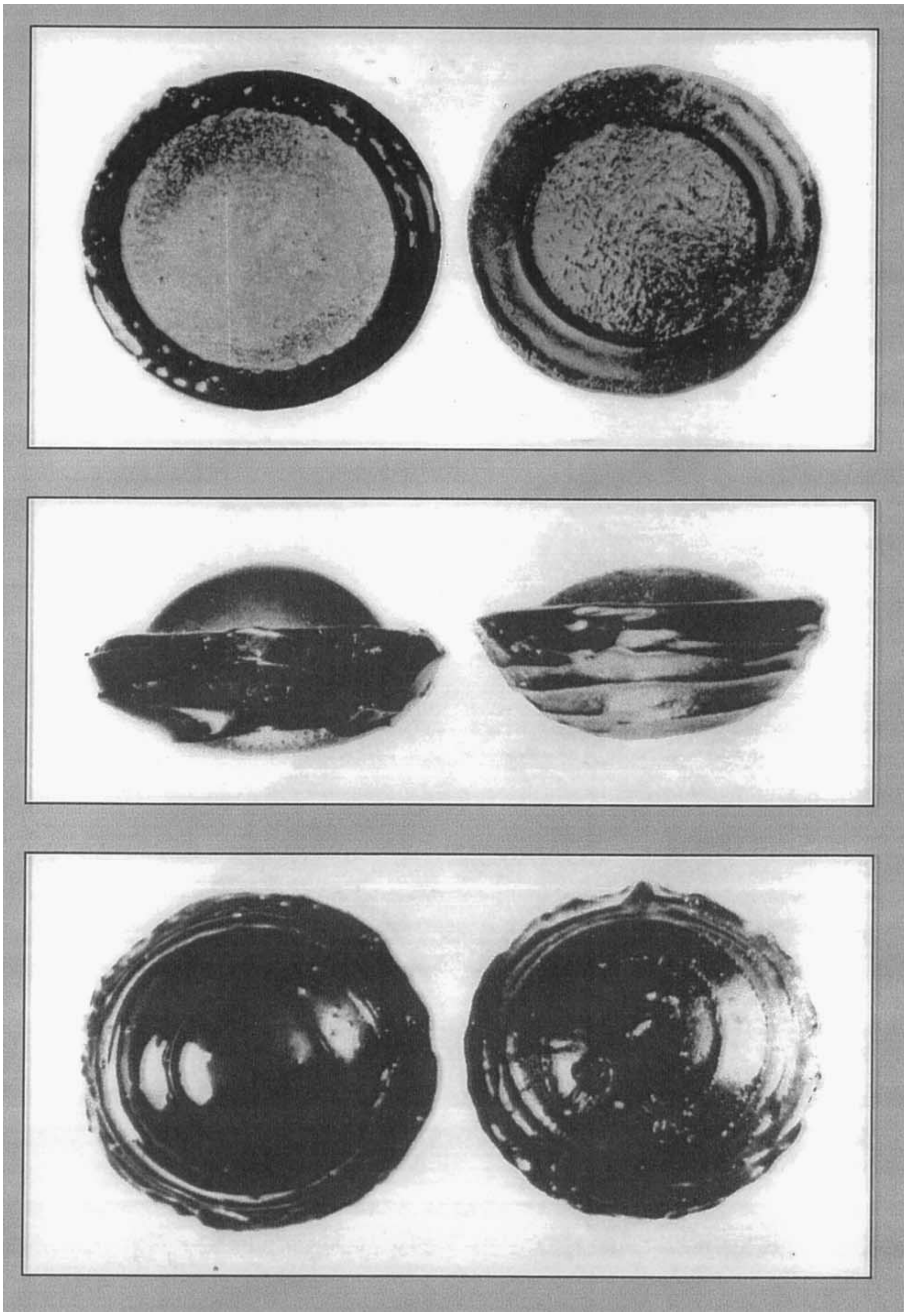


Figure 11 Comparison between a flanged australite from Port Campbell, Victoria (right) and an artificial product of a wind tunnel experiment (left) by DR Chapman and HK Larsen on tektite glass (NASA photograph, 1963). Reproduced with permission from McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

cores close to the impact melt layer associated with a chaotic formation within the structure. The forms resemble microtektites in size and shape (spheres, teardrops), but the chemistry is quite different (silica average, 45%).

Late Eocene Microspherules

A microspherule layer has been recognized just below the North American microtektite layer in Caribbean cores and in the Weddell Sea core. These

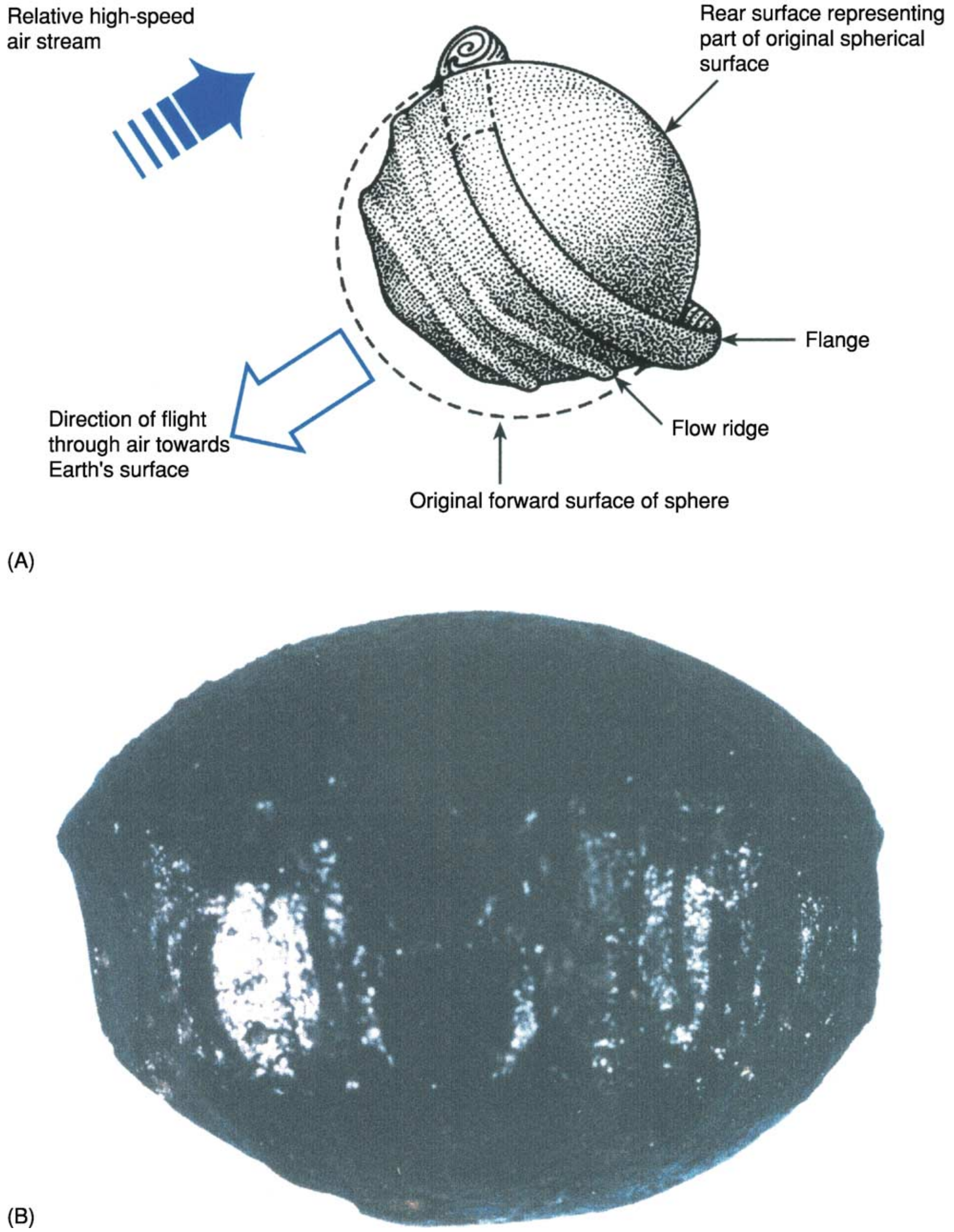

Relative high-speed air stream
Rear surface representing part of original spherical surface
Flange
Flow ridge
Direction of flight through air towards Earth's surface
Original forward surface of sphere
(A)
(B)

Figure 12 (A) Development of a spherical splash-form australite during atmospheric ablation on descent. (B) The common 'core' form produced on separation of the flange, showing equatorial chatter marks where it has separated. Reproduced with permission from McNamara K and Bevan A (2001) *Tektites*. Perth: Western Australian Museum.

spherules differ from the Eltanin glasses in that they show a palimpsest of microcrysts, as well as glass (Figure 10D), and the chemical composition is much more varied. The mineral in the microcrysts is sometimes preserved and X-ray diffraction has shown it to be pyroxene. This layer has been equated with the Popigai impact structure in North Russia and the impact layer reported from Massignano, Italy, and other European sites. The average silica content of 64% is lower than in any tektites and matches closely the Popigai impactite glasses.

K–T Boundary Glass Bodies

Carbonate sediments at the K–T boundary in Haiti (see Figure 2) contain tektite-like bodies of glass enveloped in smectite; the glass objects have the usual splash forms (spheres, ellipsoids, teardrops, and elongate and dumbbell shapes). The glass is vesicular and mostly crystal free. Similar bodies are reported from clastic sediments at Beloc, Mexico (Figure 2), where foraminifera indicate the K–T boundary (*see* **Mesozoic:** End Cretaceous Extinctions). Spheres and dumbbells are represented. In both cases, removal of smectite coating reveals the sculpturing seen on the surface of microtektites. Compositions are very different compared to those of tektites or microtektites: silica contents are about 63%, similar to melt rocks of andesitic composition revealed at the favoured source impact site, the 170-km-diameter Chicxulub Crater impact structure on land and offshore of Yucatan, Mexico. The contrast to the tektite composition is consistent with the country rock makeup at Chicxulub, and these glass bodies, though not fitting strictly the common definition of tektites, are clearly a form of tektite *sensu lato*. Microspherules are reported from the K–T boundary at Petruccio, Italy, and there are many examples of shocked quartz with planar structures at the same boundary in the United States (*see* **Impact Structures**).

Late Devonian Glass Bodies

Microspherules have been reported from the Senzeilles Shale, Belgium, a quiet-water deposit of Late Devonian age and very close above the Frasnian–Famennian stage boundary, the site of another, smaller scale extinction. The horizon is at the top of the *Palmatolepis triangularis* conodont zone, at the bottom of which is the extinction boundary. The spherules, up to 1 mm in diameter, are wholly of glass; they have been preserved from devitrification perhaps because of the anhydrous nature of the glass. Most of the bodies are spheres, but elongate, teardrop, and dumbbell shapes are present. The silica content varies from 38 to 80% and the chemical composition is variable. There are several possible source impact structures, though the Siljan, Sweden, 55-km structure is favoured.

Similar spherules occur at Qidong in Hunan Province, South China, in the *Palmatolepsis crepida* conodont zone of the Famennian. They reach up to 0.160 mm in diameter and are mostly spheres, though teardrop and peardrop forms are recorded. Lechatelierite is present in them. The chemistry is variable and the silica content ranges from 62 to 99%. They have been tentatively related to a possible impact structure at Taihu, south-west of Shanghai, but little is known of this structure. As for the Senzeilles Shale microspherules, these spherules do not correspond exactly to the extinction horizon.

Archaean Spherules

Microspherules in which no glass is preserved are known from the Wittenoom Formation in Western Australia (dating to 2500–2600 Ma) and also from rocks of the Onverwacht and Fig Tree Groups (3100–3500 Ma) near Barberton, South Africa. These microspherules have been attributed to impact processes and are related to microtektites.

Research Directions

Many state-of-the art techniques have been applied to studies of tektites, resolving many of the early questions about these objects. Geochemical, geophysical, isotopic, and statistical analyses that have been applied to tektites in the past 50 years have resolved questions concerning their source, the manner of formation of their primary and secondary flanged shapes, and the age of the four events forming the strewn fields in which they are found. The connection between tektites and microtektites has also been established. However, there are still questions that have not been fully resolved. Further research may reveal the process of melting at the target, and ejection from the target; the reason for the restricted nature of the target rocks involved in generating tektites; the reason for the restriction of tektite associations to a handful to the ~170 terrestrial megaimpact sites known; the exact relationship of tektites to microtektites in the processes occurring at the target, in flight, and during transport and dispersion to the strewn field; and the source of the Australasian strewn field.

See Also

Analytical Methods: Fission Track Analysis; Geochronological Techniques. **Gemstones**. **Igneous Rocks:** Obsidian. **Impact Structures**. **Mesozoic:** End Cretaceous Extinctions. **Shock Metamorphism**. **Solar System:** Asteroids, Comets and Space Dust; Meteorites; Moon. **Ultra High Pressure Metamorphism**.

Further Reading

Barnes VE and Barnes MA (eds.) (1973) *Tektites*. Stroudsburg, PA: Dowden, Hutchinson and Ross.

Chapman DR and Larsen HK (1963) On the lunar origin of tektites. *Journal of Geophysical Research* 64: 4305–4368.

Glass BP (1968) Glassy objects (microtektites) from deep-sea sediments off the Ivory Coast. *Science* 161: 861–862.

Glass BP (1990) Tektites and microtektites: key facts and inferences. *Tectonophysics* 171: 393–404.

Glass BP and Koeberl C (1999) Ocean Drilling Project Hole 689B spherules and Upper Eocene microtektite and clinopyroxene-bearing spherule strewn fields. *Science* 34: 197–208.

Glass BP, Chapman DR, and Prasad S (1996) Ablated tektite from the central Indian Ocean. *Meteoritics and Planetary Science* 31: 365–369.

Koeberl C (1986) Geochemistry of tektites and impact glasses: an overview. *Annual Reviews of Earth and Planetary Sciences* 14: 325–350.

Koeberl C (1987) *Geochemistry of Muong Nong-Type Tektites: A Review,* Proceedings of the 2nd International Conference on Natural Glass, Prague, pp. 371–377.

Koeberl C and Shirey MB (1993) Detection of a meteorite component in Ivory Coast tektites with rhenium–osmium isotopes. *Science* 261: 595–598.

McCall GJH (2000) The age paradox revisited. *Journal of the Royal Society of Western Australia* 83: 83–92.

McCall GJH (2001) *Tektites in the Geological Record: Showers of Glass from the Sky*. Bath: Geological Society Publishing House.

McNamara K and Bevan A (2001) *Tektites*. Perth: Western Australian Museum.

Melosh HJ (1998) Impact physics constraints on the origin of tektites. *Meteoritics and Planetary Science* 33(Supplement): A104.

O'Keefe J (1963) *Tektites*. Chicago: University of Chicago Press.

Taylor SR (1962) The geochemical composition of australites. *Geochimica et Cosmochimica Acta* 26: 685–722.

Taylor SR (1969) Criteria for the source of australites. *Chemical Geology* 4: 451–459.

TERRANES OVERVIEW

L R M Cocks, The Natural History Museum, London, UK

Introduction

The word 'terrane' is used in a specialised sense by geologists, and should not be confused with the same-sounding 'terrain', which is used by many people, particularly the military, to denote characteristics of the countryside in a particular area. To a geologist, terrane is used for a discrete block of continental crust that is moving or has moved in relation to those blocks that surround it.

Definition

The Earth is today, and through geological time, made up of a number of moving plates (*see* **Plate Tectonics**). Each plate consists of heavier oceanic crust underlying lighter continental crust. Plates are constantly being enlarged through ocean-floor spreading, reduced by subduction or obduction, or displaced laterally by transform faulting, all of which processes affect both oceanic and continental crusts. However, because of its lighter density, continental crust tends to remain at the Earth's surface for far longer periods than does oceanic crust; consequently, very often the continental crust of an old plate remains at the surface today long after the oceanic crust on which it once rested has disappeared within Earth's interior, perhaps to be later remobilized into fresh crust. The oldest ocean crust known today in its original position is only about 160 million years old (Jurassic), whereas the continental crust includes rocks from modern times to more than 3 billion years ago, the oldest known. Terranes can be of varied size, ranging today from the vast Eurasian–African block down to the relatively small microplates found in the south-west Pacific within the East Indies. The difference between a 'continent' (as strictly defined) and a 'terrane' is that the former is invariably bounded by one or more oceans, whereas the latter is defined by its surrounding structural discontinuities. An accreted terrane is one that has been added to the margin of a larger one. Many areas may or may not have been real (i.e., separate) terranes in the past, and geological opinions can often differ widely as to their reality and status. When this uncertainty exists, the area is referred to as a 'suspect terrane'.

Boundaries of Terranes

The marginal boundaries of old terranes are termed 'sutures': when exposed, they are usually faults or fault systems, with the obvious characteristic that the rocks and stratigraphy are completely different on the opposite sides of the faults. Because movement of the crust is principally dominated by horizontal components, the suture faults are usually strike-slip or transform. For example, a major suture is the Tornquist–Teisseyre Line, or the Trans-European Suture Zone (TESZ), which stretches from the North Sea to the east of Aberdeen, through southern Denmark, north-eastern Germany, south-central Poland, Slovakia, Hungary, and Romania to the Black Sea. That suture represents what remains of the south-eastern margin of the old terrane of Baltica and separates that terrane from Avalonia (see later), Perunica (often termed Bohemia), and others to its south. The suture was originally formed during the Variscan Orogeny in Late Palaeozoic time, but movements along the TESZ area of crustal weakness have been reactivated during several subsequent geological periods and continue sporadically to the present day.

Principal Terranes

At two recognizable times in Earth history, at about 1000 Ma and 250 Ma, most of the continental crust was together, forming vast supercontinents named Rodinia (*see* **Precambrian:** Overview) and Pangaea (*see* **Pangaea**), respectively. Prior to the aggregation of Rodinia, little is known of the preceding terranes, and so they are characterised and named only by the major earlier Precambrian shield areas, such as the Canadian Shield, and their positions relative to each other are currently poorly constrained and open to much scientific debate. However, after the breakup of Rodinia, which was well under away by 850 Ma, larger terranes have separate names, and the principal terranes were identified as follows:

- Gondwana. Easily the largest terrane (*see* **Gondwanaland and Gondwana**), comprising South America, Africa, India, Antarctica, and Australia, as well as a number of peripheral areas that formed parts of this huge terrane at different times.
- Siberia (otherwise known as Angara). An area that included only part of the modern political area of Siberia, but that was nevertheless very substantial.
- Laurentia. Most of North America and Greenland, and then adjacent areas, including Spitzbergen, northern Ireland, and Scotland. It was separated from Baltica and Gondwana by the Iapetus Ocean in the Lower Palaeozoic.
- Baltica. The northern part of mainland Europe eastward to the Ural Mountains and northward to include Novaya Zemlya and Franz Joseph Land in the Arctic. It was separated from Avalonia by the Tornquist Ocean and from Siberia by the Aegir Sea during the Lower Palaeozoic.
- Avalonia. An area including the western coast of the United States, the Maritime Provinces of Canada, Newfoundland, southern Ireland, Wales, England, and Belgium, which formed part of Gondwana until the Early Ordovician (about 490 Ma). It was a separate terrane only in the Ordovician. As Avalonia left Gondwana, the widening ocean between it and Gondwana is termed the Rheic Ocean.
- Laurussia. The terrane formed by the amalgamation of Laurentia, Baltica, and Avalonia during the Silurian, and which continued until the formation of Pangaea in the Late Palaeozoic. During the Upper Palaeozoic, Armorica, Perunica, the Rheno-Hercynian Terrane, and others drifted across the Rheic Ocean to become accreted to Laurussia, leaving a widening Neotethys Ocean behind them.
- Armorica-Iberia. This is sometimes termed the Armorican Terrane Assemblage (ATA); it consists of most of the western part of southern Europe, including Spain, Portugal, France, Sardinia, and parts of Germany. Some regard Perunica (Bohemia) as part of the ATA, but it moved independently of the ATA after its separation from Gondwana in the Ordovician.
- North China. The southern part of Siberia and the Korean Peninsula as well as northern China.
- South China. Most of southern China.
- Annamia. The Indochina Peninsula and adjacent areas.
- Sibumasu. The area running from eastern Burma (Myanmar) through Thailand, south-western China, and western Malaysia to Sumatra.

In addition to these named terranes, more than 50 additional terranes of variable size have been identified and named as existing during the Palaeozoic, as well as numerous discrete geological entities such as island arcs, which were independent units for differing geological times. Some of the many terranes that make up Eurasia today are shown in Figure 1. After the mutual accretion of the various terranes during the Palaeozoic, progressively forming Pangaea during the Upper Palaeozoic, the process of splitting and disintegration of Pangaea began in the Early Mesozoic to form the larger terranes that are known by their modern continental names today. The large area eastward of Pangaea was occupied first by the Neotethys Ocean and subsequently by the Tethys Ocean. The Mediterranean Sea can be considered as a remnant of the Tethys Ocean today. However, there are in addition many much smaller terranes; for example, it has been suggested that there are as many as 100 Early Mesozoic terranes in the collage that makes up the present-day Cordillera of North America.

Identifying the Positions of Old Terranes

To reconstruct the geography of Earth at different times in the geological past, it is necessary to locate the former position of the different terranes, and to understand how each moved with time. The ways of doing this using current knowledge are through evidence of ocean-floor magnetic stripes, movement over hotspots, palaeomagnetism, faunal provinces, distribution of sediments, and positioning of tectonic belts.

Ocean-floor magnetic stripes are studied by mapping out the modern ocean floor and its magnetic anomaly stripes, dating the stripes, and then progressively removing them so that it can be seen how the oceans have widened with time (*see* **Palaeomagnetism**). Magnetic field reversals and related issues such as the age of onset, the duration, and the frequency of superchrons (long periods of constant magnetic polarity) are now reasonably documented to the beginning of the Cretaceous. This is the only objective

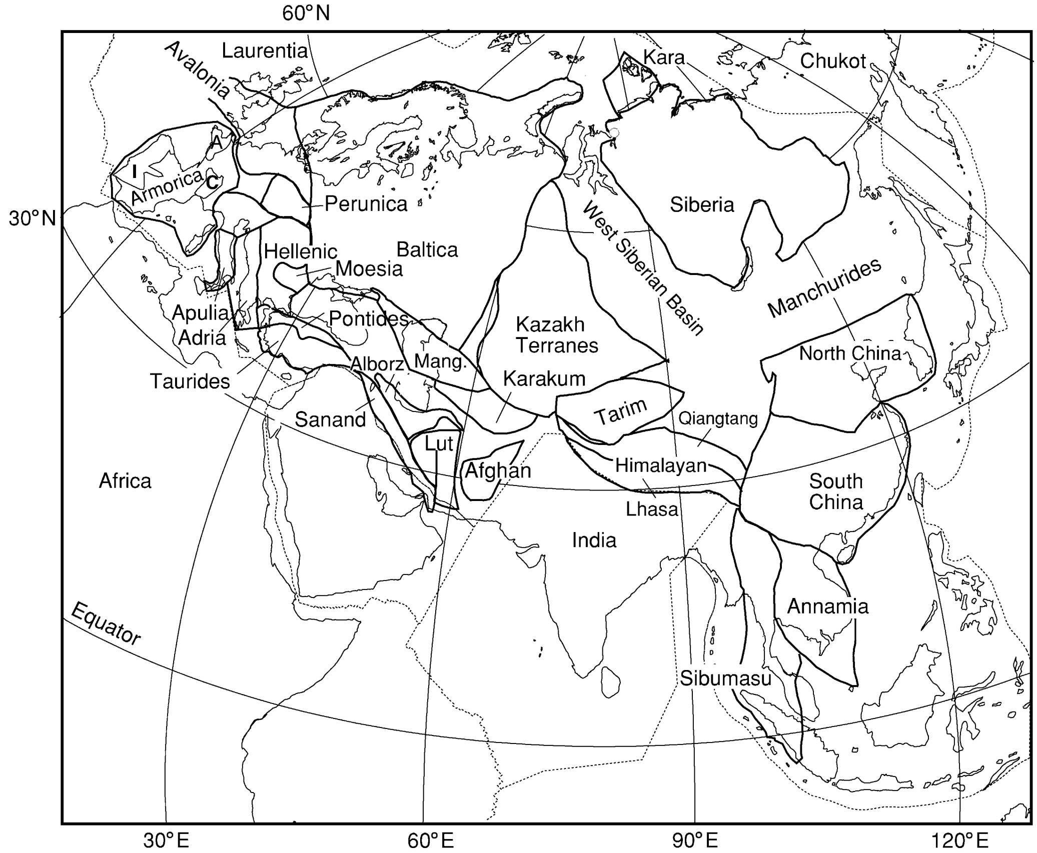


Figure 1 The boundaries of the major Palaeozoic terranes that have united to make up Eurasia today. C, Central France; I, Iberia; Mang., Mangyshlak Terrane. The large areas labelled West Siberian Basin and Manchurides are those occupied largely by continental crust that is post-Palaeozoic in origin, as is the large area of south-east Asia south-eastwards of the Sibumasu and Annamia terranes. Thin dotted lines are modern plate boundaries. Modified by Trond Torsvik, Trondheim, from Torsvik and Cocks (2004).

method of discovering the previous position of terranes through time. Unfortunately, however, because the oldest ocean-floor crust known is of Jurassic age (and most of Earth's oceans floors are very much younger than that), this method is available only for determining terrane positioning in Tertiary and Late Mesozoic times.

Some terranes can be seen to have moved over hotspots with time, which thus gives progressive and definitive positioning of the terranes. This is unique in providing objective (as opposed to subjective) longitudinal control of terrane movement. Unfortunately, as with the plotting of the magnetic stripes, this method applies only to those few terranes that are drifting over active hotspots, and no consensus on results has been obtained for positions much older than the Tertiary. The only Mesozoic data are for the Tristan da Cunha and Great Meteor hotspots in the South Atlantic, which are traceable back to 130 Ma.

Palaeomagnetism occurs when an igneous rock is emplaced, then the magnetic (largely iron) constituents within it cool with their magnetic direction pointing towards the pole. This remnant magnetism is set after cooling, and so the study of an ancient igneous rock indicates the pole position at the time of deposition of the rock. Thus two things can be calculated: the palaeolatitude of the rock at the time of cooling and the subsequent rotation of the terrane. This is the best method of calculating the position of old terranes. There are, however, two drawbacks: first, there is no determination of the palaeolongitude of the terranes, and second, a great many igneous rocks have their original palaeomagnetism completely reset by subsequent tectonic events that involved enough heating to reset the magnetisation of the rock.

If the palaeoecology and the age of the fossils contained within an individual terrane are known, then some marine benthic or terrestrial fossils are found to be specific to one or more terranes, and quite different from those in other terranes that may be close to it today. These differences are often recognised as faunal provinces. The data from these fossil distributions provide terrane affiliations and positions that are reached completely independently of palaeomagnetic methods, and the two methods have been used effectively together to discover where most of the terranes lay in Cambrian to Jurassic times. In addition, bioherms such as coral reefs are usually restricted to within 30° north and south of the equator, ancient or modern. Coals are most commonly found in two belts occurring in low temperate latitudes.

Distribution of sediments may be examined to locate the former position of the different terranes. Most clastic rocks give little indication of the climates within which they were deposited, but carbonates increase in abundance from high to low latitudes as the average surface temperature increases. Evaporites are seldom equatorial but are most common in two bands centring at about 20° north and south of the equator or palaeoequator. Glaciogenic sediments, such as tillites, are almost invariably found at high latitudes.

Positioning of tectonic belts may provide evidence of the position of old terranes. Some substantial mountain belts can be traced from one terrane to the adjoining one, but, in the past, several such correlations made by geologists have been shown to be incorrect. However, in the Precambrian, when there were no terrane-diagnostic fossils, mountain belts have proved to be the only indicators apart from palaeomagnetism.

In positioning old terranes, a key underlying precept must be kinematic continuity. This means that it must be remembered that terranes never leapt around the globe like sodium on water. Thus, if a certain terrane (for example, Baltica) appears to have been in one place, then 2000 km away 10 million years later, and then close to its original position 10 million years after that, it is probable that one or more of those postulated positions are not correct!

Ever since the acceptance of modern plate tectonic theory in the mid-1960s, geologists have realised that plates and their terranes have not been in a single position during geological time. Tuzo Wilson, in 1966, suggested that there was a substantially different terrane pattern and therefore geography before the supercontinent of Pangaea came together in the Late Palaeozoic. Since Wilson's observations,

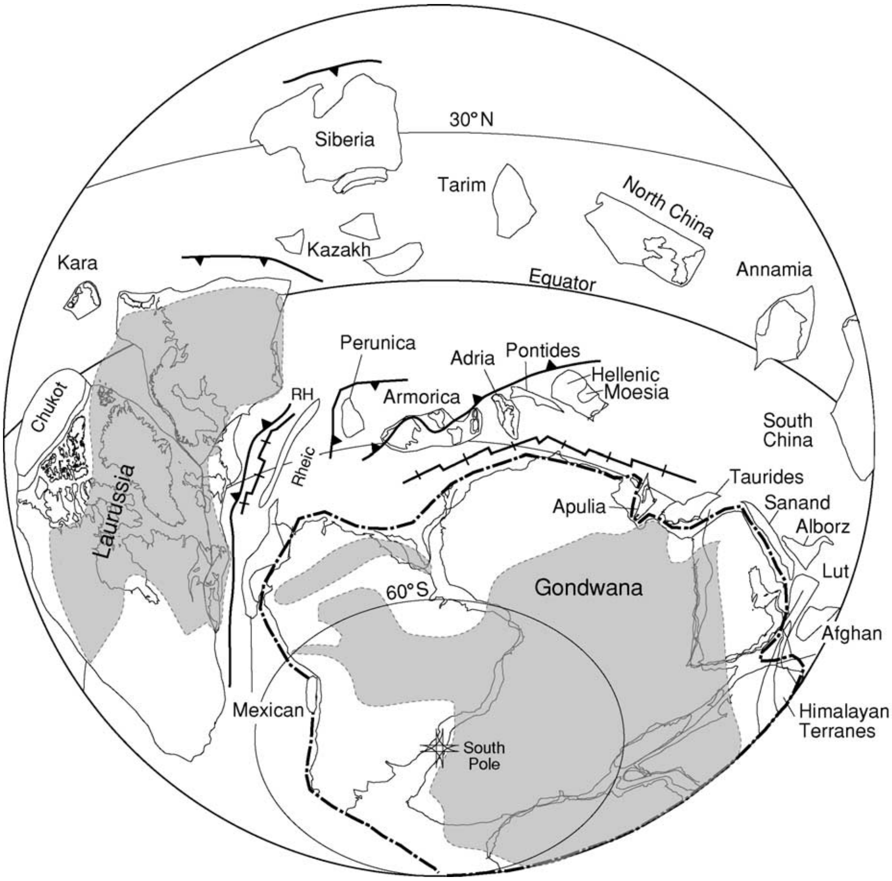


Figure 2 The major terranes and Earth geography 400 Ma (the Early Devonian), assuming that Earth's magnetic field was a simple geocentric axial dipole. The dot-dash line represents the margin of Gondwana. RH, Rheno-Hercynian Terrane. Modified by Trond Torsvik, Trondheim, from Torsvik and Cocks (2004).

there have been many published models of where the various terranes lay through geological time. There is now considerable agreement on the identity, positions, and progression, through the Palaeozoic, of terranes surrounding what is today the North Atlantic area, around which the majority of academic geologists work. However, the many terranes that make up Central and South America, Africa, Asia, and Australasia are, in many cases, rather poorly defined and recognised, and their relative positioning through the Phanerozoic (let alone the Precambrian) is a matter for unresolved debate and geological argument. Figure 2 shows a possible terrane reconstruction for half the globe at about 400 Ma (the Early Devonian), at a time when Laurentia, Baltica, and Avalonia had fused to form Laurussia and when various terranes such as Armorica, Adria, the Pontides of Turkey, and the Hellenic Terrane (including Moesia) had all left the Gondwana superterrane following the opening of the Palaeotethys Ocean to their south. The other half of the globe, not shown in Figure 2, was largely occupied by the vast Panthalassic Ocean.

See Also

Gondwanaland and Gondwana. **Palaeomagnetism**. **Pangaea**. **Precambrian:** Overview. **Volcanoes**.

Further Reading

Cocks LRM and Torsvik TH (2002) Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. *Journal of the Geological Society, London* 159: 631–644.

Leitch EC and Scheibner G (eds.) (1987) *Terrane Accretion and Orogenic Belts*. Geodynamics Series 19. Washington DC: American Geophysical Union.

Stampfli GM and Borel GD (2002) A plate tectonic model for the Paleozoic and Mesozoic constrained by dynamic plate boundaries and restored synthetic oceanic isochrons. *Earth and Planetary Science Letters* 196: 17–33.

Torsvik TH and Cocks LRM (2004) Earth geography from 400 to 250 million years ago: a palaeomagnetic, faunal and sedimentological review. *Journal of the Geological Society, London* 161: 348–361.

Windley BF (1995) *The Evolving Continents,* 3rd edn. New York: John Wiley.

TERTIARY TO PRESENT

Contents

Paleocene

J J Hooker, The Natural History Museum, London, UK

Introduction

The Paleocene Epoch/Series is the first of the Cenozoic Era/Erathem. It is the first of five epochs in the Tertiary Period and the first of three in the Paleogene, which is treated either as a period in its own right or as a subdivision of the Tertiary. The Paleocene succeeds the Cretaceous Period/System and precedes the Eocene Epoch. The Paleocene lasted nearly 10 million years, from 65.5 till 55.8 Ma, and is divided approximately equally into three ages/stages (in order of decreasing age): the Danian, the Selandian, and the Thanetian (**Figure 1**). The naming of the Paleocene follows the earlier procedure of adding a prefix denoting degree of antiquity or modernity, in this case 'paleo', from the Greek *palaios*, meaning 'ancient', and 'cene', from the Greek *kainos*, meaning 'recent'. The Paleocene was the last of the Cenozoic epochs to be named, originally being proposed by Schimper in 1874. Schimper was a palaeobotanist and, in contrast to the definitions of the other Cenozoic epochs,

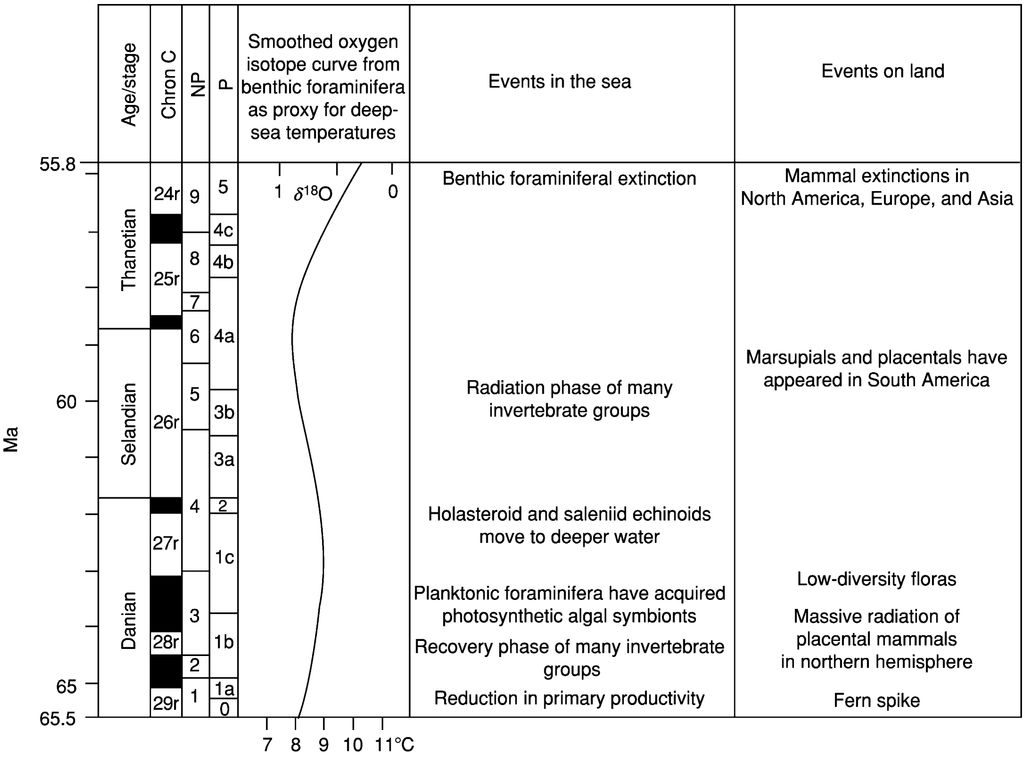


Figure 1 Time chart of the Paleocene, showing how it is divided up by ages/stages, magnetochrons (Chron C), global calcareous nannoplankton (NP), and planktonic foraminiferal (P) biozones. Also shown are an isotope proxied temperature curve and the main biotic and physical events in the sea and on land. Magnetochrons are divided into normal (black) and reversed (white) intervals; the normals, often composite, are younger than the reversals (r) that bear the same number. Data for the isotope curve from Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686–693.

he based his concept of the Paleocene on floras. Thus, he noted the presence in the Paris Basin of distinctive assemblages of plants that occur in strata overlying the Cretaceous and underlying the Eocene.

Unfortunately, some of the strata Schimper regarded as Paleocene are now known to be contemporaneous with those he regarded as Eocene. Partly as a consequence of this confusion, the Paleocene nomenclature did not gain worldwide acceptance until nearly a century after Schimper's work. The Paleocene was largely ignored by marine workers, but since the early years of the twentieth century, it has been championed by vertebrate palaeontologists, especially in North America, where it was recognized to be characterized by highly distinctive faunas of land mammals (Figure 2). Over the decades, the concept of its time-span has changed considerably and has also meant different things to palaeontologists in different fields. In Schimper's time, the first Paleocene age/stage, the Danian, was regarded as latest Cretaceous, mainly because of the continuation in northwest Europe of the typically Cretaceous chalk facies (Figure 3). The Danian was moved into the Paleocene when it was realized that it postdated a major extinction event in the sea and on land, which is now taken to mark the end of the Cretaceous Period and indeed of the Mesozoic Era. The official Global Stratotype Section and Point (GSSP) for the beginning of the Paleocene is at El Kef, Tunisia. The end of the Paleocene was also the subject of much discussion until a decade ago, when a sharp anomaly in the carbon isotope curve, known as the carbon isotope excursion (CIE), was recognized in a borehole core from the Weddell Sea, Antarctica. This has been interpreted as a climatic warming event, known as the Paleocene–Eocene thermal maximum (PETM). The PETM had profound and widespread effects on marine benthos and land mammal faunas, and is now accepted to mark the boundary between the Paleocene and Eocene epochs.

Plate Tectonics and Other Physical Phenomena

In many ways, the Paleocene represents a continuation of processes begun in the Cretaceous. Thus, Pangaea continued to fragment. Falling sea-levels since the beginning of the Late Cretaceous appear to

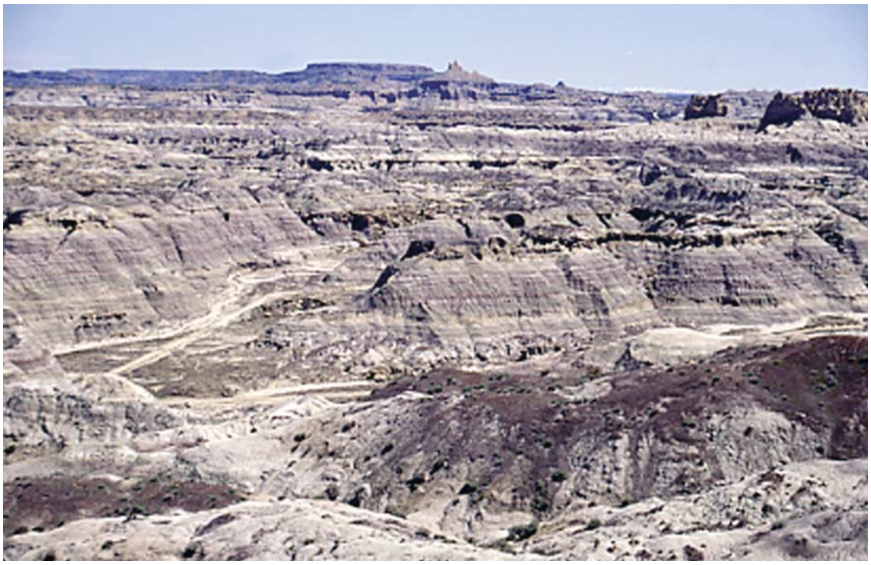


Figure 2 Kutz Canyon area, San Juan Basin, New Mexico, USA., showing the thick sequence of non-marine fluvial mudstones, spanning much of the Paleocene. This is the basin that originally yielded distinctive Paleocene mammal faunas.

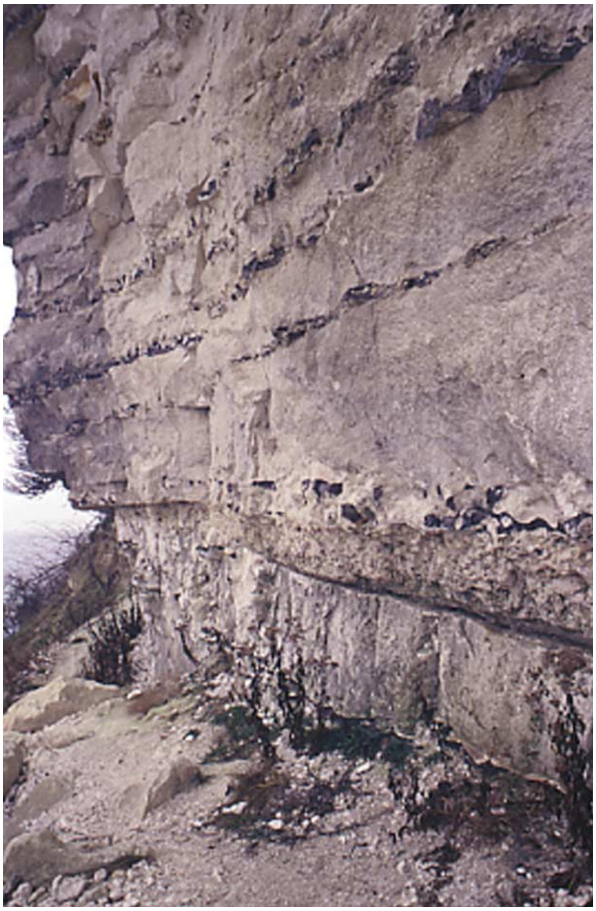


Figure 3 Stevns Klint, Denmark, showing marine Danian chalk facies, with basal fish clay unit, resting on Maastrichtian (latest Cretaceous) chalk.

have reached a threshold in the Paleocene, draining the very extensive Late Cretaceous marine carbonate platforms and dramatically increasing the non-marine area of the globe. A phase of the North American Laramide Orogeny also took place during the epoch. This involved continued sliding of the Farallon Plate under the western margin of the North American continent to produce early phases of the Rocky Mountains uplift. As a result, the North American Western Interior Seaway, which in the Late Cretaceous had split the continent from north to south, almost completely filled with sediment shed from the rising Rocky Mountains. The remnant Paleocene elongate inlet that had only a southerly opening is known as the Cannonball Sea; and the series of basins that formed in the west of the area were non-marine. The marine area of the Gulf Coast and Florida continued to subside.

The Farallon Plate also continued to be subducted beneath the western margin of South America, resulting in early phases of the Andes uplift. North America and South America were slowly moving away from each other at this time, the intervening ocean allowing deep-water circulation between the Pacific and Atlantic. Subduction in the Pacific continued on the eastern side (Kula Plate) beneath Kamchatka and Sakhalin. A land bridge across the Bering Straits was also intermittently developed between North America and Asia in response to fluctuating sea-levels.

On the other side of the world, the Atlantic Basin was continuing to spread. This particularly involved an extension of rifting northwards between Labrador and West Greenland and between East Greenland and Europe. A hotspot formed beneath Greenland, producing outpourings of lava that intensified at the end of the epoch as the East Greenland–Europe area drifted over the hotspot. Between Europe and Asia, the epicontinental West Siberian Sea, although more restricted than in either earlier or later times, extended southwards from the Arctic Ocean, reaching the northeastern part of the Tethys Ocean (known as the Peritethys) separating Asia from Europe near the end of the epoch.

Africa moved and rotated north, pushing Apulia (comprising Italy, the former Yugoslavia, and western Greece) towards the main European craton, and producing the initial phases of the Alpine Orogeny. Eastward extension of this structural belt through Asia Minor and southern Iran partially isolated the Peritethys. India was an island continent still moving north towards Asia. The massive outpourings of hotspot-related basaltic lava that occurred in the Deccan region of West India during the latest Cretaceous continued for a brief interval in the earliest Paleocene. The Aluk Plate in the South Pacific continued to be subducted beneath the western edge of the Antarctic Peninsula. Rifting occurred between Antarctica and Australasia, but the two continents did not separate. In fact, during the Paleocene, the three major elements of Gondwana (South America, Antarctica, and Australasia) remained in contact. At the opposite pole, the Arctic Ocean was, for most of the epoch, an enclosed water body, separated by land from the rest of the world's oceans.

Biota

After the End-Cretaceous extinctions (*see* **Mesozoic:** End Cretaceous Extinctions), the earliest Paleocene biota was notable for the absence of such major and formerly diverse groups as ammonites, belemnites, rudists, plesiosaurs, mosasaurs, and non-avian dinosaurs, as well as for the low abundance and diversity of brachiopods, bivalve and gastropod molluscs, and marine reptiles. There is generally low abundance and diversity of marine life at the beginning of the Paleocene. In fact, some groups (both in the sea and on land) show a low-diversity recovery phase followed by radiation. However, the pattern is different for other groups of organisms. Key biotic events are discussed below.

Marine Realm

Calcareous nannoplankton Few members of the calcareous nannoplankton, a group of microscopic calcifying algae (coccolithophores and their possible relatives), survived the end of the Cretaceous. Paleocene nannofloras comprise these relict survivors plus an array of rapidly radiating new taxa (24 genera in the course of the epoch). By the end of the Paleocene, another major turnover resulted in over half of the Cretaceous relict species, and nearly a third of the newly evolved genera, becoming extinct. The rapidity of this evolutionary turnover and the widespread occurrence of these fossils in marine strata have resulted in the establishment of nine globally recognized Paleocene biozones (Figure 1).

Dinoflagellates The dinoflagellates, a group of cyst-forming unicellular algae, exhibit a stepwise origination pattern in the Early Paleocene. Nevertheless, rapid evolution and widespread occurrence make dinoflagellates important zone fossils. In particular, the biostratigraphically important genus *Apectodinium* originates during this epoch. A nearly worldwide acme of the genus occurs at the very beginning of the succeeding Eocene and this is one of the primary markers used for recognizing the boundary between the two epochs.

Foraminifera Rapid radiation of planktonic foraminifera typical of the Cenozoic continued in the earliest Paleocene from its beginnings in the last few hundred thousand years of the Cretaceous. The result was almost complete replacement of latest Cretaceous species by Cenozoic ones in an interval of less than 1.5 million years. It is thought that some planktonic foraminiferal species acquired photosynthetic algal symbionts during the Paleocene, which may have allowed them to spread into oligotrophic environments. Benthic foraminifera fared better than did the planktonics, and low-oxygen-tolerant species increased to dominate early in the epoch. Notable newcomers were the textulariids among the agglutinated-shelled forms and the nummulitids among the calcareous-shelled forms, these latter representing one of the best known groups of larger benthic foraminifera of the Cenozoic. A major extinction, the benthic foraminiferal extinction (BFE), affected benthic foraminifera at the end of the Paleocene. The evolutionary and cosmopolitan attributes of planktonic foraminifera, like those of the calcareous nannoplankton, have resulted in planktonic foraminifera being used to divide the Paleocene into eight global biozones (Figure 1).

Coelenterata and bryozoans The millepore hydrozoans made their appearance during the Paleocene, as did two families of octocorals. There were no innovations at family level within the scleractinian corals. However, this group does show a marked low in terms of diversity. Based on gross morphology, there is no evidence that algal symbiosis collapsed at the end of the Cretaceous. Although the Paleocene reef recovery does not involve a return to Mesozoic levels of diversity, it does mark the first appearance of microbially cemented reefs since the Jurassic Period. This suggests that the Paleocene marks the emergence of modern coral reef communities, rather than a recovery from a eutrophically driven collapse. The Paleocene radiation of cheilostome and ascophoran bryozoans increased its pace; especially the cheilostomes, the family diversity of which increased exponentially.

Molluscs Cephalopods scarcely recovered from the end-Cretaceous extinctions, which eliminated ammonites and belemnites, although a new family of nautiloids, the Aturiidae, arose at the beginning of the Paleocene. Bivalve and gastropod recovery was more dramatic. Typically, the initial recovery phase in which diversity remained low varied in length and was followed by several pulses of increasing diversity separated by lows (initial radiation phase). Several families (Ostreidae, Carditidae, and Turritellidae), however, show a different pattern. In these cases, diversity increased or remained high at the beginning of the epoch, then suffered a decline. This means that these families formed a much higher percentage of the mollusc faunas early on than later in the epoch. The speciose nature of these families and their mixture of planktotrophic and brooding larval development mechanisms may have enhanced the Paleocene survival and early success of these opportunists. Once new forms arose later in the Paleocene, however, their competitiveness was low. Survival of Mesozoic taxa was more marked in high northern and southern latitudes than elsewhere.

Echinoids The end-Cretaceous extinctions resulted in a drop in diversity in the Early Paleocene and a change from roughly equal representation of regular and irregular groups to dominance by irregular echinoids. Irregular forms were also affected at the end of the Cretaceous, with holasteroids being decimated and survivors moving from shelf to deep-water environments as the Danian chalk facies disappeared. The regular family Saleniidae also shifted into deeper waters at the same time.

Vertebrates Marine fishes appear to have suffered little at the Cretaceous–Tertiary boundary. Among the cartilaginous elasmobranchs (sharks and rays), one family (the Torpedinidae) originated at the beginning of the Paleocene and four more (the Lamnidae, Otodontidae, Carcharinidae, and Mobulidae) first appear later in the epoch. Six teleost families appeared at the beginning of the Paleocene and a further 10 families appeared later in the epoch. Despite a patchy record, the Paleocene appears to mark the meagre beginning of a major Cenozoic radiation. In contrast, the only marine reptiles to survive into the Paleocene were turtles and dyrosaurid crocodilians.

Continental Realm

Land plants Schimper, when basing his Paleocene on distinctive floras, was aware that what he was observing might only be a local phenomenon of north-western Europe. In fact, Paleocene seed plant taxa and floral composition seem to represent a segment of a modernization trend, the origins of which lay in the latest Cretaceous, when angiosperms (flowering plants) became dominant over gymnosperms (broadly, conifers and cycads). There was turnover of seed plant taxa across the Cretaceous–Tertiary boundary and low-diversity opportunistic floras (particularly ferns) suggestive of abrupt ecological disruption in the very earliest Paleocene in western North America and probably elsewhere. Diversity, however, increased later in the epoch. Angiosperm fruits in the Paleocene were mainly small and dry.

Invertebrates Non-marine (mainly pulmonate) gastropods show no particular effect from end-Cretaceous events and the Paleocene saw essentially the beginning of a Cenozoic radiation of terrestrial families. For insects, there is little evidence of extinction, at least at family level, at the end of the Cretaceous, although there are few Paleocene sites yielding members of this group. Nevertheless, diversification that began in the Cretaceous seems to have increased in the Paleocene. Evidence from leaf damage by herbivorous insects suggests that early Paleocene insect herbivores were generalists. A recovery phase is lacking for at least the first million years of the Paleocene.

Vertebrates The pattern of events from freshwater teleost fishes is similar to that of the marine pattern. Five families have first records at the beginning of the Paleocene and a further 10 occur later in the epoch. Similarly, there appears to have been little effect of end-Cretaceous events on Paleocene amphibians, lizards, snakes, crocodilians, or turtles. The freshwater champsosaurs also survived into the Paleocene. The record of birds is sparse, but there is as yet no evidence of a major turnover at the end of the Cretaceous and no undoubted appearances of modern bird families in the Paleocene. There was, in contrast, a major change in mammals between the Cretaceous and the Paleocene. Although a boundary sequence exists only in western North America, Paleocene mammals differ radically from latest Cretaceous ones in every continent where they are known. This implies a major turnover at or near the boundary. In North America, the Cretaceous marsupial versus placental-dominated fauna was replaced in the Paleocene by an almost exclusively placental fauna (*see* **Fossil Vertebrates:** Placental Mammals). In Europe (poorly known in the Late Cretaceous) and Asia, the Paleocene faunas are also dominated by placentals, but the closeness of their phylogenetic relationships with Late Cretaceous placental groups is disputed. The greatest difference is in South America, where primitive non-therian mammals are replaced by a diversity of marsupials and placentals. Prominent features of Paleocene mammal faunas are (1) their dominance by archaic types not closely related to modern orders and (2) their strong, continent-specific endemism. Nevertheless, the following modern orders do have their first fossil appearances during the Paleocene: carnivorans, edentates, rodents, and probably also lipotyphlans (typical insectivorans) and macroscelideans (elephant shrews). Mammals underwent rapid recovery and massive radiation following the end-Cretaceous extinction. Extinction of many of these archaic types took place at the end of the Paleocene in the northern hemisphere continents when major dispersal of more modern types displaced them. Various endemic groups, however, evolved in South America, which became isolated, first from North America and later from Antarctica. These groups survived long after the Paleocene.

Climate and Environments

After the end of the Cretaceous, oxygen isotopic records from benthic foraminiferal tests document a gradual warming of the world's oceans during the first third of the Paleocene. This warming reversed the overall but fluctuating cooling trend of the latest Cretaceous (Maastrichtian), but was minor compared to the brief warming event near the end of the Maastrichtian. A cooling in the Late Paleocene was followed by a major warming, punctuated at the very end of the epoch by the beginning of a brief intense warming episode that marks the onset of the Eocene (Figure 1). During the Paleocene, there was likely to have been little or no polar ice. The impact of climate and other factors on Paleocene environments are examined in the following sections.

Marine Environments

Following the end of the Cretaceous, it has been suggested that primary productivity in the oceans was strongly reduced. This is thought to explain the general change from infaunal to epifaunal dominance among benthic foraminiferal communities at the Cretaceous–Tertiary boundary. In some deep-water sites, the change was briefly delayed at the very beginning of the Paleocene by an opportunistic low-diversity infaunal assemblage that may have been responding to a large but short-term flux of organic matter to the seafloor, from the mass mortality of microplankton. Moreover, radiolarians and associated biosiliceous oozes, which infer high oceanic productivity, are generally rare across the Cretaceous–Tertiary transition. However, evidence of rich radiolarian assemblages across the boundary in New Zealand suggest that enhanced upwelling caused by climatic cooling characterized this southern high-latitude area at the beginning of the Paleocene. Carbon isotope studies of foraminifera also indicate stability of surface productivity at high, in contrast to low, latitudes, with resultant lower extinction levels. Another phenomenon is the apparent enhanced survival of both mollusc and ostracod taxa from the Cretaceous at high, compared to low, latitudes, with subsequent spread of the ostracods to lower latitudes later in the Paleocene.

Later Paleocene oceans had more normal higher primary productivity according to their greater diversity of planktonic microbiota. They are also likely to have had modern rates of thermohaline circulation and thus of nutrient flux from subsurface to surface waters. Major warming at the end of the epoch caused a slowing of the thermohaline circulation, with resultant reduction in rate of nutrient flux, which in turn expanded the geographic range of oligotrophic habitats. This trend eventually involved a selective warming of the deep ocean by 4°–6°C due to changed circulation and a much reduced pole-to-equator temperature gradient, resulting in reduced wind intensity. The sudden warming at the Paleocene–Eocene boundary is attributed to a massive injection of CO_2 into the oceans and atmosphere by thermal dissociation of methane hydrates and their release from marine sediments. This event is judged to be the cause of the contemporaneous major extinction of benthic organisms, mainly foraminifera (BFE) and ostracods.

Continental Realm

Leaf physiognomy climate proxies produce a land-based climate curve resembling that for the marine realm. Vegetation shows a shift from open-canopy broad-leaved evergreen woodland in the Late Cretaceous to rainforest in the Early Paleocene. This marks the first appearance of such a vegetation type, which extended to higher latitudes than the present day because of the absence or near absence of polar ice. The highest (polar) latitudes were occupied by broad-leaved deciduous forests. Despite the enhanced equability, these floras had to survive extended winter darkness. Accordingly, their structure was open and their diversity low. The very beginning of the Paleocene is also locally marked by a dominance of the spores of ferns. This phenomenon, known as the 'fern spike', is recorded from regions as far apart as western North America and New Zealand and suggests that these plants were the first colonizers of a denuded landscape.

The mammals that lived in these Paleocene habitats were mainly small. Their Cretaceous ancestors had been mainly insectivorous and to a certain extent carnivorous, and they were only now expanding their dietary spectrum to include fruit. Larger size, which went hand in hand with leaf eating, was rare and began to be evolved late in the epoch. The ecological composition of most well-known faunas supports the presence of widespread forested environments. Land connections between continents, even if fleeting during times of low sea-level, allowed marsupials and placentals to disperse from North America to South America around the end of the Cretaceous or the beginning of the Paleocene. The marsupials went on to colonize Australasia via Antarctica before Australasia finally broke free.

See Also

Fossil Vertebrates: Fish; Placental Mammals. **Mesozoic:** Cretaceous; End Cretaceous Extinctions. **Microfossils:** Foraminifera. **Palaeoclimates**. **Sedimentary Environments:** Reefs ('Build-Ups'). **Tertiary To Present:** Eocene.

Further Reading

Aubry M-P, Lucas SG, and Berggren WA (eds.) (1998) *Late Paleocene-Early Eocene Climatic and Biotic Events in the Marine and Terrestrial Records*. New York: Columbia University Press.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Chaloner WG, Harper JL, and Lawton JH (eds.) (1991) The evolutionary interaction of animals and plants. *Philosophical Transactions of the Royal Society of London, Series B* 333: 177–288.

Collinson ME (1990) Plant evolution and ecology during the early Cainozoic diversification. *Advances in Botanical Research* 17: 1–98.

Culver SJ and Rawson PF (eds.) (2000) *Biotic Response to Global Change. The last 145 million years*. Cambridge: Cambridge University Press.

Friis EM, Chaloner WG, and Crane PR (eds.) (1987) *The Origins of Angiosperms and their Biological Consequences*. Cambridge: Cambridge University Press.

Hansen TA (1988) Early Tertiary radiation of marine molluscs and the long-term effects of the Cretaceous-Tertiary extinction. *Paleobiology* 14: 37–51.

Hartman JH, Johnson KR, and Nichols DJ (eds.) (2002) The Hell Creek Formation and the Cretaceous-Tertiary boundary in the northern Great Plains: an integrated continental record of the end of the Cretaceous. *Geological Society of America Special Papers* 361: 1–520.

Khain VE and Balukhovsky AN (1997) *Historical Geotectonics, Mesozoic and Cenozoic*. Russian Translation Series 117. Rotterdam: AA Balkema.

MacLeod N, Rawson PF, Forey PL, *et al.* (1997) The Cretaceous-Tertiary biotic transition. *Journal of the Geological Society, London* 154: 265–292.

Scotese CR and Golonka J (1992) *Paleogeographic Atlas, PALEOMAP Progress Report 20-0692*. Dept. of Geology, University of Texas at Arlington.

Smith AG, Smith DG, and Funnell BM (1994) *Atlas of Mesozoic and Cenozoic Coastlines*. Cambridge: Cambridge University Press.

Thomas DJ, Zachos J, Bralower TJ, Thomas E, and Bohaty S (2002) Warming the fuel for the fire: evidence for the thermal dissociation of methane hydrate during the Paleocene-Eocene thermal maximum. *Geology* 30: 1067–1070.

Wing SL, Gingerich PD, Schmitz B, and Thomas E (eds.) (2003) Causes and consequences of globally warm climates in the early Paleogene. *Geological Society of America Special Papers* 369: 1–614.

Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686–693.

Eocene

J J Hooker, The Natural History Museum, London, UK

Introduction

The Eocene epoch/series is the oldest of the four original subdivisions of the Tertiary period/system proposed in 1833 by Sir Charles Lyell in his *Principles of Geology*. The name derives from the Greek '*eos*', meaning dawn, and '*kainos*', meaning recent. This is "… because the very small proportion of living species contained in those strata indicates what may be considered the first commencement, or *dawn*, of the existing state of the animate creation" (vol. 3, p. 55). The time covered by Lyell's 'existing state' is what we now understand to be the Cenozoic era/erathem, which itself is broadly divided into either Tertiary and Quaternary or Paleogene and Neogene periods. Lyell subdivided the Tertiary into four epochs based on the proportions of living and extinct species of shelled organisms (molluscs and foraminifera) encountered as fossils in different strata. He recognized 1238 Eocene species, of which he considered only 42 (or 3.5%) remain alive today. Lyell's species are about equivalent to what modern marine biologists would rank as genera or even subfamilies. For this reason, no modern species of mollusc or foraminifer is currently recognized as occurring as far back as the Eocene.

The Eocene as recognized today has changed considerably in its definition and time-span since 1833. Its earliest parts have become the Paleocene and its later parts the Oligocene, both epochs that were described after 1833. The Eocene thus succeeds the Paleocene and precedes the Oligocene. It is currently estimated to have lasted nearly 22 million years, from 55.8 to 33.9 Ma. The Eocene itself is divided into four ages/stages (in order of decreasing age): the Ypresian, the Lutetian, the Bartonian, and the Priabonian (Figures 1 and 2). The beginning and end of the Eocene have only recently been stabilized by the Paleogene Subcommission of the International Union of Geological Sciences (IUGS). Its beginning is marked by a sharp dip in the carbon isotope curve, named the Carbon Isotope Excursion (CIE), interpreted as a global warming event, the Paleocene–Eocene Thermal Maximum (PETM). This climate event sparked major changes in both marine and continental biotas. The Global Stratotype Section and Point (GSSP) for the beginning of the Eocene is placed at Dababiya, Egypt. This geological section is the best available for demonstrating the boundary criteria and acts as a global reference. The end of the Eocene is marked by extinction of the planktonic foraminiferal family Hantkeninidae, representing the last in a cumulative series of extinctions caused by long-term global cooling. The GSSP for the Eocene–Oligocene boundary is at Massignano, Italy.

Plate Tectonics and Other Physical Phenomena

The processes of subduction around the Pacific rim continued from the Paleocene, with the Kula Plate disappearing beneath the Aleutian Arc of Alaska. In North America, the remnants of the Late Cretaceous Western Interior Seaway, still present in the Paleocene, disappeared completely. Caribbean, deep-water circulation between the Pacific and Atlantic began to be cut off late in the Eocene as the Central American Isthmus formed a structural unit, but remained submerged.

Seafloor spreading in the Atlantic continued to migrate northwards on either side of Greenland. Thus, a seaway formed early in the Eocene between Greenland and Europe, linking the earlier landlocked Arctic Ocean once again with the Atlantic. At the same time, the maximum outpourings of lava associated with the Iceland hotspot occurred in East Greenland and the then adjacent Hebridean Province. The Labrador Seaway between Greenland and North America also widened and the land connection between Europe and North America via Greenland finally severed. A land bridge developed intermittently between North America and Asia across the Bering Straits as a result of sea-level changes.

For the entire epoch, the epicontinental West Siberian Seaway linked the Arctic Ocean with the Peritethys, maintaining the isolation of Europe from Asia. This seaway narrowed at its southern end to form the Turgai Straits, which may occasionally have dried out at times of low sea-level. In fact, except for the first million or so years of the epoch, when Europe was connected to North America via Greenland, Eocene Europe comprised several island masses isolated from the rest of the world's continents. The Peritethys continued to be partially separated from the main mass of the Tethys Ocean by the narrow orogenic belt stretching intermittently from Italy through Asia Minor to southern Iran. In this complex area of the ancient Mediterranean, the islands of Corsica, Sardinia, and the Balearics were still part of the Iberian Peninsula. During the course of the Eocene,

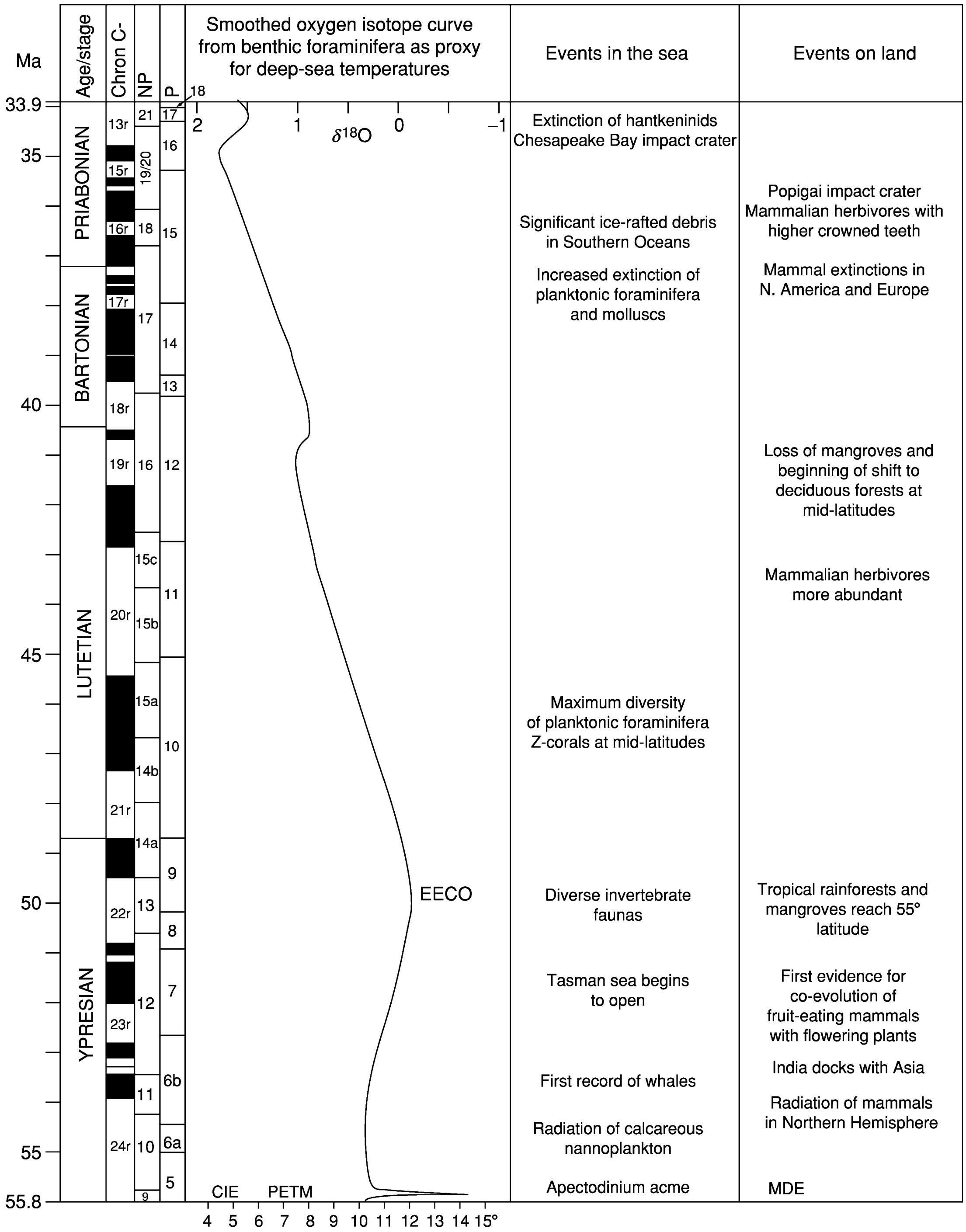


Figure 1 Time chart of the Eocene, showing how it is divided up by ages/stages, magnetochrons (chron C), and global calcareous nannoplankton (NP) and planktonic foraminiferal (P) biozones. Also shown are an isotope proxy temperature curve and the main biotic events in the sea and on land. Magnetochrons are divided into normal (black) and reversed (white) intervals; the normals (n), often composite, are younger than the reversals (r) that bear the same number. CIE, Carbon Isotope Excursion; EECO, Early Eocene Climatic Optimum; MDE, Mammalian Dispersal Event; PETM, Paleocene–Eocene Thermal Maximum. Data for the isotope curve from Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686–693. Absolute dates from Gradstein F, Ogg J, and Smith A (in press) *A Geological Timescale 2004.* Cambridge University Press, Cambridge.

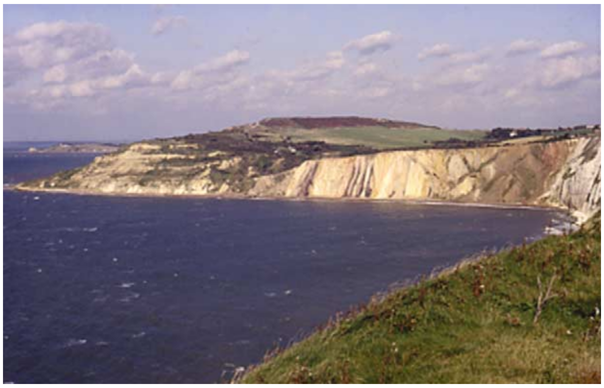


Figure 2 Coastal section in Alum Bay, Isle of Wight, UK, showing strata spanning almost the entire Eocene. Oldest is to the right, youngest to the left. The red and brown are Ypresian, the yellow with dark intercalations is Lutetian, the grey below the house is Bartonian (all vertical), and the white (mainly horizontal) is Priabonian.

anticlockwise rotation of Iberia resulted in initial uplift of the Pyrenees.

Further south, India completed its northward drift and docked with Asia in the Early Eocene, laying the foundations for the Himalayan uplift. Rifting between Australasia and Antarctica widened and developed into an (at first) narrow Tasman Sea. Australasia began its 50-million-year-long trek to lower latitudes at this time. Subduction beneath the west side of the Antarctic Peninsula led to its uplift and shallowing of the back-arc basin to the east. South America remained connected to the Antarctic Peninsula for most of the epoch. However, development of the Scotia Arc and formation of the Scotia Sea (Drake Passage) meant that, by the end of the Eocene, all major elements of Gondwana had separated and the modern arrangement of the continents was essentially in place.

Two important impact-ejecta strewn fields have also been recognized within Late Eocene sediments. It has been suggested that two large impact craters, Popigai in Siberia and Chesapeake Bay in the USA, were the sources for these (Figure 1).

Biota

The Eocene biota is characteristically diverse and abundant. Most surviving groups had recovered from the Early Paleocene diversity low by this time and had undergone or were undergoing radiation. Marine groups, such as molluscs, crustaceans, and echinoids, had a familiar modern appearance. On land, the same was true of many reptile and amphibian groups, but Eocene mammals differed radically from their living relatives. Key biotic events are discussed below.

Marine Realm

Calcareous nannoplankton An important turnover in this group of microscopic algae occurred during the first million years of the Eocene with the origination of many new taxa. The genera that had radiated and dominated in the Paleocene became extinct. Survivors were the previously low-diversity genera of modern aspect. This turnover is attributed to the PETM (see below). Later in the Eocene, diversity reduced regionally and deeper water habitats were vacated because of progressive cooling. New taxa evolved in response to the temperature changes and increased eutrophication. The rapid evolution and widespread occurrence of nannoplankton at this time have enabled the Eocene to be divided into 12 global biozones (Figure 1).

Dinoflagellates The PETM resulted in a near-global spread of dinoflagellate species belonging to the genus *Apectodinium* through middle and high latitudes. This event, known as the *Apectodinium* acme, is an important biostratigraphical marker for the Paleocene–Eocene boundary. Continued radiation produced increasingly diverse cyst assemblages during the Early and Middle Eocene, especially of the genus *Wetzeliella* and its relatives. Diversity remained high for most of the Eocene, despite the later cooling of sea surface temperatures. The very end of the epoch, though, bears witness to a reduction in abundance of low-latitude taxa, as well as the invasion of these areas by formerly higher latitude groups.

Foraminifera The initial Eocene turnover in planktonic foraminifera at the PETM was minor. Radiation continued from the Paleocene to reach a diversity maximum in the Middle Eocene. Then, deteriorating temperatures caused long-term stepwise extinctions, of which the most intense was at the end of the Middle Eocene (Bartonian). Surface-dwelling species were gradually replaced by cold-tolerant, subsurface species. Benthic foraminifera fared much worse at the PETM, suffering 30–50% extinctions for middle bathyal through abyssal forms. This is known as the Benthic Foraminiferal Extinction (BFE). Shallower water taxa fared better. The latter include the best-known Eocene calcareous foraminifera, some of which (e.g., the nummulites) became important rock formers and reached maximum diameters of 10 cm. Benthic foraminifera, like the planktonics, underwent stepwise extinctions throughout the Middle and Late Eocene. The evolutionary and cosmopolitan attributes of planktonic foraminifera, like those of the calcareous nannoplankton, have resulted in their use for dividing the Eocene into 13 global biozones (Figure 1).

Coelenterates and bryozoans Two families of octocorals originated at the beginning of the Eocene. Radiation of scleractinian corals occurred, with increasing species numbers throughout the epoch. Interestingly, there was no reduction in diversity when climates cooled in the Middle and Late Eocene. Amongst the bryozoans, the rapid diversification of cheilostome families in the Paleocene continued into the Eocene, but then began to slow, although there was a peak diversity of species in the Priabonian.

Molluscs After the terminal Cretaceous extinctions of ammonites and belemnites, the surviving coleoids (squids and cuttle fishes) were slow to recover and have left little fossil evidence in the Paleocene. Their appearance in the Eocene is therefore virtually as 'Lazarus' taxa. Most of the spirulid, and all of the sepiid, radiation took place in this epoch, with several new families appearing. Diversification in bivalves and gastropods continued from the Paleocene into the Eocene. A study in the Gulf Coast, USA, has shown that species numbers reached a peak in the Bartonian, when climates were already beginning to deteriorate, and plummeted during the Priabonian. The origination of many new families in the epoch gave these faunas a more modern aspect.

Echinoids The main Eocene evolutionary events surround the irregular group. This epoch saw a decline in cassiduloids and a corresponding rise in diversity of clypeasteroids, with five new families appearing. Clypeasteroids were the last group to evolve non-planktotrophic lecithotrophic or brooding development for their larvae, and this development occurred in the Eocene. Although this adaptation resulted in reduced dispersal ability, it is thought to have enhanced resistance to the rigours of increased seasonality associated with the deteriorating Middle and Late Eocene climate.

Vertebrates Amongst cartilaginous fishes (Elasmobranchii), there was a significant increase in rays (Rajiformes), with three new families making their first appearance at the beginning of the epoch. Eocene marine teleosts are characterized by a continuation of their Cenozoic radiation, particularly in tarpon and eels (Elopomorpha) and massively in spiny-ray fishes (Acanthomorpha). A caveat, however, is provided by the breadth of diversity of the many newly appearing acanthomorph families, implying an earlier record of these, which we have yet to discover.

Marine mammals are first recorded from the Eocene in the form of primitive whales (Cetacea) and sea cows (Sirenia). These represent the first marine tetrapod evolutionary innovation since the extinction of plesiosaurs and mosasaurs at the end of the Cretaceous. The earliest known cetaceans are from the Ypresian of Pakistan. These early members retained well-developed walking limbs, both fore and hind, and are interpreted to have been amphibious inhabitants of both freshwater and marine habitats. Limb structure demonstrates a close relationship with the entirely land-based artiodactyls (cloven-hoofed mammals) which appeared at about the same time.

Continental Realm

Land plants In the Eocene, angiosperms (flowering plants) diversified and became more modern in appearance. Many modern genera (although no modern species) can be recognized within families that are well represented today. Examples are magnolias (Magnoliaceae), grape vines (Vitaceae), citrus and allies (Rutaceae), spurges (Euphorbiaceae), dogwoods (Cornaceae), custard apples (Anonaceae), laurels (Lauraceae), moonseeds (Menispermaceae), icacinas (Icacinaceae), and palms (Arecaceae). Grasses, on the other hand, were notably rare.

Invertebrates Land gastropods underwent a major Eocene radiation with two new prosobranch and 16 new pulmonate families appearing. This gave a more modern aspect to land snail and slug faunas. The poor Paleocene record of insects makes it difficult to judge specific Eocene innovations. By the end of the Eocene, though, 223 families had appeared since the Cretaceous. In particular, there is a surge in origination of dragonflies (Odonata), flies (Diptera), butterflies and moths (Lepidoptera), and beetles (Coleoptera). Compared with the Paleocene, leaf damage indicative of insect herbivory shows an overall increase in abundance and diversity, which is expressed particularly by more specialist damage types, such as leaf mines and leaf galls.

Vertebrates Eight families of freshwater teleost fishes have their first records in the Eocene. These include such well-known families as carp (Cyprinidae), salmon (Salmonidae), and perch (Percidae). Five modern families of freshwater and terrestrial turtles appeared in the epoch. These included the well-known terrapins (Emydidae) and tortoises (Testudinidae). Nine modern families of birds have their first records in the Eocene, with doubtful records of a further 14. The undoubted records include parrots (Psittacidae), owls (Strigidae), nightjars (Caprimulgidae), swifts (Apodidae), colies (Coliidae), and rollers (Coraciidae). None, however, belong to the passerines, which were either very rare at this time or had not yet evolved.

Land mammals underwent their greatest innovation of the Cenozoic in the Eocene. Most modern placental orders appeared at the beginning of the epoch in the northern hemisphere in what is known as the Mammalian Dispersal Event (MDE). The sudden appearance of so many specialized morphotypes implies an earlier evolution extending back into the Paleocene, of which we have no record. These newcomers were bats (Chiroptera), primates (Primates), cloven-hoofed mammals (Artiodactyla), odd-toed ungulates (Perissodactyla), and elephants (Proboscidea). Other modern orders appeared later in the Eocene, including hyraxes (Hyracoidea), pangolins (Pholidota), and tree-shrews (Scandentia). Many early members of these groups were small. Moreover, few hoofed plant-eaters developed a large size or an exclusively leaf-eating diet until relatively late in the epoch.

Climate and Environments

Eocene climates include the warmest of the entire Cenozoic era (early in the epoch), deteriorating later to lead eventually to the first Cenozoic ice build-up in Antarctica (Figure 1). Carbon and oxygen isotope records based on the analysis of foraminiferal tests, mammalian dental enamel, soil carbonates (Figure 3), and lignites all show a short, sharp, 200 000-year-long, negative anomaly (CIE) at the very beginning of the Eocene. This indicates an extreme warming perturbation (PETM), when deep-sea temperatures rose by 4–6°C. This spike was superimposed on a long-term warming trend that began in the Late Paleocene and culminated late in the Ypresian with a comparable, but longer duration, warming peak, known as the Early Eocene Climatic Optimum (EECO). During the PETM and EECO, there was probably no polar ice, pole to equator temperature gradients were much reduced, and overall wind circulation slowed. Subsequently, temperatures declined, with small warming interruptions near the end of the Lutetian and in the Priabonian (Late Eocene). Significant amounts of ice-rafted debris, indicative of glacial activity, first occur in southern high-latitude sediments at the end of the Bartonian.

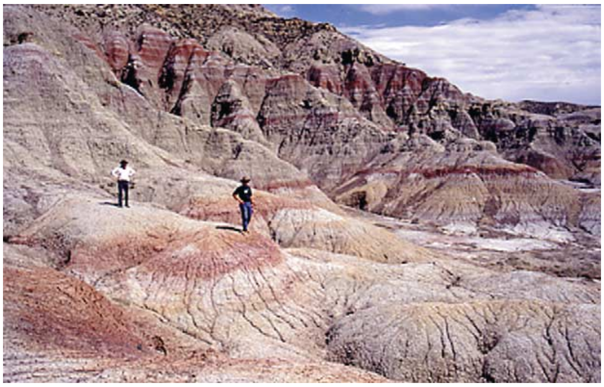


Figure 3 Polecat Bench, Wyoming, USA, showing a thick sequence of latest Paleocene and earliest Eocene fluvial mudstones and palaeosols. The base of the Carbon Isotope Excursion (CIE) is recorded from mammalian dental enamel and soil nodules at a point intermediate between the two distinct red bands. The upper red band records the Mammalian Dispersal Event (MDE).

Marine Environments

The PETM perturbation is thought to have been caused by a massive injection of CO_2 into the oceans and atmosphere by the thermal dissociation of methane hydrates and their release from marine sediments. This event, in turn, caused a slowing of the thermohaline circulation with a resultant reduction in the rate of vertical marine nutrient flux and an expansion in the geographical range of oligotrophic habitats. The major rapid extinction of benthic foraminifera (BFE) and ostracods that resulted (*see* **Tertiary To Present:** Paleocene) meant low diversity for such organisms in bathyal and abyssal regions early in the Eocene. Post-extinction benthic foraminiferal faunas vary widely with geography compared with pre-extinction faunas. This is thought to be a phenomenon of highly perturbed communities. Early Eocene planktonic foraminiferal communities also show a greater diversity of oligotrophic forms than do those of the Late Paleocene. This is consistent with the proposed model of a uniformly warm Early Eocene ocean with reduced rates of circulation. Widespread deep-water anoxia, with associated calcium carbonate dissolution at the PETM, appears to have favoured the latitudinal spread of dinoflagellates of the genus *Apectodinium* (the *Apectodinium* acme) as far north as the Barents Sea and as far south as New Zealand. *Apectodinium* is thought to have been partially or fully holozoic and replaced earlier dominant photosynthetic dinoflagellate communities.

This period of warming and equability promoted high taxonomic diversity for many groups of organisms at much higher latitudes than today. Thus, tropical-type molluscs, zooxanthellate-like corals, and giant nummulites are well represented in northern Europe, a fact that impressed Lyell as long ago as 1833. The subsequent cooling reversed processes that took place early in the epoch. Thermohaline circulation rates accelerated, ocean mixing increased, low-nutrient, surface water habitats were reduced, and instability was created in high-nutrient, surface water areas through seasonal production. This increase in seasonality, documented from oxygen isotopes in mollusc shells and fish otoliths (ear stones), is considered to be an important feature of the Late Eocene cooling. Extensive investigation of the two Late Eocene impacts has found no important, associated biotic extinctions. However, carbon and oxygen isotope analyses and microfossil assemblage composition from strata overlying the ejecta layers suggest that a minor, but prolonged, cooling followed at least one of these impact events. This prolongation of the environmental effect may have been caused by an increase in the Earth's albedo from ice-sheets that were extended in response to the cooling.

Continental Environments

Leaf physiognomy climate proxies give a land-based climate curve that, like the marine curve, shows cooling from an Early Eocene maximum, but declines in more widely fluctuating steps. Moreover, the study of leaf floras across the Paleocene–Eocene transition in the Bighorn Basin, Wyoming, USA, indicates an initial warming (PETM), followed by a cooling, and then renewed warming up to the EECO. The long-term warming phase that began in the latest Paleocene and culminated in the EECO had important effects on vegetation. During its later stages, highly diverse, multistratal rainforest with many lianas extended to latitudes of 55° or 60° north and south. This was accompanied by coastal mangrove vegetation, often dominated by the mangrove palm *Nypa*. These widespread rainforests were similar to those of south-east Asia today. They appear to have differed, however, in lacking epiphytic flowering plants and in having only very rare representatives of some taxa (e.g., dipterocarps) that are highly significant in modern south-east Asian forests. In addition to the multistratal nature of these Early Eocene forests, the plants differed from those of the Paleocene in their fruiting strategies. In the Paleocene, fruits were mainly small and dry with some drupes and nutlets. Fleshy fruits were rare. In contrast, in Eocene forests, fleshy fruits were abundant and larger nuts were also present. This change represents the beginning of the co-evolution of fleshy fruits with fruit-eating primates and of nuts with scatterhoarding rodents, that facilitates plant dispersal. At higher, polar latitudes, these rainforests were replaced by broad-leaved deciduous forests. There is no evidence in the Eocene for the coniferous forests that today clothe Earth's boreal regions. The later Eocene cooling is tracked by a shift to less diverse, more open, deciduous broad-leaved forests in middle and high latitudes, with the loss of many of the tropical taxa that dominated earlier in the epoch. Despite the development of more open vegetation, this did not include grassland, grasses being virtually absent from all Eocene floras.

The MDE that introduced many new mammals into the northern hemisphere continents at the time of the CIE (Figure 3) is partly attributed to the warm temperatures. These may have facilitated the well-documented dispersals through northern high latitudes (across the Greenland and Bering land bridges) through the extension of vegetation zones. However, minor dispersal also appears to have taken place across the Turgai Straits at lower latitudes. Low sea-levels in the vicinity of the Paleocene–Eocene boundary are also likely to have had a strong influence on dispersal. Mammal communities in the Early Eocene were dominated by small animals, many of which fed on insects and fruit and were adapted for life in the trees. Their patterns of ecological diversity resemble those of south-east Asian evergreen forests and thus support the plant evidence for the nature of the Eocene vegetation. As climates cooled later in the epoch, fruit-eating and climbing mammals diminished and larger ground-dwelling herbivores evolved. The patterns of ecological diversity that characterize these later faunas fit closely with more open wooded habitats as the plant fossils indicate. The structure and wear patterns of the teeth of these later Eocene herbivorous mammals indicate that they were browsers not grazers.

See Also

Biozones. **Fossil Invertebrates:** Insects. **Fossil Plants:** Angiosperms. **Fossil Vertebrates:** Placental Mammals. **Magnetostratigraphy**. **Microfossils:** Foraminifera; Palynology. **Palaeoclimates**. **Plate Tectonics**. **Tertiary To Present:** Paleocene; Oligocene. **Time Scale**.

Further Reading

Aubry M-P, Lucas SG, and Berggren WA (eds.) (1998) *Late Paleocene–Early Eocene Climatic and Biotic Events in the Marine and Terrestrial Records*. New York: Columbia University Press.

Benton MJ (ed.) (1993) *The Fossil Record 2*. London: Chapman & Hall.

Chaloner WG, Harper JL, and Lawton JH (eds.) (1991) The evolutionary interaction of animals and plants. *Philosophical Transactions of the Royal Society of London, Series B*, 333, pp. 177–288.

Collinson ME (1990) Plant evolution and ecology during the early Cainozoic diversification. *Advances in Botanical Research* 17: 1–98.

Culver SJ and Rawson PF (eds.) (2000) *Biotic Response to Global Change. The Last 145 Million Years*. Cambridge: Cambridge University Press.

Friis EM, Chaloner WG, and Crane PR (eds.) (1987) *The Origins of Angiosperms and Their Biological Consequences*. Cambridge: Cambridge University Press.

Khain VE and Balukhovsky AN (1997) *Historical Geotectonics, Mesozoic and Cenozoic*. Russian Translation Series 117. Rotterdam: A.A. Balkema.

Kobashi T, Grossman EL, Yancey TE, and Dockery DT III (2001) Reevaluation of conflicting Eocene tropical temperature estimates: molluskan oxygen isotope evidence for warm low latitudes. *Geology* 29: 983–986.

Prothero DR and Berggren WA (eds.) (1992) *Eocene-Oligocene Climatic and Biotic Evolution*. Princeton: Princeton University Press.

Scotese CR and Golonka J (1992) *Paleogeographic Atlas*. Arlington, TX: PALEOMAP Project, Department of Geology, University of Texas.

Smith AG, Smith DG, and Funnell BM (1994) *Atlas of Mesozoic and Cenozoic Coastlines*. Cambridge: Cambridge University Press.

Thomas DJ, Zachos J, Bralower TJ, Thomas E, and Bohaty S (2002) Warming the fuel for the fire: evidence for the thermal dissociation of methane hydrate during the Paleocene–Eocene thermal maximum. *Geology* 30: 1067–1070.

Vonhof HB, Smit J, Brinkhuis H, Montanari A, and Nederbragt AJ (2000) Global cooling accelerated by early late Eocene impacts? *Geology* 28: 687–690.

Wing SL, Gingerich PD, Schmitz B, and Thomas E (eds.) (2003) Causes and consequences of globally warm climates in the early Paleogene. *Geological Society of America Special Papers 369,* pp. 1–614. Boulder: Geological Society of America.

Zachos J, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686–693.

Oligocene

D R Prothero, Occidental College, Los Angeles, USA

Introduction

The Oligocene Epoch was defined by Heinrich Ernst von Beyrich in 1854. This interval of geological time was based on marine strata in Belgium and Germany, thought to be younger than the Lyell's classic upper Eocene (*see* **Tertiary To Present:** Eocene) strata of the Paris Basin, but older than Lyell's (*see* **Famous Geologists:** Lyell) concept of Miocene rocks. Von Beyrich's original list of 'Oligocene' rocks contained a wide spectrum of units of varying ages, including those that are now clearly referable to the Eocene or Miocene. For example, one unit (the bone sand of Eppelsheim) produced a Late Miocene *Hipparion* fauna. In addition, the type strata of von Beyrich's Oligocene in Belgium and Germany do not overlie the type strata of the Paris Basin or Italian Eocene, so the Eocene–Oligocene boundary cannot be recognised in either area. As is true of the rest of the European Cenozoic, the type sections of the stages within the Oligocene are shallow-water deposits bounded by unconformities, and represent only a small portion of its duration.

For 130 years after von Beyrich's establishment of the Oligocene, there was considerable confusion over what was Eocene and what was Oligocene, not only in the western European type areas, but especially in other regions which could only be correlated indirectly to the stratotypes. For example, in North America, the Duchesnean land mammal age was thought to be late Eocene or Oligocene (it is now considered middle Eocene), the Chadronian land mammal age was correlated with the Early Oligocene (it is now known to be Late Eocene in age), and the Orellan and Whitneyan land mammal ages were thought to be Middle and Late Oligocene (they are now both regarded as Early Oligocene in age). Although these problems made the type Oligocene stages hard to correlate to other regions, the use of planktonic microfossils and magnetic stratigraphy has allowed geologists to correlate the classic shallow-marine European stratotypes and terrestrial sections to the global deep-marine standard (Figure 1). As a result, the Oligocene is now securely correlated around the world.

In 1989, the Eocene–Oligocene boundary was formally established at the last appearance of the spiny planktonic foraminiferal genus *Hantkenina* in a quarry section near Massignano, Italy. However, later work has since shown that part of the type upper Eocene Priabonian Stage is Early Oligocene by this definition, so there are still problems with this criterion. Most of the important climatic events that many scientists believe should mark the beginning of the Oligocene (e.g., the global oxygen isotope shift indicating the expansion of Antarctic glaciers, and related events such as the cooling on North America) are earliest Oligocene (magnetic Chron C13N, about 33 million years ago) using the hantkeninid criterion. Thus, there are grounds for revising

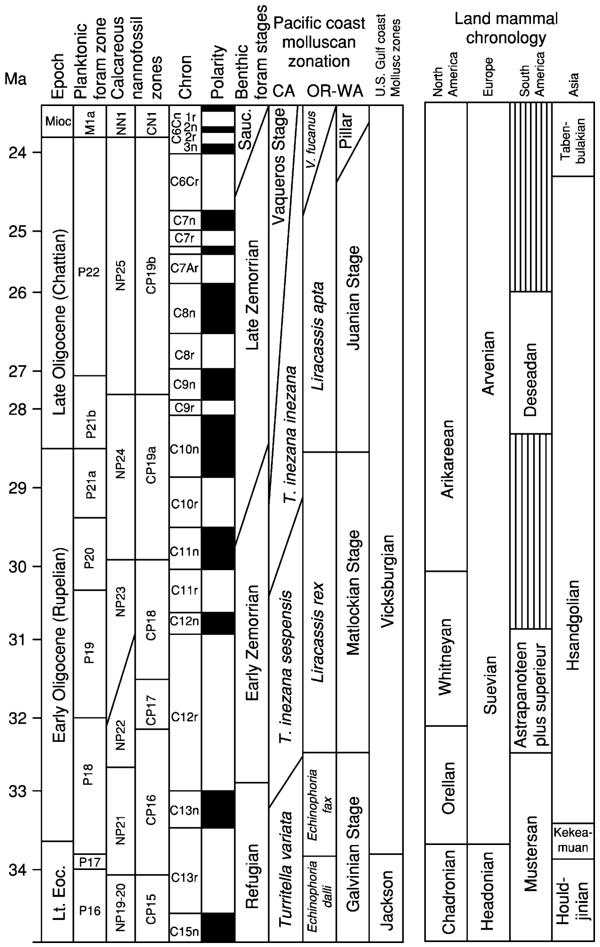


Figure 1 Correlation of various Oligocene biostratigraphic units to the global time-scale (left) and magnetic polarity time-scale (middle). Global time-scale and planktonic zonation (after Berggren *et al.* (1995)). Pacific Coast marine zonation after Prothero (2001). US Gulf Coast molluscan zonation (after Prothero *et al.* (2003)). North American land mammal chronology (after Prothero and Emry (1996)). Asian land mammal chronology (after Meng and McKenna (1998)). European land mammal chronology (after Barbera *et al.* (2001)).

the 1989 definition to a more 'natural' boundary. This would also place the beginning of the Oligocene after the end of the type Late Eocene (Priabonian Stage). However, no such revision has been formally proposed to date.

Only two stages are recognized in the 11-million-year (34–23 Ma) span of the Oligocene. The Early Oligocene Rupelian Stage includes the interval from 34–28.5 Ma. The Late Oligocene Chattian Stage is dated between 28.5–23.8 Ma. There is no formally recognized Middle Oligocene.

Oligocene Climate

The 11 million years of the Oligocene marked an important climatic transition in Earth history. The latest Paleocene and Early Eocene (55–50 million years ago) was the peak of global warming, a 'greenhouse' climate that exhibited the warmest global conditions since the Late Cretaceous. Climates were so warm and mild that crocodilians and temperate plants lived above the Arctic Circle in regions that experienced six months of darkness. Beginning in the Middle Eocene, though, this balmy climate began to transform, with the greenhouse climate gradually changing to a colder, more extreme climate. The end of the Middle Eocene (37 Ma) was marked by a major cooling event, which caused the extinction of many marine organisms adapted to warm, tropical waters.

During the three million years of the Late Eocene (37–33 Ma), there was a slight warming and recovery from the long-term cooling trend. At least three major comet or asteroid impacts struck the Earth in the Middle of the Late Eocene (35.5–36.0 Ma), but these caused no significant changes in climate, nor extinction of any importance. As noted above, the Eocene–Oligocene boundary is now formally recognized by the extinction of hantkeninid foraminifera, although no other major climatic or extinction events occurred at this time. (Note that this invalidates an old idea from the 1970s and 1980s that a 'Terminal Eocene Event' – comparable to the event that ended the Cretaceous – also marked the end of the Eocene).

The most significant climatic event of this interval occurred in the earliest Oligocene (as currently defined, using the hantkeninid criterion), at about 33 Ma. This is now known as the Oi1 event. In the marine record, both benthic and planktonic foraminiferal oxygen isotopic ratios show about a 1.3 per mil increase (Figure 2). It was calculated that about 0.3–0.4 per mil of the change was due to a major expansion of Antarctic ice-sheets that lowered global sea-level by at least 30 m. The remaining 0.9–1.0 per mil is explained by about 5–6°C global cooling, which lowered global mean temperature from as high as 13°C in the Early Eocene and 7°C in the latest Eocene to values just a few degrees above freezing (as a global average – the poles were well below freezing for the first time, while the tropics remains relatively unchanged).

Abundant data suggests that this global cooling event was due largely to the growth of the first major Antarctic ice-sheet since the Permian (over 250 Ma). Drilling on the margin of the Antarctic continent and in oceanic plateaus in the Southern Ocean (e.g., Maud Rise and the Kerguelen Plateau) have produced unmistakable evidence of ice-sheet growth. Not only do the isotopic records show its effect, but many of the sediments drilled from the

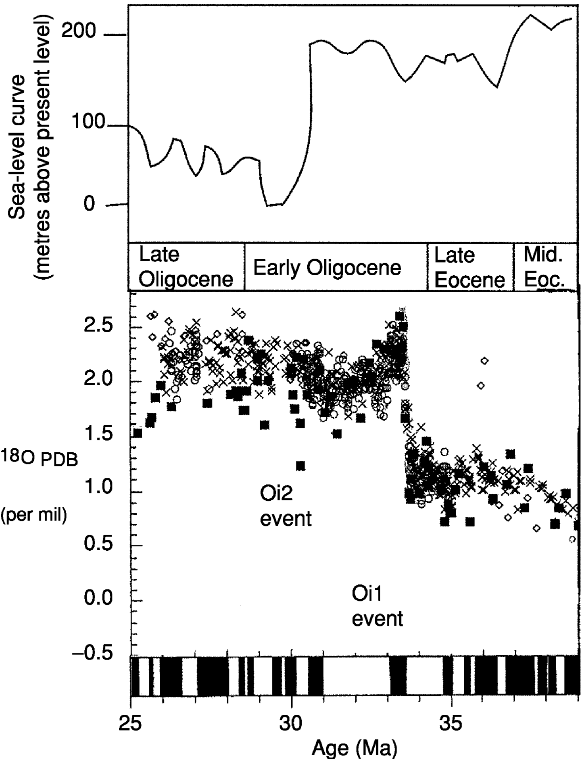


Figure 2 Details of oxygen isotope and sea-level record in the Oligocene (modified from Prothero and Dott, 2003, and Zachos *et al.*, 1999).

Antarctic margin are glacial in origin. In addition, there are even observations of sediments dropped by melting icebergs well out into the Southern Ocean.

What caused this global cooling and the extinctions in the Early Oligocene? A few geologists have suggested that the Late Eocene impact events, or major volcanic eruptions in the Ethiopian Plateau, might have been responsible, but these ideas are challenged by the stratigraphical sequence of events. As noted above, the impacts occurred in the Middle of the Late Eocene, about two million years before the Early Oligocene cooling and two million years after the End–Middle Eocene cooling. Likewise, the volcanic eruptions that formed the Ethiopian traps are now dated in the Late Oligocene, when no significant extinctions are recorded. For more than 30 years, the primary mechanism responsible for the Early Oligocene cooling has been identified as the development of the Circum-Antarctic current. Today, this current circulates in a clockwise direction around the Antarctic continent, forming a 'refrigerator door' that locks the cold temperatures formed on the poles. This current is one of the largest in the ocean, moving as fast as 25 cm sec. The volume of water that passes between Antarctica and Australia is about 233 million cubic metres per second, or more than 1000 times the flow of the largest river on Earth, the Amazon. The 'refrigerator door' also separates the polar currents from subpolar and temperate currents, so that each is isolated from the other. By contrast, in the Early Eocene tropical waters in the Atlantic and Pacific mixed all the way to the poles, decreasing the difference in temperature between the poles and the equator.

As these cold waters circulate around the poles, they sink, generating the Antarctic bottom waters. These cold, oxygenated waters then flow along the bottoms of the world's oceans, all the way to the northern hemisphere. The effect of this on global oceanic circulation and climate is enormous. Antarctic bottom waters contain up to 59% of the world's marine water, and transport this cold water all over the bottom of the ocean. This, in turn, increases the stratification of shallow- and mid-level currents in the ocean, further accentuating the climatic differences between pole and equator.

So what triggered the development of the Circum-Antarctic current? The most obvious factor is plate Tectonics (*see* **Plate Tectonics**). In the Late Cretaceous, Australia and South America were still attached to Antarctica as remnants of the ancient Gondwana supercontinent. As noted above, this caused the tropical currents to mix with polar currents along a much broader front since there were barriers between the Atlantic, Indian, and Pacific oceans, and no southern ocean. The climatic effect was expressed by moderate global temperatures. Geophysical evidence shows that these three southern continents began their separation in the latest Cretaceous, but they were not far apart enough to allow deep-ocean currents to pass through until the latest Eocene or Early Oligocene. By the Early Oligocene, deep-sea cores south of New Zealand reveal a blast of cold water was passing between Australia and East Antarctica. As the Oligocene progressed, the between-continent separation grew wider, resulting in the development of larger and more powerful cold-water currents. Originally, geologists thought that the separation between the tip of South America and the Antarctic Peninsula did not occur until the end of the Oligocene, but recent evidence has suggested that this passage also opened in the Early Oligocene. This implies that the entire Circum-Antarctic current developed in a relatively short period of geological time.

In addition to these important currents, it is also thought that another body of water, the North Atlantic Deep Water, which flows out of the Arctic Ocean past Greenland into the bottom of the North Atlantic, originated some time in the Oligocene. Thus, the global ‘icehouse’ conditions of the Oligocene can be largely attributed to the development of modern oceanic stratification and circulation patterns.

It has been suggested that a decline in greenhouse gases – especially CO_2 – might be a more important factor than the effects of oceanic circulation changes in steepening Oligocene climatic gradients. This interpretation is based on global circulation models of the atmosphere and oceans, coupled with simulation experiments in which the amount of atmospheric CO_2 was varied in the computer models. Certainly, there must have been a decline in CO_2 from the greenhouse world of the Cretaceous through Early Eocene to the icehouse of the Oligocene. But important questions have not yet been answered. Where is the global reservoir for all this carbon? There were no great bodies of unoxidized carbon locked up in coal deposits such as the one that terminated the Mid-Paleozoic greenhouse. Nor are there extensive Oligocene carbonates that might have locked up atmospheric CO_2. Those who favour this hypothesis suggest that the carbon was locked up in frozen methane hydrates on the sea-floor. Since the evidence of this frozen hydrate might not be preserved in the stratigraphical record, it is difficult to test this idea directly. Several other empirical studies also contradict this computer model. Chemical analyses of foraminifera suggests that Eocene CO_2 concentrations were not much higher than they are today. The number of stomata on the bottom of a leaf is strongly correlated with atmospheric CO_2 concentration, and it is observed that there is no evidence of higher CO_2 levels in Eocene leaf data. It has been argued that methane (CH_4), rather than CO_2, might have been the greenhouse gas responsible for Eocene warming. Given the clear evidence of oceanic circulation changes directly tied to sedimentological changes on Antarctica, and isotopic changes in the oceanic waters, there must have been a significant effect from the opening oceanic gateways and circulation changes.

The effects of these oceanic temperature changes are critical, not only to marine climates and organisms, but also to terrestrial climates as well. The most complete land-based record comes from North America, where palaeobotanical records show that mean annual terrestrial temperatures dropped 7–11°C in the earliest Oligocene. This is true of palaeofloral records all the way from Alaska and the Pacific North-west to the Gulf Coast. In addition to this rapid cooling, the record of ancient plants and soils also suggests that the continent underwent significant drying, with the establishment of much more seasonal, drought-prone climates. In the Big Badlands of South Dakota, Upper Eocene palaeosols (*see* **Soils:** Palaeosols) suggest over a metre of annual rainfall, supporting a dense forest. By contrast, in the Early Oligocene, mean annual rainfall was less than half a metre, and the vegetation was patchy scrubland with limited riparian forests. The land snails from the Badlands also change, from Late Eocene forms like those found today in tropical Central America, to Early Oligocene forms that are smaller and more drought-tolerant, and found today in Baja, California. In addition, the Late Eocene reptilian fauna that was dominated by crocodilians and pond turtles was replaced by dry land tortoises in the Early Oligocene.

Once the Early Oligocene climatic deterioration was completed, the Earth remained in this icehouse mode through the remainder of the Oligocene. The other significant Oligocene climatic event was several pulses of glaciation that occurred during the middle part of the Oligocene, about 30 Ma (the Oi2 event). Thick, extensive Mid-Oligocene glacial deposits are found throughout the Antarctic region, and benthic foraminiferal oxygen isotopes shifted by 1.6 per mil, suggesting another increase in ice volume and drop in global temperatures (Figure 2). As these ice-sheets grew, they pulled water out of the oceans, resulting in the largest drop in sea-level in the past 100 million years. Originally, it was suggested that sea-level dropped by almost 150 m, although more recent estimates suggest it was only half that amount (Figure 2). Whatever its magnitude, the Mid-Oligocene regression had a major effect on the shallow-marine realm, causing the continental shelves to become deeply incised once they were exposed to the subaerial erosion, and producing huge Mid-Oligocene unconformities in most marine rocks around the world.

The effect of the cooling and regression in the Middle Oligocene on land climates was less obvious. Sensitive tropical floral elements were already gone by the Mid-Oligocene, so the land plant record shows only minor cooling effects. The record of ancient soils from the Big Badlands of South Dakota shows that the climate became cooler and much drier, so that sand dune deposits became common in the Mid-west in the late Oligocene. These same soils suggest that vegetation was a mixture of scrublands and grasses, with few trees, by the Late Oligocene.

Oligocene Life

As the Eocene–Oligocene climatic deterioration began, the total diversity of land animals and both marine organisms decreased significantly from Eocene levels, reaching a Phanerozoic low in the Late Oligocene. The forests and jungles of the Early

Eocene were rapidly disappearing by the Late Eocene, so that by the Oligocene most of the temperate latitudes were covered by a mixture of forest and scrubland vegetation. This change in vegetation triggered by this cooling and drying was a major change in many of the land organisms. The Oligocene land-mammal fauna was dominated by primitive members of most living families. These included three-toed horses (which began to radiate into multiple lineages by the Late Oligocene), three different lineages of rhinoceroses, early camels, deer, and peccaries, as well as a handful of archaic land-mammal groups left over from the Eocene. Numerous modern carnivoran groups (especially early dogs, and the cat-like nimravids, as well as primitive members of the bear, weasel, and raccoon families) became the dominant predators as the last of the archaic carnivorous mammals (*see* **Fossil Vertebrates:** Placental Mammals), the creodonts, straggled on. On all the northern continents and Africa, rodents and rabbits both underwent a huge diversification as the niches for ground-dwelling seed-eaters increased, and the habitat for squirrel-like nut and fruit eaters diminished.

In Eurasia, many of the same trends were apparent. In the Early Oligocene, the Turgai Strait across the Obik Sea between Europe and Asia opened up, allowing Asian mammals (such as rhinoceroses and ruminants) to immigrate to Europe and drive many of the endemic natives to extinction. This Early Oligocene event is known as the Grande Coupure. However, there was only limited migration between Asia and North America via the Bering Strait. In Eurasia, the Oligocene saw a similar diversification of rhinoceroses (including one group, the giant indricotheres found in Mongolia and Pakistan, which reached 6 m at the shoulder and weighed 20 tonnes), plus some of the earliest members of the deer, giraffe, pig, and cattle families. Tree-dwelling mammals became much less common and vanished from many continents. For example, primates once flourished on all the northern continents during the Early and Middle Eocene. By the Oligocene, though, they became restricted to Africa and South America, where they evolved into Old World and New World monkeys, respectively. The remainder of the African fauna was also endemic to this island continent, which was not connected to Eurasia at the Arabian Peninsula until the Early Miocene. Instead, the African fauna was populated by archaic mastodonts, a wide diversity of hyraxes, and other peculiar endemic forms, such as the horned arsinoitheres. South America and Australia were also island continents, unconnected to the rest of the world, and each developed their own endemic faunas.

In the marine realm, the Early Oligocene extinctions triggered by global cooling were severe, causing major extinction in the planktonic and benthic foraminifera, and even in the planktonic algae (*see* **Fossil Plants:** Calcareous Algae) such as diatoms and coccolithophores. In the Gulf Coast of the US, 97% of the marine snail species and 89% of the clam species found in the Late Eocene did not survive into the late Early Oligocene, and over 50% of the sea urchins and sand dollars also became extinct. However, the overall taxonomic composition of the marine fauna remained essentially the same, with new species of clams, snails, and sea urchins replacing the extinct species (but at lower diversity), and making up the bulk of the fossilisable organisms in the Oligocene. By the end of the Early Oligocene, diversity was at an all-time low. Marine faunas were composed of groups tolerant of the cooler waters that began in the Oligocene. This is true especially in the molluscs (*see* **Fossil Invertebrates:** Molluscs Overview) of the Pacific Rim, which are mostly cold-water tolerant forms that migrated south to California from Alaska and Siberia during the Oligocene. Planktonic organisms were not only low in diversity, but occupied relatively few, simple biogeographic realms (since the area of the tropics had decreased), and evolved relatively slowly during the Oligocene.

Palaeogeography

With the Eocene separation of Australia from Antarctica and the collision of India with Asia, by the Oligocene most of the continents were approaching their present configuration. South America, however, would not finally separate from Antarctica until the beginning of the Oligocene, completing the breakup of Gondwana and opening the gateway for the full development of the Circum-Antarctic current. As discussed above, these continental movements brought about major changes in oceanic circulation, with the Circum-Antarctic current locking in cold conditions over Antarctica, initiating the first Antarctic ice-sheets, and also stimulating the flow of cold Antarctic bottom waters, which today control much of the world's oceanic circulation.

The growth of Antarctic glaciers meant that the high sea-levels of the Eocene greenhouse world were gone, and much of the seawater locked up in ice. The Late Oligocene regression turned the drowned coastal plains of the Atlantic and Gulf coast of North America into emergent floodplains, and the European archipelago largely dried up. This regression also dried up the Obik Sea and ended the separation of Europe and Asia. The Tethys Seaway was already partially disrupted by the collision of India with Asia, but global regression destroyed the remaining vestiges of this seaway and its unique tropical biota.

On land, the Himalayas continued to develop, and the Alps began to rise rapidly as Africa began to collide with Europe and close the Mediterranean. The Andes began to erupt huge volumes of volcanic rocks, forming a mountain chain for the first time. In North America, the Rocky Mountains were no longer rising, but they continued to soar high above the western part of North America. The basins between the ranges began to fill up with sediments and volcanic debris erupted from the arc volcanoes to the west. Volcanic activity on the western edge of the continent, which had ceased when the Laramide Orogeny shut off the Sierran-Sevier arc, resumed in the Oligocene. The new arc was much further east than the previous arc, running in an irregular belt from central Mexico to New Mexico (the Mogollon-Datil volcanics) to south-west Colorado (the San Juan volcanics), Utah, and Nevada, and then up through Oregon and Washington (the ancestral Cascades), and British Columbia. These explosive volcanic centres erupted huge amounts of ash, much of which blew east and helped bury the Laramide uplifts of the Rocky Mountains. Much of this ash and sediment also spilled over onto the High Plains, forming the thick Oligocene deposits of the White River and Arikaree Groups (entombing their excellent record of fossils (*see* **Fossil Invertebrates:** Echinoderms (Other Than Echinoids)) and climates (*see* **Palaeoclimates**) in the Big Badlands of South Dakota and adjacent states. At about 30 Ma, the corner of the Pacific Plate was subducted under California, so the San Andreas transform fault began its activity. This, in turn, set off a wide variety of geologic events, including the beginning of the spreading of the Basin and Range Province in Arizona, Utah, and Nevada; the end of the eruptions in Nevada and California; the clockwise rotation of the Sierra-Cascade arc to the south-west by over 400 km; and the clockwise rotation of many tectonic blocks, including the Transverse Ranges of California. By the Miocene, all of these regional events (Basin and Range extension, Sierran rotation; cessation of southern volcanism; and Transverse Ranges rotation) were well developed and approaching their modern condition.

See Also

Famous Geologists: Lyell. **Fossil Invertebrates:** Echinoderms (Other Than Echinoids); Molluscs Overview. **Fossil Plants:** Calcareous Algae. **Fossil Vertebrates:** Jawless Fish-Like Vertebrates; Placental Mammals. **Palaeoclimates**. **Plate Tectonics**. **Soils:** Palaeosols. **Tertiary To Present:** Eocene.

Further Reading

Barbera X, Carbrera L, Marzo M, Pares JM, and Agusti J (2001) A complete terrestrial Oligocene magnetobiostratigraphy from the Ebro Basin, Spain. *Earth and Planetary Sciences Letters* 187: 1–16.

Berggren WA, Kent DV, Swisher CC III, and Aubry M-P (1995) A revised Cenozoic geochronology and chronostratigraphy: *SEPM Special Publication* 54: 129–212.

Davies R, Cartwright J, Pike J, and Line C (2001) Early Oligocene initiation of North Atlantic deep water formation. *Nature* 410: 917–920.

Diester-Haass L and Zahn R (1996) The Eocene-Oligocene transition in the Southern Ocean: history of water masses, circulation, and biological productivity inferred from high resolution records of stable isotopes and benthic foraminiferal abundances (ODP Site 689). *Geology* 24(2): 16–20.

DeConto RM and Pollard D (2003) Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO_2. *Nature* 421: 245–249.

Exon N, *et al.* (2002) Drilling reveals climatic consequences of Tasmanian gateway opening. *EOS* 83(23): 253–259.

Meng J and McKenna M (1998) Faunal turnovers of Palaeogene mammals from the Mongolian plateau. *Nature* 394: 364–367.

Miller KG (1992) Middle Eocene to Oligocene stable isotopes, climate and deep-water history: the Terminal Eocene Event? In: Prothero DR and Berggren WA (eds.) *Eocene-Oligocene Climatic and Biotic Evolution,* pp. 160–177. Princeton, NJ: Princeton University Press.

Pearson PN and Palmer MR (1999) Middle Eocene seawater pH and atmospheric carbon dioxide concentrations. *Science* 284: 1824–1826.

Prothero DR (1994) *The Eocene-Oligocene Transition: Paradise Lost.* New York: Columbia University Press.

Prothero DR (ed.) (2001) *Magnetic Stratigraphy of the Pacific Coast Cenozoic.* Pacific Section SEPM Special Publication 91.

Prothero DR and Berggren WA (eds.) (1992) *Eocene-Oligocene Climatic and Biotic Evolution.* Princeton, NJ: Princeton University Press.

Prothero DR and Dott RH Jr (2003) *Evolution of the Earth,* (7th edn.) New York: McGraw-Hill.

Prothero DR and Emry RJ (eds.) (1996) *The Terrestrial Eocene-Oligocene Transition in North America.* New York: Cambridge University Press.

Prothero DR, Ivany LC, and Nesbitt ER (eds.) (2003) *From Greenhouse to Icehouse: The Marine Eocene-Oligocene Transition.* New York: Columbia University Press.

Retallack GJ (1983) Late Eocene and Oligocene palesols from Badlands National Park, South Dakota. *Geological Society of America Special Paper* 193.

Royer DL, Wing SL, Beerling DJ, Jolley DW, Koch PIL, Hickey LJ, and Berner RA (2001) Paleobotanical evidence for near present-day levels of atmospheric CO_2 during part of the Tertiary. *Science* 292: 2310–2313.

Sloan LC, Walker JCG, Moore TC Jr, Rea DK, and Zachos JC (1992) Possible methane-induced polar warming in the early Eocene. *Nature* 357: 320–322.

Wolfe JA (1978) A paleobotanical interpretation of Tertiary climates in the Northern Hemisphere: *American Scientist* 66: 694–703.

Wolfe JA (1994) Tertiary climatic changes at middle latitudes of western North America: *Palaeogeography, Palaeoclimatology, Palaeoecology* 108: 195–205.

Zachos JC, Opdyke BN, Quinn TM, Jones CE, and Halliday AN (1999) Eocene-Oligocene climate and seawater $^{87}Sr/^{86}Sr$: is there a link? *Chemical Geology* 161: 165–180.

Miocene

J M Theodor, Illinois State Museum, Springfield, IL, USA

Introduction

The Miocene (23.8–5.3 Ma) is the interval during which the world began to assume much of the configuration and topography we know today. Major tectonic changes in North America uplifted the Coast Ranges, formed the Basin and Range, and saw the evolution of the San Andreas Fault Zone. In Eurasia, the ongoing collision of India and Asia elevated the Tibetan Plateau, drastically altering global atmospheric circulation patterns. In South America, global atmospheric and oceanic circulation were greatly affected by the uplift of the Andes and the closure of the Panamanian seaway, in ways that persist to this day. The Miocene separation of Australia and Antarctica altered oceanic circulation patterns, changing the prevailing currents off Europe. In Africa, the development of the Great Rift Valley formed the environments that were home to the early hominids (*see* **Fossil Vertebrates:** Hominids). Much of the flora and fauna of the Miocene belong to groups that are still living today, but several groups experienced large radiations during this time interval, such as the whales and the snakes. Two major ecosystems – grasslands and kelp forests – evolved during the Miocene. These changes form the foundation for modern conditions.

The Miocene Epoch is the fourth subdivision of the Tertiary Period. In 1828, Sir Charles Lyell (*see* **Famous Geologists:** Lyell) noticed that different layers of rock in Europe contained different proportions of living and extinct species of marine molluscs. Concluding that the variation in this proportion indicated that the rock layers were formed at differing times, and assuming that the higher the proportion of living to extinct species, the more recently the layer was laid down, he defined three epochs – the Eocene (Dawn Recent), Miocene (Middle Recent), and Pliocene (More Recent), where the transitions between epochs were specified as a given percentage of living to extinct species.

Because the species composition of a rock layer can vary depending on the type of depositional environment, geologists no longer use this definition, and rely on other methods. Two specific, internationally recognized localities (also known as stratotypes) are currently designated as being the boundaries of the Miocene for international correlation. The base, or start, of the Miocene is defined at the Lemme Carrosio section in the upper part of the Rigoroso Formation in northern Italy. The base of the Pliocene – which is also, by definition, the end of the Miocene – is defined at the base of the Trubi Formation in the Evaclea Minoa section, on the southern coast of Sicily (*see* **Tertiary To Present:** Pliocene).

Geochronology

The European Miocene is generally divided into six marine stages (oldest to youngest): the Aquitanian, Burdigalian, Langhian, Serravallian, Tortonian, and Messinian (Figure 1). Terrestrial fossils in Europe have generally been associated with these marine stages as European marine rocks interfinger with terrestrial deposits, thus allowing a clear correlation between terrestrial and marine units. However, a number of more isolated fissure-fill localities are more difficult to correlate. As a result, two other systems based exclusively on terrestrial mammals have been developed, the European Land Mammal Ages (ELMAs) and the Mammalian Neogene Reference Level system (MN Zones).

In North America, there are fewer localities with interfingered marine and non-marine beds, but many units are associated with strata that can be radioisotopically dated. The majority of these deposits are correlated using mammalian biostratigraphic units. The North American Land Mammal Ages in use for the Miocene include the Arikareean, Hemingfordian, Barstovian, Clarendonian, and the Hemphillian. Most of these are well associated with radioisotopic dates and/or paleomagnetic assessments.

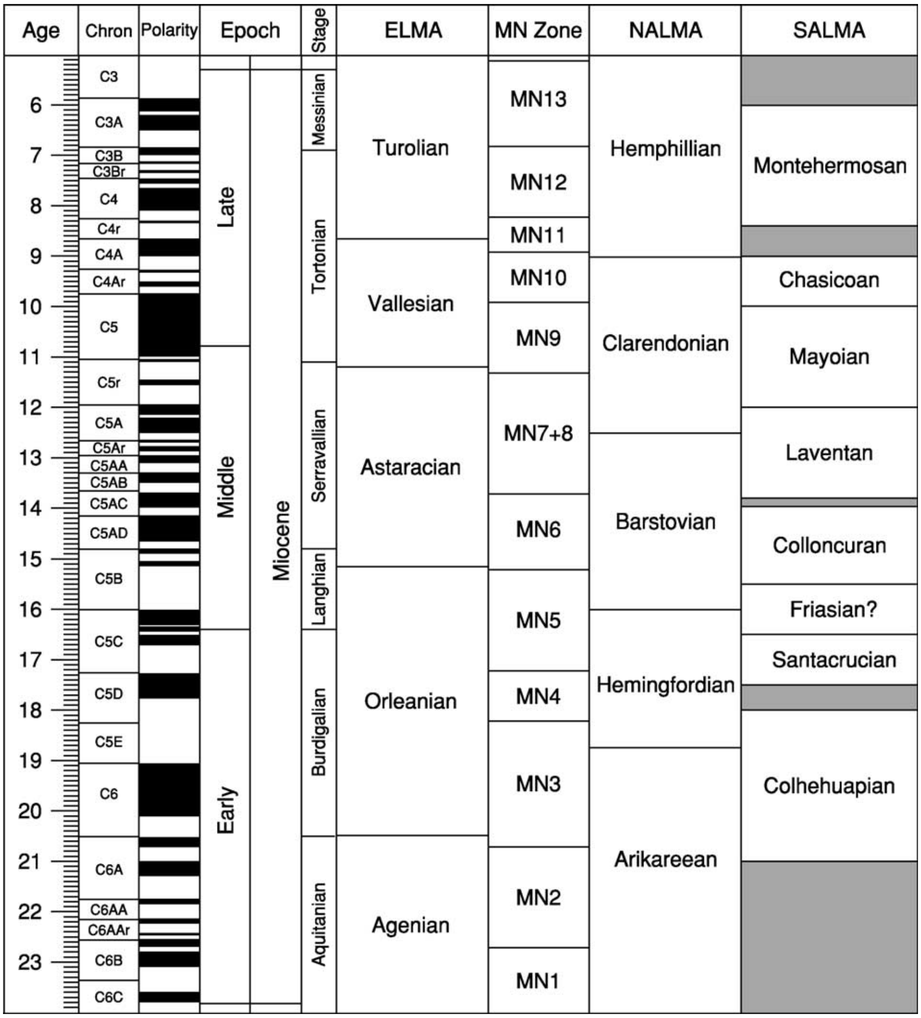


Figure 1 Chronostratigraphical correlation chart for the Miocene. ELMA = European Land Mammal Age, MN = Mammalian Neogene Reference Level Age, NALMA = North American Land Mammal Age, SALMA = South American Land Mammal Age.

Many regions are not as well described as Europe and North America. The South American Miocene Land Mammal Ages include the Colhehuapian, Santacrucian, Colloncuran, Friasian, Laventan, Mayoan, Chasicoan, Huayquerian, and Montehermosan. Some of these intervals are poorly constrained in terms of absolute ages. Asian and African faunas do not yet have a well-established system of named ages in use, but much work is being done to establish such a framework. The Australian Miocene can be correlated to the other continents using pollen and marine fossils, but the mammals are so different from those on other continents that they are not used in correlation outside of that continent.

Tectonics

Continental position During the Miocene, the continents moved into positions very close to their modern configuration (Figure 2). By the mid-Miocene, as Australia moved away from Antarctica, the modern oceanic circulation pattern was established. Miocene tectonic changes resulted in the creation of numerous important features of modern continental topography.

North America

North America underwent a complex sequence of tectonic events throughout the Cenozoic. A number of events occurred during the Miocene which combined to shape much of the landscape familiar to us today.

East Pacific Rise Spreading at the East Pacific Rise continued during the Oligocene, causing the complete subduction of the Farallon Plate near Los Angeles by 30.0 Ma, leaving two remnants to the north and south still undergoing subduction. By the Early

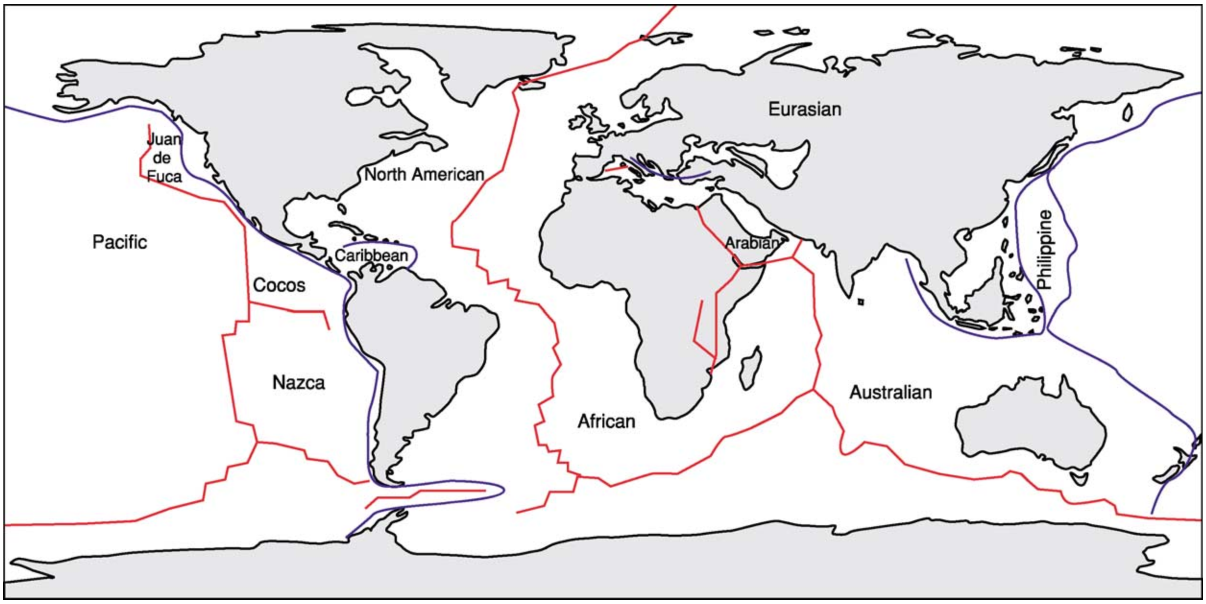


Figure 2 Palaeogeographical reconstruction of Miocene plate configurations and continent dispersions.

Miocene, most of the remaining Farallon Plate had been subducted under the North American Plate, bringing the Pacific Plate into contact with the North American Plate in southern California. The remaining two sections could no longer be considered a single plate by this time, and are henceforth regarded as independent plates with unique motions. The Juan de Fuca Plate is found to the north off the coast of Oregon and Washington and the Cocos Plate to the south off the coast of Mexico. Both continue to be subducted under the North American Plate. The contact of the Pacific, North American, and Juan de Fuca plates formed a triple junction at Mendocino, California by 20.0 Ma. During this transition, the subducting plate margin disappeared and the contact between the Pacific and North American plates developed into a transform fault system because the Pacific Plate is moving to the north, while the subducting Farallon Plate was moving more directly eastward. This transform fault system, which separates Baja California from the rest of California, runs under the continent between San Francisco and Los Angeles. The transform faulting formed along the Central Valley of California is known today as the San Andreas Fault Zone, and was in place by about 10.0 Ma.

Rocky Mountains The subduction of both the Farallon Plate and the more northerly Kula Plate are partially responsible for the character of the Rocky Mountains, which began their final uplift phase during the Miocene. The Farallon Plate was subducted under the North American Plate at an unusually shallow angle. This caused uplift of the Rockies in the western US and Mexico by compression, and probably caused some extension in the crust as the slab sank. By contrast, the Juan de Fuca and Kula plates were subducted at a much steeper angle (closer to 45°), resulting in more typical thrusted sedimentary sheets through the Canadian Rockies.

Basin and Range Changes in the continental crust to the west of the developing Rocky Mountains are poorly understood, but the thin crust in this region is evidence that this area became extended between 15.0 and 8.0 Ma. Crustal extension resulted in a series of faults that trend north–south, with dropped fault blocks (grabens) forming flat valley floors in between raised blocks (horsts). Erosion of the nearly 400 alternating mountain blocks formed the characteristic basin and range topography of the modern western United States.

Columbia River Flood Basalts From 17.0 to 15.5 Ma, a tremendous volume of volcanic rock was deposited in north-eastern Oregon, south-eastern Washington, and western Idaho, from at least 300 lava flows, some of which reached as far as the Pacific Ocean. The Columbia River Flood Basalts cover 164 000 km^2 and are situated between the Rocky Mountain and Cascade Ranges. The cause of these volcanic eruptions is complex and is tied to: (i) extension and thinning of the crust; (ii) formation of the Cascade Range; and (iii) the position of the Yellowstone hotspot (a mantle plume). Most flood basalt provinces are directly tied to a mantle hotspot, but the Yellowstone hotspot was 300–400 km south of the flood basalt vents, and the basalt composition differs from classic hotspot flood basalts, indicating a more complex tectonic origin. The numerous basalt beds allow for precise radioisotopic age assessments for this area during the Middle Miocene.

Cascade and Sierra Nevada Ranges There is some evidence that these volcanic mountain ranges experienced a pulse of uplift beginning in the Late Miocene. Because there is also some evidence of uplift in the Coast Ranges at this time, it is possible that there was uplift of the entire north-western Cordilleran coast range. This uplift might have been a result of changes in plate motion, or of plate delamination and subsequent crustal rebound.

Central and South America

Two major events affected the topography of Central and South America during the Miocene, both of which had profound climatic and biogeographic effects: the beginning of closure of Panamanian seaway and the uplift of the Andes.

Isthmus of Panama The closure of the Panamanian seaway began at almost 13.0 Ma, with the gradual emergence of the Isthmus of Panama as a result of ongoing changes in tectonics and sea-level. The emergence of the isthmus resulted in separation of North and South America, and resulted in profound changes in Caribbean oceanic circulation in the Pliocene, cutting off circulation between the Atlantic and Pacific oceans. These drastic changes in oceanic circulation resulted in massive changes in oceanic heat transfer. The isolation of the Atlantic Ocean also helped produce the warm Gulf Stream current that today maintains warmer temperatures in northern Europe relative to other locations at the same latitude. It is probable that this change also affected northern-hemisphere glaciation.

Andes The major topographic feature of South America is the Andes Range, which provides a barrier to atmospheric circulation and creates a large rain shadow desert on its eastern margin. The Andes are unique because they formed at a non-collisional plate boundary. Crustal compression of the Andes region began in the Triassic, but by in the Early Miocene (~20.0 Ma) the majority of the range had reached only 25% of current elevation. By the later Miocene, the range had reached ~50% of the modern elevation, and uplift continued. The tectonic history of the Andes is complex and not yet well understood (*see* **Andes**).

Eurasia

The tectonic history of Eurasia is somewhat less complex than that of North America. Nevertheless there were several important Eurasian tectonic events that occurred during the Miocene.

Himalayas and the Tibetan Plateau Earlier in the Cenozoic, the northward plate motion of India brought that continental plate into contact with Eurasia (*see* **Indian Subcontinent**). The resulting continent-continent collision continued through the Miocene, causing uplift of the Himalayas, and, beginning in the Early–Middle Miocene, resulted in uplift of the Tibetan Plateau. This uplift exposed a considerable volume of carbonate rock to weathering, which is thought have affected global atmospheric CO_2 levels.

Messinian salinity crisis Considerable evidence from thick deposits of halite and anhydrite beneath the Mediterranean Sea indicate that, beginning at 5.96 Ma, the Mediterranean basin was repeatedly cut off from oceanic input and became a drying salt lake (Lago Mare) system. This dry-Mediterranean interval is termed the Messinian Salinity Crisis. It continued until 5.33 Ma, when the modern opening at the Strait of Gibraltar restored marine conditions. The cause of this Messinian Salinity Crisis is unclear. The likely major causes include several tectonic changes and a global decline in sea level of ~60 m as a result of glaciation. The most recent data favour a tectonic explanation, in which changes in the subduction of oceanic lithosphere of the Tethyan Ocean caused upwelling in the asthenosphere, resulting in uplift and consequent isolation of the Mediterranean Basin (*see* **Europe:** Mediterranean Tectonics).

Africa

During the Miocene, Africa, which had previously been relatively tectonically stable, underwent a series of tectonic changes as a result of the formation of a new spreading ridge running through the eastern edge of that continent. In the early part of the Miocene, the Afro-Arabian Plate moved north, contacting the Eurasian Plate. This contact reduced the size of the Mediterranean and may have contributed to the Messinian Salinity Crisis by reducing the amount of water available for rainfall and thus increasing aridity.

By the Early Miocene, uplift had begun in southern and eastern Africa, creating a rain shadow over eastern Africa. Around 18.0 Ma crustal extension from the spreading ridge began to take place along the East African Rift Valley, which forms one arm of the Afar Triple Junction – the other arms being the Gulf of

Aden and the Red Sea. Extension caused faulting and the formation of a horst-graben topography in the rift valley. This process fragmented the habitats of the rift valley, breaking up forest, increasing the diversity of the terrain and vegetation, and leading to the formation of numerous lakes along the valley floor. These lakes, such as Lake Turkana, provided drinking water and habitat for the early ancestors of humans. The lakes also provided better opportunities for preservation in the fossil record, which is much less likely in densely vegetated forest.

Climate

Sea-level Miocene sea-levels were generally higher than those of the modern day, and are tied to glaciations in Antarctica. During the Miocene, oceanic circulation patterns were strongly affected by the evolution of the circum-Antarctic current as Antarctica separated from Australia. This separation thermally isolated Antarctica and drastically altered global marine and atmospheric circulation.

In North America, much of the south-eastern United States was covered by shallow seas. The Early Miocene saw three major episodes of deposition of oceanic phosphorus, especially in eastern North America, when much of the eastern coast was underwater. These periods (21.0 to 20.0, 19.0 to 18.0, and 17.0 to 16.0 Ma) are likely to have been responses to global changes in sea-level, oceanic upwelling, and deep water currents, and may be related to the ongoing uplift and erosion of the Tibetan Plateau. Later in the Miocene (~6.0 Ma) sea levels dropped as much as 60 m because of glaciation, a drop which probably contributed to the Messinian Salinity Crisis.

Atmospheric CO_2 Early models of atmospheric CO_2 indicated that levels in the earlier Miocene were roughly double those of modern, pre-industrial times. However, more recent measurements of geochemical signatures (e.g., boron isotopes, plant alkenones) from marine sediments indicate that Early Miocene CO_2 levels were much lower. This is important because the hypothesised decline in the middle Miocene CO_2 levels have been thought responsible for the rapid spread of grasses that use the C4 photosynthetic pathway at that time. Plants that use only one enzyme to bind CO_2 in the chemical reactions of photosynthesis are termed 'C3' plants. However, this enzyme will also bind O_2, forming a compound that is harmful to the plant. At low concentrations of atmospheric oxygen this is not a significant disadvantage to C3 plants. The C4 pathway prevents these compounds from forming, but produces slightly less energy for the plant. At lower CO_2 levels, the C4 photosynthetic pathway is favoured, and it is generally assumed that the C4 pathway would only have evolved in intervals with lowered CO_2. The evolution of the C4 pathway allowed grasses to spread greatly during the Miocene, forming extensive grassland ecosystems for the first time.

Atmospheric CO_2 is a product of a number of different factors. One factor that is hypothesised to have affected global Miocene CO_2 levels is the uplift and subsequent erosion of the Tibetan Plateau, which is thought to have consumed massive amounts of atmospheric CO_2 as a result of limestone weathering.

Temperature and seasonality Global temperatures and seasonality are inferred from three sources: the ratio of oxygen and carbon isotopes measured in the skeletal fossils of benthic and microscopic planktonic organisms living in the ocean, and from the types of land plant fossils and the shapes of their leaves. Ice-rafted debris in the South Pacific indicates that Antarctic glaciation intensified at ~24.0 Ma. This was followed by a long warming interval, reaching a climax (Mid-Miocene Climatic Optimum) at ~15.0 Ma, with some brief episodes of glaciation in Antarctica. The warm interval was generally a continuation of Oligocene warm, temperate to sub-tropical conditions. Evidence from terrestrial floras indicates a mean annual temperature range from 11°C to 17°C in the northern United States during the Climatic Optimum, with the first grass macrofossils being found in the Great Plains. Following the Climatic Optimum, global temperatures decreased as the Antarctic ice sheets returned, and aridity increased in terrestrial ecosystems. In North America, temperatures warmed again ~8.0 Ma. A major glacial advance marked the end of the Miocene, resulting in lowered sea-levels.

Marine Life

Although the species found in the Miocene are largely extinct, most of the residents of Miocene oceans would be familiar to us, as they belong to modern groups of organisms (e.g., corals, bivalves, sharks, whales, sea cows). One major change in the oceans during the Miocene was the evolution of the kelp forest ecosystem. Fossil kelps and associated animals are known from the mid-Miocene Monterey Formation in southern California, their first appearance in the fossil record. Another important ecological development was the evolution of algal ridge formation on coral reefs by coralline algae. Earlier, coralline algae had formed parts of coral reefs, but had not formed these ridges, which protect living reef systems from heavy surf conditions. The appearance of the algal ridge systems allowed the expansion of coral reef systems into more coastal environments. During the Miocene, there was also a large radiation of species of whales and dolphins, including the first appearance of the sperm and baleen whales.

Terrestrial Life

Most of the animals present in the Miocene in North America and Eurasia, and to some extent, Africa, would be generally familiar to the modern eye, as they also belong to extant groups, even though the particular species are extinct. However, many groups were living in regions they no longer occupy today. For example, although North America lacked many groups of mammals currently found there (e.g., pigs), various species of camels and proboscideans were common components of North American Miocene faunas. These three continents had sufficient connections so that animals and plants could migrate between them at various times during the Miocene. By contrast, Australia and South America were island continents during this time interval, and their vertebrate faunas evolved in virtual isolation. Large groups of species that evolved on these continents are extinct, with no extant relatives.

Plants

In the earlier part of the Miocene, up to the time of the mid-Miocene Climactic Optimum, broad-leafed evergreen vegetation and coniferous forest were broadly distributed, and broad-leafed deciduous forest was reduced in area. Cooling and drying in global climates resulted in the establishment and spread of savannahs and grasslands at the expense of forested ecosystems.

By 20.0 Ma, all the families of plants present in the fossil record belong to families still living today. By 10.0 Ma all the modern genera were present. Thus, the Miocene vegetation was composed of familiar elements, although the vegetational structure and distribution were different from their modern states.

Early in the Miocene, forests were widespread. As climate warmed, up until the time of the mid-Miocene Climactic Optimum, broad-leafed evergreen forests (similar to the type found in modern tropical forests) spread to higher latitudes. During the subsequent cooling, broad-leafed evergreen forests contracted their range, giving way to the spread of broad-leafed deciduous forests, and in higher latitudes, evergreen coniferous forests.

Although grass pollen is present in the fossil record much earlier than the Miocene, fossil remains of grass plants are rare elements of the flora until the Middle Miocene. During the Middle and Late Miocene, species diversity in the grasses increased dramatically, and evidence of more open savannah ecosystems appeared in arid regions. The greatest increases in grass diversity were among the groups of species utilising a C4 photosynthetic pathway. This has been interpreted as response to increasing aridity during the later Miocene and possibly as a response to decreasing levels of atmospheric CO_2. The spread of grasslands was facilitated by climactic cooling and drying, particularly as seen in rain shadows created by uplift in the Cordilleran regions of the Americas.

Animals

The fossil record of terrestrial animals in the Miocene is variable. The limited fossil record of insects shows that most of the modern genera were present by the Miocene.

The reptile fauna worldwide was generally similar to that of today, although the Miocene is characterized by a major radiation in the snakes. This snake radiation seems to be a response to the diversification of small mammals, especially murid rodents (rats and mice) during this time.

The Miocene also saw diversification of many groups of birds, especially the dominant modern group of songbirds, the passerines. The modern genera of owls also appeared during the Miocene, along with the appearance of daytime predators such as falcons, hawks, and eagles. Several species of very large vulture-like predators (vultures and teratorns) appeared in the Miocene in the New World, probably evolving in response to the evolution of large herbivores living in open grasslands.

In the island continents of South America and Australia, a number of flightless forms evolved, and in South America one family, the phorusrhachids, was an important terrestrial predator. By the Middle Miocene, the modern ostrich *Struthio* had appeared in Africa.

North America The Miocene mammalian history of North America is documented by a rich and well-dated fossil record, including 31 families of mammals. In the Early Miocene, faunas are dominated by ungulates: the even-toed artiodactyls, including oreodonts, camels, and peccaries, and a new family, the cervids (deer and their allies); and the odd-toed perissodactyls, with numerous species of horses, rhinos, tapirs, and a now-extinct family, the large, clawed chalicotheres. Carnivores increased in diversity in the Early Miocene, with increases in the canid (dog), ursid (bear), and amphicyonid (extinct bear-dog) families. Rodent diversity increased in the geomyids (gophers), sciurids (squirrels), and castorids (beavers). At the time of the Middle Miocene Climatic Optimum, ungulate diversity was much higher than it is today, with a much higher proportion of browsing (leaf-eating) species than known in any ecosystem today, where ungulate faunas are dominated by grazing (grass-eating) mammals. By the Late Miocene, many browsing ungulate species had become extinct.

The Miocene was also punctuated by interchanges with Eurasia. In the Early Miocene, the first of two pulses of immigration across the Bering Land Bridge brought antilocaprids (pronghorn) and mustelids (weasels and skunks) to North America. The first proboscideans (elephant relatives) appeared ~16.0 Ma, and 11.0 Ma the horse *Cormohipparion* arrived in Europe from North America.

Eurasia The Early Miocene mammalian faunas of Europe are not well differentiated from earlier Oligocene faunas. Proboscideans immigrated to Europe from Africa by 18.0 Ma, and by 17.5–15 Ma dispersal routes connected Europe with Asia, Africa, and North America, bringing the horse *Anchitherium* and the chalicotheres to Europe in the Early Miocene. A major immigration (~15.0 Ma) brought the bovids (cattle) and suids (pigs) to western Europe from western Asia and Africa. The later Miocene of Eurasia is characterized by a diverse open country woodland fauna, which included horses, rhinos, bovids, felids (cats), hyaenids (hyaenas), and proboscideans.

The fauna from southern Asia is known primarily from the Siwalik sequence of Pakistan, and differs from European and western Asian faunas. The Pakistani faunas contain more primates and omnivorous artiodactyls, and fewer browsing ungulates than in western Asia, and a number of archaic groups persisted there. About 9.5 Ma, large giraffids (giraffes), suids (pigs), and new horses emigrated from Eurasia. Around 7.4 Ma, the fauna experienced turnover, and at the same time, soil carbonate measurements indicate a shift in the ecosystem away from plants using the C3 metabolic pathway to grasses using C4 pathways, indicating a shift in environment to a drier climate.

Africa The Early Miocene of Africa was a time of faunal transition, as 29 new families appeared during this time interval. The fauna at this time consisted of endemic proboscideans, creodont carnivores, hyraxes, giraffids, bovids, and aardvarks evolved early in the Miocene. Perissodactyls arrived from Eurasia in the form of rhinos and chalicotheres; archaic suid artiodactyls (pigs) represented the artiodactyls, with viverrids (civets) and felids representing the carnivores.

During the Middle Miocene, the bovids radiated and came to dominate the African ungulate fauna, and the hippos evolved. The first horse, *Anchitherium*, emigrated from Eurasia, along with mustelid and hyaenid carnivores. The Late Miocene saw extensive interchanges with Eurasia.

South America The faunas from South America are comparatively alien, consisting of diverse marsupials, including large marsupial carnivores (borhyaenids), xenarthrans (armadillos, glyptodonts, sloths) and a large number of extinct endemic animals, mostly ungulates, including the large toxodonts, which resemble hippos, rabbit-like hegetotheres, and the camel and horse-like litopterns. In the Early Miocene, two groups entered South America, probably by waif dispersal from Africa. These were the New World monkeys and the caviomorph rodents (guinea pigs, porcupines). Rodents were the only placental mammals in South America until the Pliocene, and they radiated into 16 families. Six to eight million years ago the first true carnivore, a procyonid (raccoon) entered South America, and around the same time, a sabre-toothed marsupial carnivore, *Thylacosmilus*, evolved. This fauna was to change tremendously in the Pliocene, when full-scale interchange began with North American faunas.

Australia As in South America, the Australian fauna was entirely isolated from other continents, and a highly endemic fauna evolved from the original mammalian stock, consisting of marsupials and monotremes. During the Miocene, the climate in Australia was much wetter than it is today, as evidenced by the Riversleigh fauna. Riversleigh dates to ~15.0 Ma, and the fauna includes the marsupial wolf (thylacine) and marsupial lion *Wakaleo*, quolls, possum-like marsupials, early kangaroos and diprotodonts, a group of large herbivorous marsupials. Wombats and koalas are first known from the Miocene. Around 6.0 Ma, Australian environments became much more arid, and many animals like the koalas adapted to eating tougher vegetation.

Glossary

C4 photosynthetic pathway This photosynthetic pathway is used by many grass species. In most plants, called C3 plants, CO_2 is fixed in a chemical reaction using an enzyme called Rubisco. This enzyme binds CO_2 into a compound used in forming sugars during photosynthesis. Rubisco will also bind O_2, forming a compound that is toxic to the plant, a process known as photorespiration. When atmospheric oxygen levels were relatively low, the poor selectivity of Rubisco did not pose a problem for C3 plants, but as carbon dioxide levels declined relative to oxygen levels, photorespiration would have posed a challenge. C4 plants have evolved a mechanism where an additional reaction using a more selective enzyme binds CO_2 into another compound, which is sequestered in the plant tissues. This compound can then be broken down and the CO_2 passed to Rubisco away from the presence of oxygen, so that little photorespiration can occur.

Chalicotheres An extinct group of perissodactyls, in the family Chalicotheriidae. These animals were very large, with clawed feet. Some had very long forelimbs, and were probably bipedal, feeding on leaves from tall trees. They are known from the latest Eocene up to the Pleistocene.

Creodont carnivore The creodonts are an extinct group of carnivorous mammals. It was long thought that they were ancestral to the modern order Carnivora, but more recent work indicates they may not be close relatives. They looked somewhat dog-like, with slicing carnassial teeth similar to those of Carnivora. However, in Carnivora the carnassial teeth are the last upper premolar and first lower molar, and in creodonts the carnassial teeth are the upper first and/or second molars and the lower second or third molar.

Rain shadow Dry region on the leeward side of a mountain or mountain range. The rain shadow effect is caused when moist air rises against the windward slope and is lost to precipitation on the windward side, leaving little moisture on the leeward side.

Ungulate Hoofed mammals, such as the ariodactyls (cows, pigs, sheep, deer, etc.) or perissodactyls (horses, rhinos). The last phalanx, or ungual phalanx, of each toe in ungulates is modified with a thick hoof instead of a claw.

Viverrids The members of the extant carnivoran family Viverridae, the civets and mongoose, known today in Africa and Asia. Viverrids are small, arboreal and generally fruit-eaters.

Waif dispersal An unusual form of inter-continental dispersal across large bodies of water by terrestrial animals. In an episode of waif dispersal, one or several animals inadvertently travels between two continents on vegetation rafts, and populates the new region on arrival. Waif dispersal may be responsible for the arrival of monkeys into South America.

See Also

Analytical Methods: Geochronological Techniques. **Andes**. **Atmosphere Evolution**. **Europe:** Mediterranean Tectonics. **Famous Geologists:** Suess. **Fossil Vertebrates:** Mesozoic Mammals; Placental Mammals; Hominids. **Indian Subcontinent**. **Stratigraphical Principles**. **Tertiary To Present:** Oligocene; Pliocene.

Further Reading

Berggren WA, Kent DV, Swisher CC III, and Aubry M-P (1995) A revised Cenozoic geochronology and chronostratigraphy. In: Berggren WA, Kent DV, Aubry M-P, and Hardenbol J (eds.) *Geochronology, Time Scales, and Global Stratigraphic Correlation: Unified Temporal FrameWork for an Historical Geology,* pp. 129–212. Tulsa: Society for Sedimentary Geology. SEPM Special Publication 54.

Flower BJ and Kennett JP (1993) Middle Miocene ocean-climate transition: high resolution oxygen and carbon isotopic records from Deep Sea Drilling Project site 588A, southwest Pacific. *Paleoceanography* 8: 811–843.

Jackson JBC, Budd AF, and Coates AG (eds.) (1996) *Evolution and Environment in Tropical America.* Chicago: University of Chicago Press.

Jacobs BF, Kingston JD, and Jacobs LL (1999) The Origin of Grass-Dominated Ecosystems. *Annals of the Missouri Botanical Garden* 86: 590–643.

Janis CM, Scott KM, and Jacobs LL (eds.) (1998) *Evolution of Tertiary Mammals of North America: Volume 1 Terrestrial Carnivores, Ungulates, and Ungulatelike Mammals.* Cambridge: Cambridge University Press.

Miller KG, Wright JD, and Fairbanks RG (1991) Unlocking the icehouse: Oligocene-Miocene oxygen isotopes, eustasy, and margin erosion. *Journal of Geophysical Research* 96(B4): 6829–6848.

Pazzaglia FJ and Kelley SA (1998) Large-scale geomorphology and fission-track thermochronology in topographic exhumation reconstructions of the Southern Rocky Mountains. *Rocky Mountain Geology* 33(2): 229–257.

Raymo ME (1994) The Himalayas, organic carbon burial, and climate in the Miocene. *Paleogeography* 9: 352–404.

Rössner GE and Heissig K (eds.) (1999) *The Miocene Land Mammals of Europe.* Munich: Verlag Dr. Friedrich Pfeil.

Scotese CR, Gahagan LM, and Larson RL (1989) Plate tectonic reconstructions of the Cretaceous and Cenozoic ocean basins. *Tectonophys.* 155: 27–48.

Vickers-Rich P, Monaghan JM, Baird RF, and Rich TH (eds.) (1991) *Vertebrate Paleontology of Australasia.* Melbourne: Pioneer Design Studio.

Woodburne MO (ed.) (1987) *Cenozoic Mammals of North America: Geochronology and Biostratigraphy.* Berkeley: University of California Press.

Zachos JC, Pagani M, Sloan L, Thomas E, and Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292: 686–693.

Pliocene

C Soligo, The Natural History Museum, London, UK

Introduction

The Pliocene is the second and terminal epoch of the Neogene period. It marks the end of the Tertiary and precedes the Pleistocene, the first epoch of the Quaternary period. The term Pliocene (from the Greek *pleiõn* 'more' + *kainos* 'new') was first proposed in 1833 by Sir Charles Lyell (*see* **Famous Geologists:** Lyell) in his *Principles of Geology*. His *Newer Pliocene* was characterized by a molluscan fauna, of which 90% of species were then extant, contrasting with the *Older Pliocene,* from which only 33–50% of molluscan species had survived. Lyell subsequently renamed his *Newer Pliocene* the Pleistocene.

The Pliocene was a time of substantial climatic change and, most significantly, saw the onset of northern hemisphere glaciation, marking the start of the modern-day cycles of glacials and interglacials. In contrast, the Pliocene was also the last period that sustained global temperatures that were (over a prolonged period of time) higher on average than at present. Fauna, flora and geography were also similar to the modern day. As such the Pliocene epoch holds particular promise for the modelling of potential effects of global warming on the modern world.

The tectonic processes and climatic variability that characterized the Pliocene had a lasting impact on the distribution and evolution of plants and animals. The emergence of a land bridge between the North- and South-American continents resulted in extensive biotic movements and profoundly changed the floral and faunal characteristics of South America. In Africa, meanwhile, the Pliocene saw the diversification of our evolutionary lineage, the Homininae, and the emergence of our own genus, *Homo* (*see* **Fossil Vertebrates:** Hominids).

Definition

The Mediterranean region has traditionally been the focus of stratigraphic research aimed at defining the boundaries of the Pliocene Series and its stages. Sometime towards the end of the Miocene, the Mediterranean Sea became isolated form the Atlantic Ocean. As a result, most of the Mediterranean basin dried out, an event that has been termed the Messinian salinity crisis. The start of the Pliocene is defined to coincide with the time when oceanic waters burst through the area of today's straits of Gibraltar, refilling the Mediterranean basin in a sudden catastrophic flooding event. The abrupt transition from evaporites, precipitated from the hypersaline waters of the drying Mediterranean, back to marls, representative of deposition under normal open-marine conditions, marks the stratigraphic start of the Pliocene series. The series is thus unusual in that its base is not located within a continuous marine sedimentation record, but rather marks the sudden return of open-marine sedimentation following a period of non-marine sedimentation. Since the formal acceptance and ratification of the *Gelasian* in 1996, the Pliocene series now consists of three stages (Table 1).

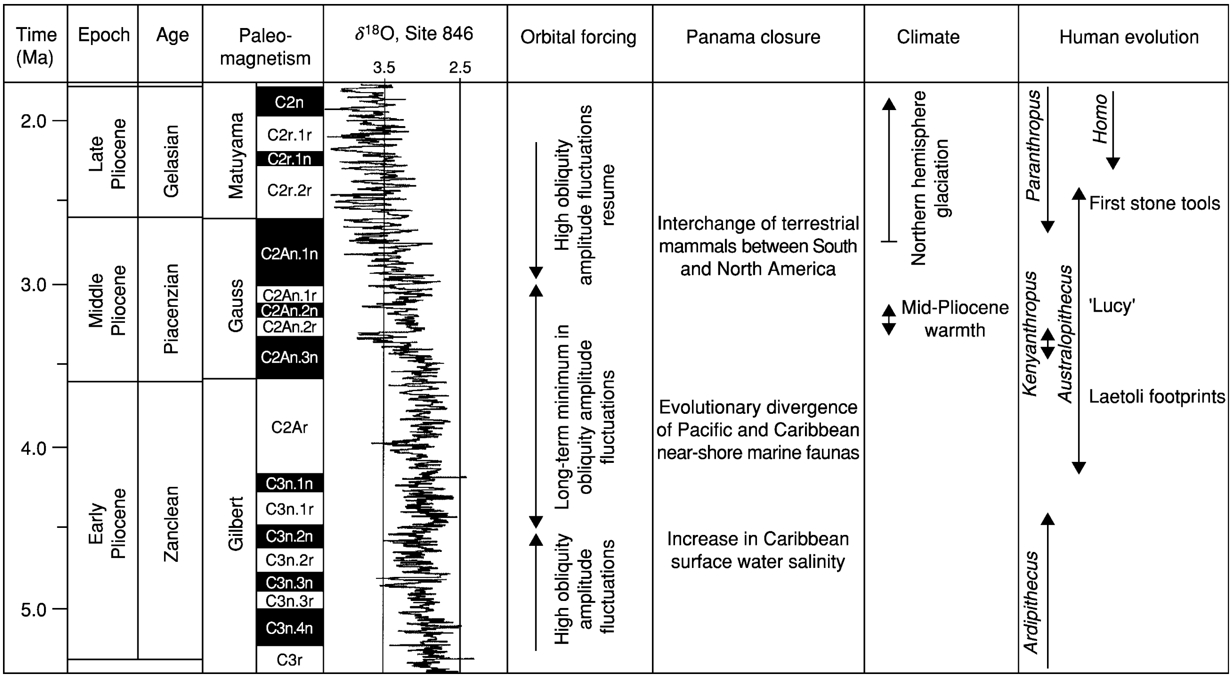
Zanclean (Lower Pliocene)

The Zanclean (from Zanclea, the classical name of Messina, Sicily) is the lowest stage of the Pliocene series. The Global Standard Stratotype section and Point (GSSP) for the Zanclean is located in the Eraclea Minoa section on the southern coast of Sicily (37°23′30″N; 13°16′50″E). The basal contact of this stage is where white Trubi Marl rests on dark brown Arenazzolo sands and marls. The base of the carbonate bed marking the small-scale Cycle 1 at the base of the stage corresponds to insolation Cycle 510 counted from the present, with an astrochronologic age of 5.33 Ma. Other tools for the global correlation of the base of the Zanclean stage include the base of the Thvera magnetic event (C3n.4n), dated to 5.236 Ma, coccoliths (the first occurrence of *Ceratolithus acutus* 5.37 Ma, the last occurrence of *Triquetrorhabdulus rugosus* 5.23 Ma and the last occurrence of *Discoaster quiqueramus* 5.537 Ma), and foraminifers (the first occurrences of *Globorotalia tumida* and *G. sphericomiozea* 5.6 Ma).

Piacenzian (Middle Pliocene)

The GSSP for the Piacenzian stage (after the town of Piacenza, Northern Italy) is located in the Punta Piccola section on the southern coast of Sicily (37°17′20″N; 13°29′36″E). The base of this stage is the base of the beige marl bed of the small-scale carbonate Cycle 77 and corresponds to the precessional excursion 347 counted from the present. Its astrochronologic age is estimated at 3.60 Ma. Further tools for the global correlation of the Piacenzian base include the Gilbert–Gauss magnetic reversal dated to 3.596 Ma and the location of the base within the obliquity-related δ^{18}O stage MG8 (O–176). Other correlation tools are geographically more restricted: within the Mediterranean region

Table 1 Composite table of Pliocene trends and events. $\delta^{18}O$, Site 846, from Shackleton *et al.*, 1995. *Proc ODP, Sci Results* 138: 337–355



the temporary disappearance of *Globorotalia puncticulata* (3.57 Ma), the first influx of *G. crassaformis*, the end of the paracme interval of *Discoaster pentaradiatus* (3.56 or 3.61 Ma) and the last occurrence of *Sphenolitus* spp. (3.73 or 3.70 Ma); in low- and mid-latitudes outside the Mediterranean, the last occurrences of *Globorotalia margaritae* (3.58 Ma) and of *Pulleniatina primalis* (3.65 Ma).

Gelasian (Upper Pliocene*)*

The Gelasian (after the town of Gela, Sicily) has only recently been ratified as the uppermost stage of the Pliocene series thereby establishing a threefold division of the Pliocene. The GSSP for the Gelasian stage is located in the Monte San Nicola section around 10 km N-NW of Gela in southern Sicily (Italy). The defined base of the Gelasian is the base of the marly layer overlying the Mediterranean Precession Related Sapropel (MPRS) 250 at 62 m in the Monte San Nicola section. The astrochronologic age of MPRS 250 (corresponding to precessional cycle 250 from the present) is 2.588 Ma. Other correlation tools include the Gauss–Matuyama boundary 20 Ky below the Gelasian base, nannofossils (last occurrences of *Discoaster pentaradiatus* and *D. surculus* 80 Ka above the base in most low- and mid-latitude areas) and foraminifers (last occurrence of *Globorotalia bononiensis* [*G. puncticulata* of some authors] 140 Ka above the boundary in the Mediterranean and North Atlantic). The last occurrence of *Stichocorys peregrina* (Radiolaria), the first occurrence of *Nitzschia joussaea* (Bacillariophycaceae/Diatomeae) in low-latitudes and the last occurrence of *Denticulopsis kamtschatica* (Bacillariophycaceae/Diatomeae) in the North-Pacific mid- and high-latitudes all approximate the Gauss–Matuyama boundary. The base of the Gelasian predates by 60 Ky the isotopic cold stage 100, which is marked by an increase in ice-rafted detritus in northern latitude oceanic sediments and the beginning of loess sedimentation in China. This base is also marked by changes in vegetation distribution patterns and continental vertebrate migrations (see below).

The end of the Pliocene series, of the Neogene and of the Tertiary period is determined by the Pliocene–Pleistocene boundary. The formal definition of that boundary has been the source of intense debate amongst stratigraphers. The currently validated Global Standard Stratotype Sections and Points (GSSP) for the base of the Pleistocene is located in the Vrica Section 4 km south of Crotone in southern Italy, just above the Olduvai subchron at an age of ~1.8 Ma.

Tectonics

The gradual formation of the isthmus of Panama(the land bridge connecting South and North America) cut off the Pacific Ocean from the Atlantic during the course of the Pliocene. The resulting rearrangement of ocean currents profoundly influenced global climates. The isthmus formed as a result of a complex combination of lithospheric plate movements, which involved a north-western drift of the South American Plate and eastern drift of the Caribbean Plate. The gradual closure of the isthmus started during the Middle Miocene. By the Early Pliocene (4.6 Ma), it was advanced enough to affect deep-ocean circulation significantly. By 3.5 Ma the closure had progressed to the extent that evolutionary divergence had taken place between the Pacific and Caribbean shallow-water mollusc faunas. Evidence from nannofossils indicates that, between 3.65 and 2.76 Ma, the westward current between Caribbean and eastern Pacific ceased to exist while a northward intra-Caribbean current was established (Figure 1). However, the interchange of terrestrial mammals between the North and South American continents may have been delayed to 2.7 Ma, coinciding with the intensification of northern hemisphere glaciation and the associated sea-level drop. The substantial impact that the gradual shoaling of the Panama isthmus had on marine biota is exemplified by the observation that extinction rates of Pliocene molluscs in the western Atlantic and Gulf of Mexico were, on average, twice those in the eastern Pacific.

Other tectonic activities continued previous trends. In Africa-Arabia relative movements of the Nubian, Somalian and Arabian plates caused continued rifting along East Africa and initiated sea-floor spreading in the Red Sea. Many areas worldwide are thought to have undergone significant uplift throughout the Pliocene. Some of the most prominent areas being the South American Andes, the North American Sierra Nevada and Cascade Range, the African Atlas, the European Pyrenees and Alps, and the Himalayas and Tibetan Plateau in Asia.

The increased height of a large number of mountain ranges across the globe had a significant impact on local climates and may have influenced global climate. For example, the uplift of the Himalayas and Tibetan plateau, caused by the northward drift of the Indian plate and its crashing into the Asian plate, has been instrumental in shaping Asian climates at least since the Miocene. Continued uplift of northern and eastern parts of the Tibetan plateau have been linked to an intensification of both the summer and winter East Asian monsoons during the Middle Pliocene, 3.6–2.6 Ma. Increased volcanism is also likely to result from an increase in mountain building activity and a distinct increase in the frequency and thickness of tephra layers in ocean sediments from the Pliocene and the Quaternary support the notion of widespread orogenic activity during those periods. However, the absence of an obvious mechanism to underlie a global uplift of distinct

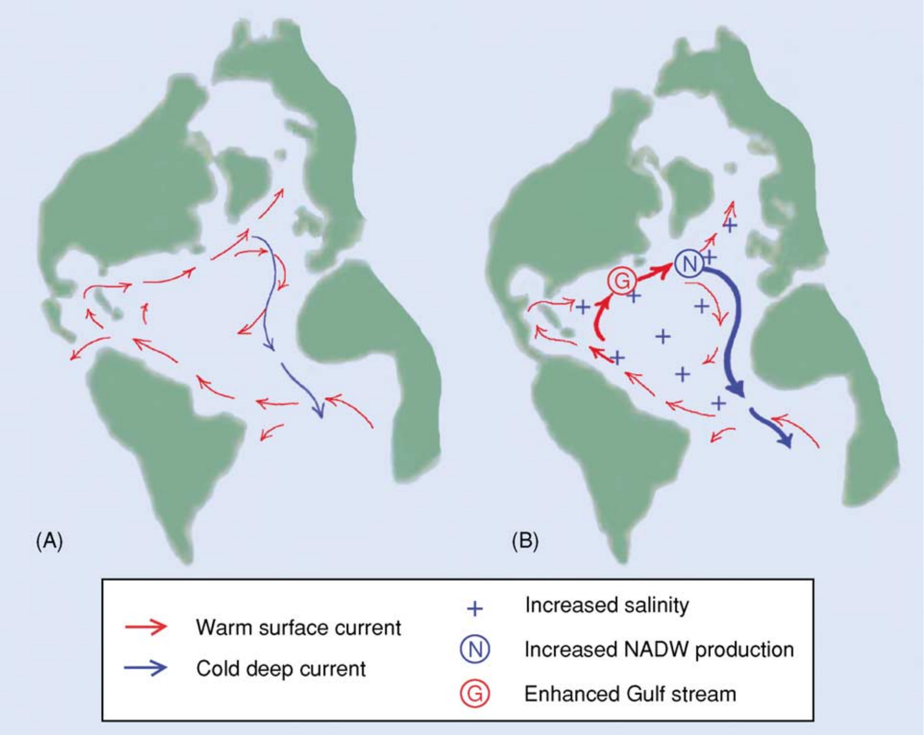


Figure 1 Atlantic marine currents before (A) and after (B) the emergence of the isthmus of Panama, showing intensified thermohaline circulation after the closure of the central American seaway. NADW: North Atlantic deep water.

mountain ranges of very varied ages has led some experts to question the existence of wide-ranging Plio-Quaternary uplift and to evoke the possibility that changes in climate have instead created the illusion of widespread uplift. Increased rates of erosion are widely cited as evidence for increased elevation gradients, but increased rates of erosion have also been associated with changes in global climate patterns, in particular with the enhancement of rainfall seasonality or the onset of rapid climatic oscillations in the later Pliocene.

Climate

The Pliocene was a time of profound climatic change. Starting with generally warm conditions up to the end of the Zanclean when global temperatures declined, resuming the trend of overall Cenozoic cooling. A subsequent phase of global warming, the 'mid-Pliocene warmth' (3.3–3.15 Ma) has been well documented and is a continued focus of intensive research due to its relevance to the prediction of future global warming events. The cooling that followed led to the northern hemisphere glacial-interglacial cycles, which continue to date. The onset of northern hemisphere glaciation was not synchronous. The first ice-sheets covered the Eurasian Arctic and north-east Asia by 2.75 Ma. Alaska became glaciated by 2.65 Ma, followed by the northeast American continent by 2.54 Ma. In terms of their potential to contribute to our understanding of the forces that drive and the consequences of global climatic change, the mid-Pliocene warm period and the onset of the modern ice age are of particular interest.

Mid-Pliocene Warming

With the exception of the last interglacial during which temperatures may have been briefly higher than today's, the period of mid-Pliocene warming (3.3–3.15 Ma) represents the most recent geological time span during which average global temperatures were higher than today's. Geography, flora and fauna of the mid-Pliocene were more similar to today's than during any other period of prolonged global warming. The mid-Pliocene, therefore, has the potential of providing the best indication of how the modern world may respond to future increases in global temperatures. Observational studies suggest that during the mid-Pliocene warmth sea levels were 20–40 m higher than today, the result of a reduced Antarctic ice-sheet area and the near complete absence of ice in the Arctic. Sea-surface temperatures were also substantially higher than today in the middle and higher latitudes. Warming was most pronounced in the north-eastern North Atlantic. Tropical sea-surface temperatures, however, were similar to those seen today.

Two basic factors have been suggested to explain the nature of Pliocene warmth: increased levels of atmospheric CO_2 (the greenhouse effect) and increased ocean heat transport. As increased atmospheric CO_2 is expected to increase temperatures globally and increased heat transport should result in a cooling of tropical waters, the reconstructed distribution of mid-Pliocene sea surface temperatures suggests that both mechanisms contributed to mid-Pliocene warming.

On land, there was a general shift of modern vegetation zones towards higher latitudes. Boreal forests occurred as far north as the Arctic coast, grading into temperate mixed-conifer and conifer-hardwood forest in the northern middle latitudes, and indicating the presence of temperatures that were much warmer than today in the Arctic regions of Iceland, Russia and North America as well as in western and central Europe. In contrast, some lower latitudinal areas may have seen temperatures drop below those of the modern day. The continental interiors of Africa, Asia and North America were more humid than they are today. Remains of pollen, leaves and wood of southern beech (*Nothofagus*) on Antarctica suggest a climate substantially warmer than at present. In the tropics, vegetation data point at temperatures lower than today in central America and East Africa, but warmer than at present in northern South America.

The Ice Age

The overwhelming climatic signature of the Pliocene is the increase in the levels of glaciation, and in particular the onset of large-scale glaciation of the northern hemisphere during the Middle Pliocene. At least three factors combined to initiate and sustain this glaciation. First, the overall trend of global cooling that prevailed through the Late Cenozoic ensured that precipitation over northern areas fell as snow rather than rain. Second, increased thermohaline circulation and Gulf Stream flow (a consequence of the closure of the isthmus of Panama) introduced the necessary moisture into far northern latitudes. Finally, a favourable Milankovitch orbital configuration between 3.1 and 2.5 Ma resulted in summers that were cool enough to prevent the snow from melting.

The onset of large-scale glaciation of the northern hemisphere is marked by an increase of $\delta^{18}O$ in benthic foraminifers between 3.0 Ma and 2.5 Ma and the appearance of substantial amounts of ice-rafted debris in the northern oceans from 2.7 Ma. The closure of the isthmus of Panama is likely to have played a major role in the onset of northern hemisphere glaciation by redirecting surface currents within the Atlantic Ocean since 4.6 Ma and by influencing salinity levels. The resulting intensification of the Gulf Stream increased the transport of warm saline surface water to northern high latitudes and, consequently, the formation of North Atlantic deep water, which in turn, ensured increased import of warmer South Atlantic surface water into the northern hemisphere (**Figure 1**). Intensification of this marine 'conveyor belt' increased the potential for evaporation at high northern latitudes, thus supplying the additional moisture that would facilitate ice-sheet growth.

The gradual shoaling of the Panama isthmus and the associated changes in ocean currents and salinity gradients did not in themselves cause the onset of large-scale northern hemisphere glaciation. Initially, the enhanced Gulf Stream and increased transport of warm surface water into the high northern latitudes pushed global climates towards warmer conditions, culminating in the mid-Pliocene warm interval. The emergence of the Panama isthmus contributed significantly to creating the necessary preconditions for the initiation of large-scale northern hemisphere glaciation. However, other factors may have been involved (e.g. deepening of the Bering Straits, increased river discharge of fresh water into the Arctic Ocean from the Palaearctic). Continued Pliocene uplift or rejuvenation of various mountain ranges and plateaus across the world has also been suggested as a cause for the deterioration of Late Neogene climates. Invoked mechanisms include the influence of high elevation areas on atmospheric circulation as well as the reduction of atmospheric CO_2 concentrations through the weathering of newly exposed silicates and the resulting formation of carbonates, the increased burial of organic carbon due to increased rates of erosion, or increased ocean productivity through an increased delivery of phosphorus to the ocean, again as the result of increased erosion.

It is likely that several of these factors combined to create the preconditions necessary for the onset of northern hemisphere glaciation. The ultimate trigger, however, lay with the nature of fluctuations in the Earth's obliquity (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). Low tilt angles of the Earth's rotational axis result in cold northern hemisphere summers. Just prior to the onset of northern hemisphere glaciation, 4.5–3.1 Ma, there was a prolonged period during which the amplitude of Earth's obliquity fluctuated only minimally. Records of $\delta^{18}O$ indicate that during this time, limited glaciations of the Arctic were initiated several times, but failed to substantiate. From around 3.0 Ma the gradual increase of the amplitude of obliquity fluctuations set the final stage for the current pattern of pronounced glacial and inter-glacial periods, which have prevailed since the end of the Middle Pliocene.

Biotic Events

The Great American Interchange

The emergence of the isthmus of Panama in the Middle Pliocene resulted in the first contact of South America with another continent since the opening of the Drake Passage finalized its separation from West Antarctica toward the end of the Eocene. Having been rifted from Gondwanaland ~100.0 Ma and spent most of the Cenozoic in total isolation, South America was inhabited by a distinct biota with a large proportion of endemic taxa. In contrast, the North American continent had previously had a long history of intermittent connections and resulting floral and faunal interchange with the Old World.

The emergence of a permanent land connection between these two sub-continents resulted in a substantial level of terrestrial biotic interchange. However, not all taxa proved equally successful at dispersing and the biotas of the two sub-continents were not equally affected. With respect to long-term effects, the influence that southern taxa have had on North American biotas is near negligible. In contrast, the long-term effect that the great American interchange has had on the South American mammal fauna is substantial. Today nearly half of the families and genera of South American mammals are members of groups that emigrated from North America since the emergence of the Panama land bridge. This does not appear to be due to a substantially larger number of species initially dispersing from north to south as opposed to from south to north. In fact, when differences in source area are taken into account, the extent of north to south dispersal was similar to that from south to north. The difference in the proportions of modern representatives arose through the fact that North American immigrants to the south underwent substantial diversification in their new environments, whereas South American immigrants in the north did not. The reasons for the substantial differences in success rates between North American and South American immigrants are not clear.

Movements across the Panamanian isthmus were not restricted to mammals, but data for other groups are scarcer. For birds the available data suggest a net movement of taxa from north to south, whereas the majority of dispersals of reptiles and amphibians appear to have been from south to north with significant levels of dispersal already occurring prior to the final emergence of the isthmus. In addition, while more than 90% of the angiosperm species of lowland rainforests in central America are estimated to be of South American origins, most montane plant species in South America have North American roots.

The Trans-Arctic Interchange

The Bering Strait between Siberia and Alaska initially opened during or just before the Early Pliocene. This new connection between the Arctic-Atlantic and the North-Pacific basins resulted in biotic movements referred to as the 'marine trans-Arctic interchange'. Data from invertebrates and algae suggest that, overall, and in line with the present prevailing currents, the trans-Atlantic interchange heavily favoured invasion from the Pacific to the Atlantic. However, data from before 4.8 Ma reveal the presence of Atlantic bivalves in the Pacific, whereas after 3.6 Ma North-Pacific molluscs suddenly became widespread in the North Atlantic. This confirms that, in accordance with existing models, currents through the Bering Straits flowed from north to south during the Early Pliocene, but were reversed some time between 4.8 and 3.6 Ma, probably as a consequence of the gradual closure of the central American seaway.

Other Biotic Movements

The emergence of the Panama land bridge and the opening of the Bering Strait resulted in a general intensification of marine currents. From a biotic perspective these also resulted in interchange between northern and southern temperate biota in the eastern Pacific and between temperate biota of the eastern and western Atlantic.

Plants

The Pliocene's climatic trends were reflected in the nature and distribution of plants worldwide. During periods of cooling, many areas saw a transition from broadleaved to coniferous woods or from closed to more open, grassy vegetations. Climate change evidently influenced the patterns of distribution of plants, but changes in the nature of the vegetation may in turn have influenced Pliocene climates. Around the Miocene to Pliocene transition, the biomass of plants using C_4 as opposed to C_3 photosynthetic pathways increased in Africa, Asia, North and South America, probably as a result of decreasing concentrations of atmospheric CO_2. The increase took place during the Late Miocene in the lower latitudes and during the Early Pliocene in the higher latitudes. However, some local environments remained C_3-dominated and no increase in the biomass of C_4-plants appears to have taken place in western Europe. The continued spread of C_4-plants during the Pliocene may have had a direct effect on global climates. Due to their shallower roots and their increased efficiency at CO_2 fixation, C_4-plants return less water from the soil to the atmosphere than C_3-plants. They thereby affect the hydrologic cycle and generally promote drier conditions downwind.

Hominin Diversification

In Africa, the Pliocene epoch saw the evolutionary diversification of the human lineage, the Homininae (*see* **Fossil Vertebrates:** Hominids). Only three species of hominin are known from the Late Miocene. In contrast, up to 12 additional species (almost all from the eastern rift of the African rift valley and from South African cave sites) have been

Table 2 Hominin diversity during the Pliocene

Species	*Age*	*Localities*
Ardipithecus ramidus White *et al.*, 1994	5.8–4.4 Ma	Middle Awash (Ethiopia)
Australopithecus anamensis Leakey *et al.*, 1995	4.2–3.9 Ma	Kanapoi, Allia Bay (Kenya)
Australopithecus afarensis Johanson *et al.*, 1978	3.6–2.9 Ma	Laetoli (Tanzania); Koobi Fora, West Turkana (Kenya); Omo, Middle Awash, Hadar (Ethiopia)
Australopithecus bahrelghazali Brunet *et al.*, 1996	3.5–3.0 Ma	Bahr El Ghazal (Chad)
Kenyanthropus platyops Leakey *et al.*, 2001	3.5–3.3 Ma	Lomekwi (Kenya)
Australopithecus africanus Dart, 1925	3.0–2.4 Ma	Taung, Makapanskat, Sterkfontein (South Africa)
Australopithecus garhi Asfaw *et al.*, 1999	2.5 Ma	Bouri (Ethiopia)
Paranthropus aethiopicus Arambourg & Coppens, 1968	2.7–2.3 Ma	West Turkana (Kenya); Omo (Ethiopia)
Paranthropus boisei Leakey, 1995	2.3–1.4 Ma	Olduvai, Peninj (Tanzania); Chesowanja, Koobi Fora, West Turkana (Kenya); Omo, Konso-Gardula (Ethiopia)
Paranthropus robustus Broom, 1938	1.9–1.4 Ma	Kromdraai, Swartkrans, Drimolen, Gondolin (South Africa)
Homo habilis Leakey *et al.*, 1964	2.3–1.6 Ma	Omo, Hadar (Ethiopia); Olduvai (Tanzania); East Lake Turkana (Kenya); Sterkfontein (South Africa)
Homo rudolfensis Alexeev, 1986	1.9 Ma	East Lake Turkana (Kenya)

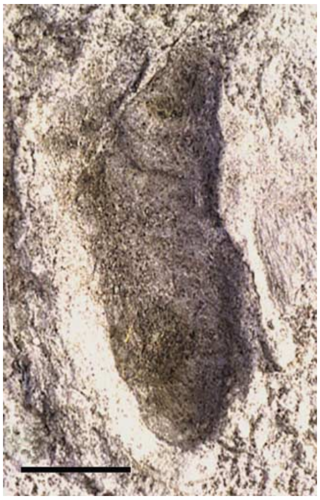


Figure 2 One of a series of footprints left behind in volcanic ash at the Early hominin site of Laetoli (Tanzania) some 3.6 Ma. Scale is 5 cm. Photograph by Peter Schmid.

described from the Pliocene to date (Table 2). Around 4.0 to 3.5 Ma a shin bone, probably of *Australopithecus anamensis*, and a series of fossilized footprints (Figure 2) provide the first clear evidence of bipedal walking. The most famous hominin fossil ('Lucy', an *Australopithecus afarensis)* lived just over 3.0 Ma and the oldest stone artifacts are known from the Middle to Late Pliocene transition, ~2.5 Ma. The Pliocene diversification of early hominins coincided with climatic fluctuations and a general trend towards more arid and less wooded habitats in eastern and southern Africa. The traditional view, however, that typical hominin features such as bipedality evolved as a direct response to the increased aridity and opening up of the East-African landscape is no longer considered valid. Recent palaeoecological reconstructions of early hominin environments generally point at wooded and well-watered habitats. Only with the advent of the genus *Homo*, around 2.5 Ma, can hominins be considered to have been fully adapted for life in open and arid environments. This is confirmed by data from marine sediments. Marine records show that the African climate saw an increase in aridity after 2.8 Ma, coinciding with the onset of large-scale glaciation in the northern hemisphere.

Before 2.8 Ma, variations in the intensity of the African monsoon were mainly the result of variations in low-latitude insolation due to Earth orbital precession and caused 19 Ka–23 Ka cyclical alternations between dry and wet conditions. After 2.8 Ma the African climate became sensitive to the increased amplitude of high northern latitude climate variations and the cycles changed to 41 Ka, paralleling the periodicity of the Earth's orbital obliquity variation and the Late Pliocene cycles of northern hemisphere glacials and interglacials. From 2.8 Ma, periodically cooler and drier conditions and the transition towards overall lower and more seasonal precipitation strongly favoured open savannah vegetation in east Africa. It is clear that the appearance and spread of open savannah cannot be considered to have contributed to shape the earliest evolution and diversification of the hominin lineage. Instead, it is more likely, that the gradual reduction of areas of humid and wooded habitat towards the Middle to Late Pliocene transition contributed to the disappearance of many taxa. In contrast, the emergence of the genus *Homo* coincided with substantial aridification of much of the African continent.

See Also

Atmosphere Evolution. **Carbon Cycle**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Famous Geologists:** Lyell. **Fossil Vertebrates:** Hominids. **Tectonics:** Mountain Building and Orogeny; Rift Valleys. **Tertiary To Present:** Miocene; Pleistocene and The Ice Age.

Further Reading

Castradori D, Rio D, Hilgen FJ, and Lourens LJ (1998) The Global Standard Stratotype-section and Point (GSSP) of the Piacenzian Stage (Middle Pliocene). *Episodes* 21: 88–93.

Dowsett HJ, Barron JA, Poore RZ, Thompson RS, Cronin TM, Ishman SE, and Willard DA (1999) *Middle Pliocene Paleoenvironmental Reconstruction: PRISM2*. U.S. Geological Survey Open File Report, pp. 99–535, http//pubs.usgs.gov.

Driscoll NW and Haug GH (1998) A short circuit in thermohaline circulation: a cause for northern hemisphere glaciation? *Science* 282: 436–438.

Haug GH and Tiedemann R (1998) Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation. *Nature* 393: 673–676.

Hay WH, Soeding E, DeConto RM, and Wold CN (2002) The Late Cenozoic uplift – climate change paradox. *International Journal of Earth Sciences* 91: 746–774.

Jackson JBC, Budd AF, and Coates AG (eds.) (1996) *Evolution and Environment in Tropical America*. Chicago: University of Chicago Press.

Poore RZ and Sloan LC (eds.) (1996) *Climates and Climate Variability of the Pliocene*. Marine Micropaleontology 27.

Reed KE (1997) Early hominid evolution and ecological change through the African Plio-Pleistocene. *Journal of Human Evolution* 32: 289–322.

Rio D, Sprovieri , Castradori D, and Di Stefano E (1998) The Gelasian Stage (Upper Pliocene): a new unit of the global standard chronostratigraphic scale. *Episodes* 21: 82–87.

Van Couvering JA, Castradori D, Cita MB, Hilgen FJ, and Rio D (2000) The base of the Zanclean Stage and of the Pliocene Series. *Episodes* 23: 179–187.

Vrba ES, Denton GH, Partridge TC, and Burckle LH (eds.) (1995) *Paleoclimate and Evolution with Emphasis on Human Origins*. New Haven: Yale University Press.

Wrenn JH, Suc J-P, and Leroy SAG (eds.) (1999) *The Pliocene: Time of Change*. Dallas: American Association of Stratigraphic Palynologists Foundation.

Pleistocene and The Ice Age

A Currant, The Natural History Museum, London, UK

Introduction

Pleistocene is the name given to the geological epoch succeeding the Pliocene and preceding the Holocene. It is usually taken as beginning about 1.8 million years ago and ending with the termination of the last major cold stage and the beginning of the Holocene warm stage, about 10 000 years ago, but the start of the Pleistocene has proved difficult to define. Together, the Pleistocene and the Holocene form the Quaternary Period. The Pleistocene is characterized by great global climatic instability, particularly from the onset of major, cyclic, northern hemisphere, terrestrial ice accumulations about 700 000 years ago. This later period is popularly known as the Ice Age.

The term 'Pleistocene' was first used by Charles Lyell in 1839 to denote the age of sediments in and around the London and Paris basins, in which 70% of the molluscan fossils were considered to represent species that are still alive today. The deposits in question were clearly related to the ancient courses of the Thames and the Somme. Lyell's original definition was of limited value. It was very much at the mercy of changes in opinion on molluscan identification and taxonomy and had extremely restricted geographical utility, but the name remained and has been the subject of almost continual redefinition since Lyell's time. The current global stratigraphic section and point (GSSP) for the start of the Pleistocene at Virca in Italy still has unresolved dating problems. The start of the epoch is widely believed to be between 1.65 and 1.8 million years ago, although some authors push this back as far as 2.5 million years. The environmental distinction between the Late Pliocene and the Early Pleistocene is actually quite difficult to pin down, and there was considerable faunal and floral continuity between the two.

Historical Studies

In global terms, the Pleistocene was dominated by the changing extent of ice volume. Although sea-floor spreading and continental drift continued throughout the period, the effect of these longer term processes has been dwarfed by ice-related phenomena. The history of our changing and growing understanding of the Pleistocene is important to anyone examining the older literature related to the period, particularly that which deals with the 'Ice Age'. The pioneer geologist and geomorphologist Louis Agassiz published works on his studies of Alpine glaciers and glacial processes, *Étude sur les Glaciers* (1840) and *Système Glaciare* (1847), in which he put forward the idea of there having been a great Ice Age in which large parts of Europe had been covered by ice and the surface of the land was greatly modified under its influence. Agassiz then went on to make similar observations based on his travels in North America. The subsequent history of Ice Age studies has been one of very gradual recognition that there were probably multiple Ice Age events, widely referred to as 'glacials', punctuated by warmer phases, with temperatures as warm as or warmer than those now prevailing, known as 'interglacials'. Numerous models of Pleistocene climatic events have been put forward, but the one that held sway for longest and was most widely adopted was that proposed in 1909 by Penck and Bruckner in *Die Alpen im Eiszeitalter*, in which the authors suggested there had been four major Ice Age stages, the Gunz, Mindel, Riss, and Wurm. This system, based on the terrace deposits of Alpine rivers, became adopted all over the world, and although local names often replaced the Alpine originals, the fourfold division of Pleistocene ice ages was accepted almost everywhere. The intervening interglacial stages became named after the cold stages that delimited them, i.e., Gunz-Mindel, Mindel-Riss, and Riss-Wurm.

It is important to recognize that the early understanding of Pleistocene events was based on hybrid data. Evidence for the cold glacial stages was largely derived from characteristic sediments and landforms whereas evidence for the warmer interglacial stages came mainly from organic sediments and the fossils they contained. The stratigraphic relationship between glacial and interglacial sequences was often obscure and the means of correlating between isolated terrestrial sequences was poorly defined. The Alpine model of Ice Age events became to a very large extent a self-perpetuating system. Although a number of lines of evidence began to emerge suggesting much greater overall complexity, the lack of clarity in the original stage definitions and the virtual impossibility to establish secure interregional correlations led to such evidence being subsumed into the model. Against the background of purely geological interpretations, there was also a growing body of archaeological evidence charting human prehistory through the Pleistocene in various parts of the world. It was not until the 1860s that it became widely accepted that the human race had any great antiquity, with stone tools that were clearly made by early people being found in undisturbed contexts along with the remains of extinct animals. Such a likelihood was first noted by the French antiquarian Jacques Boucher de Perthes, who from 1846 to 1857 published his findings of stone tools from the high-level (and therefore very ancient) terrace deposits of the River Somme near Abbeville. There was considerable opposition to the idea that there were human populations of great age represented in the fossil record, primarily from religious groups and individuals, and to some extent this opposition still survives today. For most people, the clinching discovery was that of prehistoric cave art in France and Spain, where people living in the last glaciation had painted pictures of contemporary woolly mammoths, woolly rhinos, reindeer, bison, and horses in the caves they also used for shelter – direct evidence with a ancient human signature. It quickly became clear that the record of human antiquity was longest in Africa, Europe, and southern Asia and relatively short in northern Asia, Australia, and the Americas.

There is probably no other period of Earth history that has been so intensively studied yet so badly misinterpreted than the Pleistocene, and all because the construction of theoretical models has tended to run ahead of the collection and interpretation of hard evidence. Far too little attention has been paid to testing models and far too much to reinforcing them. The Alpine Model of Pleistocene climatic events eventually fell victim to new evidence from the oceans. Cesare Emiliani and colleagues, working in the late 1950s and early 1960s on cores taken from deep-ocean sediments, found evidence from the shells of buried foraminifera that their oxygen isotope content varied through time. Emiliani had already worked on the palaeoclimate signals recovered from microfossils in older deposits, and he recognized his new data as being a direct record of changing global ice volume. The two significant isotopes of oxygen, ^{16}O and ^{18}O, behave differently during evaporation, leaving the oceans enriched with the heavier isotope. When water evaporates from the oceans and falls as rain or snow on the land, it is isotopically ‘light’, and if sufficient frozen water remains on land in large, stable ice-sheets, then the isotopic composition of the oceans will change to a measurable degree. When the ice melts, eventually the meltwater will ultimately return to the oceans, thus recentring the isotopic signal. Foraminifera living in the sea will deposit shells that reflect the chemistry of the seawater during their lifetime, and when they die, their shells fall to the seafloor and become incorporated in the sediments being deposited there, creating a retrievable record of changing global ice volume through time. Emiliani found evidence for many more episodes of global ice accumulation than were allowed for in the Alpine Model.

When plotted onto a graph, the marine isotope record bore a very strong resemblance to the theoretical curve of variations in the predicted amount of solar energy reaching a particular point on the planetary surface; this is determined by known variations in Earth’s orbit. This curve had been calculated by the Serbian astrophysicist Milutin Milankovitch in the early twentieth century to explain long-term climatic variability, but, in the absence of any evidence to back his theory, Milankovitch’s work remained largely ignored. Emiliani’s recognition of the resemblance between the Milankovitch solar radiation curve and the oceanic oxygen isotope record has led to a completely new understanding of the driving forces behind global climatic change, now extending far back beyond the Pleistocene. Alongside plate tectonics, this must number as one of the greatest breakthroughs in geological science.

Milankovitch theory examines the combined effects of three known variables in Earth’s orbit: the eccentricity of Earth’s orbit around the Sun, the obliquity of Earth’s axis relative to the Sun, and the precession of the equinoxes that changes the season at which Earth’s axis is most tilted towards the Sun. Each of these has respective cycles of approximately 100 000, 41 000 and 23 000 years. In the past 700 000 years, the 100 000-year cycle has become dominant, giving major global cold stages at about this interval. Prior to this, there was considerable climatic instability during the earlier part of the Pleistocene, but not on anything like this dramatic scale. Amazingly detailed climatic signals for the last major climatic cycle recovered by drilling through the Greenland ice-cap confirm and amplify the resolution of the marine isotope record for this later period.

It is now recognized that there have been about seven major global Ice Age events in the past 700 000 years, though these vary to some degree in intensity and progress. Armed with this information, it is now possible to re-examine the geological record and reinterpret the evidence for the relative order and extent of ice-sheet advance and for the intervening warmer interglacial stages. The modern terminology of Pleistocene subdivision is based on the marine isotopic signal. The isotope stages are numbered backwards from 1, with Marine Isotope Stage (MIS) 1 representing the Holocene. Warm stages have odd numbers and cold stages have even numbers. A historical anomaly in the last cold cycle has been divided up into three isotope stages (2, 3, and 4), so the last interglacial period is MIS 5, but otherwise the stages appear to represent single major climatic events (see Figure 1).

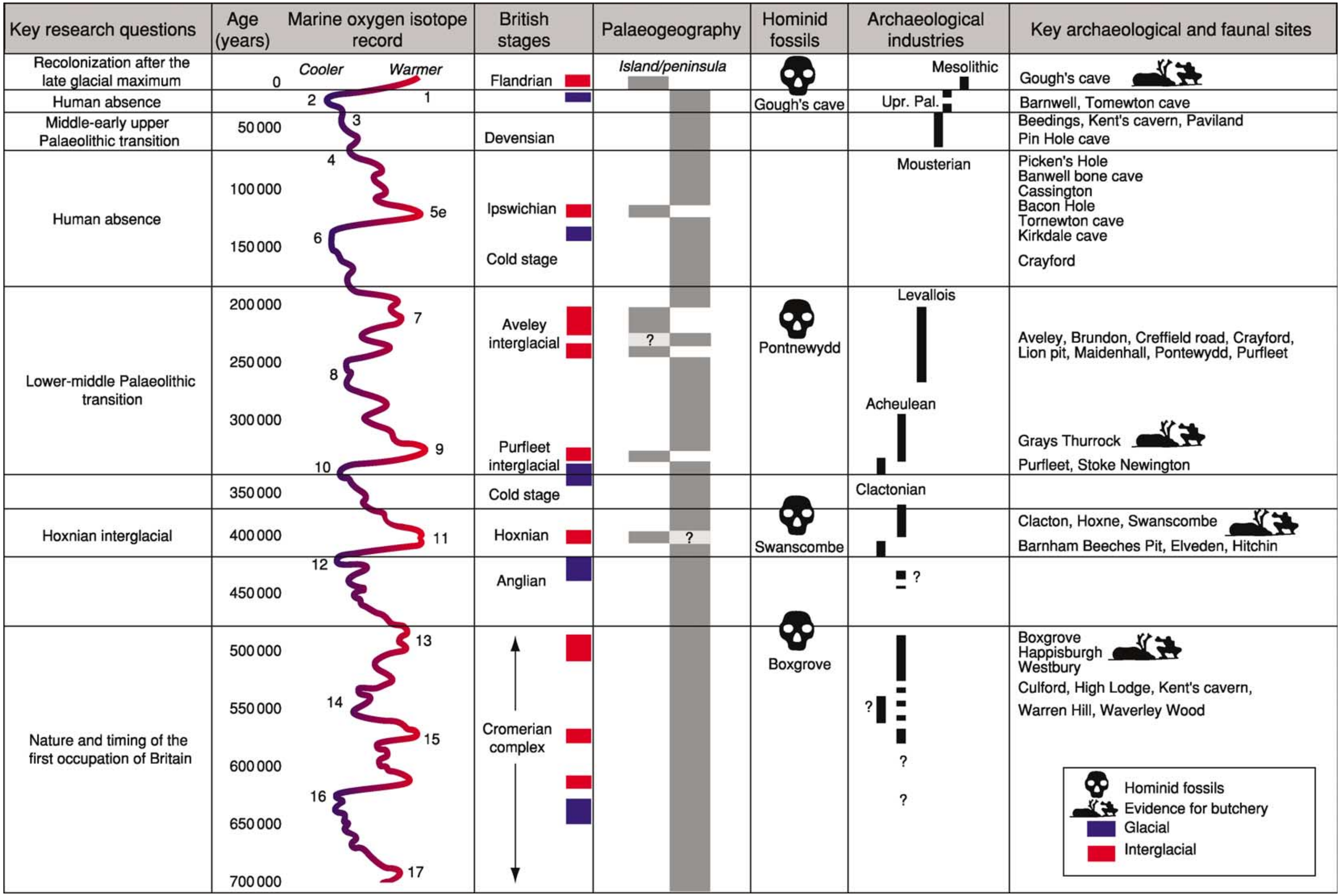
Life on Earth in the Pleistocene

The distribution of plants and animals on Earth's surface has been hugely influenced by large-scale high- to mid-latitude ice accumulation. Not only have ice-sheets periodically excluded life from large areas of land, but the effects of climate change have radically changed the normal distribution of many species, while significant areas of new land have become available for colonization at times of lowered sea-level. Interglacial sea-levels have generally been as high as those of today, or sometimes a few metres higher, but during cold stages, the amount of water locked up on land as ice has reduced global sea-levels by up to 120 m, exposing vast areas that are now continental shelf. Such drops in sea-level would unite the islands of south-east Asia into a single landmass, greatly extend the eastern seaboard of southern South America, make the British Isles a peninsula of north-west Europe, and connect Asia to North America by a broad northern land bridge (Beringia). In the southern hemisphere, the scope for significant increase in terrestrial ice cover is limited by the disposition of the continents. Antarctic ice masses grew and waned, but it was only in the Andean region of South America that very large, new, non-polar ice-sheets developed. The primary effect of Pleistocene ice accumulation is seen in the northern hemisphere, where the Hudson Bay area and Scandinavia became the central foci for the growth of major new ice-caps and large mountain glaciers developed in the Rockies and in the Alpine mountain belt.

The influence of Pleistocene cold-stage events is still very much with us today. Large regions of permanently frozen ground or permafrost in the arctic regions are attributable to the deep cold experienced by both glaciated and unglaciated higher latitudes, and the crustal distortion caused by the massive weight of former ice-sheets can be witnessed in a recovery phase around the Baltic Sea and Hudson Bay regions, where isostatic uplift is still very active. The Baltic Sea and Hudson Bay are no more than the slowly draining crustal depressions caused by former ice-sheets.

Rich sources of evidence for the faunal and floral history of the Pleistocene are to be found in the deposits laid down by major river systems. It is probably fortunate that the study of such deposits began in north-western Europe, where long-term regional uplift has created staircase-like terrace systems representing the former levels at which rivers ran as they progressively cut down to new base levels. The Pleistocene deposits of south-east England are an extremely good illustration of this kind of sequence, particularly those related to the Thames drainage. A fossil landscape is preserved beneath eastern England, buried beneath the debris of ice-sheets that invaded the area about 400 000 years ago. This glacial event (the Anglian glaciation, believed to represent Marine Isotope Stage 12) obliterated some of the major river systems of central England and diverted the Thames to something like its present course. Beneath the glacial tills, fossiliferous deposits representing these extinct river systems provide information on life in the area in the early Pleistocene and early middle Pleistocene.

From the Anglian glaciation onwards, the Lower Thames Valley preserves fragments of river deposits representing all the major climatic stages revealed by the oceanic isotope record. Because the terrace fragments can be related to the existing topography, it has been possible to build up a clear picture of the post-Anglian evolution of this particular landscape. The fact that the modern city of London is sited on the Thames plays no small part in unravelling the story. Almost every pocket of sand, gravel, and brickearth has at some stage in the recent past been excavated as an economic resource, and many of these sections have been carefully recorded and their fossils collected. It is a sad fact that the economic development that created the opportunity to understand the later Pleistocene history of the Thames has now destroyed much of the evidence on which that understanding was based, and it is one of the responsibilities of present-day geographers and Pleistocene specialists to make certain that what little remains is well documented and, where possible, conserved. This particular case history underlines the point that much of the evidence for terrestrial



Key research questions	Age (years)	Marine oxygen isotope record	British stages	Palaeogeography	Hominid fossils	Archaeological industries	Key archaeological and faunal sites
Recolonization after the late glacial maximum	0	Cooler Warmer	Flandrian 🟥	Island/peninsula	☠	Mesolithic	Gough's cave
Human absence		2 1	🟦		Gough's cave	Upr. Pal.	Barnwell, Tomewton cave
Middle-early upper Palaeolithic transition	50 000	3	Devensian				Beedings, Kent's cavern, Paviland Pin Hole cave
Human absence	100 000 150 000	4 5e 6	Ipswichian 🟥 🟦 Cold stage			Mousterian	Picken's Hole Banwell bone cave Cassington Bacon Hole Tornewton cave Kirkdale cave Crayford
Lower-middle Palaeolithic transition	200 000 250 000 300 000	7 8 9 10	Aveley interglacial 🟥 🟥 Purfleet interglacial 🟥 🟦	?	☠ Pontnewydd	Levallois Acheulean	Aveley, Brundon, Creffield road, Crayford, Lion pit, Maidenhall, Pontewydd, Purfleet Grays Thurrock Purfleet, Stoke Newington
	350 000		Cold stage			Clactonian	
Hoxnian interglacial	400 000	11	Hoxnian 🟥	?	☠ Swanscombe		Clacton, Hoxne, Swanscombe Barnham Beeches Pit, Elveden, Hitchin
	450 000	12	Anglian 🟦			?	
Nature and timing of the first occupation of Britain	500 000 550 000 600 000 650 000 700 000	13 14 15 16 17	Cromerian complex 🟥 🟥 🟥 🟦		☠ Boxgrove	? ? ?	Boxgrove Happisburgh Westbury Culford, High Lodge, Kent's cavern, Warren Hill, Waverley Wood

☠ Hominid fossils
Evidence for butchery
🟦 Glacial
🟥 Interglacial

Figure 1 The relationship between apparent human occupation of the British Isles during the Pleistocene and climate change, based on the work of the Ancient Human Occupation of Britain research team directed by Professor C B Stringer FRS.

Pleistocene events lies buried within existing landscapes and can only usually be seen when it is artificially excavated.

In the Thames Valley, cold-stage deposits are usually poor in fossils, but the interglacial stages are represented by exceptionally rich organic levels containing bones, molluscs, beetles, and plants. Five distinct post-Anglian interglacials can be recognized, representing Marine Isotope Stages 11, 9, 7, 5, and 1 (stage 3 is not an interglacial, but is nonetheless well represented by cold-stage deposits laid down in the middle of the last glacial stage). Most of the known river terrace deposits have been found to have a cold–warm–cold sandwichlike structure, with an organic interglacial filling of varied lithology underlain and overlain by cold-stage gravels. The relationship and order of these terrace deposits can be worked out from their topographic position, but this interpretation has been greatly enhanced by study of their fossil content. Biostratigraphic data have been recovered from all of the major groups examined, but it has been found that the vertebrate faunas, particularly the mammals, give the most distinctive signature to each interglacial event. It has proved possible to use the distinctive mammalian faunas recovered from the Thames terrace sequence to correlate with other river and lake sequences around the country and even with fossiliferous cave deposits. The importance of cave sequences cannot be overstated, because they currently offer the only reliable means of direct dating (uranium series age determinations on clean stalagmite) that is available for Pleistocene sequences beyond the range of radiocarbon.

Many other river systems around the world are now being studied systematically to understand their distinctive depositional cycles and the ways in which these reflect Pleistocene climatic variation. Just as ice-sheets were restricted to particular areas of Earth's surface, so drainage systems have their own dynamics, according to the regions they drain. It is essential to understand those dynamics before trying to interpret the palaeoclimatic evidence that they may contain, in particular their fossil content. Returning briefly to the Thames, the great paucity of fossils relating to cold periods before the last one is a rather odd feature, particularly given the richness of interglacial sequences within the same region, but this is the signature of this particular river system. In other parts of the world, different parts of the climatic cycle may be better represented, and it is certainly the case that other European drainage systems have quite different depositional histories.

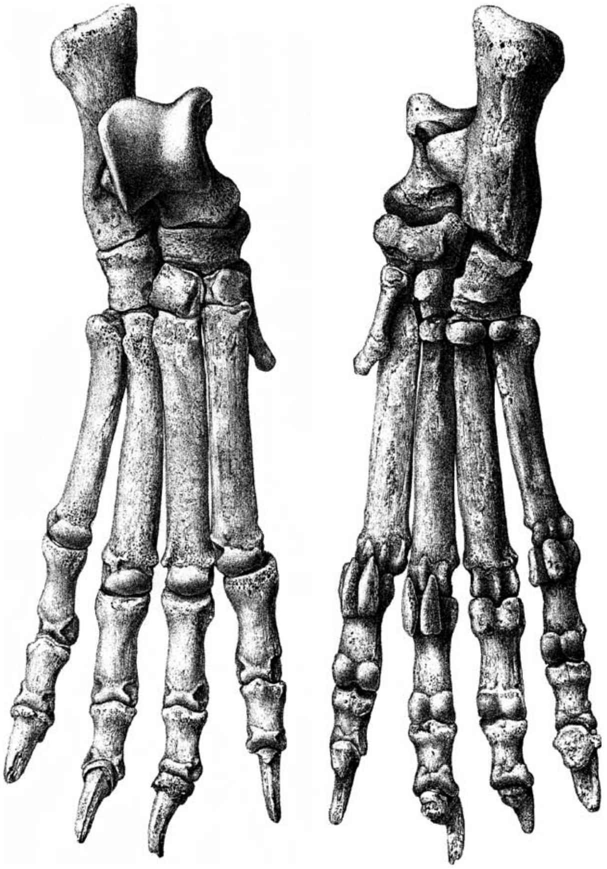


Figure 2 Engraving of the hind foot bones of a Pleistocene spotted hyena, *Crocuta crocuta*, from Tornewton Cave, Devon. Caves often contain exceptionally well-preserved fossilized mammalian bones, particularly those of the carnivores that have used them as lairs or dens.

Historically, caves have been a great focus of interest. Caves are natural dustbins, anomalous sites for the temporary deposition of detritus that would otherwise be destroyed by natural processes (Figure 2). They are a natural focus for biological activity and wide range of animals use them for shelter and to raise their young. This includes amphibians, bats, a variety of mammalian and avian carnivores, and humans. Bone tends to be preserved preferentially in limestone environments, and most caves are to be found in limestones. Many caves have been found in which animal bone is one of the primary constituents of the cavern infilling. One of the most significant features of caves is that they are one of the few places in the terrestrial environment where there is at least the possibility of preserving quite long sequences of deposits representing multiple climatic cycles. Long stratigraphic sequences are comparatively rare on land in the Pleistocene.

One of the great biological phenomena of the Pleistocene is the terminal Pleistocene extinction event, and a huge body of literature has been devoted to its likely causes. Somewhere around the very end of the Pleistocene and extending in numerous cases into the early Holocene, a large number of large to medium-sized mammals died out in a relatively short time. It is informative to look at the anatomy of this extinction event. The areas in which the extinctions took place were Europe, northern Asia, Australia, and the Americas. Generally speaking, the faunas of southern Asia and Africa survived intact. There had previously been numerous rapid transitions from late glacial to early temperate environments, but no other event during the Pleistocene had seen global-scale extinctions. Some have suggested that environmental change alone caused mass extinction, but the evidence on the ground does not support this case very well.

Southern Asia and Africa both seem to have a long and continuous record of human activity, extending well back into the Pleistocene, and in Africa earlier still. The mammals of these areas are likely to have had long-standing adaptive responses to the presence of dangerous and effective bipedal predators. In Europe, the co-existence of people and native mammals was markedly sporadic, particularly during the later Pleistocene, whereas in much of northern Asia and in Australia and the Americas co-existence was either very limited or perhaps even completely absent. When anatomically modern humans began to spread across Europe, Asia, and beyond, the animals living in these regions would probably never have encountered people before and had little or no time to adapt to their presence. If this natural naivety of the animal populations was coupled with the great environmental stress of rapid environmental change, undoubtedly causing a natural crash in population numbers, then this is more likely to account for the widespread but highly regionalized pattern of extinction.

One particular case history is illuminating. By the time modern human populations had spread along the newly formed Arctic Ocean shoreline at the end of the Pleistocene, sea-levels were already beginning to rise, and a population of mammoths had become isolated beyond human reach on what is now Wrangel Island. Though mammoth populations on the European and Asian mainland and throughout North America dwindled to extinction by the end of the Pleistocene, the isolated animals on Wrangel Island survived to as recently as 4500 years ago, in one of the most inhospitable environments on Earth. It was only when humans eventually reached the island that the mammoths finally vanished. So when the great megalithic monument at Stonehenge was under construction on the grasslands of southern England, there were still woolly mammoths alive on an island in the Arctic Ocean.

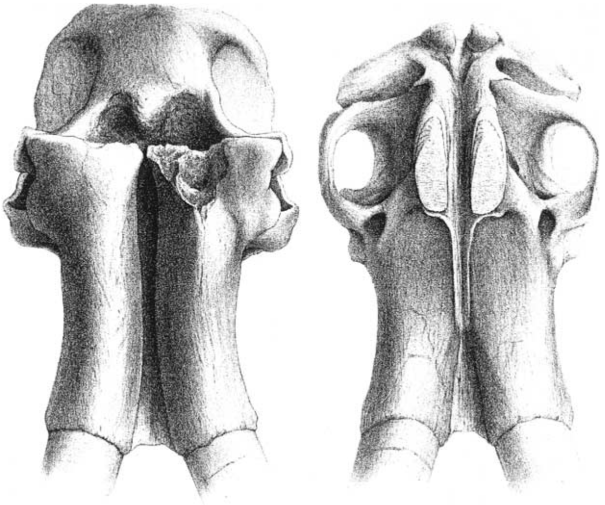


Figure 3 Engraving of the skull of a Pleistocene mammoth, *Mammuthus* cf. *primigenius*, from Ilford, Essex. Megaherbivores such as the mammoth were significant creators and modifiers throughout the Pleistocene and their activities helped maintain the preferred habitats of many other plants and animals.

The effects of mammalian extinctions in the Americas and in Australia were particularly severe, with the loss of all of the megaherbivores (animals weighting a tonne or more; Figure 3) as well as a large number of medium-sized species. Mammoths, mastodonts, giant sloths, glyptodons, native horses, notoungulates, litopterns, and many other forms were completely wiped out in the Americas, and Australia lost all of its larger marsupials. Large animals with long regeneration times and slow-growing, vulnerable young seem to have been pushed to extinction by the rapid spread of modern humans to new parts of the globe. The timing of both events coincides perfectly and it seems strange not to accept a direct link between the two phenomena. The significance of the loss of the Pleistocene megaherbivores cannot be overstated, for these animals were the environment creators of the higher latitudes. The vast cold, dry grasslands of Eurasia and North America were created and maintained by creatures such as mammoths throughout the Pleistocene. Once these animals became extinct, the environment that they had helped to create disappeared with them, destroying the habitat that had supported many smaller species, and the northern coniferous woodlands took over. In the longer term, human populations have effectively replaced the larger Pleistocene herbivorous mammals in living primarily on cultivated grasses such as wheat, maize, rice, and millet. In summary, the Pleistocene was a period of large-scale, cyclic environmental change. Plants and animals showed considerable adaptive responses to these changes, but in the case of the environmentally important megaherbivores, their progress was cut short, probably by the rapid spread of modern humans.

Further Reading

Delcouty HR and Delcourt HA (1991) *Quaternary Ecology.* London: Chapman & Hall.

Lowe JJ and Walker MJC (1984) *Reconstructing Quaternary Environments.* England: Longman Scientific and Technical.

THERMAL METAMORPHISM

R Abart and R Milke, University of Basel, Basel, Switzerland

Introduction

In the course of geological processes, rocks may be subject to changing pressure-temperature and chemical conditions. This may induce mineral transformations and re-crystallization. If the mineralogical and (micro)structural changes take place, when the rock is in the solid state, this is referred to as metamorphism. Metamorphic transformations, which are primarily driven by elevated temperature at constant (low) pressure, may be referred to as thermal metamorphism. In this chapter, we summarize the settings in which thermal metamorphism may occur. We then describe characteristic mineralogical patterns of thermal metamorphism. We address the timing of thermal metamorphism and its implications on the extent of equilibration and the development of microstructures and reaction textures. Finally, we address fluid flow and associated chemical transport during thermal metamorphism.

Geological Settings of Thermal Metamorphism

Both pressure and temperature increase with depth. Whereas isobars, i.e., loci of equal pressure, constitute approximately horizontal surfaces, the thermal structure of the Earth's crust may be rather complex (Figure 1). Heat is constantly liberated by radioactive decay and by slow cooling and solidification of the Earth's interior. Heat is transported to the relatively cool surface primarily through thermal conduction. In the absence of mass flow associated with tectonic and magmatic activity and with fluid circulation, a time invariant relation between depth and temperature would be established in the subsurface, which may be referred to as a stable geothermal gradient. Conduction of heat is a relatively slow process, and the geothermal gradient may be modified, if heat is transported passively with moving matter. There are several processes which may cause deviations from a stable geothermal gradient towards high temperatures.

Processes, which may cause thermal perturbations on a regional scale

Large fragments of the lithosphere may be displaced by tectonic or gravity mechanisms. If the movement includes a vertical component, the uplifted lithosphere transports heat to shallower depths, thereby raising and condensing the isotherms in the crust. An efficient mechanism to transport thermal energy is magmatic underplating. This is the formation of extended mafic intrusions at deep-crustal or subcrustal level. Seismic signals indicate that underplating melt bodies may have thicknesses of several kilometres. Such large magma bodies increase the heat flow to the Earth's surface on a regional scale. Due to the lower density of the underplating magma compared to the lithospheric mantle, it may also give rise to surface uplift. The elevation of the highest cenozoic mountain ranges cannot be explained by isostatic movement of the post-collisional thickened lithosphere alone. They require a combination of isostatic uplift, magmatic underplating, and thinning of the lithospheric mantle from its root. Areas of increased heat flow are also found in zones of extensional geotectonic regime. Thinning of the lithosphere may occur by horizontal stretching or movement on normal faults. At the base of the lithosphere, mantle material rises to compensate for the lost volume. Since the rate of tectonic transport is generally faster than the rate of heat transfer, each point of the pre-extensional geotherm moves vertically to establish the new geotherm.

All regional scale processes that lead to thermal perturbations, increased heat flow, and raised geotherms involve vertical displacement of the crust

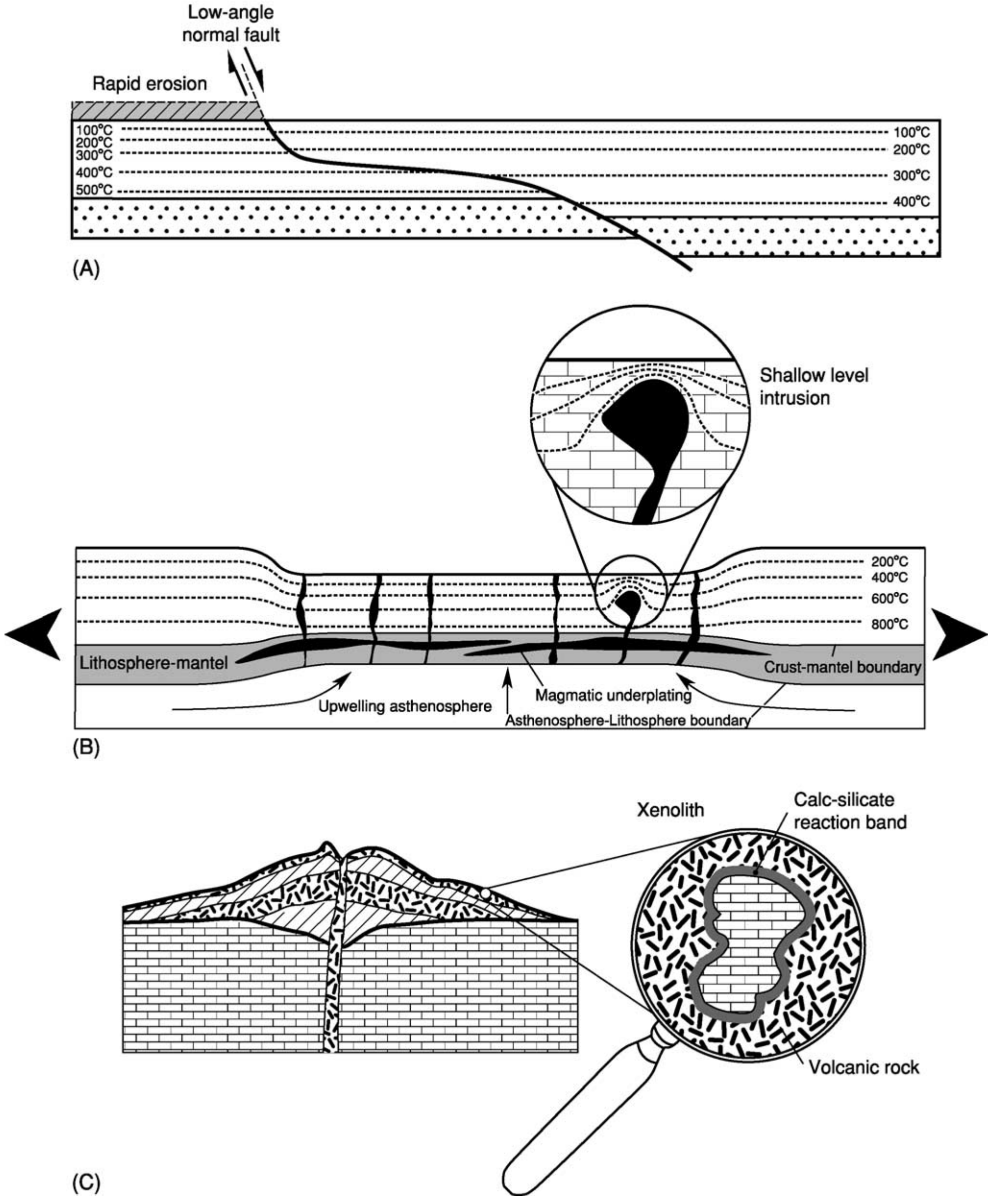


Figure 1 Settings of thermal metamorphism: (A) condensation of isotherms during tectonic exhumation of crustal fragments along low-angle normal fault; (B) condensation of isotherms in the course of lithospheric thinning in an extensional regime; also shown is the effect of magmatic underplating, which is often associated with lithospheric thinning; insert shows the effect on the thermal structure of a shallow level intrusion; (C) thermal overprint of xenoliths in a volcanic setting.

and thus induce both changes in temperature and pressure conditions. Thermal metamorphism in a narrow sense, that is at constant pressure, is restricted to processes that operate on a local scale.

Processes, which may cause thermal perturbations on a local scale

Magmatic intrusions Large amounts of heat may be transported into shallow levels of the crust via the intrusion of magmas. Heat is liberated during solidification (latent heat of crystallization) and during cooling of the magma intrusion and it is transferred into the country rocks via thermal conduction and possibly via fluid circulation. The domain of thermal perturbation surrounding a shallow intrusion is referred to as the thermal aureole or the contact aureole and the associated metamorphic transformations are referred to as contact metamorphism. Within a thermal aureole peak metamorphic temperatures increase exponentially with decreasing distance from the intrusive contact, from background temperatures at the outer limit to maximum temperatures at the intrusive

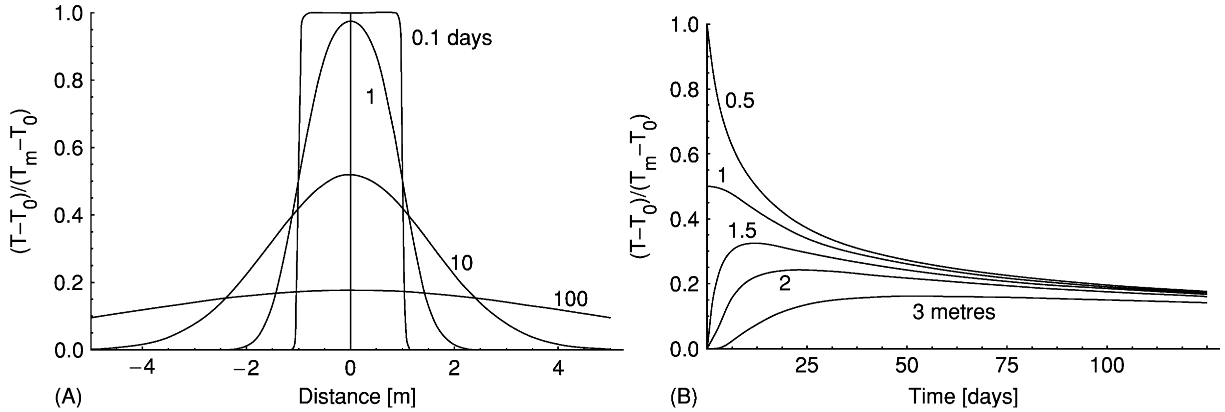


Figure 2 Temperature distribution around a dyke intrusion; for construction of this figure it was assumed that the magma intruded at a temperature T_m and the country rock was at a uniform temperature T_0 prior to magma intrusion; the magma and the country rock were assigned similar thermal diffusivities of 10^{-6} m^2/s; (A) shows the temperature distribution across the dike for different times after intrusion; (B) shows the evolution of temperature with time at different distances from the dyke centre.

contact. This gives rise to characteristic sequences of metamorphic parageneses (see below). Peak metamorphic temperatures at the intrusive contact, among others, depend on the temperature of the intruding magma, the temperature of the country rocks prior to magma intrusion, and on the rate and extent of heat production during crystallisation. The temperature pattern produced by the intrusion of a magmatic dike and its evolution with time are illustrated in Figure 2. It is important to note that the thermal evolution of a rock depends on its distance from the dike. Whereas the dyke material is subject to strictly monotonic cooling after intrusion the country rocks go through an initial heating stage before they start cooling. Peak metamorphic temperatures and heating rates are high close to the intrusive contact and relatively low further out in the country rock. Penetrative deformation associated with magma intrusion is usually confined to the immediate contact region and metamorphic crystallization is usually static in large portions of the contact aureole.

Volcanic activity Effusion of lavas in the course of volcanic activity is an obvious setting of thermal metamorphism. On the one hand, lavas may substantially heat the material over which they flow and cause sharp thermal aureoles of limited extent. On the other hand, on their way to the surface, magmas may incorporate fragments of country rocks, which then are quickly heated to magmatic temperatures. Many of the calc-silicate minerals have first been described from limestone derived xenoliths, which were metamorphosed in lavas of the Monte Somma volcano in southern Italy.

Thermal metamorphism on a regional scale is associated with the production of basaltic lavas at mid-ocean ridges. Thermal metamorphism in the mid-ocean-ridge environment is usually accompanied by intense hydrothermal activity, which may lead to pronounced chemical alteration of the original mid-ocean-ridge basalts.

Natural coal seam and hydrocarbon seepage burns, the production of fulgurites from lightning strikes, the formation of pseudotachylites from frictional heating, and the transformation of kinetic into thermal energy associated with meteoritic impacts may be considered as short-term and very localised special cases of thermal metamorphism.

Mineral Zones in Thermal Metamorphism

Mineral zones are defined by the systematic appearance of new minerals in a series of rocks with identical whole rock composition. These minerals are referred to as index minerals. Mineral zones can be mapped in the field.

Isograds are lines on a map connecting points of equal metamorphic grade. The practical application of this definition to rocks is problematic since there are sets of P-T-X conditions that satisfy reaction equilibrium for a given mineral assemblage. In practice, it is therefore difficult to exactly pin down the metamorphic grade and it is more practical to use the concept of reaction isograds to describe metamorphic rock series. Reaction isograds are lines joining points that are characterized by the equilibrium assemblage of a given reaction.

In thermal metamorphism, temperature is the leading variable controlling metamorphic grade. Contact aureoles are characterised by spatial patterns of reaction isograds that parallel the intrusive contact

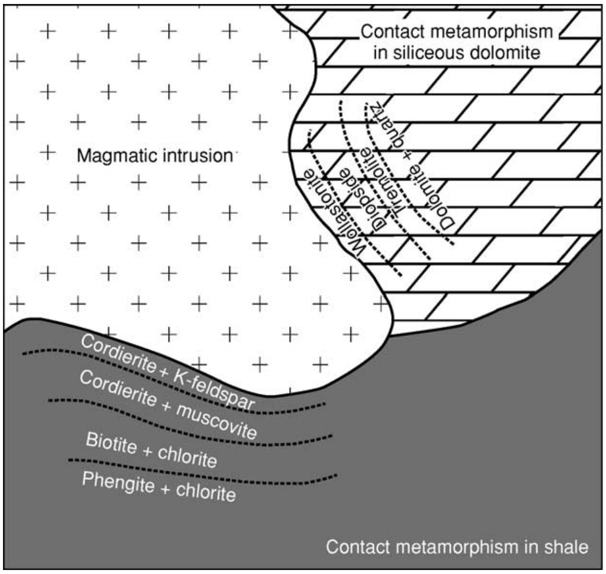


Figure 3 Schematic map of a shallow magmatic intrusion into country rocks of sedimentary origin. For a given lithology, different mineral zones characterized by index minerals evolve depending on the distance to the heat source.

and result in mineral zones (Figure 3). Contact metamorphism is most pronounced if the intruded rocks were previously unmetamorphosed or had only been subjected to low-grade regional metamorphism. The respective succession of mineral zones depends on the lithology of the metamorphosed rocks and comprises different index minerals for contact metamorphism of shales, impure carbonates, ultramafic rocks, or other country rocks. For the reconstruction of the thermal evolution in a contact aureole, simple rock compositions with clearly defined reactions are most convenient, e.g., siliceous dolomites (Figure 3).

H_2O and CO_2 are the prime constituents of metamorphic fluids. These species may be liberated or sequestred in the course of mixed-volatile mineral reactions. As a consequence, the composition of the pore fluid is influenced by mineral reactions and has an important influence on the stability of mineral parageneses.

The kinetics of mineral reactions may be slow compared to the time-scales of thermal perturbations. High heating rates in contact aureoles may lead to significant overstepping of mineral equilibria, such that the formation of mineral zones lags behind the thermal evolution.

Fluid-Rock Interaction

Thermal metamorphism, in particular contact metamorphism, is often accompanied by fluid migration. On the one hand, magmas may dissolve significant amounts of volatiles and fluid may be liberated during crystallization. On the other hand, thermal perturbations may cause bouyancy driven fluid circulation. Fluid flow tends to be pervasive during early stages of contact metamorphism, and it may become focused along joints and fractures at later stages. The fluid in the pore space of rocks is the most efficient transport medium in the solid crust. Fluid may transport dissolved species and isotopes. The chemical and isotopic signature of fluids is imprinted on the solid phases of the rocks via fluid rock interaction. If metamorphic transformations are accompanied by changes of the bulk rock chemical composition, this process may be referred to as allochemical metamorphism or metasomatism. Mineralogical and (stable)isotope alteration patterns are robust and enduring manifestations of palaeo fluid flow.

See Also

Igneous Processes. **Igneous Rocks:** Granite. **Metamorphic Rocks:** Classification, Nomenclature and Formation; Facies and Zones. **Mining Geology:** Magmatic Ores.

Further Reading

Bucher K and Frey M (1994) *Petrogenesis of metamorphic rocks*. Berlin: Springer.

Jamtveit B and Yardley BWD (1997) *Fluid Flow and Transport in Rocks*. London: Chapman and Hall.

Kerrick DM (1991) *Contact Metamorphism. Reviews in Mineralogy*. Vol. 26. Mineralogical Society of America.

Kornprobst J (2002) *Metamorphic rocks and their geodynamic significance*. Kluwer Academic Publishers.

Kretz R (1994) *Metamorphic Crystallization*. John Wiley & Sons.

Turcotte DL and Schubert G (1992) *Geodynamics*. John Wiley & Sons Inc.

Winter JD (2001) *An introduction to igneous and metamorphic petrology*. Prentice Hall.

Yardley B (1989) *An introduction to metamorphic petrology*. Longman Earth Sciences Series.

TIME SCALE

F M Gradstein, University of Oslo, Oslo, Norway
J G Ogg, Purdue University, West Lafayette, IN, USA

Introduction

Boundary stratotypes of stages, high-resolution radioisotopic dating, earth-orbit tuning of cyclic sequences, advances in biostratigraphic scaling of stages, and detailed error analysis are keys to the standard geological time-scale. Construction and assembly of the geological time-scale follow several well-defined steps, including (1) construction of an updated global chronostratigraphic scale for Earth's rock record, (2) identification of key linear-age calibration levels for the chronostratigraphic scale using high-resolution radioisotopic (or other source of) absolute age dates, (3) application of earth-orbit tuning to intervals with cyclic sediments or stable isotope sequences that have sufficient biostratigraphic or magnetostratigraphic ties, (4) interpolation of the combined chronostratigraphic and chronometric scale when direct information is insufficient, and (5) calculation or estimation of error statistics on the combined chronostratigraphic and chronometric information to obtain a geological time-scale with estimates of uncertainty on boundaries and unit durations. The International Commission on Stratigraphy (ICS) cosponsors the standard globally applicable geological time-scale.

Human Time

Time is an indispensable tool for all of us. The time kept by innumerable watches and clocks regulates our everyday life, and the familiar calendar governs our weekly, monthly, and yearly doings. These sequences eventually condense into the historical record of events over centuries. The standard unit of modern timekeeping is the second, defined as the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the caesium-133 atom. This value was established to agree as closely as possible with the ephemeris second based on Earth's motion. The advantage of having the atomic second as the unit of time in the International System of Units is the relative ease, in theory, for anyone to build and calibrate an atomic clock with a precision of 1 part per 10^{11} (or better). In practise, clocks are calibrated against broadcast time signals, with frequency oscillations in hertz being the 'pendulum' of the atomic timekeeping device.

The tick of the second paces the quick heartbeat, and traditionally was the 60th part of the 60th part of the 24th part of the 24-h day, with the minute and the hour being convenient multiples to organize daily life. The day carries the record of light and dark, the month is marked by the regularly returning shapes of the moon, and the year is represented by the cycle of the seasons and the apparent path of the sun. All of these time passages are clearly understood, and humans have long recognized the notion that time is a vector, pointing from the present to the future. Events along its path mark the 'arrow of time', and the arrow is graded either in relative 'natural' units, or in units of duration – the standard second and its multiples, such as hours and years.

Geological Time and the Rock Record

What is often less clear is the concept of geological time, the bases of its units, and how to use these units properly. A good understanding of geological time is vital for every earth scientist, especially those who strive to understand geological processes and determine rates of change. This understanding takes place in a framework called 'Earth geological history', a kind of supercalendar of local and global events. The challenge to this understanding is reading, organizing, and sorting Earth's stone calendar pages, and, as best as we can, reconstructing the content of any missing pages. Stratigraphic correlation is a vital part of this event-reconstruction process.

One of the earliest reconstructions was made by Nicolas Steno (*see* **Famous Geologists:** Steno) (1631–1687), who made careful and original stratigraphic observations. Based on these observations, Steno concluded that Earth's strata contain the superimposed records of a chronological sequence of events that can be correlated worldwide. Geological correlation formally is expressed in terms of five consecutive operations and units:

- Rock units, such as formations or well log intervals (lithostratigraphic correlation; e.g., the Kimmeridge Clay Formation of England).
- Fossil units, such as zones (biostratigraphic correlation; e.g., the *Turrilina alsatica* benthic foraminifer range zone).
- Relative time units (geochronological ('Earth time') correlations; e.g., Jurassic Period, Eocene Epoch, Oxfordian Age, Magnetic Polarity Chron C29R).

- Rocks deposited during these time units (chronostratigraphic (time–rock) correlation; e.g., Jurassic System, Eocene Series, Oxfordian Stage, Magnetic Polarity Zone C29R).
- Linear time units or ages (geochronological correlation; e.g., 150 Ma, 10 ka).

Without correlation to a global reference scale, successions of strata (i.e., events in time derived in one area) are unique and contribute nothing to an understanding of Earth history elsewhere. The rules of hierarchy in geological correlation—from rocks and fossils to relative and linear time—are carefully laid down in the *International Stratigraphic Guide*. An abbreviated copy of this 'rule book' with further references may be found on the website of the International Commission on Stratigraphy (http://www.stratigraphy.org).

Before dealing further with linear geological time, consider the common geological calendar built from relative age units. This chronostratigraphic scheme is not unlike a historical calendar in which societal periods (e.g., the Minoan Period, the reign of Louis XIV, or the American Civil War) are used as building blocks, devoid of a linear scale. Archaeological relicts deposited during these intervals (the Palace of Minos on Crete, Versailles, or spent cannon balls at Gettysburg, respectively) comprise the associated physical chronostratigraphic record. A chronostratigraphic scale is assembled from rock sequences stacked and segmented in relative units based on their unique fossil and physical content. When unique local fossil and physical records are matched with those of other rock sequences across the globe – in a process known as stratigraphic correlation – a relative scale can be assembled; when calibrated to stage-type sections, this scale becomes a chronostratigraphic scale.

The standard chronostratigraphic scale, in downloadable graphics format, is available from the ICS website. This time-scale is made up of up of successive stages in the rock record (e.g., Cenomanian, Turonian, and Coniacian within the Cretaceous System). Originally, each stage unit was a well-defined body of rocks at a specific location of an assigned and agreed upon relative age span, younger than typical rocks of the underlying stage (Jurassic) and older than the typical rocks of the next higher stage (Paleogene). This is the concept of defining stage units with stage-type sections, commonly referred to as stratotype sections. The principles and building blocks of this chronostratigraphy were slowly established during centuries of study in many discontinuous and incomplete outcrop sections. Inevitably, lateral changes in lithology between regions and lack of agreement on criteria (particularly, which fossils were characteristic of which relative unit of rock), have always resulted in considerable confusion and disagreement with respect to stage nomenclature and use. Almost invariably, classical stage stratotypes turned out to represent only parts of stages. Hence, a suite of global subdivisions with precise correlation horizons was required.

Global Stratotype Section and Point

Now, relatively rapid progress is being made with definition of Global Stratotype Sections and Points (GSSPs) to fix the lower boundary of all geological stages, using discrete fossil and physical events that correlate well in the rock record. For the ladder of chronostratigraphy, this GSSP concept switches the emphasis from marking the spaces between steps (stage stratotypes) to fixing the rungs (stage boundaries).

Each progressive pair of GSSPs in the rock record also precisely defines the associated subdivision of geological time. Hence, philosophical arguments are being heard to cut out time–rock units, and deal only with rock, fossil, and time units *sensu strictu*. Essentially, the argument is that dual systems of precisely defined subdivisions of geological time and of parallel similarly defined subdivisions of the time–rock record are redundant. Or, even more radical, why not replace the hundreds of melodious, but confusing, 'ian' subdivisions of stages (Gelasian, Sinemurian, Spathian, Emsian, etc.) with simple 'real' ages – who needs the 'Victorian Age' when there is a 'nineteenth century'?

It is now 25 years ago that a 'golden spike' struck the first GSSP. This event (of historic proportions for development of the geological time-scale) involved the boundary between the Silurian and Devonian periods, or rather the lower limit of the Devonian, at a locality called Plonk in Czechoslovakia. The problem of the Silurian–Devonian boundary and its consensus settlement in the Klonk section hinged on a century-old debate, known as the 'Hercynian Question', that touched many outstanding geoscientists of the previous century. The issue came to the foreground after 1877, when Kaiser stated that the youngest stages (étages) of Barrande's 'Silurian System' in Bohemia correspond to the Devonian System in the Harz Mountains of Germany and other regions. Kaiser's findings contrasted with the conventional nineteenth-century wisdom that graptolites became extinct at the end of the Silurian. Eventually, it became clear that so-called Silurian graptolites in some sections occur together with so-called Devonian fossils in other sections, leading to the modern consensus that graptolites are not limited to Silurian strata.

A bronze plaque in the Plonk outcrop shows the exact position of the modern Silurian–Devonian Boundary, which is taken at the base of the Lochkovian Stage, the lowest stage of the Devonian. The base of the Lochkovian Stage is defined by the first occurrence of the Devonian graptolite *Monograptus uniformis* in Bed #20 of the Klonk Section, northeast of the village of Suchomasty. The Lower Lochkovian index trilobites with representatives of the *Warburgella rugulosa* group occur in the next younger limestone Bed #21 of that section.

The concept of the GSSP has gained acceptance among those stratigraphers who consider it a pragmatic and practical solution to the common problem that conventional stage-type sections inevitably leave gaps, or lead to overlap between successive stages. The boundary stratotype very much relies on the notion that it is possible to arrive at accuracy in correlation through the use of events (e.g., a geomagnetic reversal, a global change in a stable isotope value, or the evolutionary appearance of one or more prominent and widespread fossil taxa). Thus, the limits of a stage can now be defined with multiple event criteria that, using the best of current knowledge, are synchronous over the world. Delimiting successive stages in a clear and practical manner enhances their value as standard units in chronostratigraphy and ultimately in geochronology. Without standardised units, neither the (relative) stratigraphic scale nor the (absolute) time-scale can exist.

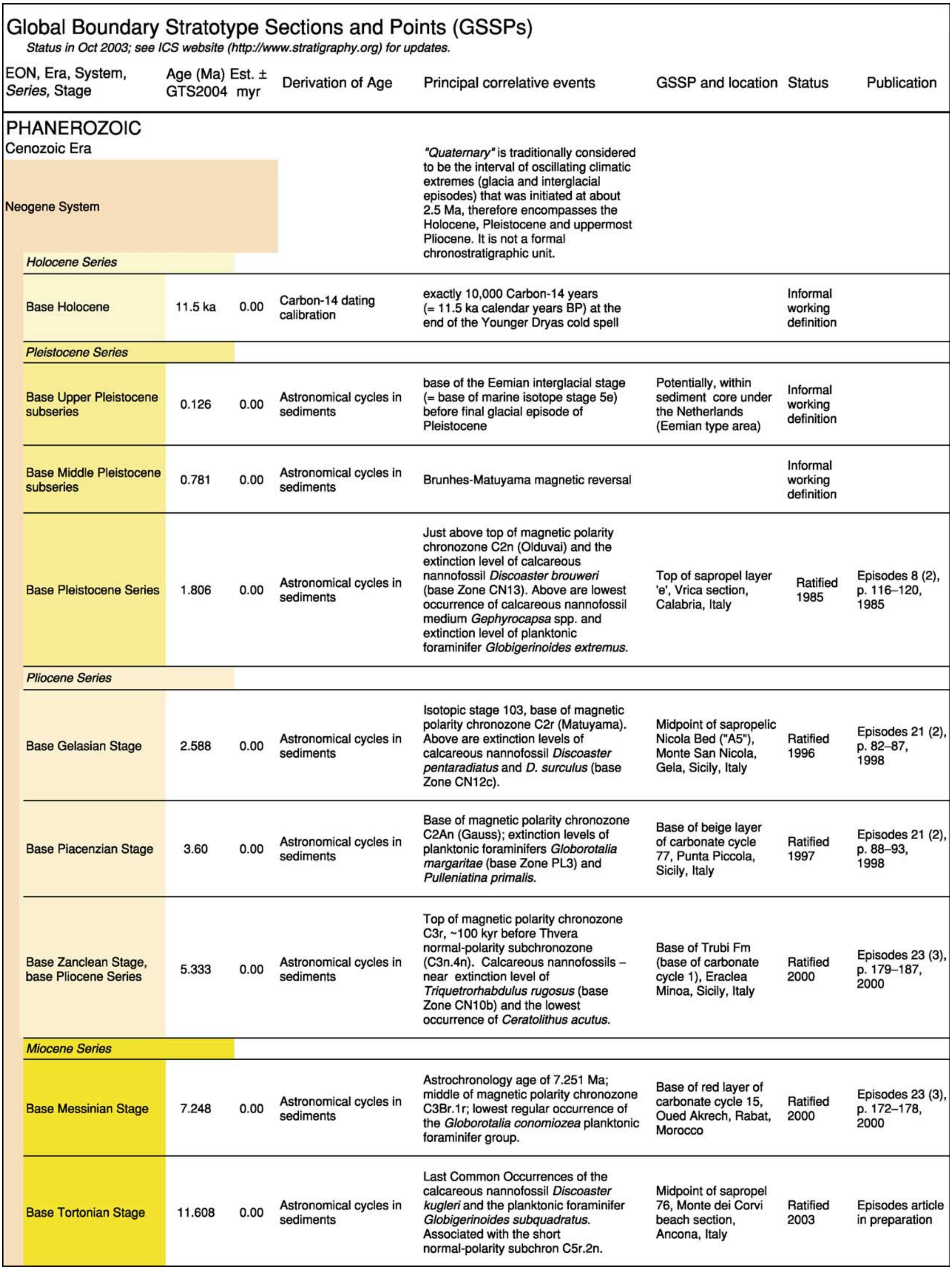
This is not to say that the classical concept of the stage stratotype has suddenly become obsolete, and should be abolished. Although the GSSP concept is not ruled by priority regulation, going back through the historical notions of how and where a stage was originally conceived, defined, and correlated sheds valuable light on the geological meaning and correlative content of stages and the historical notion of their boundaries. Nevertheless, it is clear that, at a time when scientific viewpoints are becoming increasingly more global, stage stratotypes have more regional rather than global significance. At present, nearly 50 GSSPs have been defined (Figures 1 and 2). Over 50 more Phanerozoic stages are in need of base definition, and ICS has set the year 2008 as the completion date for all remaining GSSPs.

Due to the fact that most of the Proterozoic lacks adequate fossils for correlation, a different type of boundary stratotype, the Global Standard Stratigraphic Age (GSSA), is in use for this interval of Earth history. The definition of a boundary by its linear age is the consequence of the fact that the Proterozoic now recognizes units of global stratigraphic subdivision, with the boundaries being defined in terms of the abstract age in millions of years. Summaries of ratified GSSAs may be found in Figure 2. Although there appears to be consensus that the subdivision of the Proterozoic in three eras (Palaeoproterozoic, Mesoproterozoic, and Neoproterozoic) is excellent, the finer subdivisions often contain no dated rocks, which makes their use haphazard. An intensive search is going on for physical events in the Proterozoic rock record suitable to qualify as 'golden spikes'.

It is desirable to add some documentation to a GSSP proposal about the historical significance of a new-stage GSSP in the hierarchy of stratigraphic units of higher rank. Careful consideration needs to be given to the fact that, for example, the GSSP for the base of the Triassic Period in Meishan, China, automatically defines the boundary of the Permian–Triassic, and the base of the Induan Stage. Because the Triassic is considered the lowest period in the Mesozoic, the golden spike in the Meishan section also separates the Palaeozoic from Mesozoic. In the same vein, the base of the Cambrian Period as defined in a GSSP in the Fortune Head section in south-west Newfoundland also defines the base of the Nemakitian-Daldynian Stage, often considered to be the lowest stage in the Cambrian. Because the Cambrian, by consensus, is considered to be the lowest period in the Palaeozoic Era and the lowest major unit in the Phanerozoic Aeon, the Fortune Head GSSP also defines the lower boundary of these major chronostratigraphic units. This leads to correlation in linear time units, called geochronological correlation, with reference to the geochronological calendar of Earth events. Whereas the chronostratigraphic scale is a convention to be agreed upon rather than discovered, calibration of the scale in seconds and (mega-) years is a matter for discovery and estimation rather than agreement. Like human time, linear geological time is expressed in units of standard duration – the second, and hence (thousands or millions of) years.

Building a Geological Time-Scale

The ideal time-scale is built from accurate radioisotopic ages, taken precisely at stage boundaries throughout the stratigraphic column in the Phanerozoic. There are two basic measuring (dating) tools to build the linear geological time-scale: (1) stratigraphically meaningfull readiometric dates in millions of years and (2) Earth-orbit-tuned sedimentary cycles in thousands of years. Both types of measurements are achieved in rocks that are stratigraphically discontinuous. Also required is a certain amount of luck in finding core or outcrop sections suitable for dating with index fossils, geomagnetic reversals, stable isotope anomalies, etc.



Global Boundary Stratotype Sections and Points (GSSPs)

Status in Oct 2003; see ICS website (http://www.stratigraphy.org) for updates.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
PHANEROZOIC Cenozoic Era Neogene System				*"Quaternary"* is traditionally considered to be the interval of oscillating climatic extremes (glacia and interglacial episodes) that was initiated at about 2.5 Ma, therefore encompasses the Holocene, Pleistocene and uppermost Pliocene. It is not a formal chronostratigraphic unit.			
Holocene Series							
Base Holocene	11.5 ka	0.00	Carbon-14 dating calibration	exactly 10,000 Carbon-14 years (= 11.5 ka calendar years BP) at the end of the Younger Dryas cold spell		Informal working definition	
Pleistocene Series							
Base Upper Pleistocene subseries	0.126	0.00	Astronomical cycles in sediments	base of the Eemian interglacial stage (= base of marine isotope stage 5e) before final glacial episode of Pleistocene	Potentially, within sediment core under the Netherlands (Eemian type area)	Informal working definition	
Base Middle Pleistocene subseries	0.781	0.00	Astronomical cycles in sediments	Brunhes-Matuyama magnetic reversal		Informal working definition	
Base Pleistocene Series	1.806	0.00	Astronomical cycles in sediments	Just above top of magnetic polarity chronozone C2n (Olduvai) and the extinction level of calcareous nannofossil *Discoaster brouweri* (base Zone CN13). Above are lowest occurrence of calcareous nannofossil medium *Gephyrocapsa* spp. and extinction level of planktonic foraminifer *Globigerinoides extremus.*	Top of sapropel layer 'e', Vrica section, Calabria, Italy	Ratified 1985	Episodes 8 (2), p. 116–120, 1985
Pliocene Series							
Base Gelasian Stage	2.588	0.00	Astronomical cycles in sediments	Isotopic stage 103, base of magnetic polarity chronozone C2r (Matuyama). Above are extinction levels of calcareous nannofossil *Discoaster pentaradiatus* and *D. surculus* (base Zone CN12c).	Midpoint of sapropelic Nicola Bed ("A5"), Monte San Nicola, Gela, Sicily, Italy	Ratified 1996	Episodes 21 (2), p. 82–87, 1998
Base Piacenzian Stage	3.60	0.00	Astronomical cycles in sediments	Base of magnetic polarity chronozone C2An (Gauss); extinction levels of planktonic foraminifers *Globorotalia margaritae* (base Zone PL3) and *Pulleniatina primalis.*	Base of beige layer of carbonate cycle 77, Punta Piccola, Sicily, Italy	Ratified 1997	Episodes 21 (2), p. 88–93, 1998
Base Zanclean Stage, base Pliocene Series	5.333	0.00	Astronomical cycles in sediments	Top of magnetic polarity chronozone C3r, ~100 kyr before Thvera normal-polarity subchronozone (C3n.4n). Calcareous nannofossils – near extinction level of *Triquetrorhabdulus rugosus* (base Zone CN10b) and the lowest occurrence of *Ceratolithus acutus.*	Base of Trubi Fm (base of carbonate cycle 1), Eraclea Minoa, Sicily, Italy	Ratified 2000	Episodes 23 (3), p. 179–187, 2000
Miocene Series							
Base Messinian Stage	7.248	0.00	Astronomical cycles in sediments	Astrochronology age of 7.251 Ma; middle of magnetic polarity chronozone C3Br.1r; lowest regular occurrence of the *Globorotalia conomiozea* planktonic foraminifer group.	Base of red layer of carbonate cycle 15, Oued Akrech, Rabat, Morocco	Ratified 2000	Episodes 23 (3), p. 172–178, 2000
Base Tortonian Stage	11.608	0.00	Astronomical cycles in sediments	Last Common Occurrences of the calcareous nannofossil *Discoaster kugleri* and the planktonic foraminifer *Globigerinoides subquadratus.* Associated with the short normal-polarity subchron C5r.2n.	Midpoint of sapropel 76, Monte dei Corvi beach section, Ancona, Italy	Ratified 2003	Episodes article in preparation

Figure 1 Overview of Global Stratotype Sections and Points in the Mesozoic and Cenozoic. The Internet version (see http://www.stratigraphy.org) is being updated regularly to reflect the growing number of ratified stratigraphic boundaries. Reprinted from the International Commission on Stratigraphy (2003), with permission.

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Serravallian Stage	13.64	0.00	Astronomical cycles in sediments	Near lowest occurrence of nannofossil *Sphenolithus heteromorphus,* and within magnetic polarity chronozone C5ABr		GSSP anticipated in 2004	
Base Langhian Stage	15.97	0.0	Calibrated magnetic anomaly scale	Near first occurrence of planktonic foraminifer *Praeorbulina glomerosa* and top of magnetic polarity chronozone C5Cn.1		GSSP anticipated in 2004	
Base Burdigalian Stage	20.43	0.0	Calibrated magnetic anomaly scale	Near lowest occurrence of planktonic foraminifer *Globigerinoides altiaperturus* or near top of magnetic polarity chronozone C6An		Guide event is undecided	
Base Aquitanian Stage, base Miocene Series, base Neogene System	23.03	0.0	Astronomical cycles in sediments	Base of magnetic polarity chronozone C6Cn.2n; lowest occurrence of planktonic foraminifer *Paragloborotalia kugleri*; near extinction of calcareous nannofossil *Reticulofenestra bisecta* (base Zone NN1).	35 m from top of Lemme-Carrosio section, Carrosio village, north of Genoa, Italy	Ratified 1996	Episodes 20 (1), p. 23–28, 1997
Paleogene System							
Oligocene Series							
Base Chattian Stage	28.4	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n. Arbitrary 100 kyr uncertainty assigned.	Planktonic foraminifer, extinction of *Chiloguembelina* (base Zone P21b)	Probably in Umbria-Marche region of Italy	GSSP anticipated in 2004	
Base Rupelian Stage, base Oligocene Series	33.9	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Planktonic foraminifer, extinction of *Hantkenina*	Base of marl bed at 19m above base of Massignano quarry, Ancona, Italy	Ratified 1992	Episodes 16 (3), p. 379–382, 1993
Eocene Series							
Base Priabonian Stage	37.2	0.1	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Near lowest occurrence of calcareous nannofossil *Chiasmolithus oamaruensis* (base Zone NP18)	Probably in Umbria-Marche region of Italy		
Base Bartonian Stage	40.4	0.2	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Near extinction of calcareous nannofossil *Reticulofenestra reticulata*			
Base Lutetian Stage	48.6	0.2	Calibrated magnetic anomaly scale relative to base-Miocene and C24n.	Planktonic foraminifer, lowest occurrence of *Hantkenina*	Leading candidate is Fortuna section, Murcia province, Betic Cordilleras, Spain	GSSP anticipated in 2004	
Base Ypresian Stage, base Eocene Series	55.8	0.2	Astronomical cycles in sediments scaled from base-Paleocene	Base of negative carbon-isotope excursion	Dababiya section near Luxor, Egypt	Ratified 2003	Episodes article in preparation
Paleocene Series							
Base Thanetian Stage	58.7	0.2	Astronomical cycles in sediments scaled from base Paleocene, using base of magnetic polarity chronozone C26n. Arbitrary 0.1 (2 precession cycles, plus the base-Paleogene radiometric) uncertainty assigned to all estimates.	Magnetic polarity chronozone, base of C26n, is a temporary assignment	Leading candidate is Zumaya section, northern Spain	Guide event is undecided	
Base Selandian Stage	61.7	0.2	Astronomical cycles in sediments scaled from base Paleocene, using magnetic polarity chronozone placement of C27n.9	Boundary task group is considering a higher level - base of calcareous nannofossil zone NP5 – which would be ~1 myr younger.	Leading candidate is Zumaya section, northern Spain	Guide event is undecided	

Continued

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Danian Stage, base Paleogene System, base Cenozoic	65.5	0.3	Ar-Ar and U-Pb age agreement	Iridium geochemical anomaly. Associated with a major extinction horizon (foraminifers, calcareous nannofossils, dinosaurs, etc.);	Base of boundary clay, El Kef, Tunisia *(but deterioration may require assigning a replacement section)*	Ratified 1991	
Mesozoic Era							
Cretaceous System				*Most substages of Cretaceous also have recommended GSSP criteria*			
Upper							
Base Maastrichtian Stage	70.6	0.6	Estimated placement relative to Ar-Ar calibrated Sr-curve	Mean of 12 biostratigraphic criteria of equal importance. Closely above is lowest occurrence of ammonite *Pachydiscus neubergicus.* Boreal proxy is lowest occurrence of belemnite *Belemnella lanceolata.*	115.2 m level in Grande Carrière quarry, Tercis-les-Bains, Landes province, SW France	Ratified 2001	Episodes 24 (4), p. 229–238, 2001; Odin (ed.) IUGS Spec. Publ. Series, v.36, Elsevier, 910pp.
Base Campanian Stage	83.5	0.7	Spline fit of Ar-Ar ages and ammonite zones.	Crinoid, extinction of *Marsupites testudinarius*	Leading candidates are in southern England and in Texas		
Base Santonian Stage	85.9	0.7	Spline fit of Ar-Ar ages and ammonite zones.	Inoceramid bivalve, lowest occurrence of *Cladoceramus undulatoplicatus*	Leading candidates are in Spain, England and Texas		
Base Coniacian Stage	89.3	1.0	Spline fit of Ar-Ar ages and ammonite zones.	Inoceramid bivalve, lowest occurrence of *Cremnoceramus rotundatus* (*sensu* Tröger *non* Fiege)	Base of Bed MK47, Salzgitter Salder Quarry, SW of Hannover, Lower Saxony, northern Germany	GSSP anticipated in 2004	
Base Turonian Stage	93.6	0.8	Spline fit of Ar-Ar ages and ammonite zones.	Ammonite, lowest occurrence of *Watinoceras devonense*	Base of Bed 120, Rock Canyon Anticline, east of Pueblo, Colorado, west-central USA	Ratified 2003	Episodes article in preparation
Base Cenomanian Stage	99.6	0.9	Spline fit of Ar-Ar ages and ammonite zones, plus monitor standard correction. Then cycle stratigraphy to place foraminifer datum relative to ammonite zonation.	Planktonic foraminifer, lowest occurrence of *Rotalipora globotruncanoides*	36 m below top of Marnes Bleues Formation, Mont Risou, Rosans, Haute-Alpes, SE France	Ratified 2002	Episodes article in preparation
Lower							
Base Albian Stage	112.0	1.0	Estimated placement relative to bases of Cenomanian and Aptian, with large uncertainty due to lack of GSSP criteria. Ar-Ar age of 114.6 +/– 0.7 Ma from *Parahoplites nutfieldensis* below.	Calcareous nannofossil, lowest occurrence of *Praediscosphaera columnata* (= *P. cretacea* of some earlier studies), is one potential marker.		Guide event is undecided	
Base Aptian Stage	125.0	1.0	Base of M0r, as recomputed from Ar-Ar age from MIT guyot	Magnetic polarity chronozone, base of M0r	Leading candidate is Gorgo a Cerbara, Piobbico, Umbria-Marche, central Italy		
Base Barremian Stage	130.0	1.5	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at M5n.8.	Ammonite, lowest occurrence of *Spitidiscus hugii – Spitidiscus vandeckii* group	Leading candidate is Río Argos near Caravaca, Murcia province, Spain		
Base Hauterivian Stage	136.4	2.0	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at base M11n.	Ammonite, lowest occurrence of genus *Acanthodiscus* (especially *A. radiatus*)	Leading candidate is La Charce village, Drôme province, southeast France		
Base Valanginian Stage	140.2	3.0	Pacific spreading model for magnetic anomaly ages (variable rate), using placement at M14r.3 (base *T. pertransiens*).	Calpionellid, lowest occurrence of *Calpionellites darderi* (base of – Calpionellid Zone E); followed by the lowest occurrence of ammonite *"Thurmanniceras" pertransiens*	Leading candidate is near Montbrun-les-Bains, Drôme province, southeast France		

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Berriasian Stage, base Cretaceous System	145.5	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base of *Berriasella jacobi* zone (M19n.2n.55)	Maybe near lowest occurrence of ammonite *Berriasella jacobi*		Guide event is undecided	
Jurassic System							
Upper							
Base Tithonian Stage	150.8	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M22An	Near base of *Hybonoticeras hybonotum* ammonite zone and lowest occurrence of *Gravesia* genus, and the base of magnetic polarity chronozone M22Ar		Guide event is undecided	
Base Kimmeridgian Stage	155.7	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M26r.2 (Boreal ammonite definition)	Ammonite, near base of *Pictonia baylei* ammonite zone of Boreal realm	Leading candidates are in Scotland, SE France and Poland	GSSP anticipated in 2004	
Base Oxfordian Stage	161.2	4.0	Pacific spreading model for magnetic anomaly ages (variable rate), assigning to base M36An	Ammonite, *Brightia thuouxensis* Horizon at base of the *Cardioceras scarburgense* Subzone (*Quenstedtoceras mariae* Zone)	Leading candidates are in SE France and southernEngland	GSSP anticipated in 2004	
Middle							
Base Callovian Stage	164.7	4.0	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of the genus *Kepplerites (Kosmoceratidae)* (defines base of *Macrocephalites herveyi* Zone in sub-Boreal province of Great Britain to southwest Germany)	Leading candidate is Pfeffingen, Swabian Alb, SW Germany	GSSP anticipated in 2004	
Base Bathonian Stage	167.7	3.5	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of *Parkinsonia (G.) convergens* (defines base of *Zigzagiceras zigzag* Zone)			
Base Bajocian Stage	171.6	3.0	Equal subzones scale Bajo-Bath-Callov	Ammonite, lowest occurrence of the genus *Hyperlioceras* (defines base of the *Hyperlioceras discites* Zone)	Base of Bed AB11, 77.8 m above base of Murtinheira section, Cabo Mondego, western Portugal	Ratified 1996	Episodes 20 (1), p. 16–22, 1997
Base Aalenian Stage	175.6	2.0	Duration of Aalenian-Toarcian from cycle stratigraphy	Ammonite, lowest occurrence of *Leioceras* genus	base of Bed FZ107, Fuentelsalz, central Spain	Ratified 2000	Episodes 24 (3), p. 166–175, 2001
Lower							
Base Toarcian Stage	183.0	1.5	Duration of Aalenian-Toarcian from cycle stratigraphy	Ammonite, near lowest occurrence of a diversified *Eodactylites* ammonite fauna; correlates with the NW European *Paltus* horizon.		Guide event is undecided	
Base Pliensbachian Stage	189.6	1.5	Cycle-scaled linear Sr trend	Ammonite, lowest occurrences of *Bifericeras donovani* and of genera *Apoderoceras* and *Gleviceras.*	Wine Haven section, Robin Hood's Bay, Yorkshire, England, UK	GSSP anticipated in 2003	
Base Sinemurian Stage	196.5	1.0	Cycle-scaled linear Sr trend	Ammonite, lowest occurrence of arietitid genera *Vermiceras* and *Metophioceras*	0.9 m above base of Bed 145, East Quantoxhead, Watchet, West Somerset, SW England, UK	Ratified 2000	Episodes 25 (1), p. 22–26, 2002
Base Hettangian Stage, base Jurassic System	199.6	0.6	U-Pb age just below proposed GSSP for base-Jurassic	Near lowest occurrence of smooth *Psiloceras planorbis* ammonite group		Guide event is undecided	
Triassic System							
Upper							
Base Rhaetian Stage	203.3	1.5	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Near lowest occurrence of ammonite *Cochlocera*,conodonts *Misikella* spp. and *Epigondolella mosheri*, and radiolarian *Proparvicingula moniliformis.*	Key sections in Austria, British Columbia (Canada), and Turkey	Guide event is undecided	

Continued

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Norian Stage	216.5	2.0	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Base of *Klamathites macrolobatus* or *Stikinoceras kerri* ammonoid zones and the *Metapolygnathus communisti* or *M. primitius* conodont zones.	Leading candidates are in British Columbia (Canada), Sicily (Italy), and possibly Slovakia, Turkey (Antalya Taurus) and Oman.	Guide event is undecided	
Base Carnian Stage	228.0	2.0	Magnetostratigraphic correlation to cycle-scaled Newark magnetic polarity pattern	Near first occurrence of the ammonoids *Daxatina* or *Trachyceras*, and of the conodont *Metapolygnathus polygnathiformis*	Candidate section at Prati di Stuores, Dolomites, northern Italy. Important reference sections in Spiti (India) and New Pass, Nevada (USA).	Guide event is undecided	
Middle							
Base Ladinian Stage	237.0	2.0	U-Pb array by Mundil *et al.* on levels near *Nevadites* (= *Secedensis*) ammonite zone in Dolomites, plus placement relative to magnetostratigraphy corrlations to cycle-scaled Newark magnetic polarity pattern	Alternate levels are near base of *Reitzi, Secedensis,* or *Curionii* ammonite zone; near first occurrence of the conodont genus *Budurovignathus.*	Leading candidates are Bagolino (Italy) and Felsoons (Hungary). Important reference sections in the Humboldt Range, Nevada (USA).	Guide event is undecided	
Base Anisian Stage	245.0	1.5	Proportional subzonal scaling	Ammonite, near lowest occurrences of genera *Japonites, Paradanubites, and Paracrochordiceras;* and of the conodont *Chiosella timorensis*	Candidate section probable at Desli Caira, Dobrogea, Romania; significant sections in Guizhou Province (China).	GSSP anticipated in 2004	
Lower							
Base Olenekian Stage	249.7	0.7	Composite standard from conodonts scaled to base-Anisian and base-Triassic	Near lowest occurrence of *Hedenstroemia* or *Meekoceras gracilitatis* ammonites, and of the conodont *Neospathodus waageni.*	Candidate sections in Siberia (Russia) and probably Chaohu, Anhui Province, China. Important sections also in Spiti.	Guide event is undecided	
Base Induan Stage, base Triassic System, base Mesozoic	251.0	0.4	U-Pb ages bracket GSSP (Bowring *et al.*, 1998)	Conodont, lowest occurrence of *Hindeodus parvus*; termination of major negative carbon-isotope excursion. About 1 myr after peak of Late Permian extinctions.	Base of Bed 27c, Meishan, Zhejiang, China	Ratified 2001	Episodes 24 (2), p. 102–114, 2001

The steps involved in modern time-scale construction can be summarized as follows:

Step 1. Construct an updated global chronostratigraphic scale for Earth's rock record.

Step 2. Identify key linear-age calibration levels for the chronostratigraphic scale using radioisotopic age dates.

Step 3. Apply Earth-orbit tuning to intervals with cyclic sediments or stable isotope sequences that have sufficient biostratigraphic or magnetostratigraphic ties.

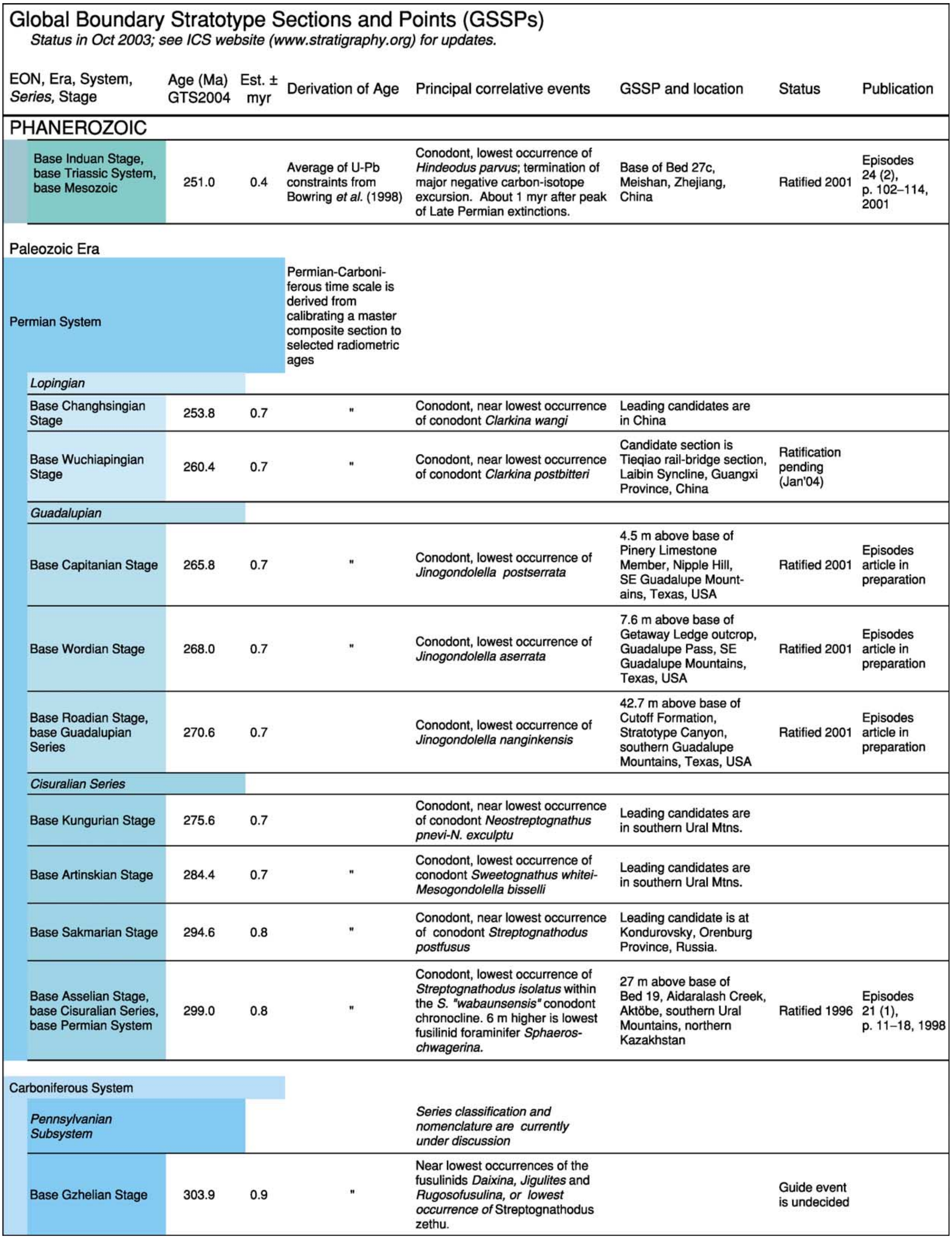
Step 4. Interpolate the combined chronostratigraphic and chronometric scale when direct information is insufficient.

Step 5. Calculate or estimate error bars on the combined chronostratigraphic and chronometric information to obtain a geological time-scale with estimates of uncertainty on boundaries and on unit durations.

Step 6. Peer-review the resultant geological time-scale.

The first step, integrating multiple types of stratigraphic information in order to construct the chronostratigraphic scale, is the most time-consuming; it summarizes and synthesizes centuries of detailed geological research and tries to understand all relative correlations and calibration to the standard chronostratigraphic scale.

The second step, identifying which radiometric and cycle-stratigraphic studies to use as the primary constraints for assigning linear ages, is the one that has much evolved. Historically, Phanerozoic time-scale building went from an exercise with very few and relatively inaccurate radioisotopic dates, as available to the pioneer of the geological time-scale Arthur Holmes, to one with many dates with greatly varying analytical precision, as in the mid-1980s of the past century. Next, time-scale studies started



Global Boundary Stratotype Sections and Points (GSSPs)

Status in Oct 2003; see ICS website (www.stratigraphy.org) for updates.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
PHANEROZOIC							
Base Induan Stage, base Triassic System, base Mesozoic	251.0	0.4	Average of U-Pb constraints from Bowring *et al.* (1998)	Conodont, lowest occurrence of *Hindeodus parvus*; termination of major negative carbon-isotope excursion. About 1 myr after peak of Late Permian extinctions.	Base of Bed 27c, Meishan, Zhejiang, China	Ratified 2001	Episodes 24 (2), p. 102–114, 2001
Paleozoic Era							
Permian System			Permian-Carboniferous time scale is derived from calibrating a master composite section to selected radiometric ages				
Lopingian							
Base Changhsingian Stage	253.8	0.7	"	Conodont, near lowest occurrence of conodont *Clarkina wangi*	Leading candidates are in China		
Base Wuchiapingian Stage	260.4	0.7	"	Conodont, near lowest occurrence of conodont *Clarkina postbitteri*	Candidate section is Tieqiao rail-bridge section, Laibin Syncline, Guangxi Province, China	Ratification pending (Jan'04)	
Guadalupian							
Base Capitanian Stage	265.8	0.7	"	Conodont, lowest occurrence of *Jinogondolella postserrata*	4.5 m above base of Pinery Limestone Member, Nipple Hill, SE Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Base Wordian Stage	268.0	0.7	"	Conodont, lowest occurrence of *Jinogondolella aserrata*	7.6 m above base of Getaway Ledge outcrop, Guadalupe Pass, SE Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Base Roadian Stage, base Guadalupian Series	270.6	0.7		Conodont, lowest occurrence of *Jinogondolella nanginkensis*	42.7 m above base of Cutoff Formation, Stratotype Canyon, southern Guadalupe Mountains, Texas, USA	Ratified 2001	Episodes article in preparation
Cisuralian Series							
Base Kungurian Stage	275.6	0.7		Conodont, near lowest occurrence of conodont *Neostreptognathus pnevi-N. exculptu*	Leading candidates are in southern Ural Mtns.		
Base Artinskian Stage	284.4	0.7	"	Conodont, lowest occurrence of conodont *Sweetognathus whitei-Mesogondolella bisselli*	Leading candidates are in southern Ural Mtns.		
Base Sakmarian Stage	294.6	0.8	"	Conodont, near lowest occurrence of conodont *Streptognathodus postfusus*	Leading candidate is at Kondurovsky, Orenburg Province, Russia.		
Base Asselian Stage, base Cisuralian Series, base Permian System	299.0	0.8	"	Conodont, lowest occurrence of *Streptognathodus isolatus* within the *S. "wabaunsensis"* conodont chronocline. 6 m higher is lowest fusilinid foraminifer *Sphaeroschwagerina.*	27 m above base of Bed 19, Aidaralash Creek, Aktöbe, southern Ural Mountains, northern Kazakhstan	Ratified 1996	Episodes 21 (1), p. 11–18, 1998
Carboniferous System							
Pennsylvanian Subsystem				*Series classification and nomenclature are currently under discussion*			
Base Gzhelian Stage	303.9	0.9	"	Near lowest occurrences of the fusulinids *Daixina, Jigulites* and *Rugosofusulina, or lowest occurrence of* Streptognathodus zethu.		Guide event is undecided	

Figure 2 Overview of Global Stratotype Sections and Points and Global Standard Stratigraphic Ages in the Palaeozoic and Precambrian, respectively. The Internet version (see http://www.stratigraphy.org) is being updated regularly to reflect the growing number of ratified stratigraphic boundaries. Reprinted from the International Commission on Stratigraphy (2003), with permission.

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Kasimovian Stage	306.5	1.0	"	Near base of *Obsoletes obsoletes* and *Protriticites pseudomontiparus* fusulinid zone, or lowest occurrence of *Parashumardites* ammonoid.		Guide event is undecided	
Base Moscovian Stage	311.7	1.1	"	Near lowest occurrences of *Declinognathodus donetzianus* and/or *Idiognathoides postsulcatus* conodont species, and fusulinid species *Aljutovella aljutovica.*		Guide event is undecided	
Base Bashkirian Stage, base Pennsylvanian Subsystem	318.1	1.3	"	Conodont, lowest occurrence of *Declinognathodus nodiliferus* s.l.	82.9 m above top of Battleship Wash Fm., Arrow Canyon, southern Nevada, USA	GSSP ratified 1996. Subsystem rank of Mississippian and Pennsylvanian names ratified 2000.	Episodes 22 (4), p. 272–283, 1999
Mississippian Subsystem				*Series classification and nomenclature are currently under discussion*			
Base Serpukhovian	326.4	1.6	"	Near lowest occurrence of conodont, *Lochriea crusiformis.*		Guide event is undecided	
Base Visean	345.3	2.1	"	Foraminifer, lineage *Eoparastaffella simplex* morphotype 1/morphotype 2	Leading candidate is Pengchong, south China		
Base Tournaisian, base Mississippian Subsystem, base Carboniferous System	359.2	2.5	"	Conodont, above lowest occurrence of *Siphonodella sulcata*	Base of Bed 89, La Serre, Montagne Noir, Cabrières, southern France	Ratified 1990	Episodes 14 (4), p. 331–336, 1991
Devonian System			Devonian time scale is a statistical fit of a composite biostratigraphic zonation (based on Figure 8 of Williams *et al.*, 2000) to selected radiometric ages				
Upper							
Base Famennian Stage	374.5	2.6	"	Just above major extinction horizon (Upper Kellwasser Event), including conodonts *Ancyrodella* and *Ozarkodina and goniatites of Gephuroceratidae and Beloceratidae*	base of Bed 32a, upper Coumiac quarry, Cessenon, Montagne Noir, southern France	Ratified 1993	Episodes 16 (4), p. 433–441, 1993
Base Frasnian Stage	385.3	2.6	"	Conodont, lowest occurrence of *Ancyrodella rotundiloba* (defines base of Lower *Polygnathus asymmetricus* conodont Zone)	Base of Bed 42a', Col du Puech de la Suque section, St. Nazaire-de Ladarez, SE Montagne Noir, southern France	Ratified 1986	Episodes 10 (2), p. 97–101, 1987
Middle							
Base Givetian Stage	391.8	2.7	"	Conodont, lowest occurrence of Polygnathus hemiansatus, near base of goniatite Maenioceras Stufe	Base of Bed 123, Jebel Mech Irdane ridge, Tafilalt, Morocco	Ratified 1994	Episodes 18 (3), p. 107–115, 1995
Base Eifelian Stage	397.5	2.7	"	Conodont, lowest occurrence of *Polygnathus costatus partitus*; major faunal turnover	Base unit WP30, trench at Wetteldorf Richtschnitt, Schönecken-Wetteldorf, Eifel Hills, western Germany	Ratified 1985	Episodes 8 (2), p. 104–109, 1985
Lower							
Base Emsian Stage	407.0	2.8	"	Conodont, lowest occurrence of *Polygnathus kitabicus* (= *Po. dehiscens*)	Base of Bed 9/5, Zinzil' ban Gorge, SE of Samarkand, Uzbekistan	Ratified 1995	Episodes 20 (4), p. 235–240, 1997
Base Pragian Stage	411.2	2.8	"	Conodont, lowest occurrence of *Eognathodus sulcatus*	Base of Bed 12, Velká Chuchle quarry, southwest part of Prague city, Czech Republic	Ratified 1989	Episodes 12 (2), p. 109–113, 1989

EON, Era, System, *Series,* Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base Lochkovian Stage, base Devonian System	416.0	2.8	base-Devonian from scale in Cooper (this volume), which is 1 myr younger than Tucker et al (1998) estimate.	Graptolite, lowest occurrence of *Monograptus uniformis*	Within Bed 20, Klonk, Barrandian area, southwest of Prague, Czech Republic	Ratified 1972	Martinsson (ed.), The Silurian-Devonian Boundary, IUGS Series A, no.5, 349 pp., 1977
Silurian System			Silurian and Ordovician time scales are from calibrating a CONOP composite graptolite zonation to selecte radiometric ages				Holland and Bassett (eds), *A Global Standard for the Silurian System,* Nat. Mus. Wales, Geol. Series No.10, Cardiff, 325 pp., 1989
Pridoli Series							
Base Prídolí Series *(not subdivided in stages)*	418.7	2.7	"	Graptolite, lowest occurrence of *Monograptus parultimus*	Within Bed 96, Pozáry section near Reporje, Barrandian area, Prague, Czech Republic	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Ludlow Series		2.6					
Base Ludfordian Stage	421.3	2.6	"	*Imprecise.* May be near base of *Saetograptus leintwardinensis* graptolite zone.	Base of lithological unit C, Sunnyhill Quarry, Ludlow, Shropshire, southwest England, UK.	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Base Gorstian Stage	422.9	2.5	"	*Imprecise.*Just below base of local acritarch *Leptobrachion longhopense* range zone. May be near base of *Neodiversograptus nilssoni* graptolite zone.	Base of lithological unit F, Pitch Coppice quarry, Ludlow, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Wenlock Series							
Base Homerian Stage	426.2	2.4	"	Graptolite, lowest occurrence of *Cyrtograptus lundgreni* (defines base of *C. lundgreni* graptolite zone)	Graptolite biozone intersection in stream section in Whitwell Coppice, Homer, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Base Sheinwoodian Stage	428.2	2.3	"	*Imprecise.* Between the base of acritarch biozone 5 and extinction of conodont *Pterospathodus amorphognathoides.* May be near base of *Cyrtograptus centrifugus* graptolite zone.	Base of lithological unit G, Hughley Brook, Apedale, Shropshire, southwest England, UK	Ratified 1980	Lethaia 14, p.168, 1981; Episodes 5 (3), p. 21–23, 1982
Llandovery Series							
Base Telychian Stage	436.1	1.9	"	Brachiopods, just above extinction of *Eocoelia intermedia* and below lowest succeeding species *Eocoelia curtisi.* Near base of *Monograptus turriculatus* graptolite zone.	Locality 162 in transect d, Cefn Cerig road, Llandovery area, south-central Wales, UK	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Base Aeronian Stage	439.0	1.8	"	Graptolite, lowest occurrence of *Monograptus austerus sequens* (defines base of *Monograptus triangulatus* graptolite zone)	Base of locality 72 in transect h, Trefawr forestry road, north of Cwm-coed-Aeron Farm, Llandovery area, south-central Wales, UK	Ratified 1984	Episodes 8 (2), p. 101–103, 1985
Base Rhuddanian Stage, base Silurian System	443.7	1.5	"	Graptolites, lowest occurrences of *Parakidograptus acuminatus* and *Akidograptus ascensus*	1.6 m above base of Birkhill Shale Fm., Dob's Linn, Moffat, Scotland, UK	Ratified 1984	Episodes 8 (2), p. 98–100, 1985
Ordovician System							
Upper							
Base of sixth stage *(not yet named)*	450.2	1.6	"	Potentially near first appearance of the graptolite *Dicellograptus complanatus* and of the conodont *Amorphognathus ordovicicus.*		*Guide event is undecided*	

Continued

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Base of fifth stage *(not yet named)*	460.9	1.6	"	Graptolite, lowest occurrence of *Nemagraptus gracilis*	1.4 m below phosphorite in E14a outcrop, Fågelsång, Scane, southern Sweden	Ratified 2002	Episodes 23 (2), p. 102–109, 2000 *(proposal; formal GSSP publication in preparation).*
Middle							
Base Darriwilian Stage	468.1	1.6	"	Graptolite, lowest occurrence of *Undulograptus austrodentatus*	Base of Bed AEP184, 22 m below top of Ningkuo Fm., Huangnitang, Changshan, Zhejiang province, southeast China	Ratified 1997	Episodes 20 (3), p. 158–166, 1997
Base of third stage *(not yet named)*	471.8	1.6	"	Conodont, lowest occurrence of *Tripodus laevis*			
Lower							
Base of second stage *(not yet named)*	478.6	1.7	"	Graptolite, lowest occurrence of *Tetragraptus approximatus*	Just above E bed, Diabasbrottet quarry, Västergötland, southern Sweden	Ratified 2002	Episodes article in preparation
Base of Tremadocian Stage, base Ordovician System	488.3	1.7	"	Conodont, lowest occurrence of *Iapetognathus fluctivagus*;just above base of *Cordylodus lindstromi* conodont Zone. Just below lowest occurrence of planktonic graptolites. Currently dated around 489 Ma.	Within Bed 23 at the 101.8 m level, Green Point, western Newfoundland, Canada	Ratified 2000	Episodes 24 (1), p. 19–28,
Cambrian System				Potential GSSP correlation levels include *Cordylodus proavus, Glyptagnostus reticulatus, Ptychagnostus punctuosus, Acidusus atavus,* and *Oryctocephalus indicus.*		Overview of potential subdivisions in Episodes 23 (3), p. 188–195, 2000.	
Upper ("Furongian") Series							
Upper stage(s) in Furongian				*Potential GSSP levels in upper Cambrian are based on trilobites and condonts*			
Base Paibian Stage, base Furongian Series	501.0	2.0	Radiometric ages near primary marker level. Estimated age and uncertainty only.	Trilobite, lowest occurrence of agnostoid *Glyptagnostus reticulatus.* Coincides with base of large positive carbon-isotope excursion.	369.06 m above base of Huaqiao Fm, Paibi section, NW Hunan province, south China	Ratified 2003	Episodes article in preparation
Middle	513.0	2.0	Radiometric ages near primary marker level. Estimated age and uncertainty only.	*Potential GSSP levels in Middle Cambrian are based mainly on trilobites*			
Lower				*Potential GSSP levels in Lower Cambrian are based on archaeocyatha, small shelly fossils, and to a lesser extent, trilobites*			
Base Cambrian System, base Paleozoic, base PHANEROZOIC	542.0	1.0	U-Pb age from Oman coinciding with the negative carbon excursion.	Trace fossil, lowest occurrence of *Treptichnus (Phycodes) pedum.* Near base of negative carbon-isotope excursion.	2.4 m above base of Member 2 of Chapel Island Fm., Fortune Head, Burin Peninsula, southeast Newfoundland, Canada	Ratified 1992	Episodes 17 (1&2), p. 3–8, 1994.
PROTEROZOIC				*Pre-Cambrian eras and systems below Ediacaran are defined by absolute ages, rather than stratigraphic points.*			
Neoproterozoic Era							
Base Ediacaran System	600		Vague estimation from bracketing radiometric ages	Termination of Varanger (or Marinoan) glaciation.	Base of the Nuccaleena Formation cap carbonate, immediately above the Elatina diamictite in the Enorama Creek section, Flinders Ranges, South Australia.	Age-definition (650 Ma) ratified in 1990, but ICS now voting on Australian GSSP	Episodes 14 (2), p. 139–140, 1991

EON, Era, System, *Series*, Stage	Age (Ma) GTS2004	Est. ± myr	Derivation of Age	Principal correlative events	GSSP and location	Status	Publication
Cryogenian System	850		Defined chronometrically	Base = 850 Ma			
Tonian System	1000		Defined chronometrically	Base = 1000 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Mesoproterozoic Era							
Stenian System	1200		Defined chronometrically	Base = 1200 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Ectasian System	1400		Defined chronometrically	Base = 1400 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Calymmian System	1600		Defined chronometrically	Base = 1600 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Paleoproterozoic Era							
Statherian System	1800		Defined chronometrically	Base = 1800 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Orosirian System	2050		Defined chronometrically	Base = 2050 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Rhyacian System	2300		Defined chronometrically	Base = 2300 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Siderian System	2500		Defined chronometrically	Base = 2500 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
ARCHEAN							
Neoarchean Era	2800		Defined chronometrically	Base = 2800 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Mesoarchean Era	3200		Defined chronometrically	Base = 3200 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Paleoarchean Era	3600		Defined chronometrically	Base = 3600 Ma		Ratified 1990	Episodes 14 (2), p. 139–140, 1991
Eoarchean Era			Base is not defined				

to appear for selected intervals, such as Paleogene, Late Cretaceous, or Ordovician; these studies selected a small suite of radioisotopic dates with high internal analytical precision and relatively precise stratigraphic position. At the same time, a high-resolution Neogene time-scale started to take shape, using orbital tuning of long sequences of sedimentary and/or oxygen isotope cycles in the Mediterranean region and in Atlantic and Pacific pelagic sediments. The present trend for the pre-Neogene is to incorporate radioisotopic dates that have very small analytical and stratigraphic uncertainties, and pass the most stringent tests.

The third step, interpolating the stratigraphic and radiometric information, has also evolved. An early method had already constructed the basic two-way graph and was being used. This graph plotted the cumulative sum of maximum thicknesses of strata in thousands of feet per stratigraphic unit along the vertical axis and selected dates from volcanic tuffs, glauconites, and magmatic intrusives along the horizontal linear axis. This 'best-fit' line method, shown in Figure 3 as it was used in 1960, incorporated an uncertainty envelope from the errors on the radioisotopic age constraints. Despite its crudeness, the method was remarkably effective, but is a far cry from methods used today.

In the mid-1990s, Frits Agterberg and Felix Gradstein started to apply mathematical-statistical error analysis to the time-scale ages, which, for the first time, allowed them to assign fairly realistic error bars to ages of Mesozoic stage boundaries, a trend that persists today. A simplified introduction to the modern building tools is presented the following discussions, but first consider a frequently asked question: 'Which time-scale should be used?'.

Which Time Scale Should Be Used?

Of the several geological time-scales published in the past decade, a simple answer is to use the most recent one. A better answer is to investigate the pros and cons of these time-scales, the improvements that a new one might bring to a study, and to remember that it is not desirable to change units of measurement during execution of an experiment or survey or basin modelling exercise.

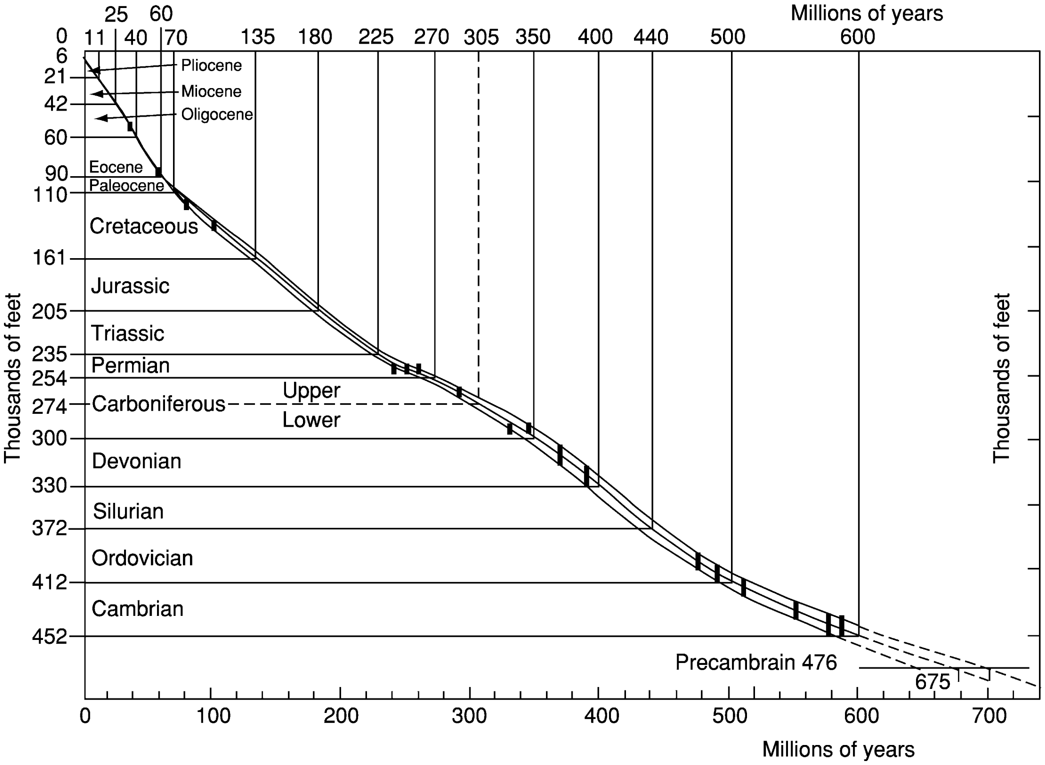


Figure 3 Scaling concept employed by Arthur Holmes in the first half of the twentieth century to construct the geological time-scale. The cumulative sum of maximum thicknesses of strata, in thousands of feet per stratigraphic unit, is plotted along the vertical axis; selected radiometric dates from volcanic tuffs, glauconites, and magmatic intrusives are plotted along the horizontal linear axis. This version incorporated an uncertainty envelope from the errors on the radiometric age constraints. Modified from Holmes A (1960) A revised geological time-scale. *Transactions of the Edinburgh Geological Society* 17: 183–216, with permission.

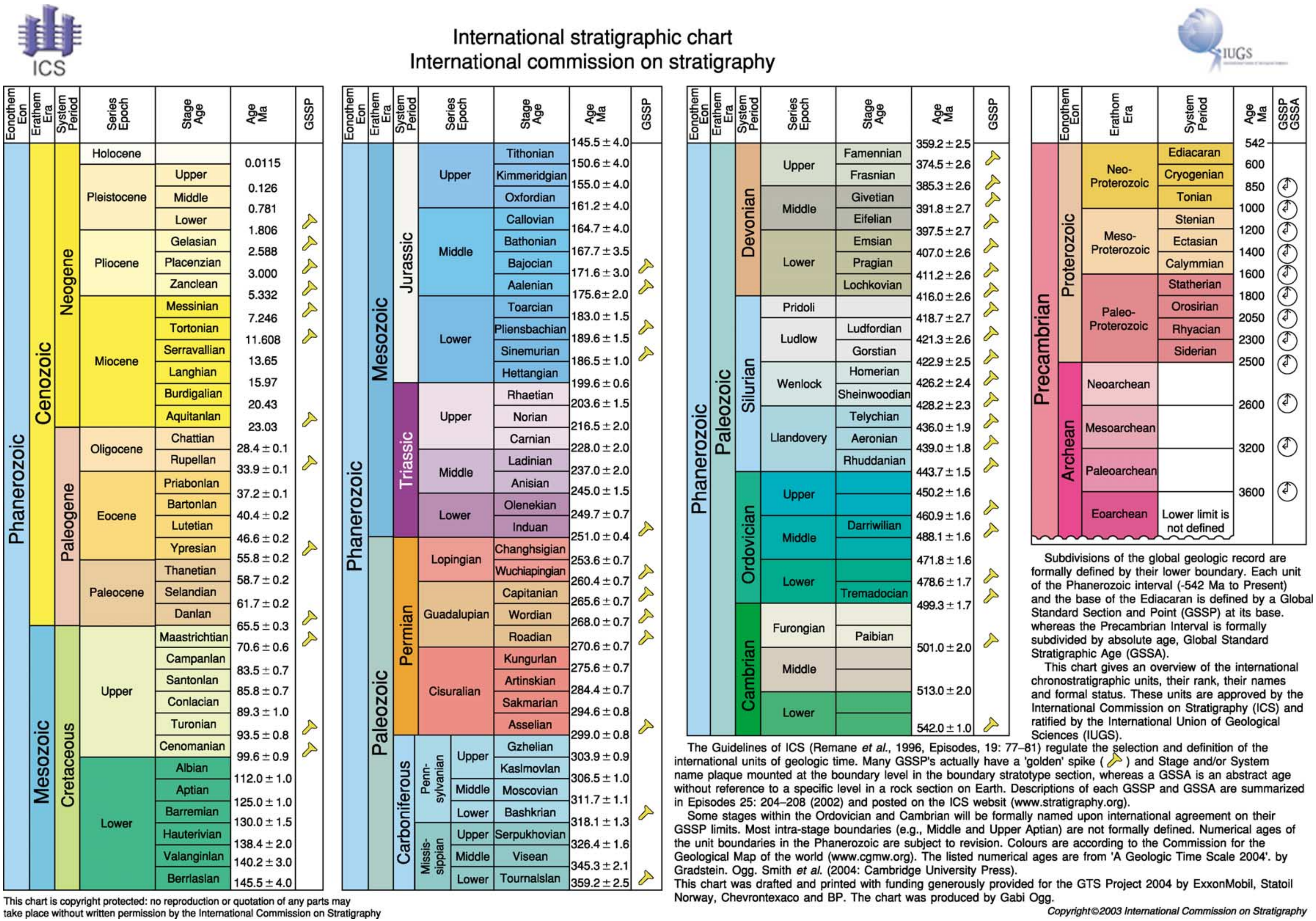
Certain qualities of a geological time-scale should be considered:

1. Does it use updated chronostratigraphic nomenclature of international standard?
2. Does it use all stratigraphic tools relevant to the interval for which the time-scale is intended, including magnetostratigraphy and its chronology, standard biozonations, stable isotope stratigraphy, cyclic stratigraphy interpolations, and/or orbital time-scale calibrations?
3. Are the stratigraphically meaningful radiometric age dates based on rigorous methodology, with the most up-to-date and accepted interlaboratory standards?
4. Does the scale bring out uncertainty in the age of individual stratigraphic boundaries?
5. Is the scale used in major geological projects and studies?

The ICS is actively engaged in time-scale construction and publication, which assists in bringing consensus and stability to the use of this international geological standard. Figure 4 shows the most up-to-date edition of the ICS scale.

Music of the Spheres

The sedimentary cycles approach to time-scale building, as is now standard for the time-scale of the past 23 My (Neogene), provides superior resolution and precision. Gravitational interactions of Earth with the sun, moon, and other planets cause systematic changes in Earth's orbital and rotational systems. These interactions give rise to cyclic oscillations in the eccentricity of Earth's orbit, and in the tilt and precession of Earth's rotational axis, with mean dominant periods of 100 000, 41 000, and 21 000 years, respectively (*see* **Earth:** Orbital Variation (Including Milankovitch Cycles)). The associated cyclic variations in annual and seasonal solar radiation falling onto different latitudes alter long-term climate in colder versus warmer, and wetter versus dryer, periods. These, in turn, create easily recognizable sedimentary cycles, such as regular interbeds of limy and shaly facies. Massive outcrops of hundreds or thousands of such cycles are observed in numerous geological basins (e.g., around the Mediterranean) and in sediment cores from ocean-based drilling sites.



International stratigraphic chart
International commission on stratigraphy

Eonothem Eon	Erathem Era	System Period	Series Epoch	Stage Age	Age Ma	GSSP
Phanerozoic	Cenozoic	Neogene	Holocene		0.0115	
			Pleistocene	Upper	0.126	
				Middle	0.781	
				Lower	1.806	✓
			Pliocene	Gelasian	2.588	✓
				Placenzian	3.000	✓
				Zanclean	5.332	✓
			Miocene	Messinian	7.246	✓
				Tortonian	11.608	✓
				Serravallian	13.65	
				Langhian	15.97	
				Burdigalian	20.43	
				Aquitanlan	23.03	✓
		Paleogene	Oligocene	Chattian	28.4 ± 0.1	
				Rupellan	33.9 ± 0.1	✓
			Eocene	Priabonlan	37.2 ± 0.1	
				Bartonlan	40.4 ± 0.2	
				Lutetian	46.6 ± 0.2	
				Ypresian	55.8 ± 0.2	✓
			Paleocene	Thanetian	58.7 ± 0.2	
				Selandian	61.7 ± 0.2	
				Danlan	65.5 ± 0.3	✓
	Mesozoic	Cretaceous	Upper	Maastrichtian	70.6 ± 0.6	✓
				Campanlan	83.5 ± 0.7	
				Santonlan	85.8 ± 0.7	
				Conlacian	89.3 ± 1.0	
				Turonian	93.5 ± 0.8	✓
				Cenomanian	99.6 ± 0.9	✓
			Lower	Albian	112.0 ± 1.0	
				Aptian	125.0 ± 1.0	
				Barremian	130.0 ± 1.5	
				Hauterivian	138.4 ± 2.0	
				Valanginlan	140.2 ± 3.0	
				Berrlaslan	145.5 ± 4.0	

Eonothem Eon	Erathem Era	System Period	Series Epoch		Stage Age	Age Ma	GSSP
						145.5 ± 4.0	
Phanerozoic	Mesozoic	Jurassic	Upper		Tithonian	150.6 ± 4.0	
					Kimmeridgian	155.0 ± 4.0	
					Oxfordian	161.2 ± 4.0	
			Middle		Callovian	164.7 ± 4.0	
					Bathonian	167.7 ± 3.5	
					Bajocian	171.6 ± 3.0	✓
					Aalenian	175.6± 2.0	✓
			Lower		Toarcian	183.0 ± 1.5	
					Pliensbachian	189.6 ± 1.5	✓
					Sinemurian	186.5 ± 1.0	✓
					Hettangian	199.6 ± 0.6	
		Triassic	Upper		Rhaetian	203.6 ± 1.5	
					Norian	216.5 ± 2.0	
					Carnian	228.0 ± 2.0	
			Middle		Ladinian	237.0 ± 2.0	
					Anisian	245.0 ± 1.5	
			Lower		Olenekian	249.7 ± 0.7	
					Induan	251.0 ± 0.4	✓
	Paleozoic	Permian	Lopingian		Changhsigian	253.6 ± 0.7	
					Wuchiapingian	260.4 ± 0.7	✓
			Guadalupian		Capitanian	265.6 ± 0.7	✓
					Wordian	268.0 ± 0.7	✓
					Roadian	270.6 ± 0.7	✓
			Cisuralian		Kungurlan	275.6 ± 0.7	
					Artinskian	284.4 ± 0.7	
					Sakmarian	294.6 ± 0.8	
					Asselian	299.0 ± 0.8	✓
		Carboniferous	Penn-sylvanian	Upper	Gzhelian	303.9 ± 0.9	
					Kaslmovlan	306.5 ± 1.0	
				Middle	Moscovian	311.7 ± 1.1	
				Lower	Bashkrian	318.1 ± 1.3	✓
			Missis-sippian	Upper	Serpukhovian	326.4 ± 1.6	
				Middle	Visean	345.3 ± 2.1	
				Lower	Tournalslan	359.2 ± 2.5	✓

Eonothem Eon	Erathem Era	System Period	Series Epoch	Stage Age	Age Ma	GSSP
					359.2 ± 2.5	
Phanerozoic	Paleozoic	Devonian	Upper	Famennian	374.5 ± 2.6	✓
				Frasnian	385.3 ± 2.6	✓
			Middle	Givetian	391.8 ± 2.7	✓
				Eifelian	397.5 ± 2.7	✓
			Lower	Emsian	407.0 ± 2.6	✓
				Pragian	411.2 ± 2.6	✓
				Lochkovian	416.0 ± 2.6	✓
		Silurian	Pridoli		418.7 ± 2.7	✓
			Ludlow	Ludfordian	421.3 ± 2.6	✓
				Gorstian	422.9 ± 2.5	✓
			Wenlock	Homerian	426.2 ± 2.4	✓
				Sheinwoodian	428.2 ± 2.3	✓
			Llandovery	Telychian	436.0 ± 1.9	✓
				Aeronian	439.0 ± 1.8	✓
				Rhuddanian	443.7 ± 1.5	✓
		Ordovician	Upper		450.2 ± 1.6	
					460.9 ± 1.6	✓
			Middle	Darriwilian	488.1 ± 1.6	✓
					471.8 ± 1.6	
			Lower		478.6 ± 1.7	✓
				Tremadocian	499.3 ± 1.7	✓
		Cambrian	Furongian			
				Paibian	501.0 ± 2.0	✓
			Middle			
					513.0 ± 2.0	
			Lower			
					542.0 ± 1.0	✓

	Eonothem Eon	Erathom Era	System Period	Age Ma	GSSP GSSA
				542	
Precambrian	Proterozoic	Neo-Proterozoic	Ediacaran	600	
			Cryogenian	850	◷
			Tonian	1000	◷
		Meso-Proterozoic	Stenian	1200	◷
			Ectasian	1400	◷
			Calymmian	1600	◷
		Paleo-Proterozoic	Statherian	1800	◷
			Orosirian	2050	◷
			Rhyacian	2300	◷
			Siderian	2500	◷
	Archean	Neoarchean		2600	◷
		Mesoarchean		3200	◷
		Paleoarchean		3600	◷
		Eoarchean	Lower limit is not defined		

Subdivisions of the global geologic record are formally defined by their lower boundary. Each unit of the Phanerozoic interval (-542 Ma to Present) and the base of the Ediacaran is defined by a Global Standard Section and Point (GSSP) at its base. whereas the Precambrian Interval is formally subdivided by absolute age, Global Standard Stratigraphic Age (GSSA).

This chart gives an overview of the international chronostratigraphic units, their rank, their names and formal status. These units are approved by the International Commission on Stratigraphy (ICS) and ratified by the International Union of Geological Sciences (IUGS).

The Guidelines of ICS (Remane *et al.*, 1996, Episodes, 19: 77–81) regulate the selection and definition of the international units of geologic time. Many GSSP's actually have a 'golden' spike (✓) and Stage and/or System name plaque mounted at the boundary level in the boundary stratotype section, whereas a GSSA is an abstract age without reference to a specific level in a rock section on Earth. Descriptions of each GSSP and GSSA are summarized in Episodes 25: 204–208 (2002) and posted on the ICS websit (www.stratigraphy.org).

Some stages within the Ordovician and Cambrian will be formally named upon international agreement on their GSSP limits. Most intra-stage boundaries (e.g., Middle and Upper Aptian) are not formally defined. Numerical ages of the unit boundaries in the Phanerozoic are subject to revision. Colours are according to the Commission for the Geological Map of the world (www.cgmw.org). The listed numerical ages are from 'A Geologic Time Scale 2004'. by Gradstein. Ogg. Smith *et al.* (2004: Cambridge University Press).

This chart was drafted and printed with funding generously provided for the GTS Project 2004 by ExxonMobil, Statoil Norway, Chevrontexaco and BP. The chart was produced by Gabi Ogg.

Figure 4 The international stratigraphic chart and time-scale for the Phanerozoic, issued by the International Commission on Stratigraphy (ICS) (available on the Internet at http://www.stratigraphy.org). Reprinted from the International Commission on Stratigraphy (2003), with permission.

Greatly detailed counts these centimetre- to metre-thick cycles over land outcrops and in ocean drilling wells, combined with the additional correlation aids provided by magnetostratigraphy, oxygen isotope stratigraphy, and biostratigraphy, have been used to produce a very detailed Neogene sediment/orbital cycle pattern. The critical step is the direct linkage of each cycle to the theoretical computed astronomical scale of the 21 000-, 41 000-, and 100 000-year palaeoclimatic cycles. This astronomical tuning of the geological cycle record from the Mediterranean and Atlantic by earth scientists at Utrecht and Cambridge universities (e.g., Luc Lourens, Frits Hilgen, and Nick Shackleton) has led to unprecedented accuracy and resolution for the past 23 My. In New Zealand, Tim Naish and colleagues have calibrated the Upper Neogene record to the standard Neogene time-scale. Using the high-resolution land-based cycle, the isotope and magnetic record in the Wanganui Basin, these authors thereby transferred precise absolute ages to local shallow marine sediments and demonstrated the link between sequence and cycle stratigraphy.

Efforts are under way to extend the continuous astrochronological scale back into the Oligocene and Eocene by applying a combination of cycle stratigraphy, improved astronomical projections, oxygen isotope stratigraphy, and magnetostratigraphy to the deep-sea stratigraphic record. A special application of orbitally tuned cyclic sediment sequences is to ‘rubber band’ stratigraphically floating units, such as parts of the Paleocene, the Albian, and parts of the Lower Jurassic. A quantitative estimation of the duration of all cycles within a stratigraphic unit allows estimating their duration.

Decay of Atoms

For rocks older than the Neogene, the derivation of a numerical time-scale depends on the availability of suitable radioisotopic ages. Radioisotopic dating generally involves measuring the ratio of the original element in a mineral, such as sanidine feldspar or zircon, to its isotopic daughter products. The age of a mineral may then be calculated by means of the isotopic decay constant. Depending on the half-life of the element, several radioisometric clocks are available; ^{40}Ar/^{39}Ar and the family of U/Pb isotopes are the most common suites currently applied to the Phanerozoic, because of analytical precision and utility with tuffaceous beds in marine or non-marine sequences.

Radioisometric dating of sedimentary rocks follows several geological strategies:

1. Dating of igneous intrusions within sediments records the time of primary cooling, when the igneous rocks were emplaced and had cooled sufficiently (to a few hundreds of degrees centigrade) to set the radiometric decay clock in action. (Note: Because of uncertainty in the relation of the intrusion to the host sediment, such dates may be of limited stratigraphic use.)
2. Dating of volcanic flows and tuffs as part of the stratified sedimentary succession.
3. Dating of authigenic sedimentary minerals (mainly involving glauconite) found commonly in many marine sediments.

Note that mild heating or overburden pressure after burial may lead to loss of argon, the daughter product measured in the ^{40}K/^{40}Ar clock in glauconite. Another problem is that glauconite also contains an abundance of tiny flakes that allow diffusion of Ar at low temperatures. The result is that glauconite dates may be too young. Because of such problems, which may be difficult to detect, modern geological time-scales avoid dates based on glauconite.

Calibration of the decay constants or measurement standards can be enhanced by intercalibration to other radioisotopic methods, or by dating rocks of a known age (for example, a volcanic ash within an astronomically tuned succession). Astrochronological and interlaboratory recalibration of the ^{40}Ar/^{39}Ar monitor standard indicates that many of the ^{40}Ar/^{39}Ar ages used in previous Phanerozoic time-scales are too young by about 0.5–1.0%. For example, the age dating to 65.0 Ma that was assigned 10 years ago to the top Cretaceous is now 65.5 Ma.

Radioisotopic dating techniques with less than 1.0% analytical error are providing suites of high-precision U/Pb and Ar/Ar dates for the Palaeozoic and Mesozoic. The integration of this level of chronometric precision with high-resolution biostratigraphy, magnetostratigraphy, or cyclic scales is a major challenge to time-scale studies. Even the most detailed biostratigraphic scheme probably has no biozonal units of less than half a million years in duration, not to speak of the actual precision in dating a particular ‘stratigraphic piercing’ point, for which a U/Pb age estimate would be available with an analytical uncertainty of 100 000 years or more. Similarly, combining analytically less precise K/Ar dates with much more precise Ar/Ar or U/Pb dates in statistical interpolations creates a strong bias towards the latter, despite the fact that both may have equal litho-, bio-, and chronostratigraphic precision. Nevertheless, the combination of precise stratigraphic definitions through GSSPs and accurate radiometric dates near these levels is paving the way for a substantial increase in the precision and accuracy of the geological time-scale. The bases of Palaeozoic, Mesozoic, and

Cenozoic are bracketed by analytically precise ages at their GSSP or primary correlation markers – 542 ± 1.0, 251.0 ± 0.4, and 65.5 ± 0.3 Ma, respectively – and there are direct age dates for basal Carboniferous, basal Permian, basal Jurassic, basal Aptian, basal Cenomanian, and basal Oligocene. Most other period or stage boundaries lack direct age control. Therefore, the third step, linear interpolation, also plays a key role for development of the time-scale.

Interpolation and Statistics

Despite progress in standardization and dating, parts of the Mesozoic and Palaeozoic have sparse radioisotopic records. Ideally, each of the 90 or more stage boundaries that comprise the Palaeozoic, Mesozoic, and Cenozoic eras of the Phanerozoic should coincide with an accurate radioisotopic date from volcanic ashes that coincide with each of the stage boundaries. However, this coincidence is rare in the geological record. The combined number of fossil events and magnetic reversals far exceeds the total number of radioisotopically datable horizons in the Phanerozoic. Therefore, a framework of bio-, magneto-, and chronostratigraphy provides the principal fabric for stretching of the relative time-scale between dated tiepoints on the loom of linear time. For such stretching, interpolation methods are employed that are both geological and statistical in nature.

The outdated method of plotting the cumulative global thickness of periods against selected linear age dates was previously mentioned. Among the modern geological scaling methods, an assumption of relative constancy of seafloor spreading over limited periods of time is a common tool for interpolating the latest Cretaceous through Palaeogene relative scale. Magnetic polarity chrons – the units of magnetochronology – can be recognized both on the ocean floor, as magnetic anomalies measured in kilometres from the midocean spreading center, and in marine sediments, as polarity zones that contain biostratigraphic events and can be linked to linear time (*see* **Magnetostratigraphy**). Knowing the linear age of a few ocean crust magnetic anomalies (Earth magnetic reversals, or magnetochrons) allows interpolation of the ages of the intervening magnetic pattern, which, in turn, can be correlated to the fossil record and geological stage boundaries. The subduction of pre-Late Jurassic oceanic crust precludes such an interpolation approach for older Mesozoic and Palaeozoic strata.

A second geological method involves building a zonal composite to scale stages. Several outstanding examples are documented in the geological time-scale built by a large team of International Geological Congress (IGC) scientists in 2004. For this scale, Roger Cooper and colleagues have built a very detailed composite standard of graptolite zones from 200 or more sections in oceanic and slope environment basins for the uppermost Cambrian, Ordovician, and Silurian intervals. With zone thickness taken as directly proportional to zone duration, the detailed composite sequence was scaled using selected, high-precision age dates. For the Carboniferous through Permian, a composite standard of conodont, fusulinid, and ammonoid events from many classical sections can now be calibrated to a combination of U/Pb and ^{40}Ar/^{39}Ar dates. A composite standard of conodont zones was used for the Early Triassic. This procedure directly scales all stage boundaries and biostratigraphic horizons.

The two-way graphs of linear age versus scaled stages require a best-fit method, and that is where statistics comes into play, with cubic spline fitting and maximum-likelihood interpolation most suitable. On the time-scale of Figure 4, a majority of Phanerozoic stage boundaries for the first time show error bars; an exception is the Neogene Period, wherein analytical errors are negligible. These error bars reflect both radioisotopic and stratigraphic uncertainty. In addition, error bars were calculated on stage duration. Uncertainty in the duration of the age units is less than the error in age of their boundaries.

See Also

Analytical Methods: Geochronological Techniques. **Biozones**. **Earth:** Orbital Variation (Including Milankovitch Cycles). **Famous Geologists:** Steno. **Magnetostratigraphy**. **Stratigraphical Principles**.

Further Reading

Agterberg FP (2004) Geomathematics. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004.* Cambridge, UK: Cambridge University Press.

Bleeker W (2004) Towards a natural Precambrian time scale. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004.* Cambridge, UK: Cambridge University Press.

Bowring SA, Erwin DH, Jin YG, Martin MW, Davidek K, and Wang W (1998) U/Pb zircon geochronology and tempo of the end-Permian mass extinction. *Science* 280: 1039–1045.

Cande SC and Kent DV (1995) Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *Journal of Geophysical Research* 100: 6093–6095.

Carter RM and Naish TR (eds.) (1999) *The High-Resolution, Chronostratigraphic and Sequence Stratigraphic*

Record of the Plio-Pleistocene, Wanganui Basin, New Zealand. Folio Series 2. Lower Hutt, NZ: Institute of Geological and Nuclear Sciences.

Gradstein FM, Agterberg FP, Ogg JG, Hardenbol J, van Veen P, Thierry J, and Huang Z (1995) A Triassic, Jurassic, and Cretaceous time scale. *SEPM Special Publication* 54: 95–126.

Gradstein FM, Ogg JG, Smith AG, *et al.* (2004) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Harland WB, Armstrong RL, Cox AV, Craig LE, Smith AG, and Smith DG (1990) *A Geologic Time Scale 1989*. New York: Cambridge University Press.

Herbert TD, D'Hondt SL, Premoli-Silva I, Erba E, and Fischer AG (1995) Orbital chronology of Cretaceous-Early Palaeocene marine sediments. In: Berggren WA, Kent DV, and Hardenbol J (eds.) *Geochronology, Time Scales and Global Stratigraphic Correlations: A Unified Temporal Framework for a Historical Geology, SEPM Special Volume, No. 54,* pp. 81–94. Tulsa, OK: SEPM.

Hilgen FJ, Krijgsman W, Langereis CG, and Lourens LJ (1997) Breakthrough made in dating of the geological record. *Eos (Transactions of the American Geophysical Union)* 78(28): 285, 288–289.

Holmes A (1947) The construction of a geological time-scale. *Transactions Geological Society of Glasgow* 21: 117–152.

Holmes A (1960) A revised geological time-scale. *Transactions of the Edinburgh Geological Society* 17: 183–216.

Kamo SL, Czamanske GK, Amelin Y, Fedorenko VA, Davis DW, and Trofimov VR (2003) Rapid eruption of Siberian flood-volcanic rocks and evidence for coincidence with the Permian-Triassic boundary and mass extinction at 251 Ma. *Earth and Planetary Science Letters* 214: 75–91.

Lourens L, Hilgen F, Shackleton NJ, Laskar L, and Wilson D (2004) The Neogene Period. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Martinsson A (ed.) (1977) *The Siluro-Devonian boundary. International Union of Geological Sciences, Series A,* vol. 5. Vienna: International Union of Geological Sciences.

Obradovich JD (1993) A Cretaceous time scale. In: Caldwell WGE (ed.) *Evolution of the Western Interior Basin, Geological Association of Canada Special Paper 39,* pp. 379–396. St. John's, NL: Geological Association of Canada.

Ogg JG (2004) The Triassic, Jurassic and Cretaceous. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge, UK: Cambridge University Press.

Renne PR, Deino AL, Walther RC, Thurrin BD, Swisher CC, Becker TA, Curtis GH, Sharp WD, and Jaouni AR (1994) Intercalibration of astronomical and radio-isotopic time. *Geology* 22: 783.

Röhl U, Ogg JG, Geib TL, and Wefer G (2001) Astronomical calibration of the Danian time scale. *Geological Society, Special Publication* 183: 163–184.

Shackleton NJ, Crowhurst SJ, Weedon GP, and Laskar J (1999) Astronomical calibration of Oligocene-Miocene time. *Philosophical Transactions of the Royal Society of London, Series A* 357: 1907–1929.

Villeneuve M (2004) Radiogenic isotope chronology. In: Gradstein FM, *et al.* (eds.) *A Geologic Time Scale 2004*. Cambridge UK: Cambridge University Press.

Weedon GP, Jenkyns HC, Coe AL, and Hesselberg SP (1999) Astronomical calibration of the Jurassic time-scale from cyclostratigraphy in British mudrock formations. *Philosophical Transactions of the Royal Society of London, Series A* 357: 1787–1813.

TRACE FOSSILS

P J Orr, University College Dublin, Dublin, Ireland

Introduction

Trace fossils and ichnofabrics offer an alternative source of data on the ecology of any palaeoecosystem to that provided by the body fossil record. Different preservational biases apply to the trace and body fossil records. 'Soft-bodied' organisms, i.e., those lacking biomineralized tissues, can produce trace fossils, yet their fossilization potential is minimal. Most trace fossils are emplaced into unconsolidated sediment, cannot survive reworking, and are thus autochthonous (in marked contrast to the vast majority of body fossils). The study of trace fossils and ichnofabrics has made a fundamental and growing contribution to our understanding of the evolutionary palaeoecology of the Earth's biosphere. Furthermore, bioturbation impacts on the physical and chemical properties of sediments, including their diagenesis. Key terms employed in the taxonomic, preservational, and ethological classification of trace fossils are defined; how trace fossils and ichnofabrics are used in palaeoenvironmental and community reconstructions is discussed.

'Trace fossil' is used herein to describe any discrete structure produced by the reworking of sediment or the bioerosion of lithic (rock) or xylic (wood) substrates by infauna or epifauna. Eggs and sedimentary structures produced by physical processes do not represent trace fossils. Some authorities include stromatolites within this term, on which basis

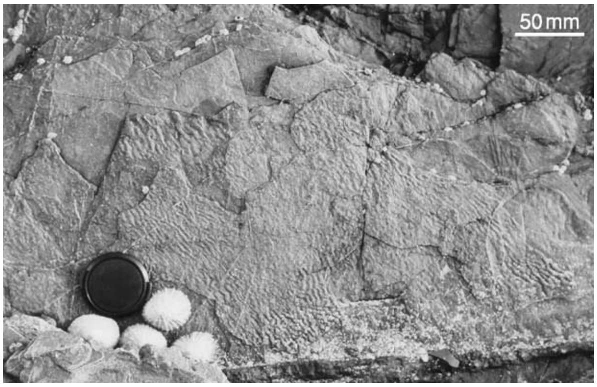


Figure 1 Wrinkle structures in siliciclastic lithologies attributed to the former presence of a microbial mat. Image courtesy of Séan Burke.

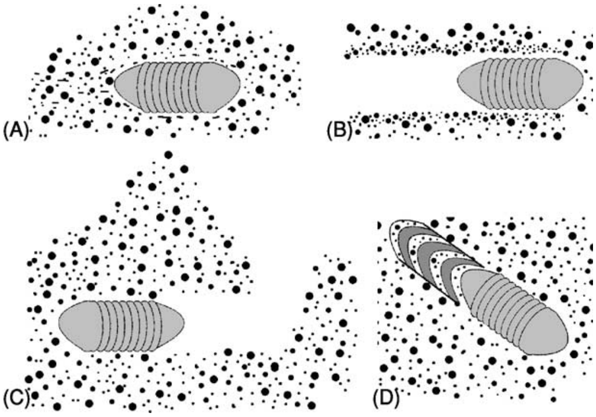


Figure 2 Style of burrowing reflects substrate consistency. (A) Intrusive burrowing typical of near-surface soupgrounds. (B) Compression burrowing generates an open burrow and no spoils. (C) Excavation, yielding an open burrow and advected spoil. (D) Excavation, producing a backfilled burrow. In the latter case, the active infill, i.e., by the producer, may be structureless, but is often sorted; examples include separation into a distinct core and marginal or (as here) meniscate structures. The geometry of individual menisci can vary from shallow and saucer-like to sharply pointed cones; successive menisci are often defined by alternations in their colour and/or composition.

'suspect microbial structures' (e.g., wrinkle structures (Figure 1), roll-up structures, and elephant skin texture) would be included; these are attributed to the presence, including plastic deformation, of microbially bound crusts or veneers on the top of the sediment column. An ichnofabric, all aspects of the sedimentary texture that result from biogenic reworking, includes any remnant sedimentary structures, trace fossils, or indistinct structures produced by macrofauna ('sediment churning', 'burrow mottling', and biodeformational structures) that cannot be accorded formal taxonomic status. The poor definition of the latter usually reflects their emplacement in sediments with a high pore water content and low sediment shear strength (soupgrounds); it may be compounded by the employment of a different burrowing strategy (Figure 2). Intrusive burrowing in uncompacted sediment involves simple deflection of the sediment around the body; the sediment collapses behind the body to occlude any void immediately (sediment swimming). In more compacted sediments (softgrounds and firmgrounds), compression burrowing and, especially, excavation (accompanied by either advection or backfilling) produce more complex, open, and actively infilled structures. Cryptobioturbation, resulting from the activities of meiofauna and microfauna, may obliterate any primary sedimentary depositional fabric and homogenize sediment without producing either discrete trace fossils or biodeformational structures.

Taxonomy, Preservation, and Ethology

Taxonomy

Trace fossils are classified using ichnogenus and ichnospecies; the prefix 'ichno' distinguishes these ichnotaxa from either living organisms or body fossils. Some higher rank categories (ichnofamilies), or informal groupings, have been established, but are not used widely. As a sedimentary structure, the detailed morphology of a trace fossil reflects both the producer's behaviour and the properties of the host sediment; characteristics of the latter (such as water content or grain size) can vary between beds, or even the laminae of a bed (Figures 3A and 3B). The abundance, probably surfeit, of ichnotaxa, particularly ichnospecies, in the geological literature has its origin in the middle to late nineteenth century, when many trace fossils were interpreted as body fossils; every subtle variation in morphology was therefore considered to be significant, and the structure worthy of separate classification. Further, and more easily resolved by comprehensive study, the morphology of a trace fossil *in partim*, and thus the ichnotaxobases (the features used to classify it), may vary depending on the orientation in which it is observed (Figures 3B and 3C).

Identifying the producer Trace fossils are very rarely preserved in association with their producer; if they are, each should be named separately. The identification of the causative organism is not a prerequisite for the naming of a trace fossil. One organism may produce more than one trace fossil depending on its behaviour (Figure 4A). A structure may also be produced by different coexisting organisms (Figure 4B), or modified later by a different organism. Potential producers may be constrained (usually at the

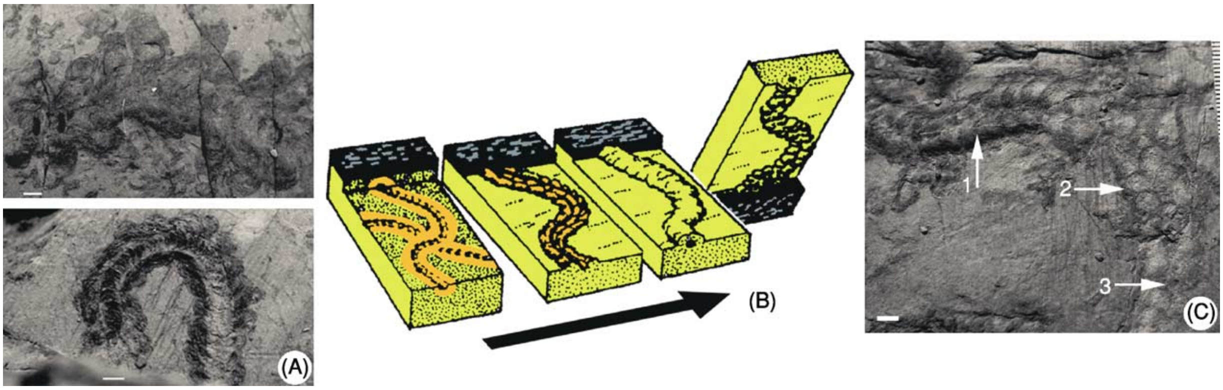


Figure 3 Key concepts in the study of trace fossils 1. (A) Preservational variants of the same ichnotaxon, reflecting emplacement in sediments of different consistency. (B) Different views of the same ichnotaxon reflect slight differences in the level of burrowing relative to the clay–sand interface. From Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists. (C) A slight angular difference between the plane of bedding/weathering and that containing the trace fossil results in different two-dimensional views of the trace fossil (1–3). Scale bars, 5 mm.

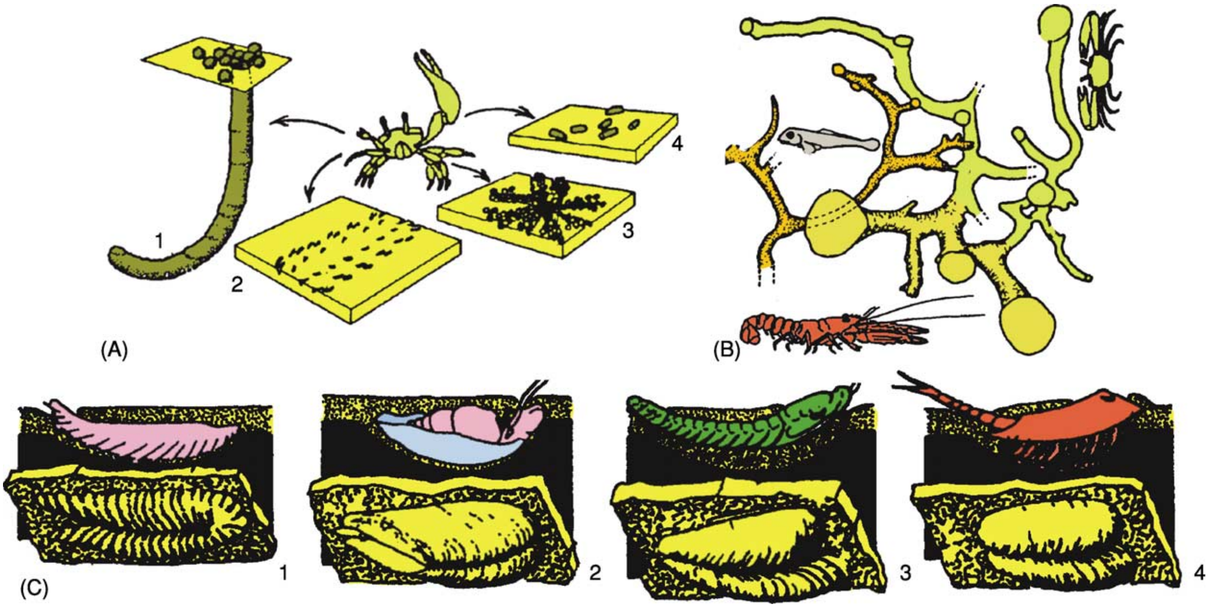


Figure 4 Key concepts in the study of trace fossils 2. (A) Open J-shaped dwelling burrow and excavated pelleted sand (1), trackway (2), feeding structure (3), and faecal pellets (4) produced by the fiddler crab, *Uca*. (B) Plan view of open burrow network produced by a crab, lobster, and fish. (C) Examples of *Rusophycus* produced by a polychaete, snail, trilobite, and notostracan. From Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

resolution of phylum or class) by the morphology and inferred function of the trace fossil (e.g., 'arthropod-produced trackway'). The body fossil content of the same, or surrounding, lithologies may be suggestive; for example, the breadth of a trackway may correspond to that of one, but not other, potential candidates. The potential producers of trace fossils, however, include organisms with minimal preservation potential. Furthermore, the same trace fossil may be produced by a variety of animals, precluding extrapolation of the producer's identity between case studies. The resting trace *Rusophycus*, examples of which in Palaeozoic marine sediments are often attributed to trilobites (*see* **Fossil Invertebrates:** Trilobites), also occurs in Mesozoic strata (after trilobites became extinct) and in sediments from (non-marine) environments that were never colonized by trilobites (Figure 4C).

Preservation

Exogenic trace fossils are emplaced at the sediment–water or sediment–air interface, and endogenic traces within the substrate; intergenic trace fossils are those emplaced endogenously at the interface between two beds (Figure 5A). Exogenic trace fossils have limited fossilization potential; they must be covered and thus cast, not eroded, during later deposition. Splitting at the original interface will yield both epirelief and hyporelief views of the structure. Loading of the substrate during emplacement of exogenic traces may depress underlying sediment; in finely laminated lithologies, splitting a short vertical distance below the original interface may reveal such undertracks, the resolution of which is often poorer than that of the corresponding exogenic expression; furthermore, only those parts of the trace fossil where loading was greater may be expressed as undertracks. Endogenic trace fossils observed in three dimensions are in full relief; those exposed on bedding-parallel surfaces

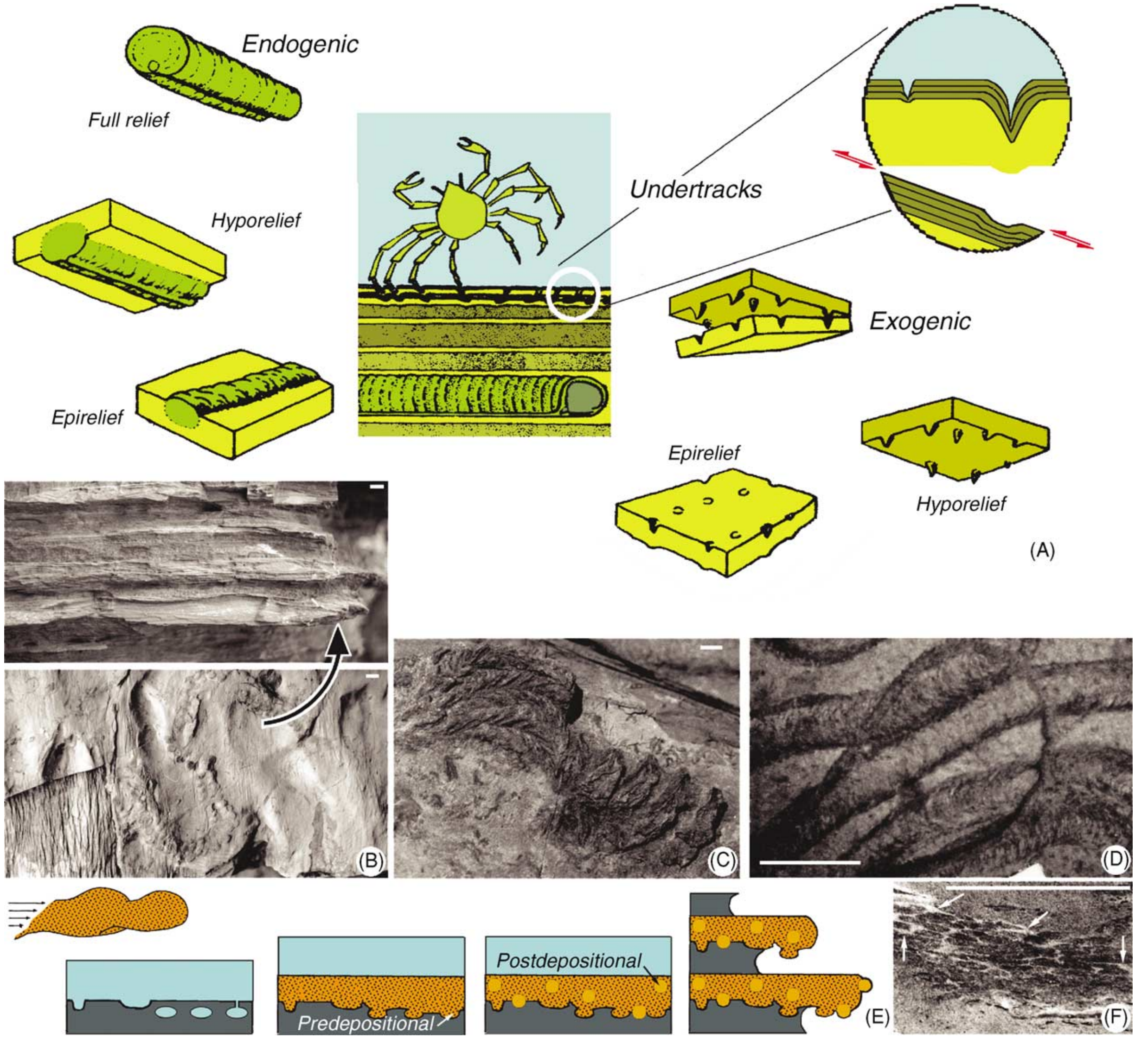


Figure 5 Key concepts in the study of trace fossils 3. (A) Preservational classification of trace fossils. Adapted from Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists. (B) The intergenic trace fossil is exposed in semirelief (positive hyporelief) on the sole of a bed of sandstone following preferential weathering of the underlying finer grained bed. (C) Positive epirelief view of *Lophoctenium* showing differential weathering of alternate menisci. (D) Polished horizontal cross-section through an actively infilled burrow comprising a thick marginal wall structure and meniscate core. (E) Formation of predepositional and postdepositional trace fossils on the sole of an event bed. (F) Although the faecal pellets are the same colour and sourced from the host lithology, the light-coloured sediment in the interstices between them indicates that the trace fossil is secondary. Scale bars, 5 mm.

are in semirelief; hyporelief and epirelief are used when the lower or upper surface, respectively, is that exposed. The orientation of specimens in the field should be determined at the time of their collection using independent criteria provided by sedimentary structures. The prefixes positive and negative describe the relief of the trace fossil relative to the host lithology. Full relief, and particularly semirelief, views can result from planes of splitting deflecting to follow the external surface of, rather than continuing across, a trace fossil. Semirelief views of intergenic structures are most common when successive lithologies have a different resistance to weathering (Figure 5B). Parts of an individual structure may weather differently; in the example in Figure 5C, alternate chevron-shaped menisci are either more sand-rich or mud-rich than, and thus have weathered positive and negative with respect to, the host lithology. Localized differences in the sedimentary fabric that result from bioturbation can be exacerbated during diagenesis. Trace fossils are often sites for the precipitation of early diagenetic minerals; this may be encouraged by the presence of mucus secreted during burrowing as an aid to locomotion, excretion, or to maintain an open structure. This can impact on economically important variables such as sediment texture, organic content, porosity, and permeability.

An open structure connected to either the air or water column may be later infilled gravitationally (passive infill) by sediment of a different composition and/or grain size. Active infill of a structure is by its progenitor. Modifications resulting from the reworking of sediment, for example during its ingestion and passage through the gut of an animal, include (1) excretion as pellets or a faecal string; (2) the exclusion or selection of grains by shape, size, or type; and (3) the reorientation of grains to produce a localized sedimentary fabric (Figure 5D).

Predepositional and postdepositional trace fossils Postdepositional trace fossils are emplaced after the deposition of a bed; the term 'predepositional' is used for trace fossils, now preserved on the sole of a bed, that were emplaced on the surface of, or within, the underlying bed. Thus, while an exogenic trackway in epirelief view is postdepositional, the secondary cast produced during deposition of the next bed, and exposed subsequently in hyporelief on the sole thereof, is predepositional. For exogenic trace fossils, the process requires there to be no erosion of the existing sediment. More extensive erosion, however, may expose endogenic open burrow systems, or wash away the active infill of burrows whilst preserving their outlines; infill of the voids created generates predepositional secondary casts (Figure 5E). The critical conditions, erosion closely followed by deposition, are often met during deposition of event beds, such as turbidites or storm deposits.

Postdepositional intergenic burrows may be emplaced later onto a surface containing predepositional secondary casts (Figure 5E). As well as the difference in their relative age, each suite may reflect different environmental conditions. Only the lower surface of a predepositional endogenic burrow system will be preserved; its infill will comprise the host lithology. Active infill of a postdepositional burrow may include modification of the host lithology (see above); in vertical cross-section, the outline of the burrow (whether infilled actively or, later, passively) will be complete.

Primary and secondary trace fossils Primary trace fossils represent either reworking of the host lithology *in situ* (but not necessarily at the time of its deposition), or the piping downwards of sediment of identical composition from one layer into another. Their identification relies on the properties of the host lithology being altered in the process (see above). Secondary trace fossils originate when sediment with different properties is 'piped' (usually downwards) into the host lithology; the evidence can be subtle (see example in Figure 5F).

Ethology

A behavioural, or ethological, classification of trace fossils is presented in Figure 6. Few trace fossils represent just one activity, and classification is thus on the basis of what is considered to be the dominant, most significant, behaviour. Cubichnia, temporary resting traces, are usually shallow, exogenic excavations (Figure 7A). Repichnia, locomotion traces, include both endogenic continuous burrows and exogenic structures. The latter may be continuous (trails), or comprise a series of discrete sequential footfalls (tracks or imprints) made by an appendage or limb (trackways) (Figure 7B). Pascichnia combine continuous locomotion parallel to bedding with feeding; emphasis is on the systematic extraction of food resources within an area (cf., strip mining). This is achieved by phobotactic (the avoidance of crossing previously formed parts of the structure), combined with strophotactic (episodic or periodic 180° turns) and/or thigmotactic (tendency to keep close to a previously formed part of the structure), behaviour patterns (Figures 7C and 7D). Trails in negative epirelief with similar geometries on the surfaces of modern ocean floors are exogenic structures with limited fossilization potential; at least, the vast majority of fossil paschichnia were actively filled endogenic burrow systems. Agrichnia are endogenic burrow

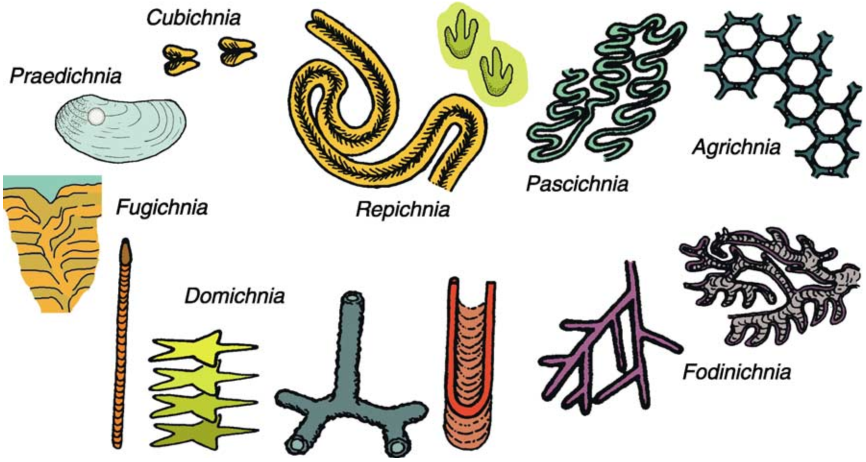


Figure 6 An ethological classification of trace fossils. Adapted from Ekdale AA, Bromley RG, and Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

systems that were maintained as open structures, and are therefore almost always preserved in positive hyporelief as predepositional secondary casts (Figures 5E and 7E). Their characteristic, complex geometries in plan view include fish scale-like, polygonal, and planispiral patterns. Spiralling patterns maintain a constant spacing between successive whorls, i.e., they do not exhibit thigmotactic behaviour. Many agrichnia are essentially two-dimensional, contained within a single plane parallel to the sediment–water interface; three-dimensional structures comprise several such levels, separated from each other vertically and connected by a more steeply inclined continuous open burrow. Agrichnia possessed single or multiple vertical connections to the overlying water column. As with domichnia (see below), the introduction of oxygenated seawater to depth within the sediment could have reduced the idealized vertical stratification of electron acceptors to a patchwork of biogeochemical microenvironments (Figure 8). Their exact function is unknown, but agrichnia may represent traps or 'microbial gardens'; microbes may be cultured on, and harvested from, the walls. Fodinichnia combine semistationary behaviour with deposit feeding. Geometries are highly variable and include two-dimensional, bedding-parallel, and three-dimensional forms. The key morphological element is the derivation of a series of branching, non-interpenetrating, straight or curved shafts, or spreiten-infilled lobes, from a common source (Figures 7F–7H). Domichnia are open burrow systems, or borings into xylic and lithic substrates (Figures 7I–7L), utilized on a semipermanent basis; most are for habitation, but structures specifically constructed for other functions (e.g., brood chambers) have been identified; some examples of the fossilized cases of caddis larvae contain pupae (Figure 7K). Domichnia in soft sediments often exhibit passive infill from the overlying layer and a thick marginal lining (to prevent collapse of the open burrow system); examples include the pelleted walls typical of *Ophiomorpha*. A domichnion in a firmground may have a bioglyph on its internal surface (Figure 7L). A 'spreite' (plural spreiten) is produced by shifting the position of a U-shaped vertical burrow in order to maintain its base a constant distance below the sediment–water interface. The spreite, which marks the former position of the basal part of the burrow, may be protrusive (occurring inside the limbs of the open burrow system in response to sediment being removed; Figure 7I), and/or retrusive (occurring below the active burrow in response to sediment aggradation). Equilibrichnia, formed by the regular, incremental shift upwards of burrows in response to the semicontinuous accretion of sediment, are considered here as a variety of domichnion. Rapid upward (attempted) escape, for example following burial by sediment, results in poorly structured escape traces, fugichnia (fugichnion). Unequivocal examples of praedichnia, trace fossils indicating predation, include the holes drilled in the shells of other molluscs by some gastropods by mechanical and/or chemical means (Figure 7M); the acid secreted by muricid gastropods can have a pH as low as 3.8. More equivocal examples in soft substrates include the intersection of two trace fossils, following which only one continues.

This classification scheme works well for the majority of trace fossils, although it has its limitations. Categories can grade into each other; for example,

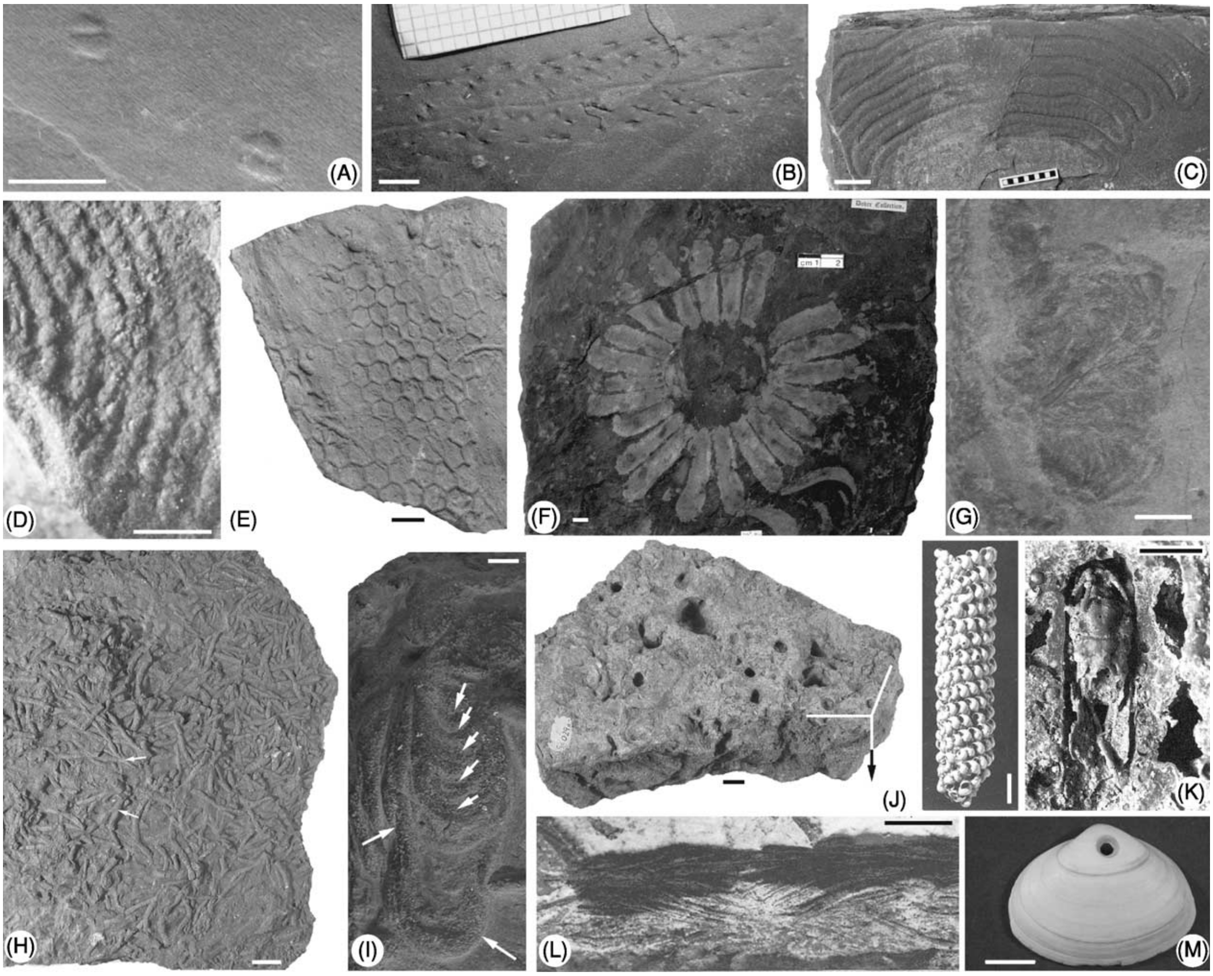


Figure 7 Variation in behaviour represented by trace fossils. (A) The cubichnion, *Rusophycus*. (B) An arthropod-produced repichnion. (C) The pascichnion, *Nereites*, exhibiting strophotactic and thigmotactic behaviour. (D) The consistent sense of braiding between successive burrows indicates thigmotactic behaviour achieved by spiralling alone. (E) The agrichnion, *Paleodictyon*. (F) Fodinichnion comprising unbranched shafts radiating parallel to bedding from a central area. (G) Fodinichnion comprising a series of lobate spreiten. (H) High-density, monospecific occurrence of the fodinichnion, *Chondrites*; each burrow system comprises a series of branching shafts (at arrows) fanning outwards and downwards from a central shaft. (I) Protrusive spreite (short arrows) contained inside a U-shaped burrow (long arrows). (J) Oblique view of block of shallow marine oolitic limestone that was later bored and encrusted during emplacement of a hardground community. (K) External view and rare example of a pupa preserved inside the case constructed by a caddis fly larva. From Hugueney M, Tachet H, and Escuillié F *Caddisfly Pupae* from the Miocene indusial Limestone of saint-Gerand-le-Puy, France (1990) *Palaeontology* 33: 495–502, Plate 1, Figures 3 and 4. (L) Bioglyph on the internal surface of a burrow excavated in semilithified sediments. Image courtesy of Richard Bromley, Copenhagen. (M) Praedichnion: a hole drilled by a gastropod through the shell of a bivalve. Scale bars, 5 mm.

at what point is a meandering pattern sufficiently regular to warrant the descriptor pascichnion, not repichnion? Opinions may vary between authors as to what the dominant behaviour is; the burrow *Planolites* has been considered a fodinichnion, pascichnion, and repichnion, as it involves reworking of the sediment whilst on the move in a straight to sinuous curve. Finally, each ichnospecies within an ichnogenus need not have the same ethology; more rarely, the ethology may vary (e.g., between a repichnion and a pascichnion) amongst a set of specimens of an ichnospecies.

Use as Environmental Indicators

Information from trace fossils and ichnofabrics can be incorporated into palaeoenvironmental reconstructions at scales ranging from the individual bed (reconstruction of a single endobenthic community) to depositional sequences tens to hundreds of metres thick (e.g., ichnofacies).

Ichnofacies

Ichnofacies are recurrent combinations of sedimentary facies and trace fossils; the Skolithos, Cruziana, Zoophycos, and Nereites ichnofacies, emplaced within marine softgrounds, characterize successively greater water depths (Figures 9 and 10). The use of trace fossils as a palaeobathymetric indicator exploits the fact that many, particularly ichnogenera, have a long time range, but are restricted to, or most common in, specific environmental conditions. These conditions include wave or current energy, temperature, chemistry (including salinity and quantity of dissolved gases such as oxygen), light penetration, nutrient supply, competition for ecospace and resources, sedimentation rate, and substrate character (including the grain size and geotechnical properties of soft sediments). Changes in these conditions tend to correlate with changes in absolute water depth, and thus the palaeobathymetry is depth related rather than depth controlled. Environmental conditions are, however, far from uniform at any given water depth. Local factors may control the distribution of ichnofacies; for example, depositional conditions in the proximal or channellized parts of a submarine fan may resemble the high-energy, mobile substrates typical of nearshore environments, and the Skolithos ichnofacies occur in each. A change in environmental conditions can produce a succession of ichnofacies that mimics, but is not the product of, significant changes in water depth; progradation of the submarine fan in Figure 9A would result in lithologies with a Skolithos ichnofacies succeeding those with a Nereites ichnofacies, unaccompanied by any significant decrease in water depth. The Glossifungites, Trypanites, and Teredolites ichnofacies are emplaced into firmgrounds, lithic substrates, and xylic substrates, respectively (Figure 10); water depth is not a controlling factor (Figure 9B). Some authors recognize the Psilonichnus (between the foreshore zone and the terrestrial realm) and Arenicolites (opportunistic colonization of newly deposited event beds) ichnofacies. Originally identified as the Scoyenia ichnofacies, the heterogeneity of non-marine environments does not lend itself to classification; several alternative detailed subdivisions have been proposed, but no consensus has emerged.

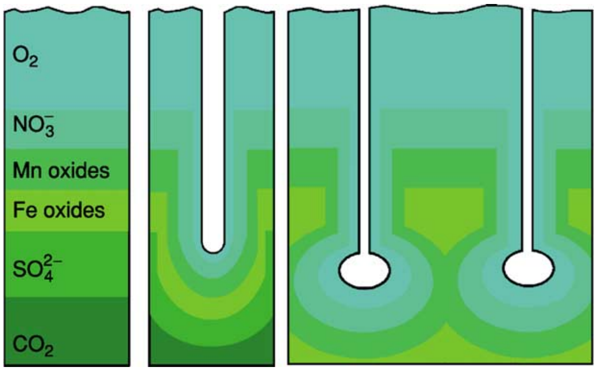


Figure 8 Modification of the idealized vertical stratification of electron acceptors via the emplacement of an open burrow structure at depth.

The presence or absence of an individual ichnogenus, even that after which the ichnofacies is named, is not strong evidence for a particular water depth. Furthermore, the bathymetric ranges of some ichnotaxa are known to have changed over

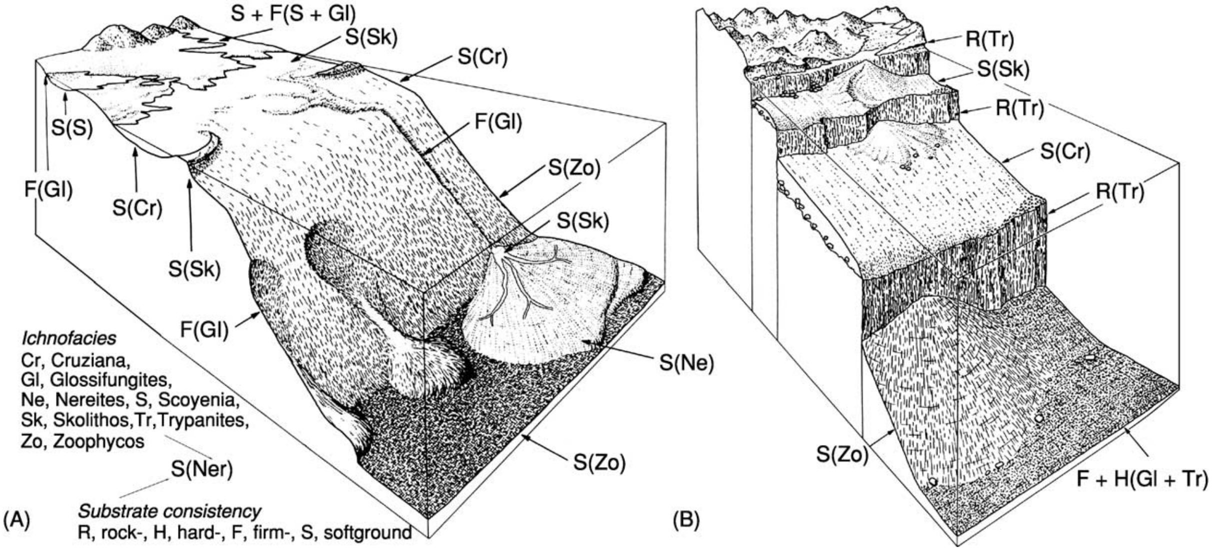
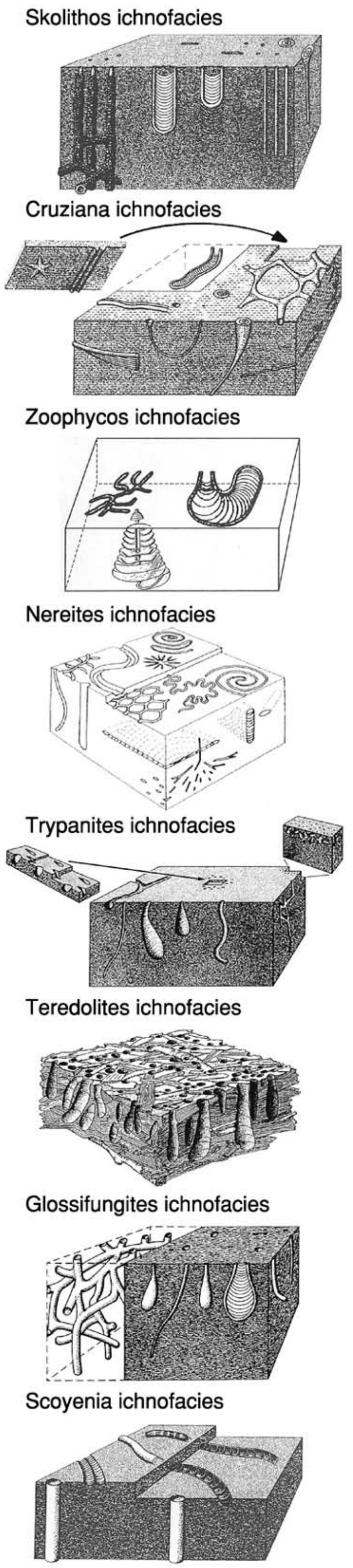


Figure 9 Schematic representation of the distribution of ichnofacies in marine environments. (A) Passive continental margin. (B) Sediment-starved active continental margin. Reprinted from Bromley RG and Asgaard U (1991) Ichnofacies: a mixture of taphofacies and biofacies. *Lethaia* 24: 153–163 (www.tandf.no/leth), by permission of Taylor and Francis AS.



Skolithos ichnofacies

- *Typical environments:* moderate to relatively high energy conditions; muddy to clean, well-sorted, shifting substrates, subject to abrupt deposition and erosion.
- *Trace fossil content:* vertical, cylindrical, or U-shaped domichnia; latter may have protrusive and/or retrusive spreiten; also fugichnia.

Cruziana ichnofacies

- *Typical environments:* infralittoral to shallow circalittoral substrates; below daily wave base, but not storm wave base to offshore shelf; moderate to low energy; well-sorted to muddy silts and sands.
- *Trace fossil content:* varied; bedding-parallel epigenic and endogenic repichnia. Domichnia often with protrusive spreiten. Fodinichnia varied. Cubichnia common.

Zoophycos ichnofacies

- *Typical environments:* circumlittoral to littoral; quiet-water, ?reduced oxygen; firm organic-rich sands and muds; sediment accretion continuous.
- *Trace fossil content:* simple to complex fodinichnia often with planar, inclined, or helicoidal spreiten; deposit feeding predominates; reduced diversity may reflect poorly oxygenated waters.

Nereites ichnofacies

- *Typical environments:* bathyal to abyssal, quiet, oxygenated waters; episodic turbiditic deposition or continually accreting pelagic surfaces.
- *Trace fossil content:* bedding-parallel pascichnia and agrichnia dominate in distal parts of turbiditic fans; their distribution can be mutually exclusive reflecting partitioning into specific environmental niches. Proximal parts of turbidite fans often with vertical, cylindrical, or U-shaped domichnia; pelagic and hemipelagic sequences including inter-trubidite deposits usually dominated by fodinichnia and actively infilled repichnia.

Trypanites ichnofacies

- *Typical environments:* lithic substrates, notably littoral omission substrates; also organic (shell, bone) substrates.
- *Trace fossil content:* cylindrical or vase-shaped endolithic domichnia; shallow cubichnia produced by epifauna; praedichnia in biomineralized tissues, e.g., shells.

Teredolites ichnofacies

- *Typical environments:* xylic substrates, including forested marginal marine settings (mangrove swamps); high-density occurrence of allochthonous bored wood can be associated with flooding surfaces.
- *Trace fossil content:* lined or unlined domichnia, diversity often low or monospecific.

Glossifungites ichnofacies

- *Typical environments:* firmgrounds, often marine littoral and sublittoral omission surface; suitable substrates can also occur following incision into partially consolidated sediments.
- *Trace fossil content:* varied domichnia, including vertical, cylindrical, tear-shaped, or U-shaped burrows, and branching and gallery-like open networks.

Scoyenia ichnofacies

- *Typical environments:* terrestrial environments are extremely varied, but preservation potential highest in softgrounds and firmgrounds at margins of aquatic environments. Subdivisions or alternatives are available for specific environmental settings.
- *Trace fossil content:* bedding-parallel, meniscate burrows; vertical or gallery-like domichnia; invertebrate, particularly arthropod-produced, cubichnia and repichnia; vertebrate-produced repichnia.

Figure 10 Summary of the ethologies, lithologies, and sedimentary processes characteristic of each of the main ichnofacies. After Frey RW and Pemberton SG (1984) Trace fossil facies models. In: Walker RG (ed.) *Facies Models*, 2nd edn., pp. 189–207. Ontario: Geological Association of Canada.

time. During most of the Palaeozoic, the ichnogenus *Zoophycos* occupied a broad range of marine water depths; since the Early Permian, it has become progressively restricted to greater water depths and is only found in continental slope and deep basin settings today.

Use of Infaunal Ecospace: Endobenthic Tiering

Bioturbation in modern, fine-grained substrates undergoing accretion that is the work of a single community can be divided into three general levels: the surficial mixed and underlying transition layers in which bioturbation occurs, and the lowest historical layer which contains the ichnofabric preserved after diagenesis and lithification (Figure 11A). As sediment properties change progressively with depth, the boundary between the mixed and transition layers is gradational, not abrupt; it is convenient, however, to model the two as distinct layers. Mixed layer sediments are often soupgrounds and thus, although bioturbated completely, the trace fossils are often poorly preserved. Bioturbation in the transition layer is heterogeneous and occurs as discrete burrows. These are usually well defined because of the high shear strength of these more dewatered sediments (softgrounds or even firmgrounds); other progressive changes with depth include reductions in both the volume and degree of oxygenation of the interstitial pore waters. Vertical partitioning, tiering, of the transition layer infauna (and thus the trace fossils they emplace) occurs in response to such changes in the physical and chemical properties of the sediments (Figure 11A). Reconstruction of the tiering profile therefore provides a measure of the community complexity; this can include the depth to which sediments were bioturbated and thus the volume of ecospace exploited. As sediment accretes, organisms will move upwards to maintain the same level or ecological niche; trace fossils produced at shallower depths will thus be cross-cut by those emplaced at successively greater depths. Dominant to unilateral cross-cutting of one trace fossil by another implies that the latter belonged to a deeper tier; the mutual intersection of two or more trace fossils implies that they are

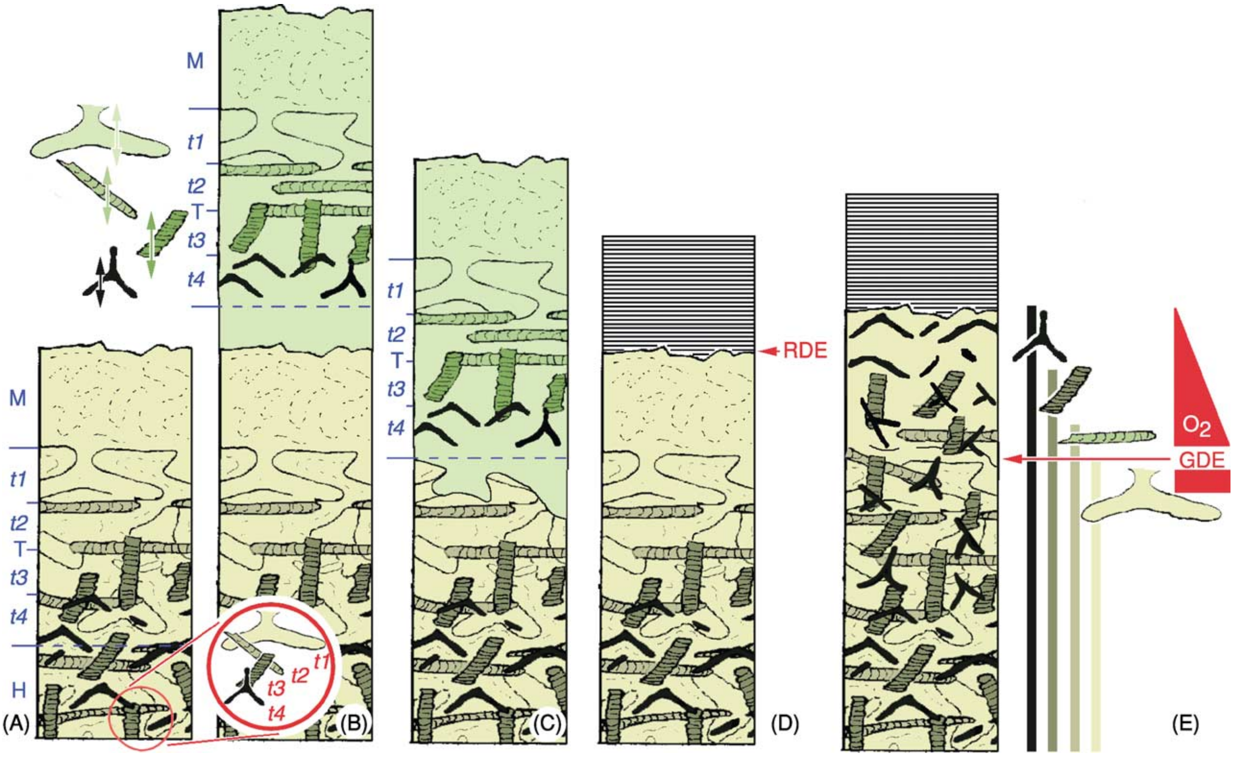


Figure 11 Endobenthic tiering revealed by the ichnofabrics produced 1. (A) Subdivision of the sediment column into the mixed (M), transition (T), and historical (H) layers; inset shows dominant to unilateral cross-cutting relationships amongst the transition layer trace fossils allowing the identification of four tiers (*t1–4*). (B, C) Rapid relocation of the infauna following event bed deposition leaves a 'frozen profile' in the older sediments, all of which (B) or, if the event bed is erosively based, only the lower part of which (C), may be preserved. (D) Production of a 'frozen profile' resulting from the evacuation of the sediment column by infauna following a rapid deoxygenation event (RDE). (E) A gradual deoxygenation event (GDE), in which the oxygen content of the interstitial pore waters declines gradually, results in exclusion of the infauna of successively deeper tiers. Adapted from Savrda CE and Bottjer DJ (1986) Trace-fossil model for reconstruction of paleo-oxygenation in bottom waters. *Geology* 14: 3–6.

components of the same tier. In practice, deviations from the random cross-cutting of older structures by younger can occur; these include preferential re-exploitation as well as avoidance of earlier formed structures. Complete reworking of sediment at depth will obliterate the record of earlier activity in shallower tiers, and thus reduce the diversity of the trace fossil community (the palaeoichnocoenosis). The preservation of a complete tiering profile requires the intensity of bioturbation to decline with depth; earlier formed parts of the ichnofabric, including occasionally the mixed layer ichnofabrics, occur as relict patches between later structures.

The deposition of an event bed will result in a rapid upward relocation of the infauna; if this event bed is sufficiently thick, the base of the transition layer will be moved above the older sediments, leaving a frozen tiering profile preserved within them; the distance from the sediment surface to the base of the transition layer indicates the thickness of sediment occupied (Figure 11B). The upper part of the sediment column is often a soupground, and thus remobilized relatively easily; the upper parts of the frozen tiering profile, notably the mixed layer, may be eroded during the deposition of the next bed (Figure 11C). Evacuation of the sediment column following a rapid deoxygenation event, in which the redox threshold boundary is moved above the sediment–water interface, will leave a frozen profile preserved below the succeeding sediments (Figure 11D). Mixed layer ichnofabrics, rare in the geological record, are more likely to be preserved in this manner than by burial below an event bed.

Other means of reconstructing the tiering profile include the identification of the depth to which an intergenic postdepositional trace fossil penetrated the substrate. In a sequence of event beds of different thickness, the components of shallower tiers occur only on the soles of thinner beds (Figure 12A). Secondary trace fossils (see above) occur within a 'piped zone' (Figures 12B and 12C). The thickness of the piped zone indicates the volume of ecospace used by the endobenthic community; components of progressively deeper tiers occur closer to its base. In the simple model in Figure 12B, the secondary trace fossils were sourced from the immediately overlying layer whilst it was being deposited, and the piped zone is contained entirely within one layer. The example in Figure 12C shows a more complex situation; the lowest bed was rebioturbated during deposition of the second, not the first, overlying layer, but the latter was too thin to contain the piped zone entirely. In this idealized example, the colour of each layer is different, and the source of sediment in the piped zone is obvious. In practice, however, many hemipelagic and pelagic sequences are characterized by alternating lithologies, the differences in colour and composition of which can be the result of orbital forcing. In such cases, trace fossils within the piped zone may be actively infilled by the same sediment as the host lithology.

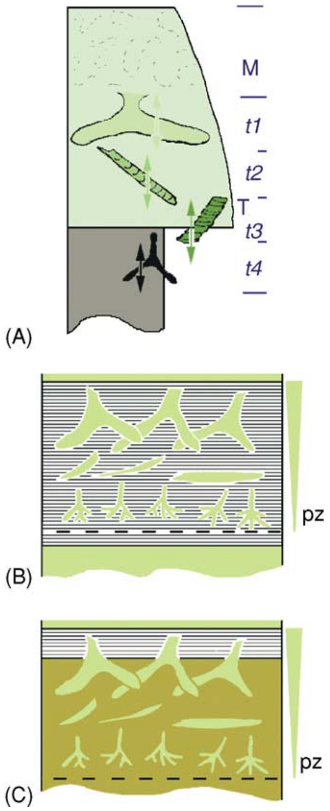


Figure 12 Endobenthic tiering revealed by the ichnofabrics produced 2. (A) The base of the event bed cannot be reached by the infauna of the shallower tiers; the trace fossils at its sole will comprise those of tier 3 and, if the accretion of sediment subsequently is gradual, tier 4. (B, C) Emplacement of secondary trace fossils defines a piped zone (pz), the thickness of which is a measure of the volume of ecospace utilized. For clarity, earlier formed parts of the ichnofabrics have been omitted.

Natural systems are obviously more complex, but more dynamic models can be produced by allowing variables, such as the rate of sediment accretion, including negative values (erosion), or the levels of oxygenation, to fluctuate over time. Not all circumstances will satisfy the assumptions within the model. Thin-bedded siliciclastic turbidites often exhibit a bipartite division into a sand-rich lower part and a mud-rich upper part, and infauna can position themselves with respect to the interface between the two, independent of its depth below the sediment–water interface. Cross-cutting relationships can also be generated by the superimposition of two successive

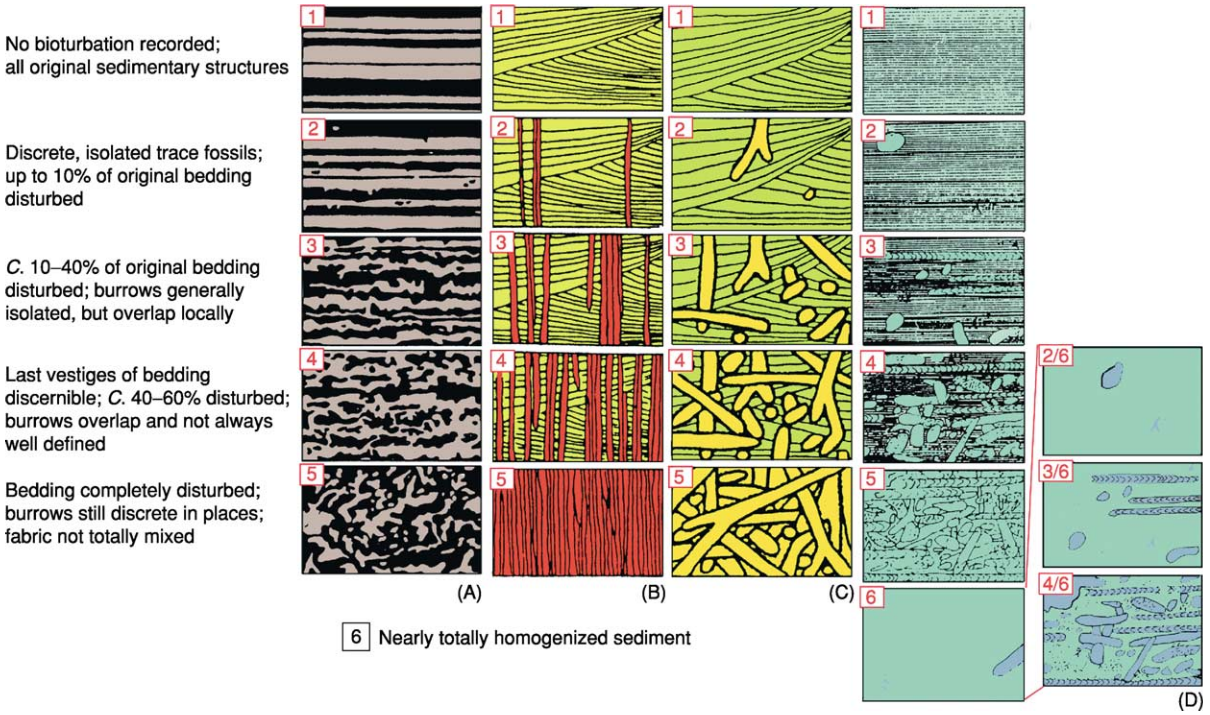

No bioturbation recorded; all original sedimentary structures
Discrete, isolated trace fossils; up to 10% of original bedding disturbed
C. 10–40% of original bedding disturbed; burrows generally isolated, but overlap locally
Last vestiges of bedding discernible; C. 40–60% disturbed; burrows overlap and not always well defined
Bedding completely disturbed; burrows still discrete in places; fabric not totally mixed
6 Nearly totally homogenized sediment
1
2
3
4
5
6
2/6
3/6
4/6
(A)
(B)
(C)
(D)

Figure 13 Definitions and schematic illustrations of ichnofabric indices for strata deposited in: (A) shelf environments; (B) high-energy nearshore sandy environments dominated by *Skolithos*; (C) high-energy nearshore sandy environments dominated by *Ophiomorpha*; (D) deep-sea deposits, including examples of homogenized sediment that has been reworked subsequently. Reprinted with permission from Droser ML and Bottjer DJ (1993) Trends and patterns of Phanerozoic ichnofacies. *Annual Review of Earth and Planetary Sciences* 21: 205–225. © 1993 by Annual Reviews www.annualreviews.org

communities; following event bed deposition, it is not unusual for a short-lived opportunistic community to exploit the new ecospace and resources, before the longer term equilibrium community is re-established.

Palaeo-oxygenation

As bottom water oxygenation declines, the general tendency is for the thickness of the mixed and transition layers, and the size (reflected in the burrow diameter) and diversity of actively infilled transition layer trace fossils, to decrease (Figure 11E). It has been suggested that ichnofaunal assemblages dominated by domichnia, pascichnia, and fodinichnia indicate a progressive decline in pore water oxygenation. Related to this, certain ichnogenera, notably the fodinichnion *Chondrites*, especially if abundant in a low-diversity or monospecific assemblage, have been considered to be diagnostic of low oxygen conditions; however, this should not be assumed without supporting ichnological and sedimentological evidence. The shaft diameter is relatively large in the monospecific assemblage of *Chondrites* emplaced in a storm bed in Figure 7H, and an interpretation as an opportunistic colonization of newly deposited sediment is favoured.

Intensity of Bioturbation

The ichnofabric index (ii) is a semiquantitative measure of the intensity of bioturbation, based on the extent to which the original stratification is disrupted; as both the nature of bioturbation and host lithology will vary, the schematic illustrations are specific to certain environmental settings (Figure 13). The ichnofabric is observed in vertical sections; a standard horizontal field of view should be used and reported (500 mm is often used in outcrop studies).

See Also

Biosediments and Biofilms. **Diagenesis, Overview**. **Fossil Invertebrates:** Trilobites. **Palaeoecology**. **Palaeontology**. **Sedimentary Environments:** Depositional Systems and Facies; Storms and Storm Deposits.

Further Reading

Bottjer DJ, Hagadorn JW, and Dornbos SQ (2000) The Cambrian Substrate Revolution. *GSA Today* 10: 1–7.

Bromley RG (1996) *Trace Fossils. Biology, Taphonomy and Applications*. London: Chapman and Hall.

Bromley RG and Asgaard U (1991) Ichnofacies: a mixture of taphofacies and biofacies. *Lethaia* 24: 153–163.

Donovan SK (ed.) (1994) *The Palaeobiology of Trace Fossils*. Chichester: Wiley.

Droser ML and Bottjer DJ (1993) Trends and patterns of Phanerozoic ichnofacies. *Annual Review of Earth and Planetary Sciences* 21: 205–225.

Ekdale AA, Bromley RG, Pemberton SG (1984) *Ichnology. The Use of Trace Fossils in Sedimentology and Stratigraphy. SEPM Short Course No. 15*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Frey RW and Pemberton SG (1984) Trace fossil facies models. In: Walker RG (ed.) *Facies Models,* 2nd edn., pp. 189–207. Ontario: Geological Association of Canada.

Frey RW, Pemberton SG, and Saunders TDA (1990) Ichnofacies and bathymetry: a passive relationship. *Journal of Paleontology* 64: 155–158.

Lockley M (1991) *Tracking Dinosaurs*. Cambridge: Cambridge University Press.

Maples CG and West RR (eds.) (1992) *Trace Fossils. Short Courses in Paleontology No. 5*. Tulsa, OK: Paleontological Society.

Pemberton SG (1992) (ed.) *Applications of Ichnology to Petroleum Exploration – A Core Workshop. SEPM Core Workshops No. 17*. Tulsa, OK: Society of Economic Paleontologists and Mineralogists.

Savrda CE (1995) Ichnologic applications in paleoceanographic, paleoclimatic and sea-level studies. *Palaios* 10: 565–577.

Savrda CE and Bottjer DJ (1986) Trace-fossil model for reconstruction of paleo-oxygenation in bottom waters. *Geology* 14: 3–6.

Taylor A, Goldring R, and Gowland S (2003) Analysis and application of ichnofabrics. *Earth Science Reviews* 60: 227–259.

Wetzel A and Aigner T (1986) Stratigraphic completeness; tiered trace fossils provide a measuring stick. *Geology* 14: 234–237.

ULTRA HIGH PRESSURE METAMORPHISM

H-J Massonne, Universität Stuttgart, Stuttgart, Germany

Introduction

Ultra high pressure (UHP) metamorphic rocks of common basic to felsic nature are defined by the occurrence of coesite, a silica polymorph that is denser than quartz. According to several experimental studies, the transition from quartz to coesite at 600°C requires a pressure (P) of around 27 kbar, a temperature (T) of conditions that occur on Earth at depths close to 100 km. Coesite in nature was detected first in rocks affected by impact metamorphism, but a coesite-bearing rock that had been subjected to regional metamorphism was described in 1983 by Chopin. Based on the coesite–quartz transition pressure and temperature curve, this rock, which was from the Dora Maira Massif of the Western Alps, must have been buried at depths of about 100 km or more. Transition curves of SiO_2 polymorphs show a moderate pressure (P) and temperature (T) slope, dP/dT, of only 10 bar °C^{-1} (Figure 1).

Prior to discovery of the Dora Maira Massif rock, it was thought that metamorphic rocks (excepting ultrabasites) that were formed during orogenic events, and now exposed at Earth's surface, represent, in general, a fossil record of pressure and temperature conditions only of Earth's crust, equivalent to maximum depths of 70 km and pressures up to ~18 kbar. Eclogites (basalts that have been metamorphosed under high pressure) were believed to have formed, depending on temperature, at pressures between 12 and 16 kbar, corresponding to the jadeite content of omphacite, the mineral that characterizes eclogites. However, these estimates are only justified when omphacite co-exists with plagioclase and quartz (but plagioclase is often only a retrograde product in eclogites, due to the breakdown of omphacite). Rare jadeite occurrences in felsic rocks with plagioclase supported the view of metamorphic pressures not exceeding 18 kbar for crustal rocks. In geodynamic models of subduction of oceanic crust under oceanic or continental plates, it was assumed that the return of subducted material was possible only for shallower regimes of the collision wedge. Such subducted material then became, for instance, part of an accretionary wedge complex. Within the framework of such a scenario, the discovery of coesite in regional metamorphic rocks was sensational and a real turning point in scientific thinking about deep burial and subsequent exhumation of crustal rocks. Soon after Chopin's report in 1983, it became evident that UHP rocks are more widespread than had been thought. Coesite was recognized in rocks of the Norwegian Caledonides, the Dabie Shan in China, and orogenic regions elsewhere. Moreover, in 1990, microdiamonds, another indicator mineral for UHP metamorphism, were reported by Sobolev and Shatsky in marbles and gneisses from the Kokchetav Massif, Kazakhstan. These diamonds provided evidence for burial of crustal rocks to depths of at least 130 km.

Identification of UHP Rocks

The discovery of UHP rocks at Earth's surface at a relatively late date in geological science history can be attributed to the nature of the processes involved. Retrogression of rocks during exhumation can result

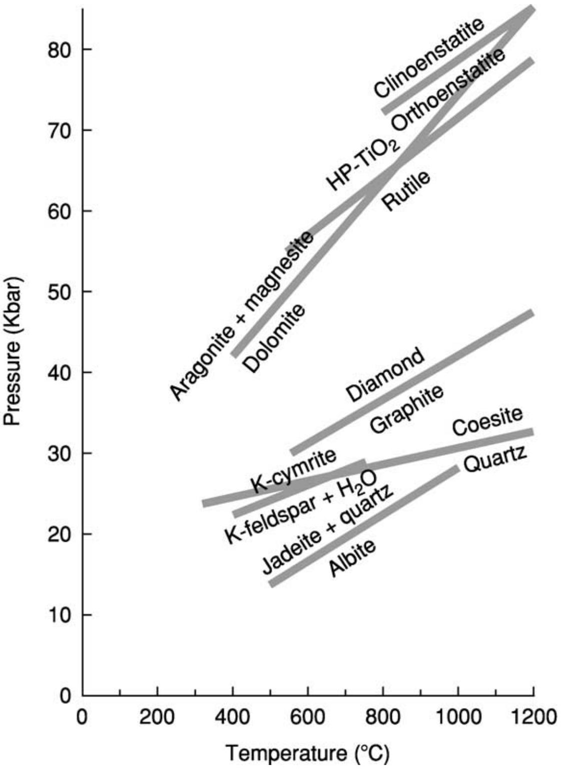


Figure 1 Pressure and temperature stability of various mineral phases that are of relevance to UHP metamorphism. Except for the transition curves below 30 kbar, the experimental error is even higher than is expressed by the thickness of the lines. The clinoenstatite–orthoenstatite transition refers to a composition with 10 mol% of ferrosilite component.

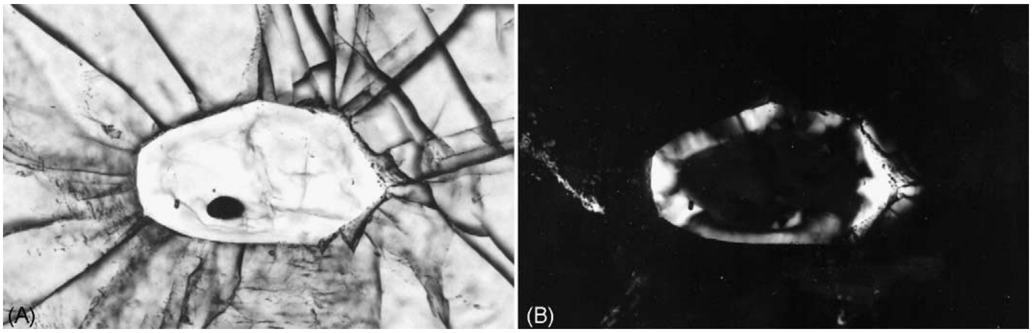


Figure 2 Photomicrograph of a coesite relic enclosed in garnet. Left-hand side shows the appearance of palisade quartz using plane-polarized light; the specimen is best seen under crossed nicols (i.e., a polarizing filter; right-hand side). Typically, coesite has been partially replaced by the palisade quartz, causing radial cracks in the host mineral. Image width is 0.6 mm. The example is from a thin section of an eclogite from the central Saxonian Erzgebirge.

in a complete overprint of the UHP mineral assemblage, thus erasing the 'memory' (record) of the UHP event in the metamorphic rock. Therefore, it is important for the geoscientist working with such rocks to look for mineralogical hints of UHP metamorphism (e.g., tiny inclusions in resistant garnet porphyroblasts and unusual microfabric textures and appearances). Coesite, the 'indicator' mineral that points to UHP, has never been found to be a major constituent of UHP rocks. On the contrary, quartz dominates among the SiO_2 polymorphs, even in the best preserved UHP rocks, and coesite commonly occurs only as inclusions in minerals such as garnet and zircon (Figure 2). Even in such inclusions, coesite is normally partly decomposed to quartz. A typical decomposition fabric is formed of palisade quartz, with lamellae perpendicular to the replaced coesite. Cracks in the host mineral around the coesite inclusion form due to the volume increase during the decomposition of coesite to quartz. Thus, even completely replaced coesite can be recognized by such features. Tiny coesite relics can be identified by confocal micro-Raman spectroscopy even when the relics are not at the surface of a rock thin section.

Identifying UHP rocks on the basis of coesites is problematic when cracks around quartz inclusions in the host mineral are discernible but no coesite relic or palisade quartz is detectable. Palisade quartz formations may have recrystallized, forming polycrystalline quartz, but this cannot be considered a clear indication for UHP metamorphism. No doubt there have been a few reports in the literature of UHP rocks as a result of mistaken identification of unusual quartz inclusions as former coesite. Symplectites consisting of K-feldspar and quartz resemble quartz pseudomorphs of replaced coesite, and there are usually cracks in the host mineral, typically eclogitic omphacite, around such inclusions. Massonne, Dobrzhinetskaya, and Green have described such symplectites as being decomposition products of K-cymrite ($KAlSi_3O_8 \cdot H_2O$). K-Cymrite has not yet been found as a relic, but it has potential for identifying UHP rocks due to its lower pressure stability limit (Figure 1), which is similar to that of coesite.

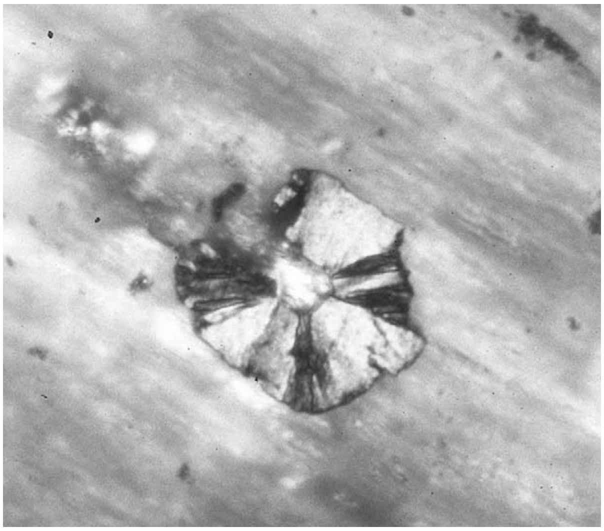


Figure 3 Inclusion of a pseudomorph of radially oriented graphite after diamond in clinopyroxene from a siliceous calcite marble of the Kokchetav Massif. A relic microdiamond is present in the centre of the pseudomorph. Image width is 0.11 mm.

Another mineral that is diagnostic for UHP is diamond. Like coesites in UHP rocks, diamond can be easily transformed. During exhumation, unless it is enclosed as tiny grains in resistant porphyroblasts, diamond is transformed to graphite. Even when enclosed in porphyroblasts, diamond rarely survives, but it can react to form pseudomorphs of radially oriented graphite (Figure 3). Although microdiamonds are not as widespread in UHP rocks as coesite is, they are found in all kinds of UHP rocks, from ultrabasic to felsic or silicate to carbonate-rich rocks, whereas coesite can occur only in felsic to basic rocks when SiO_2 is in excess. Phengite, a silicon-rich potassic white mica common in basic to felsic HP to UHP rocks, also has potential to be diagnostic of UHP rocks. In contrast to coesite and diamond, which are pure phases, potassic white mica is chemically complex. Mica compositions of the corresponding solid solution series can be stable at both high and low pressures, which is probably the reason that the potassic white mica that formed under HP to UHP conditions often remains in rocks during exhumation. However, phengite is diagnostic for UHP only if it is possible to detect its precise composition at UHP, and the compositions of its coexisting partners (e.g., garnet and omphacite). Although numerous experiments at high-pressure conditions have contributed to an understanding of the relationship between mica compositions and coexisting phases as a function of pressure and temperature, the degree of uncertainty in determining pressure conditions for phengite-bearing rocks is at less than 2 kbar, even when advanced thermodynamic calculation methods and suitable mineral paragenesis are considered. In any case, metamorphic temperature determinations are required using a well-calibrated geothermometer, such as the Fe^{2+}–Mg exchange reaction for different mineral pairs (e.g., garnet and omphacite). A problem arises when phengite inclusions or core compositions are used: for example, subduction of a basalt of the oceanic crust first reaches pressure and temperature conditions of the blueschist facies. At that stage, Si-rich phengites ($Si > 3.5$ per formula unit) can form and coexist with a typical blueschist facies assemblage. Further subduction at elevated pressure and temperature causes garnet growth. During this process, the blueschist-facies phengite is enclosed in garnet and thus survives the subsequent metamorphism, leading to a kyanite-bearing eclogite. At peak temperature conditions, Si contents of re-equilibrated phengites reach 3.3 per formula unit in the kyanite-bearing assemblage; this is typical for near-ultrahigh pressure conditions. If the blueschist-facies phengite is related to the peak temperature assemblage, pressures of the UHP regime would be indicated. Thus it is likely that in a few cases, rocks lacking coesite and diamond have been erroneously assigned to UHP metamorphism because a silicon-rich phengite was found in an ordinary high-pressure rock. This may have occurred with eclogite blocks in the low-grade metasediments of the Franciscan Formation in California, where phengites with high Si contents ($Si > 3.55$ per formula unit) were reported. Nevertheless, phengite geobarometry combined with geothermometers is a powerful tool and allows determination of pressure and temperature conditions in the UHP regime, whereas coesite and diamond can yield only minimum pressures. This is also true for tetrahedrally coordinated Al in clinopyroxene and orthopyroxene, which act as suitable geobarometers in the presence of garnet, especially for ultrabasic rocks lacking phengites.

In the known UHP regions of the world, minerals with specific exsolution fabrics have been observed and assigned to UHP conditions. Clinopyroxenes can contain rods of SiO_2. Garnets show clinopyroxene and orthopyroxene exsolution lamellae and precipitates (Figure 4). Both phenomena are explained by introduction of Si into the octahedral site of the corresponding mineral structure at UHP. Subsequent pressure release results in dissolution of the octahedrally coordinated Si and formation of specific minerals. Experimental constraints are related only to garnet, whereby small amounts of majorite component, $(Mg,Ca)_4Si_4O_{12}$, can be dissolved in the garnet structure at pressures exceeding 50 kbar. However, the dissolution products, clinopyroxene and orthopyroxene, can form from ordinary garnet by pressure release as well, but in that case, the pyroxenes should contain significant amounts of Tschermak's component, $CaAl_2SiO_6$. Titanite in marble from the Kokchetav Massif is yet another example for the likely introduction of Si into the octahedral site. This mineral contains coesite precipitates as dissolution product. Clinopyroxene from UHP areas can also show K-feldspar lamellae, which are interpreted as exsolution from K-bearing clinopyroxene. Such clinopyroxenes, with significant

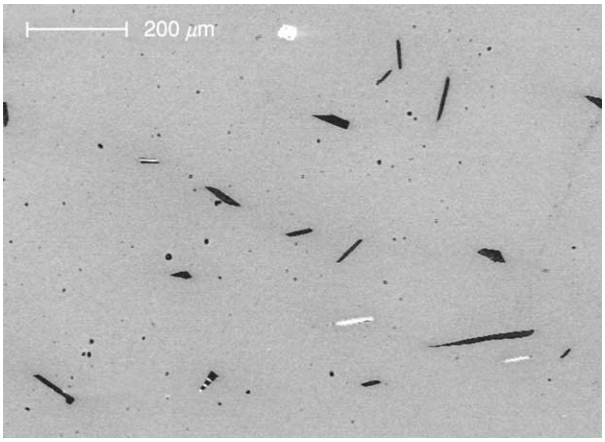


Figure 4 Exsolution of orthopyroxene and clinopyroxene from majoritic garnet, shown by a back-scattered electron image (pyroxenes are darker than garnet). This sample is from the Western Gneiss Region of the Norwegian Caledonides. Reproduced with permission from Van Roermund HLM, Drury MR, Barnhoorn A, and De Ronde AA (2000) Super-silicic garnet microstructures from an orogenic garnet peridotite, evidence for an ultra-deep (>6 GPa) origin. *Journal of Metamorphic Geology* 18: 135–147.

amounts of K_2O (>0.5 wt.%), have been reported from diamondiferous siliceous marbles of the Kokchetav Massif. Experiments have proved that K is introduced into the clinopyroxene lattice at high pressures, but conclusions on the metamorphic pressures of the Kokchetav rocks have so far been only semiquantitative. In high-pressure experiments, TiO_2 with the α-PbO_2 structure is stable at high temperatures above 60 kbar instead of rutile (Figure 1). A nanocrystal of HP-TiO_2 was observed in a diamondiferous quartzofeldspathic rock from the Saxonian Erzgebirge, Bohemian Massif. Clinoenstatite lamellae in pyroxenes were reported from the Dabie Shan–Sulu terrane, China. This would point to pressures of 80 kbar and more (see Figure 1). Magnesite and calcite, probably former aragonite, in direct contact or separated by dolomite, were observed in the Dabie Shan. This allows the conclusion that the corresponding rocks experienced pressures of at least 60 kbar (Figure 1). Ilmenite rods in olivine were reported from several ultrabasic rocks of different UHP terranes. Relatively high concentrations of these rods were found in olivine from Alpe Arami, in the central Alps. Dobrzhinetskaya, Green, and Wang argued that this feature points to depths of more than 300 km as the origin of the corresponding rock. Certainly a number of other potential candidates, either minerals or specific mineral assemblages, could serve to diagnose UHP conditions; experimental studies and calculations of mineral equilibria using thermodynamic data have indicated this possibility, but corresponding observations have not yet been made.

Global Distribution

Following Chopin's 1983 report of finding the first UHP metamorphic rock of crustal origin in the Western Alps, several other finds were reported from around the globe, relating to more than 10 different orogens or far-distant sections of an orogenic chain. The locations of these rock finds are shown in Figure 5, together with some potential UHP terranes where, so far, no coesite or diamond relic has been found. There are certainly even more UHP potential terranes than are indicated in Figure 5 (note also that neither the Franciscan Formation in California nor Neoproterozoic areas with eclogites, for which geothermobarometrically derived pressure and temperature estimates lie in the UHP field, are indicated in Figure 5). Typically, the confirmed UHP areas worldwide are parts of young orogens resulting from

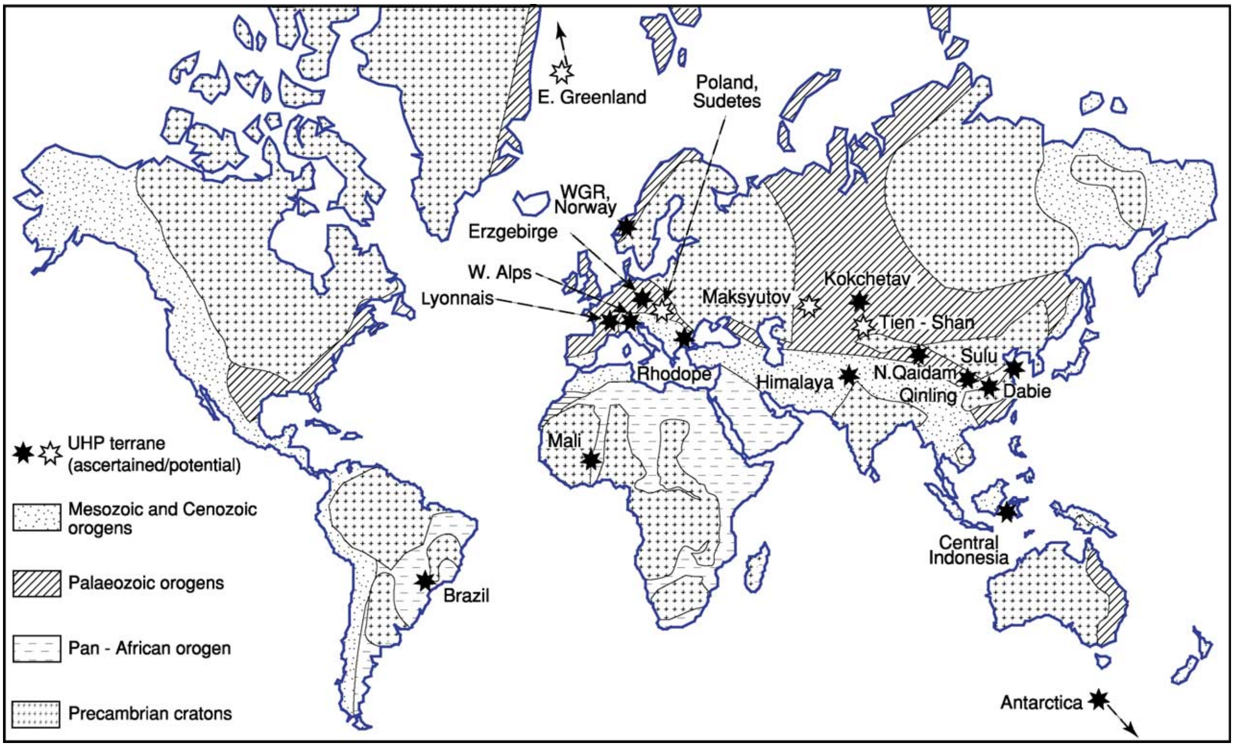


Figure 5 Occurrences of ascertained and potential UHP rocks worldwide. The group of potential UHP rocks includes pseudomorphs replacing coesite and/or diamond. WGR, Western Gneiss Region. Reproduced with permission from Chopin C (2003) Ultrahigh pressure metamorphism: tracing continental crust into the mantle. *Earth and Planetary Science Letters* 212: 1–14.

continent–continent collision, but relatively few of these areas have been well explored for UHP rocks. Those areas that have been examined closely are in the European Alps, the Norwegian Caledonides, the Chinese Qinling Shan–Dabie Shan–Sulu Terrane belt, the Kokchetav Massif in northern Kazakhstan, the Mid- and West-European Variscides, and, to some extent, the Himalaya.

European Alps

UHP rocks in the Western and Central Alps occur exclusively in the Penninic unit. Potential UHP rocks also exist in the Eastern Alps; these rocks are mainly eclogites, whereas in the Western and Central Alps, different metapelites, metagranitoids, and ultrabasic rocks are also involved in UHP metamorphism. The lateral extent of UHP units in the Alps is in the range of several to a few tens of kilometres. The thickness of these units is commonly less than 1 km, because they are part of the nappe stack of the Alpine Orogen. Age dating of the UHP event yielded values close to 35 Ma. Potential UHP eclogites in the Eastern Alps were formed in the Cretaceous. The protoliths of felsic UHP rocks of the Western Alps can be at least partially assigned to former Variscan metamorphic and plutonic rocks. This implies that Variscan continental crust was subducted to great depths during a late Alpine continent–continent collisional event.

Geothermobarometric evolution has resulted in a pressure and temperature path for the majority of the UHP rocks that is characterized by a prograde branch along geothermal gradients, decreasing from 10 to $6^\circ\mathrm{C\,km^{-1}}$ towards the pressure and temperature climax, which is settled between 30 and 40 kbar and around 750°C. The retrograde path typically shows decreasing temperatures during pressure release, but the temperatures of the retrograde path are always higher than those of the prograde path at a given pressure. Age dating, for instance, on zoned zircon grains has demonstrated that significant exhumation from great depths happened within a few million years. Thus the uplift rates must be in the range of several centimetres per year. This rate and dating to 35 Ma for UHP metamorphism were also found for ultrabasic rocks, for instance, outcropping at Alpe Arami, Central Alps.

Norwegian Caledonides

Felsic, basic, and ultrabasic UHP rocks are widespread in the Western Gneiss Region (WGR) of the Norwegian Caledonides. Similar rocks also occur in the Caledonian Orogen further north (for instance, in the Lofoten range) and at the eastern coast of Greenland. The widely scattered occurrences of UHP rocks within the WGR may indicate the existence of a coherent UHP terrane of up to $350 \times 150\,\mathrm{km^2}$. Prograde metamorphism is characterized by geotherms around $6^\circ\mathrm{C\,km^{-1}}$, reaching peak temperatures of 550°C in the south-east inland area and more than 800°C in the north-west coastal area of the WGR. In the latter area, microdiamonds were found on the island of Fjørtoft. Age data obtained from UHP metamorphic rocks scatter between about 440 and 390 Ma, but it has been concluded that UHP metamorphism happened close to 400 million years after closure of the Iapetus Ocean. Early exhumation rates of 1 cm or more per year are assumed. The coeval garnet peridotites originated at mantle depths of about 200 km (60–65 kbar) and temperatures significantly above 1000°C. However, it is possibly that they had resided there since mid-Proterozoic times, until they were uplifted by a mantle diapir in the Lower Devonian.

Chinese Qinling Shan - Dabie Shan - Sulu Terrane Belt

In central and eastern China, the Triassic orogenic belt resulting from collision of the Yangtze Craton and the Sino-Korean Craton contains sections with UHP metamorphic rocks. Although intruded by enormous volumes of Cretaceous granitoids, these areas can be traced over 100 km and more, leading to the conclusion of the existence of wide, coherent UHP terranes. Under this scenario, orthogneisses with Proterozoic protolith ages would be the dominant UHP rock type there. The pressure and temperature conditions of the Chinese UHP rocks resemble those of the WGR in regard to spread of peak pressure and temperature ($\leq$600–$\geq$800°C) conditions. However, a significant difference between the WGR and the Triassic orogenic belt is related to the peak temperature conditions of ultrabasic rocks. Garnet peridotites of the Sulu Terrane were metamorphosed at only 800°C, and those of the Dabie Shan were metamorphosed below 800°C. In spite of the relatively low temperatures, pressure conditions were estimated to be as high as 55 kbar or even more, thus yielding geothermal gradients somewhat below $5^\circ\mathrm{C\,km^{-1}}$. It is assumed that rocks as well as other UHP rocks of the Qinling Shan–Dabie Shan–Sulu Terrane belt dates close to 220 Ma.

Kokchetav Massif in Northern Kazakhstan

An enormous variety of identified UHP rocks is known to occur at local sites (e.g., close to Lake Kumdy Kol) of the Kokchetav Massif. Although the development of such rocks can be traced over several tens of kilometres, the recent view of a

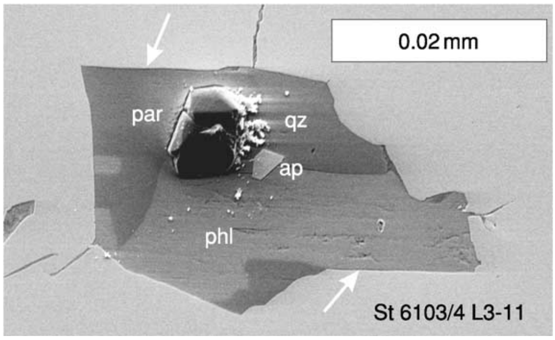


Figure 6 Scanning electron microscope image of a polyphase diamond-bearing inclusion in garnet, interpreted as trapped siliceous fluid or melt before crystallization. The example is a quartzofeldspathic rock from the Saidenbach reservoir, central Saxonian Erzgebirge. Arrows point to rational mica garnet interfaces. qz, quartz; par, paragonite; phl, phlogopite; ap, apatite.

'megamelange', including medium-pressure rocks, may best describe the geological situation. Nevertheless, two types of UHP rocks can be distinguished in terms of maximum pressure and temperature conditions. One type has experienced pressure and temperature conditions similar to those of UHP rocks of the Western Alps ($T_{\max}$ ~750°C). The other type, characterized by abundant microdiamond inclusions in various host minerals, was metamorphosed at pressures as high as 70 kbar. At these high pressures, temperatures exceeded 1000°C, probably resulting in partial melting. Polyphase aggregates (Figure 6), commonly containing microdiamonds, are enclosed in garnet and other phases. These aggregates serve as evidence for siliceous fluids or melts. The silica-rich material of these inclusions was probably trapped in growing minerals during exhumation of the (partially molten) rocks at decreasing temperatures. This happened at 530 Ma, as deduced from zircon dating, also indicating a fast uplift of the rock in the range of centimetres per year.

Mid- and West-European Variscides

An abundance of ascertained and potential UHP rocks have been detected in several crystalline complexes of the Variscides, including those later involved in the Alpine Orogeny. However, due to a significant fragmentation of specific major units by a late orogenic event, a former, possibly extended, coherency of UHP terranes has been broadly lost. Nevertheless, HP to UHP metamorphism cannot be related to a single event. For instance, in the Bohemian Massif, representing the north-eastern section of the Variscides, two major HP–UHP events can be discriminated from abundant age data. The earlier event, around 395 Ma, led to eclogites, probably former oceanic crust, which experienced maximum pressure and temperature conditions close to 700°C and 23–30 kbar. A common feature of these eclogites is the replacement of omphacite by amphibole porphyroblasts due to the influx of hydrous fluids, probably close to the pressure and temperature climax. In the younger UHP event, occurring about 340 Ma, felsic, basic, and ultrabasic rocks were involved. Peak temperature conditions can exceed 1000°C–1200°C and 80 kbar are indicated by a diamondiferous quartzofeldspathic rock from the Saxonian Erzgebirge; this rock has polyphase inclusions in garnet (Figure 6), similar to rocks from the Kokchetav Massif. The anatectic evolution of the Erzgebirge rock is inferred to be similar to that of diamondiferous rocks from the Kokchetav Massif, in addition to the fast exhumation. The diamondiferous rock from the Erzgebirge forms lenses up to 1 km long, like other UHP rocks in the Variscides. The surrounding gneisses display a medium to high pressure signature, but also a clear peak pressure contrast to the UHP rocks.

Himalaya

UHP rocks in the Himalaya are very rare, but their geodynamic context is much clearer compared to other UHP regions. The protoliths of coesite-bearing eclogites and surrounding metasediments in the north-western Himalaya were part of the continental margin of India that was subducted beneath Asia. The prograde metamorphism during this event followed a geothermal gradient of 6°C km^{-1} reaching peak pressure and temperature conditions at depths between 90 and 120 km about 50 Ma ago. Amphibole blastesis at the expense of clinopyroxene is common. The subsequent exhumation of the UHP rocks is characterized by moderate cooling at uplift rates in the range of centimetres per year, slowing down to millimetres per year for the past 40 million years.

Mechanisms

The known UHP rocks at Earth's surface can be subdivided into two groups. One group suffered from metamorphism along geothermal gradients of about 7°C km^{-1}, reaching maximum pressures generally below 40 kbar (Figure 7). Frequently observed peak pressure and temperature conditions are 30 kbar at 650–700°C. Moderate cooling and uplift rates in the range of centimetre per year characterize the exhumation of this group of UHP rocks, commonly starting with the influx of hydrous fluids (for instance, into eclogites). The subduction of oceanic crust under continental crust, including the exhumation within a subduction channel, best explains the characteristics of these UHP rocks. However,

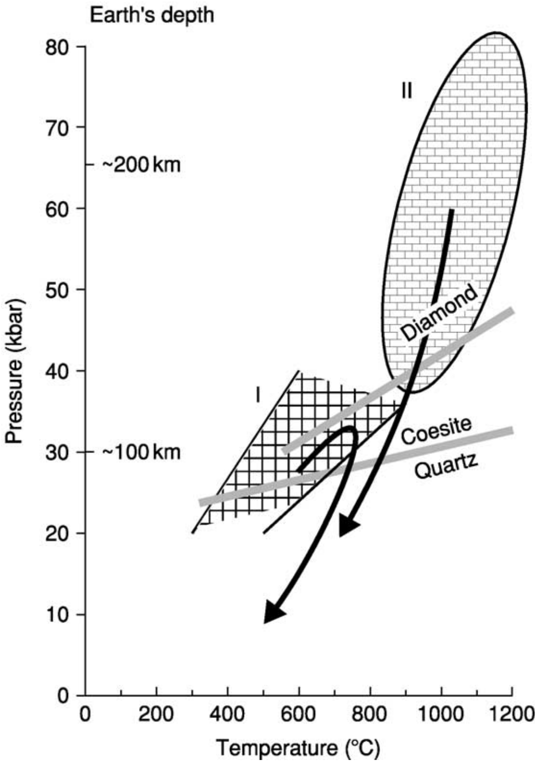


Figure 7 Range of peak pressure and temperature conditions of UHP rocks (patterned areas). The typical shapes of common pressure and temperature paths for UHP/ near-UHP (I) and UHP (II) are depicted.

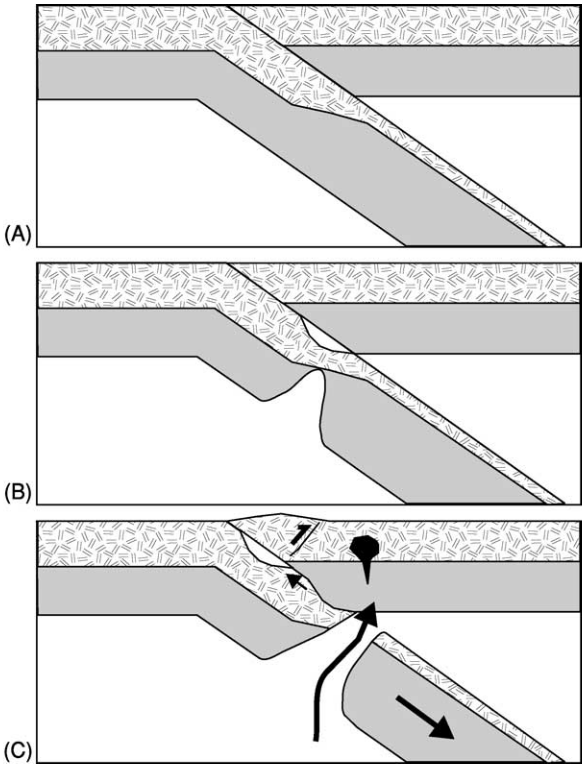


Figure 8 The slab-breakoff model, evolution (from top to bottom). (A) Subduction; (B) slab weakening and narrow rifting; (C) slab breakoff, magmatism, and uplift of UHP sheets. Reproduced with permission from Hacker BR and Liou JG (1998) *When Continents Collide: Geodynamics and Geochemistry of Ultrahigh-Pressure Rocks*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

formation of larger coherent UHP terranes consisting broadly of continental crust, as inferred from the WGR in Norway and the Triassic UHP belt in China, cannot be explained by this mechanism alone. If continental crust adherent to a slab of subducted oceanic crust is drawn into depth at the beginning of a continent–continent collision, as suggested by the slab-breakoff model (Figure 8), extended regions of continental crust can be affected by UHP metamorphism. Fast exhumation is caused by buoyancy forces exerted by continental material that is less dense than eclogites and garnet-bearing ultrabasic rocks, after the oceanic slab has been broken off to be subducted further down.

A second group of UHP rocks experienced significantly higher temperatures, or at least higher pressures (Figure 7). Often both groups of UHP rocks occur together in one crystalline complex. Moreover, both crustal material and mantle material were metamorphosed at peak pressure conditions, between 60 and 80 kbar, as has been proved at least for the UHP regions of the Kokchetav Massif and the Bohemian Massif. The slab-breakoff model may explain this situation as well, because it is believed that continental crust can be dragged to depths of about 200 km by the adherent oceanic slab, despite the buoyancy forces of the continental material. An alternative explanation, however, is delamination of continental lithosphere after continent–continent collision and significant thickening of continental crust, a process that is currently observable in the range of the Himalaya and in the Tibetan Plateau. Modelling experiments suggest that material from the continental crust can be involved in the delamination process and deeply submerged into Earth's mantle. There, anatectic processes in the continental material are caused by the hot environment before fast uplift starts. Because there is limited coherency among the hot-temperature UHP rocks of the Bohemian Massif or the Kokchetav Massif, the debate continues as to whether the slab-breakoff model, lithospheric delamination, or any other process can sufficiently explain the UHP rocks there.

Conclusions

Continent–(micro)continent collision leads to the formation of UHP rocks. Such rocks, although a minority among metamorphic rocks in crystalline basement areas, are common in denuded Phanerozoic orogens. UHP rocks that originate from crustal protoliths are difficult to detect and it seems that they did not appear at Earth's surface until about 600 Ma ago. What could be responsible for that? Crustal thickening during continent–continent collision over a wide lateral area, resulting in a crust 60–70 km thick, may have been uncharacteristic in Proterozoic and Archaean times, thus limiting lithospheric delamination with continental crust involved. The magnitude of crustal thickening might have been influenced by the thermal structure of the (lower) crust and lithospheric mantle, where geothermal gradients are presently lower than they were a long time ago. The present thermal structure could also be responsible for lower geothermal gradients in subducted and metamorphosed oceanic crust, compared to those in pre-Phanerozoic times. Thus, dipping into the UHP field before melting would have been possible for oceanic crust only since the beginning of the Phanerozoic. This would also apply to continental crust adherent to the subducted oceanic crust when the slab-breakoff model would mirror the true process in overprinting continental material at UHP conditions.

See Also

Igneous Rocks: Kimberlite. **Impact Structures**. **Minerals:** Definition and Classification; Quartz. **Regional Metamorphism**. **Shock Metamorphism**. **Solar System:** Meteorites.

Further Reading

Carswell DA (2000) Special issue: Ultra-high pressure metamorphic rocks. International Lithosphere Programme contribution 345. *Lithos* 52.

Carswell DA and Compagnoni R (2003) Ultrahigh pressure metamorphism. *EMU Notes in Mineralogy* 5.

Chopin C (1984) Coesite and pure pyrope in high grade blueschists of the Western Alps: a first record and some consequences. *Contributions to Mineralogy and Petrology* 86: 107–118.

Chopin C (2003) Ultrahigh pressure metamorphism: tracing continental crust into the mantle. *Earth and Planetary Science Letters* 212: 1–14.

Coleman RG and Wang X (1995) *Ultrahigh Pressure Metamorphism*. Cambridge, UK: Cambridge University Press.

Dobrzhinetskaya L, Green HWII, and Wang S (1996) Alpe Arami: a peridotite massif from depths of more than 300 kilometers. *Science* 271: 1841–1846.

Ernst WG and Liou JG (2000) *Ultrahigh Pressure Metamorphism and Geodynamics in Collision-type Orogenic Belts*. Columbia, MD: Bellwether Publishing.

Hacker BR and Liou JG (1998) *When Continents Collide: Geodynamics and Geochemistry of Ultrahigh-Pressure Rocks*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Liou JG, Zhang RY, Ernst WG, Rumble D III, and Maruyama S (1998) High-pressure minerals from deeply subducted metamorphic rocks. In: Hemley RJ (ed.) *Ultrahigh-Pressure Mineralogy: Physics and Chemistry of the Earth's Deep Interior. Reviews in Mineralogy,* vol. 37, pp. 33–96.

Massonne H-J (2003) A comparison of the evolution of diamondiferous quartz-rich rocks from the Saxonian Erzgebirge and the Kokchetav Massif: are so-called diamondiferous gneisses magmatic rocks? *Earth and Planetary Science Letters* 217: 1–19.

Massonne H-J, Dobrzhinetskaya L, and Green HW II (2000) Quartz–K-feldspar intergrowths enclosed in eclogitic garnet and omphacite. Are they pseudomorphs after coesite? *Extended Abstracts of the 31st International Geological Congress at Rio de Janeiro, Brazil, 6–17 August 2000* (on CD; search for Massone).

Parkinson CD, Katayama I, Liou JG, and Maruyama S (2002) *The Diamond-Bearing Kokchetav Massif, Kazakhstan*. Tokyo, Japan: Universal Academic Press.

Schreyer W and Stöckhert B (1997) Special issue: High pressure metamorphism in nature and experiment. International Lithosphere Programme contribution 327. *Lithos* 41.

Sobolev NV and Shatsky VS (1990) Diamond inclusions in garnets from metamorphic rocks: a new environment for diamond formation. *Nature* 343: 742–745.

Van Roermund HLM, Drury MR, Barnhoorn A, and De Ronde AA (2000) Super-silicic garnet microstructures from an orogenic garnet peridotite, evidence for an ultra-deep (>6 GPa) origin. *Journal of Metamorphic Geology* 18: 135–147.

UNCONFORMITIES

A R Wyatt, Sidmouth, UK

Introduction

An unconformity is a surface that separates rocks of significantly different ages. This was at one time an exposed part of the Earth's land surface or the rock surface below a body of water (for example, a lake or the sea), and the younger rocks were deposited on this surface. Juxtaposition of rocks of different ages caused by faulting does not give rise to an unconformity. An unconformity represents a substantial break or gap in the local or regional depositional record. In modern usage this break or gap may have been caused by the erosion of previously deposited rocks or by a long period of non-deposition of sediments (that is, a long enough period that the absence of sediments of the relevant age can be recognized).

Early workers confined the use of the term unconformity to places where the older rocks had been deformed and eroded, so that the unconformity cut across the truncated beds of the lower deposits. The idea that structural discordance is an essential feature of an unconformity continued for much longer in the UK than in many other parts of the world, such as, for example, the USA. Other terms were introduced for breaks where there was no structural discordance. In the Phanerozoic these would normally be identified by gaps in the expected fossil sequence. For the simple case of non-deposition, the terms diastem and non-sequence were used. Although these terms have often been considered to be synonymous, some workers have suggested that a diastem is a break of shorter duration than a non-sequence. Where the break can be shown to be associated with erosion but the upper beds are still parallel to the lower beds, the term disconformity was used. The disconformity surface is often parallel to the bedding surfaces, but it may also show major relief.

Terms such as unconformity and disconformity refer to the surface (and, by implication, the time) that separates the older from the younger rocks. Terms have also been introduced to refer to the relationship between the bedding of the upper (younger) rocks and that of the lower (older) rocks. Where there is no structural discordance, so that the attitude of the upper beds is the same as that of the lower beds, the upper beds are said to be conformable. When there is structural discordance, the upper beds are unconformable.

Where there is structural discordance, as we follow the base of the overlying bed we find that it moves from one to another member of the lower, truncated, series. This is known as overstep (see Figure 1A, side face of block). The term is chiefly used when the angular nature of the unconformity is not obvious but is made evident by detailed mapping. One of the earliest recorded examples was the overstep of the base of the Cretaceous across the underlying Jurassic

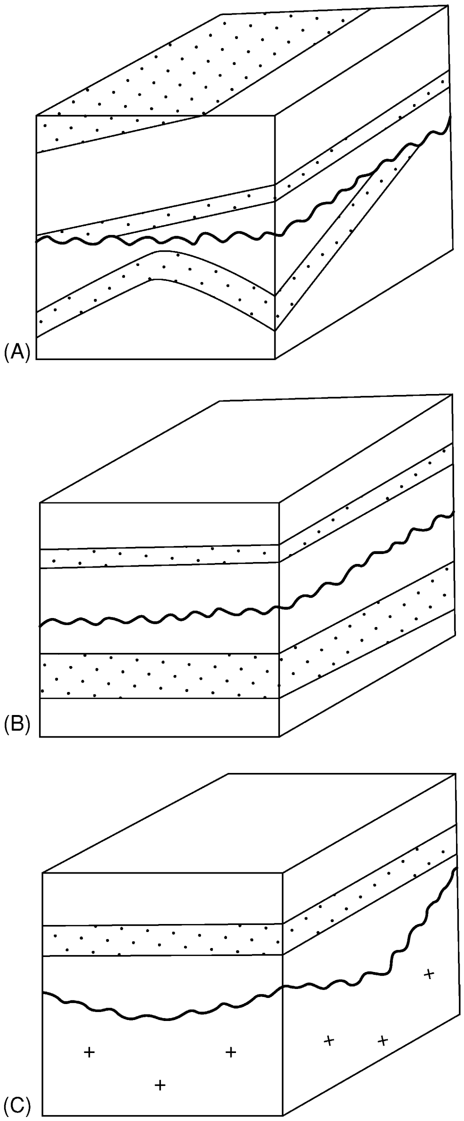


Figure 1 Types of unconformity. (A) Angular unconformity. The front face of the cube shows overlap, caused by onlap from the right. The side face shows overstep of the base of the overlying beds over the dipping lower beds. (B) Disconformity. (C) Heterolithic unconformity.

formations in Yorkshire, which was first noticed in the late eighteenth century.

The rocks overlying an unconformity often show features that record changes in the areal extent of deposition through time, particularly when looked at on a regional scale. For example, where the sea transgresses over the land each bed will cover a slightly greater area than the bed below. This process is known as onlap, and the result, as seen in the rocks, is referred to as overlap (see Figure 1A, front face of block). In recent years many authors have failed to distinguish between the process and the product, using the term onlap to describe both. In traditional usage the opposite of onlap – that is, the successive contraction in the lateral extension of beds in an upward succession – is known as offlap. It should be noted that offlap has been used in a very different sense by seismic stratigraphers (for reflection patterns generated by strata prograding into deep water).

History of the Concept

Some early workers published sketches of what would later be called unconformities, although they did not discuss their significance. Nicolaus Steno (1638–1687) (*see* **Famous Geologists:** Steno) produced a series of diagrams suggesting how unconformable beds in Tuscany could have been produced by cave formation and the subsequent collapse of the roof of the cave. John Strachey (1671–1743) published a diagram showing a sequence of horizontal Triassic and Jurassic rocks overlying inclined Carboniferous Coal Measures in Somerset. Some of his other diagrams suggest that he had a very vague understanding of what he was showing. Jean Etienne Guettard (1715–1786), working in northern France, produced some of the first geological maps, and two of the published maps included sections that clearly show unconformities. Unfortunately, no explanation of the observations was published.

James Hutton (1726–1797) (*see* **Famous Geologists:** Hutton) was the first author whose writings show that he understood the significance of unconformities (though he did not name them). In 1785 he presented his theory of the Earth at a meeting of the Royal Society of Edinburgh. Included in this theory was the concept of geostrophic cycles: the idea that the denudation of the landmasses produces sediments that are deposited on the seafloor and that these sediments are consolidated into rocks, elevated, folded, and denuded. Hutton theorized that there should be places where rocks from one cycle are overlain by rocks from a younger cycle, but there is no evidence that, at the time of giving his paper, Hutton had either seen or read about actual examples. If he had seen the work of Steno, Strachey, or Guettard, he would have been able to point to their diagrams as evidence for his theory.

Over the next few years Hutton searched for, and eventually found, field examples. In 1787 he discovered the unconformity at North Newton, near Lochranza, Arran (Figure 2). Here, reddish and yellowish sandstones, associated with some bands of caliche palaeosol, probably of Late Devonian age, rest with a marked discordance on Dalradian Schists of Late Cambrian age.

Almost immediately after finding the Arran locality, Hutton discovered horizontal beds of the Upper Old Red Sandstone lying on highly inclined (almost vertical) Silurian greywackes near Jedburgh. In the spring of the following year (1788) he found another example of almost horizontal Upper Old Red Sandstone unconformably overlying highly inclined Silurian greywackes at Siccar Point, north of Berwick.

Other examples were found by Hutton's friend John Playfair (1748–1819), whose book *Illustrations of the Huttonian Theory*, published in 1802, did much to draw attention to Hutton's work. Both Hutton and Playfair lacked a simple term to name what they were describing. The term unconformable was introduced in 1805 by one of their geological opponents, Robert Jameson (1774–1854), as an English translation of the German expression *abweichende Lagerung* ('deviating bedding or stratification') used by followers of Abraham Gottlob Werner (1749–1817). For some decades after 1805 geologists described examples of unconformable rocks, without, it appears, paying much attention to the cause of the relationship. It was only in the late 1830s, after Charles Lyell (1795–1875) (*see* **Famous Geologists:** Lyell) started publishing his extremely popular books, that the concept of, and name, unconformity really became incorporated in the thinking and language of geology.

For the rest of the nineteenth century the term unconformity was used to describe an angular discordance between two sets of strata. It was in 1905 that Amadeus William Grabau (1870–1946) extended the use of the term to include cases where there was an obvious erosional break in otherwise parallel strata. Grabau called this a case of disconformable strata, which soon became known as a disconformity.

In 1909 Eliot Blackwelder (1880–1969) suggested that the contact between sedimentary rocks and underlying igneous or metamorphic rocks should also be called an unconformity. Some authors began to apply the term nonconformity to this type of unconformity. Unfortunately, the term nonconformity was already in use as an alternative name for an angular unconformity. To try to avoid this confusion later authors used the term heterolithic unconformity. (Although at first

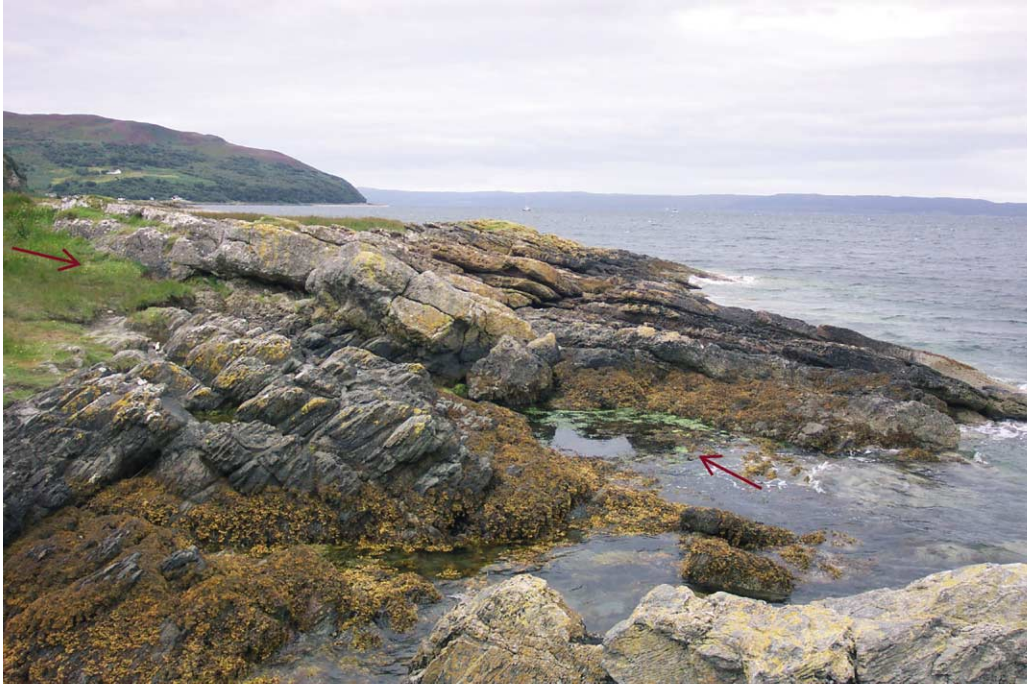


Figure 2 Hutton's unconformity at North Newton, near Lochranza, Isle of Arran, Scotland. The arrows mark the obvious change from steeply dipping Dalradian Schists in the lower part of the photograph to sub-horizontal Devonian sandstones in the upper part.

sight 'heterolithic' (Greek for unlike rocks) may appear to be an appropriate term, it could also apply to clastic rocks overlying carbonates or to marine rocks overlying non-marine rocks. In some parts of the world heterolith is used to refer to interbedded sandstones, siltstones, and mudstones).

In all of the cases discussed above there was an assumption that the surface of unconformity represented a subaerial erosion surface. If we look at many parts of the world today, we see that the land surface is not a smooth plane. It is therefore not surprising that many unconformities that originated as subaerial surfaces are also irregular. This is often referred to as a buried-landscape type of unconformity. A good example is the Torridonian unconformably overlying the Lewisian in north-western Scotland.

In 1910 Bailey Willis (1857–1949) included surfaces of non-deposition in marine sediments as a variety of unconformity. Over the next few decades several authors stressed the importance of subaqueously formed breaks, but it was not until 1957 that John Essington Sanders (1926–1999) proposed a complex Greek-based nomenclature that attempted to distinguish clearly between subaerial and subaqueous breaks. It is, perhaps, not surprising that most people were put off by the nomenclature or thought that it was all an elaborate joke. But in ignoring the terms many people also ignored the attempt to refine geological thinking.

It was inevitable that the expansion of the use of the term unconformity would give rise to some nomenclatorial confusion (Table 1). The same word was used by different authors to describe different concepts, and the same concept was given different names. In recent years there has been some convergence of views on the nomenclature. Authors have tended to use the descriptive terms angular unconformity, disconformity, and heterolithic unconformity (Figure 1). This consensus has, however, been challenged by the specific definition of unconformity that has been used by the proponents of seismic and sequence stratigraphy (see below).

Lateral Variation

One descriptive term is usually adequate to describe an unconformity at a single exposure, but when the surface is traced over wide areas it is common to see the nature of the unconformity vary. A good example is the North Sea Unconformity Complex, often called the 'base-Cretaceous unconformity' or the 'Late-Cimmerian unconformity'. This is perhaps the most easily identifiable surface of the Phanerozoic succession of the Norwegian continental shelf. It displays great local complexity and great variability on a regional scale, such that in different places it has been classified as a nonconformity (in the sense of heterolithic unconformity), a disconformity, and an angular

Table 1 Names that have been given to different types of unconformity

Angular	*Parallel*	*Non-depositional*	*Overlying igneous or metamorphic rock*
Angular discordance	Accordance	Concordant leuroatmodialeima	Heterolithic unconformity
Angular unconformity	Concordant trachyatmodialeima	Concordant leurodiscontinuity	Nonconformity
Clinounconformity	Concordant trachydiscontinuity	Concordant leurohydrodialeima	
Discordance	Concordant trachyhydrodialeima	Diastem	
Discordant atmodialeima	Disconformity	Marine unconformity	
Discordant discontinuity	Eroded surface	Nonevident disconformity	
Discordant hydrodialeima	Evident disconformity	Non-sequence	
Nonconformity	Parallel unconformity	Paraunconformity	
Unconformity	Paraunconformity	Surface of non-deposition	

unconformity. This variation in the nature of the surface reflects local differences in the processes of formation.

The unconformity complex developed during the transition from the synrift stage (active stretching) to the post-rift stage (thermal subsidence and sediment loading) in the development of the northern North Sea basin. A transgression coincided with the transition. This combination of differential subsidence, block rotation, changing patterns of sediment input, and sea-level rise caused local differences in patterns of erosion and sedimentation, which are reflected in the spatial variation of the type of unconformity.

Although varied in detail, there is a general distribution pattern of the different types of unconformity. At the rift margins the rising sea covered the previously exposed basement, producing nonconformities. On the rift flanks, where faulted blocks subsided and rotated, angular unconformities were normally developed. In the centre of the rift, subsidence dominated, generally giving rise to disconformities.

This example, produced during the development of a passive margin, helps to demonstrate that unconformities can originate in a variety of tectonic and sedimentary settings and are not just products of erosion at the end of a geostrophic cycle.

Unconformities and the Stratigraphic Record

Once the Huttonian theory of geostrophic cycles became commonly known, geologists started to apply the reasoning in their efforts to understand and classify the rock record. It became clear that there were major periods of deformation, uplift, and erosion, known as orogenies, which could be recognized over large areas, and the consequent unconformities were used to subdivide the geological column. It soon became apparent that much of Britain and Scandinavia had been affected by the Caledonian orogeny, which was originally thought to have culminated in the Late Silurian. The three classic unconformities discovered by Hutton were all produced by deposition after Caledonian deformation. South-western Britain, and much of the adjacent continent, had been affected by an orogeny that culminated in the Late Carboniferous, which was variously termed the Armorican, Hercynian, or Variscan (*see* **Europe:** Variscan Orogeny). Geologists in North America recognized a similar pattern of orogeny. It was also apparent that there is an ongoing Alpine–Himalayan orogeny.

Although the causes of these orogenic episodes were unclear and were to remain so until the development of plate-tectonic theory in the 1960s, the practical result was the rapid development of the broad outlines of the stratigraphic column. In addition to the major unconformities associated with the final phases of uplift and erosion, other unconformities were discovered that helped in the processes of subdivision and classification. Stratigraphers and palaeontologists could then look in more detail at the rocks bounded by these unconformities.

In many cases the first rocks deposited on an unconformity surface are conglomerates, often containing pebbles eroded from locally weathered rocks. These pebbles can give us information about the rocks that were exposed at that time. For example, the basal Carboniferous conglomerates that lie unconformably on Silurian shales in the east of the English Lake District contain distinctive fragments of the Shap granite. We know that the granite is intrusive into the Silurian (up to and including the Upper Ludlow), so we have some constraints on the timing of cooling, crystallization, uplift, and erosion of the granite.

At a higher stratigraphic level, although in the same area of England, we find the Lower Brockram of Permian age unconformably overlying the Carboniferous Limestone. The Lower Brockram is formed of pebbles that are mostly Carboniferous Limestone. The Upper Brockram, found slightly higher in the succession, contains a large proportion of fragments of Ordovician and Silurian sedimentary and volcanic rocks, demonstrating that by the time the Upper Brockram was deposited an area of these older rocks was exposed and being eroded.

With the development of stratigraphic thinking and knowledge it became clear, however, that unconformities were a poor choice for defining widely correlatable boundaries. An unconformity necessarily implies that there is a gap in the record. This gap may represent different durations of time in different areas. For good correlation, detailed information about fossil occurrences and other temporal markers, such as ash bands, is needed. A gap cannot provide this kind of detail. So a lot of effort was put into searching for areas where a continuous sedimentary record, deposited during the time-span represented by the gap, could be demonstrated. In many cases this involved looking outside the European area, where most of the stratigraphic units had first been defined.

Unconformities and Sequences

For many practical problems, the refinements of stratigraphy are less important than the local rock distribution, so not all geologists abandoned the use of unconformities. From the 1930s onwards Arville Irving Levorsen (1894–1965) interpreted the geology of the mid-continent region of North America in terms of large-scale unconformity-bounded tectonostratigraphic units. Levorsen did not propose names for these units, referring to them as layers of geology, but he demonstrated their importance in petroleum exploration. It was in the late 1940s that Laurence Louis Sloss (1913–96) began to use the term sequence for such major unconformity-bounded units, eventually proposing a formal name for each sequence. In order to distinguish these sequences from standard chronostratigraphic units they were given the names of Indian tribes (Table 2).

Sloss, and many other North American stratigraphers, had encountered problems when attempting to apply the mostly European-based stratigraphic divisions to North American rocks. However, a small number of major craton-wide unconformities could be recognized, based on an integrated study of outcrop and subsurface data. Sloss stressed that at the scale of an individual exposure there was no obvious characteristic by which these inter-regional unconformities, which were used to separate sequences, could be distinguished from the many local unconformities. This was graphically illustrated when the unconformity at the base of the Absaroka Sequence was redefined.

The inter-regional unconformities represented major breaks in the depositional record and were associated with a great degree of overstep and overlap. For example, the rocks at the base of the Sauk Sequence range in age from latest Precambrian to Late Cambrian. They were deposited on an unconformity that cut across rocks of a great range of Precambrian ages. The boundary at the top of a sequence was interpreted as representing a time of major regression of the sea from the continental craton, with associated subaerial erosion. The base of the next sequence represented the re-flooding of the craton. Local unconformities were thought to be produced by minor fluctuations in the rate of sea-level rise or fall, by local tectonics, or by local changes in sediment supply.

Sequences, as promoted by Sloss, are major units, with demonstrably diachronous boundaries. Although they are useful as major subdivisions of North American strata, they were not intended to replace the standard chronostratigraphic units. They were erected because of the perceived differences between the rock record in North America and that in Europe. By no means all stratigraphers agreed with this approach, arguing that better-defined chronostratigraphic boundary sections and improved techniques of correlation would eventually help to solve the difficulties.

Unconformities, Seismic Stratigraphy, and Sequence Stratigraphy

The development of high-quality seismic-reflection profiles, primarily as a result of the intensive search for hydrocarbons, and the calibration of these profiles

Table 2 North American sequences

Sequence name	*Age of rocks included in the sequence*
Tejas	Late Paleocene to Holocene
Zuni	mid-Jurassic to mid-Paleocene
Absaroka	latest Mississippian (post-Chesterian) to Early Jurassic[a]
Kaskaskia	late Early Devonian to Late Mississippian
Tippecano	mid-Ordovician to Early Devonian
Sauk	latest Precambrian to Early or possibly early mid-Ordovician

[a] The Absaroka was originally defined as from the Chesterian.

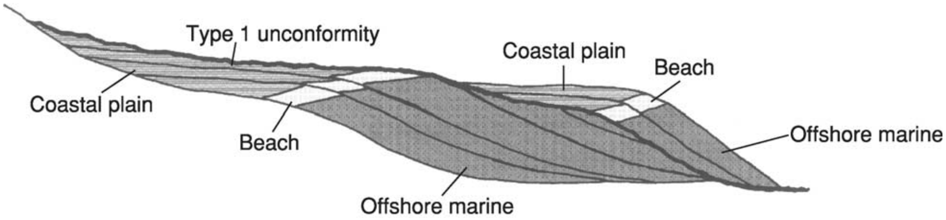


Figure 3 A type 1 unconformity, produced during a rapid relative fall in sea-level, when the rate of eustatic fall exceeded the rate of basin subsidence. The diagram shows the erosion associated with the sea-level fall, the ongoing subaerial erosion, the basinwards shift in facies, the downward shift in coastal onlap, and the abrupt change in facies. Only parts of the sequences above and below the unconformity are represented, but minor marine-flooding surfaces show how the sequences can be subdivided into parasequences.

by geophysical logging techniques applied to relevant boreholes, has stimulated a major development of interest in unconformities. Seismic stratigraphy was originally developed by members of an Exxon research team and was presented and promoted to the worldwide geological community from the mid-1970s onwards. Seismic stratigraphy is simply the geological interpretation of seismic data. Its basic premise is that primary seismic reflectors represent either major bedding surfaces (with the assumption that the reflections are following isochronous horizons) or unconformities. By the 1980s the originators had broadened their concepts and were talking about sequence stratigraphy rather than seismic stratigraphy (*see* **Sequence Stratigraphy**).

Data analysis is based on the identification of stratigraphic units composed of genetically related strata, known as depositional sequences. The lower and upper boundaries of these depositional sequences are unconformities or their correlative conformities. When sequence stratigraphy was first promoted, an unconformity was defined as a surface of erosion or non-deposition that separates younger rocks from older rocks and represents a significant hiatus. A conformity was defined as a surface along which there is no evidence of erosion or non-deposition and along which no significant hiatus is indicated. (Purists might point out the difficulty of correlating a gap with a surface or argue whether theoretically there needs to be a correlative conformity, but most workers seem to accept the overall concept). The sequences of seismic and sequence stratigraphy are much smaller units than the North American sequences named by Sloss, and the time-transgressive nature of their bounding unconformities is not considered to be significant.

By the late 1980s the originally proposed concept of sequences was being refined and extended. Two types of sequence were now recognized, which are known as type 1 and type 2. They are differentiated on the basis of their lower boundaries, which have come to be called type 1 and type 2 unconformities. A type 1 unconformity is characterized by subaerial exposure and erosion associated with stream rejuvenation, a basinwards shift in facies, a downward shift in coastal onlap, and an abrupt change in facies, for example non-marine or very shallow-water marine rocks overlying deeper-water marine rocks (Figure 3). It is interpreted to form when there is a relative fall in sea-level at the depositional shoreline break, i.e. when the rate of eustatic fall exceeds the rate of basin subsidence. A type 2 unconformity lacks both subaerial erosion associated with stream rejuvenation and a basinwards shift in facies (Figure 4). It is interpreted to form when no relative fall in sea-level occurs at the depositional-shoreline position, i.e. when the rate of eustatic fall is less than the rate of basin subsidence.

Sequences were now subdivided into parasequence sets and parasequences. A parasequence is a relatively conformable succession of genetically related beds bounded by marine-flooding surfaces and their correlative surfaces. A parasequence set is a succession of genetically related parasequences that form a distinctive stacking pattern bounded, in many cases, by major marine-flooding surfaces and their correlative surfaces.

These definitions require a clear distinction to be made between a marine-flooding surface and an unconformity. A marine-flooding surface is defined as a surface that separates younger strata from older strata across which there is evidence of an abrupt increase in water depth. The deepening is commonly associated with minor submarine erosion (but no subaerial erosion or basinwards shift of facies) and non-deposition, and a minor hiatus may be indicated. An abrupt increase in water depth implies a transgression at the basin margins, with marine sediments overlying an exposure surface.

An unconformity is now defined as a surface separating younger strata from older strata, along which there is evidence of subaerial erosional truncation (and, in some cases, correlative submarine erosion) or subaerial exposure, with a significant hiatus. This use of the term unconformity is obviously more restrictive than that used when seismic stratigraphy was

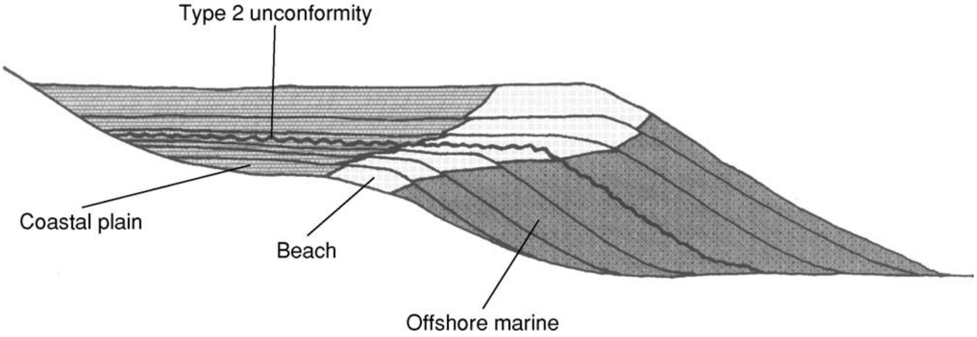


Figure 4 A type 2 unconformity, produced when the rate of eustatic fall is less than the rate of basin subsidence such that there is no relative fall in sea-level. There is no major erosion of the underlying sediments. A major feature associated with a type 2 unconformity is a change from prominent progradation below the unconformity, caused by an increasing rate of regression, to prominent aggradation above the unconformity as the rate of regression slows.

first proposed and is also more restrictive than traditional use. This definition is required to differentiate between sequence and parasequence boundaries. Although the promoters of sequence stratigraphy proposed this restricted definition only in the context of their work, their definition has been applied more widely by some workers.

Although Hutton's original concept of an unconformity was based on his geostrophic theory, subsequent workers have tended to use the term to describe what has been observed, with explanations and hypotheses being separated from the observation. To restrict the use of the term to fit a particular hypothesis, as has been done by the promoters of sequence stratigraphy, is, to some people, a retrograde step. With such a well-known term as unconformity it can also lead to confusion, as many workers continue to follow traditional usage.

One early product of the sequence-stratigraphic model was a series of charts of global cycles of relative change in sea-level through time, based on the interpretation of unconformities and marine-flooding surfaces as products of eustatic sea-level change. Much of the data from which the charts were developed came from North America, with a small number of regional studies from elsewhere. The fact that many of the unconformities on these global cycle charts do not match traditional stratigraphic-unit boundaries, many of which were originally erected based on unconformities, does suggest a rather more complex interplay between local, regional, and global events than that proposed in the model.

Despite these caveats, work on seismic and sequence stratigraphy has promoted a huge increase in our knowledge of unconformities. Although much of this data remains in confidential commercial files, there is sufficient in the public sector to keep geologists arguing and theorizing for years to come.

See Also

Europe: Variscan Orogeny. **Famous Geologists:** Hutton; Lyell; Steno. **History of Geology From 1780 To 1835**. **Seismic Surveys**. **Sequence Stratigraphy**. **Stratigraphical Principles**. **Tectonics:** Mountain Building and Orogeny.

Further Reading

Sanders JE (1957) Discontinuities in the stratigraphic record. *New York Academy of Science Transactions, Series 2,* 19: 287–297.

Sloss LL (1984) The greening of stratigraphy, 1933–1983. *Annual Reviews of Earth and Planetary Sciences* 12: 1–10.

Tomkeieff SI (1962) Unconformity – an historical study. *Proceedings of the Geologists' Association* 73: 383–416.

Vail PR, Mitchum RM, Todd RG, *et al.* (1977) Seismic stratigraphy and global changes of sea level. In: Payton CE (ed.) *Seismic Stratigraphy – Applications to Hydrocarbon Exploration,* pp. 49–212. Memoir 26. Tulsa: American Association of Petroleum Geologists.

Van Wagoner JC, Posamentier HW, Mitchum RM, *et al.* (1988) An overview of the fundamentals of sequence stratigraphy and key definitions. In: Wilgus CK, Hastings BS, Posamentier HW, *et al.* (eds.) *Sea-Level Changes – An Integrated Approach,* pp. 39–45. Special Publication 42. Tulsa: Society for Sedimentary Geology.

UNIDIRECTIONAL AQUEOUS FLOW

J Best, University of Leeds, Leeds, UK

Introduction

Unidirectional water flows are vital agents of erosion, transportation, and deposition in many Earth surface environments and can occur in a wide variety of depositional settings from continental rivers to flows in the deep sea. Unidirectional flows move in one principal direction, with no time-averaged reverse flows within the depth-averaged fluid and, apart from any local deviations caused by bed topography, experience no reverse or oscillatory motion, such as may be produced by waves and tides. Unidirectional flows can be either uniform, where the flow does not vary in velocity or cross-sectional area along its path, or non-uniform, where the fluid velocity and cross-sectional area do change spatially. Non-uniform flows show convective acceleration, where the cross-sectional area decreases and velocity increases, or convective deceleration, where the cross-sectional area expands and the flow slows. In addition to this spatial change in flow properties, unidirectional flows may vary temporally in their behaviour. Flows that show no temporal change in their behaviour are termed steady, whereas those whose velocity changes over time are termed unsteady. Unsteady flows show temporal increases and decreases in velocity, which are often related to the passage of a discrete event such as a flood.

Water flowing over a boundary, whether solid or mobile, develops a flow structure that depends on the velocity and depth of the fluid together with its density and viscosity. The surface over which the fluid moves exerts a frictional drag on the flow, and the region of flow near the bed that is retarded by this friction is termed the boundary layer. Minor friction at the upper atmospheric interface can also cause a small decrease in velocity at the top of the flow in open channels, whereas in unidirectional flows that have solid boundaries all around the flow (such as flow in ice-covered channels or flows through conduits and pipes) significant boundary layers develop from all surfaces. Additionally, unidirectional flows that propagate within another fluid, such as unidirectional density currents, experience significant mixing at their upper boundary owing to shear at this surface. The boundary-layer structure near the surface (or wall) generates a stress on the bed, which initiates and causes sediment transport and ultimately the development of bed morphology. However, the sediment in transport and exact shape and nature of the topography also exert significant feedbacks upon the flow.

Flow Types

Unidirectional flows may be either laminar or turbulent. Laminar flows are dominated by viscous forces rather than the inertial forces acting on the fluid, whereas turbulent flows are dominated by inertial forces. The laminar or turbulent state of flow is expressed by the Reynolds number, Re, where

$$\mathrm{Re} = \frac{\rho \bar{u} Y}{\mu} = \frac{\textit{inertial forces}}{\textit{viscous forces}}$$

and ρ is fluid density, $\bar{u}$ is a characteristic velocity of the flow (such as the depth-averaged mean downstream velocity), Y is a characteristic length scale (such as the flow depth), and μ is the molecular viscosity of the fluid. Flows are termed laminar when $\mathrm{Re} < 500$ and any mixing that occurs is on a molecular scale, turbulent when $\mathrm{Re} > \mathrm{O}2000$ and mixing occurs through the action of turbulent eddies or coherent flow structures at various scales, and transitional when $500 < \mathrm{Re} < 2000$. Many aqueous flows are fully turbulent in their behaviour, although it should be remembered that, owing to their small size, many organisms living in unidirectional flows may experience the overwhelming effects of viscosity and thus live in laminar worlds. Additionally, changes to the turbulent nature of flow can be caused by increasing concentrations of fine suspended sediment, which may modify the nature of the velocity profile and mechanisms of turbulence generation.

Additionally, unidirectional flows are significantly affected by the relative influence of gravitational forces compared with inertial forces, as expressed through the Froude number, Fr.

$$\mathrm{Fr} = \frac{\bar{u}}{\sqrt{gY}} = \frac{\textit{inertial forces}}{\textit{gravitational forces}},$$

where g is acceleration due to gravity. Flows are termed subcritical when $\mathrm{Fr} < 1$, supercritical when $\mathrm{Fr} > 1$, and critical when $\mathrm{Fr} \cong 1$. This dimensionless number expresses the relative celerity of a gravity wave ($\sqrt{gY}$) on the flow: subcritical flows are able to experience the upstream effects of the wave, which has a velocity greater than the flow velocity, whereas the effects of the wave are felt only downstream for

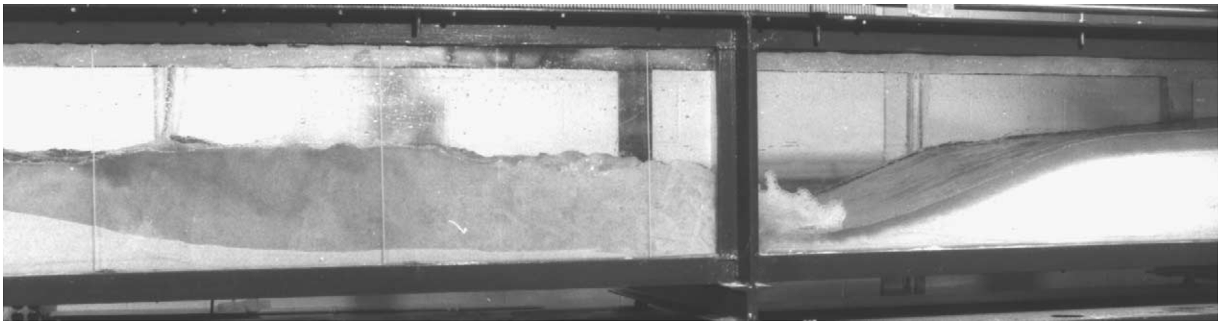


Figure 1 Flow in a laboratory channel, showing the transition from supercritical flow (Fr > 1, right) to subcritical flow (Fr < 1, left) through a hydraulic jump. Image courtesy of John Bridge.

supercritical flows. The form of the Froude number is slightly different for unidirectional density-current flows, where one fluid flows into another, as the effects of reduced gravity must be taken into account. However, the behaviour of the flow, and especially the interaction between the bed and the flow surface and hence the nature of energy loss within the flow, is strongly linked to the Froude number, as exemplified by the transition from supercritical to subcritical flow through a hydraulic jump (Figure 1).

In addition to these properties of flow, unidirectional aqueous flows can show a range of behaviours dependent on their rheology, or how their internal rate of strain responds to an applied external stress (Figure 2). Newtonian fluids, such as pure water, show a linear relation between applied shear stress and strain rate, and hence their viscosity is invariant with respect to the applied stress. Non-Newtonian fluids, however, do not behave in this manner, and either strain rate changes with applied stress (pseudoplastic and dilatant behaviours; Figure 2) or there is an initial yield stress with a subsequent linear stress–strain relationship (Bingham plastic; Figure 2). Although the majority of aqueous flows behave in a Newtonian manner, the addition of significant quantities of fine sediment can cause a change in behaviour and result in flows that have non-Newtonian characteristics, eventually leading to mud or debris flows in which the percentage of water may be very low and the rheology markedly non-Newtonian.

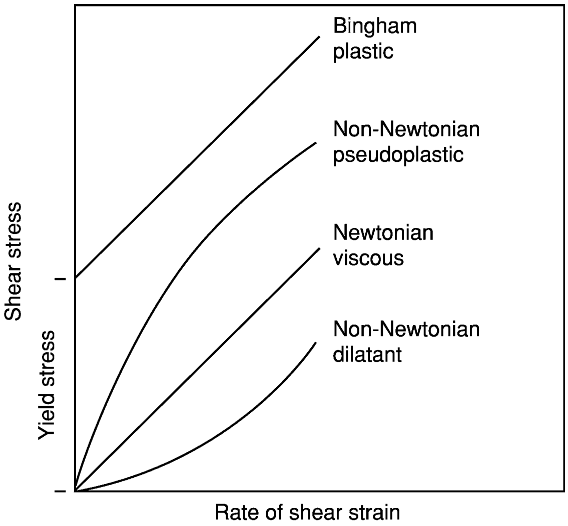


Figure 2 The relationship between the shear stress applied to a fluid and its strain rate, illustrating the various types of behaviour. The viscosity of the fluid is given by the slope of the line.

Velocity Profiles and Boundary Layer Structure

Unidirectional aqueous flows moving over a solid impermeable surface develop a distinct velocity profile away from the wall (Figure 3), and the boundary layer extends into the outer flow until the effects of wall friction become minimal (where the velocity at a point is approximately 95% of the maximum velocity

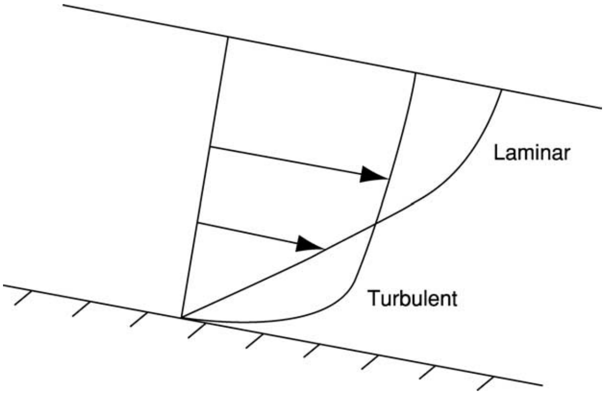


Figure 3 Comparison of a laminar flow and a turbulent flow over an impermeable flat surface. Both flows have the same mean discharge through a cross-section, but they adopt different velocity profiles. (After Middleton GV and Southard JB (1984) *Mechanics of Sediment Movement.* SEPM Short Course Notes 3. Tulsa: Society of Economic Paleontologists and Mineralogists).

in the profile). Turbulent flows, since they possess appreciable mixing between adjacent fluid layers, have a steeper velocity gradient near the wall than do laminar flows (Figure 3). The stresses generated in a laminar flow, τ, are a product of the diffusion of momentum within the fluid on a molecular level and can be expressed as

$$\tau = \mu\left(\frac{\partial \bar{u}}{\partial y}\right)$$

where $\bar{u}$ is the time-averaged velocity at a point and y is the vertical height above the boundary.

However, for turbulent fluids the flow at each point can be broken down into a mean flow velocity, $\bar{u}$, and the deviation from that mean, u' (i.e. $u = \bar{u} + u'$ for the downstream component of the flow). This decomposition of the turbulent-flow signal can be applied to all three components of velocity (i.e. $u = \bar{u} + u'$, $v = \bar{v} + v'$, and $w = \bar{w} + w'$, with u, v, and w denoting the downstream, vertical and spanwise components of velocity, respectively, in the x, y, and z directions). The stresses within turbulent flows are thus a function of both transfer of momentum on a molecular level (the viscous shear stress) and mixing caused by the movement of turbulent eddies within the flow, such that shear stress within a turbulent flow, τ, is given by

$$\tau = (\mu + \eta)\frac{\partial \bar{u}}{\partial y}$$

where η is the so-called eddy viscosity.

The effect of turbulent mixing is that packets of low-momentum fluid from near the bed may be mixed upwards in the flow and relatively faster parcels of fluid from higher in the flow may be carried downwards towards the bed. This exchange of downstream momentum is expressed by $-\rho u'$, and, for a parcel of fluid moving upwards or downwards in the flow (considering the $x{-}y$ plane with u and v as the components of velocity), the rate of change of downstream momentum through a given area is expressed by $-\rho u'v'$. Hence, the time-averaged shear stress in a turbulent flow can be given by

$$\tau = \mu\left(\frac{\partial \bar{u}}{\partial y}\right) - \rho\overline{u'v'}.$$

Similar expressions can be written for the shear stresses exerted by the combination of the velocity fluctuations between the three components of flow, and these are termed the Reynolds stresses.

The velocity profile of a turbulent boundary layer developed over a smooth surface can be divided into several distinct regions (Figure 4): the viscous sublayer, in which the influence of viscosity is dominant but in which some turbulent eddies are initiated (there is a linear increase in velocity with height within the viscous sublayer); the turbulence-generation layer (or buffer zone), which is the region of largest velocity gradient and where the majority of turbulence is generated, which is characterized by a logarithmic velocity distribution with height above the bed; and an outer zone, which is characterized by the dissipation of turbulent eddies generated near the boundary.

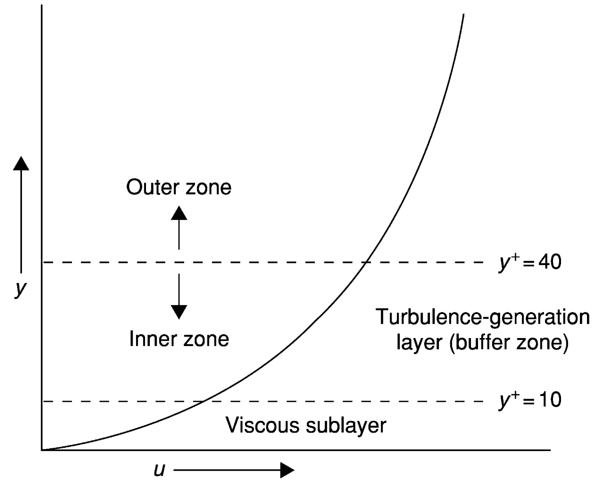


Figure 4 The different regions of a turbulent boundary layer.

The Estimation of Boundary Shear Stress

A key aim of understanding and quantifying boundary-layer structure is to enable prediction of the boundary shear stress exerted on the wall, in order that erosion thresholds and sediment transport can be estimated. Six methods are commonly used to estimate the mean value of the shear velocity, u_*, which is related to the boundary shear stress, τ_B, by $\tau_B = \rho u_*^2$.

1. The slope method uses the slope of the water surface, S, and flow depth, Y, such that $u_* = gYS^{0.5}$.
2. The best fit to the linear velocity profile within the viscous sublayer, $u^+ = y^+$, where $u^+ = U/u_*$ and $y^+ = yu_*/v$, where U is the mean velocity at a point and v is the kinematic viscosity of the fluid (μ/ρ).
3. The best fit to the logarithmic 'law of the wall', which describes the logarithmic shape of the downstream velocity profile in the lower part of a flow (up to the top of the turbulence-generation layer), such that $u^+ = (1/\kappa)\ln(y^+) + C$, where κ is the von Kárman constant (*ca.* 0.40 in clear-water flows) and C is a function of the roughness of the bed.

4. Using linear extrapolation of the Reynolds stress profile to the bed at a height of $y = 0$, such that $-(\overline{u'v'})/u_*^2 = 1 - (y/Y)$.
5. Adopting the spectral method, which uses estimated values of the turbulent dissipation rate, ε, in the inertial region of the spectral domain, such that $u_* = (\varepsilon \kappa y)^{1/3}$.
6. Using the normalized vertical flux of turbulent kinetic energy, which has been proposed to adopt a universal value irrespective of wall roughness, such that $0.5q^2v/u_*^3 \approx 0.30$, where q is the turbulent kinetic energy $(\mathrm{q} = 0.5(\overline{u'}^2 + \overline{v'}^2 + \overline{w'}^2))$.

The Structure of Turbulent Boundary Layers

The nature of mixing within a turbulent boundary layer depends on the exact nature of the turbulent eddies, or coherent flow structures, that are present within the flow. These coherent flow structures are generated within the viscous sublayer and turbulence-generation region, and can be investigated through their velocity signatures and their temporal and spatial length scales. If a two-dimensional flow is considered, four quadrants of flow behaviour can be defined based on the deviations of the u and v components of flow (downstream and vertical velocities) from their respective mean values (Figure 5). This allows recognition of regions of relatively slow downstream-momentum fluid moving upwards within the flow (quadrant 2 events; Figure 5); relatively fast downstream-momentum fluid moving downwards within the flow (quadrant 4 events; Figure 5); and outward and inward interactions of flow (quadrants 1 and 3 events, respectively; Figure 5). This simple quadrant analysis has been widely used in turbulent-boundary-layer research and is appropriate if the flow is largely two dimensional; in fully three-dimensional flows, all three components of velocity should be considered, and octant analyses may be required to characterize truly the fluctuations in fluid flow.

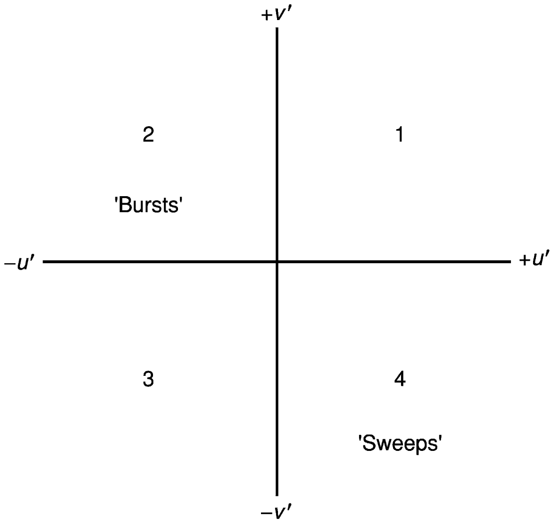


Figure 5 Quadrant classification of a turbulent flow, according to the deviations from the mean values of downstream (u') and vertical (v') velocity.

Much research over the past 40 years has been devoted to elucidating the form of coherent flow structures within turbulent boundary layers, linking these structures to their velocity signatures, and assessing their contributions to both the Reynolds stresses and turbulent kinetic energy budget. Studies have progressed from largely qualitative flow visualizations to quantitative measurements and more recent numerical simulations. Coherent flow structures within a flat-bed turbulent boundary layer are principally composed of:

- low-speed streak areas in the region $0 < y^+ \leq 10$ (where $y^+ = yu^*/\nu$), which are areas of relatively low downstream velocity that are aligned parallel to the flow and form a series of spanwise areas separated by regions of slightly higher flow velocity,
- ejections of low-speed fluid away from the wall (quadrant 2 events; Figure 5),
- sweeps of relatively high-momentum fluid towards the wall (quadrant 4 events; Figure 5), and
- vortical structures of several different kinds, including larger-scale structures, which may occupy a significant fraction of the flow depth and could be generated by amalgamations of smaller groups of vortices originating near the wall.

These structures have been visualized and modelled as a series of longitudinal vortices near the bed, which link through to the legs of ‘horseshoe’-, ‘hairpin’-, or ‘arch’-shaped vortices higher in the flow, which constitute the ejections (quadrant 2 events) (Figure 6). The majority of the Reynolds stresses, approximately 70% in a smooth-wall boundary layer, may be linked to these quadrant 2 and quadrant 4 events, which may be critical in both the suspension of sediment (quadrant 2 events or ‘bursts’) and the entrainment of sediment as bedload (quadrant 4 events), although several studies have also highlighted the significance of quadrant 1 and quadrant 3 events. The transition from laminar to turbulent flow also appears to be linked to the generation of small ‘packets’ of vortices near the wall, which first appear as ‘turbulent patches’ near the bed (Figure 7). These ‘patches’ or ‘turbulent spots’ bear a striking resemblance to the groupings of hairpin vortices generated in fully turbulent boundary layers, which extend through a significant percentage of the flow depth.

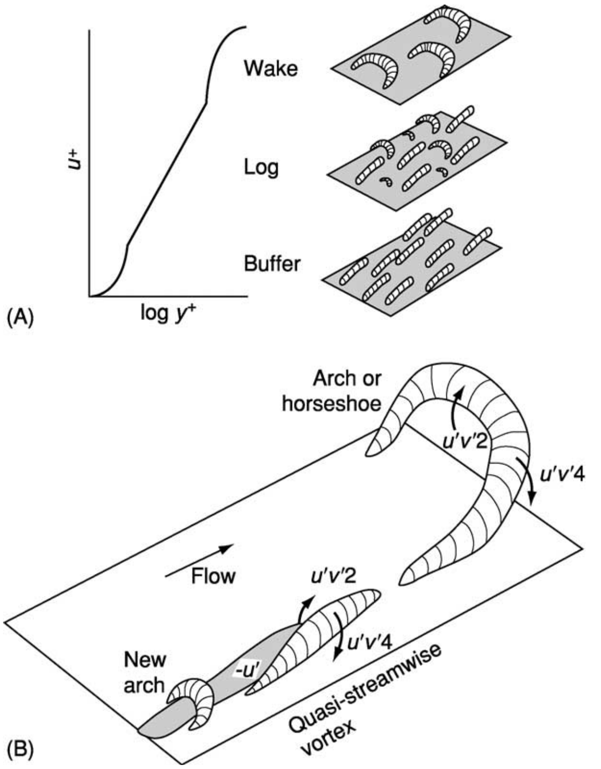


Figure 6 (A) Idealized model of populations of vortices in different regions of a turbulent boundary layer. (B) Schematic model of the links between ejection and sweep motions and streamwise vortices and 'arch'- or 'horseshoe'-shaped vortices in a turbulent boundary layer 2 and 4 refer to quadrant 2 and 4 events (see Figure 5); Reproduced from Robinson SK (1991) Coherent motions in the turbulent boundary layer, *Annual Review of Fluid Mechanics* 23: 601–639.

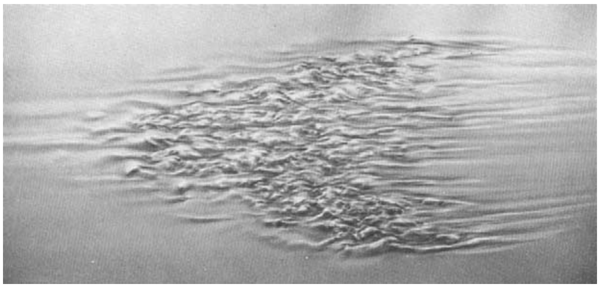


Figure 7 Photograph of a developing 'turbulent spot' at the laminar–turbulent transition, as visualized by transition, as visualized by reflective aluminum tracer particles at the base of the boundry layer and as viewed from above the flow. Flow is right to left. Reproduced from Cantwell B, Coles D, and Dimotakis P (1978) Structure and entrainment in the plane of symmetry of a turbulent spot, *Journal of Fluid Mechanics* 87: 642–672, with permission from Cambridge University Press.

Flow Separation

If a flow encounters a marked change in the gradient of the wall over which it is flowing or is subject to an adverse positive pressure gradient (for example, as a result of injection of fluid through a porous wall), then the fluid may be forced away from the wall and separate from the boundary, with subsequent reattachment of the flow to the bed at some distance downstream, creating a zone of recirculating, or separated, flow near the bed (Figure 8). Flow separation is a key process under most unidirectional water flows and frequently occurs at both positive and negative steps or changes in bed topography. Such areas of flow separation are critical in many unidirectional flows, occurring at a range of scales from, for example, separation behind individual grains, bedforms (Figure 8), and bars to larger-scale features such as those associated with changes in channel

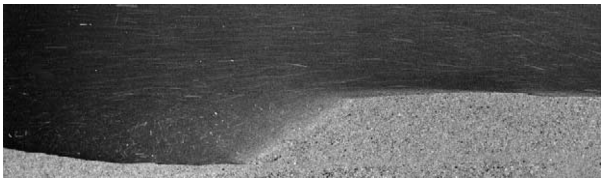


Figure 8 Flow separation visualized behind a dune bedform. Flow right to left. Flow is visualized by the path of neutrally buoyant particles within the water, which shows the recirculating flow within the dune leeside.

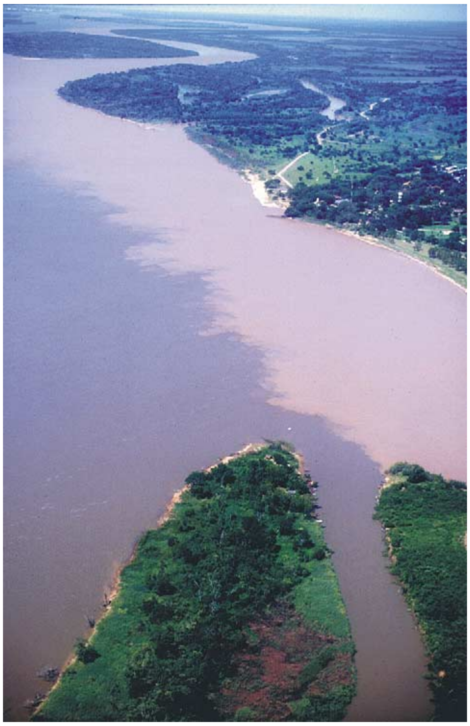


Figure 9 Large-scale Kelvin–Helmholtz instabilities generated along the mixing layer between the Rio Paraná (left, clearer water) and Rio Paraguay (right, higher sediment concentration), Argentina. Flow is away from the viewer, and the width of the image is approximately 1.5 km.

curvature, abrupt gradient changes at the edge of subaqueous slopes, and subaqueous topography. One key consequence of flow separation is that a steep velocity gradient and shear layer are generated between the separation zone and the faster free-stream fluid outside: large-scale coherent vortices, termed Kelvin–Helmholtz instabilities, are generated along this shear layer. Such large-scale vortices are highly turbulent and may be responsible for generating large instantaneous Reynolds stresses, which are often critical in erosion of the bed and sediment transport.

Free Shear Layers

In addition to shear layers associated with flow separation, zones of distinct differential velocity and rapid change in velocity may be present within the body of a unidirectional flow, owing to flow convergence around topography or in combining channels or to shear at the top of a subaqueous density current, for example. Turbulence and mixing across such 'free' shear layers depend on the velocity differential across the shear layer and the relative densities and viscosities of the two incoming flows, but these free shear layers often create large-scale Kelvin–Helmholtz instabilities (Figure 9), which dominate both fluid mixing and the instantaneous Reynolds stresses. For example, fluid mixing at channel confluences (Figure 9) has been shown to be greatly influenced by the shear-layer dynamics between the incoming flows, and interactions between free shear layer and bed topography are thought to control the downstream dispersal of suspended sediments and pollutants.

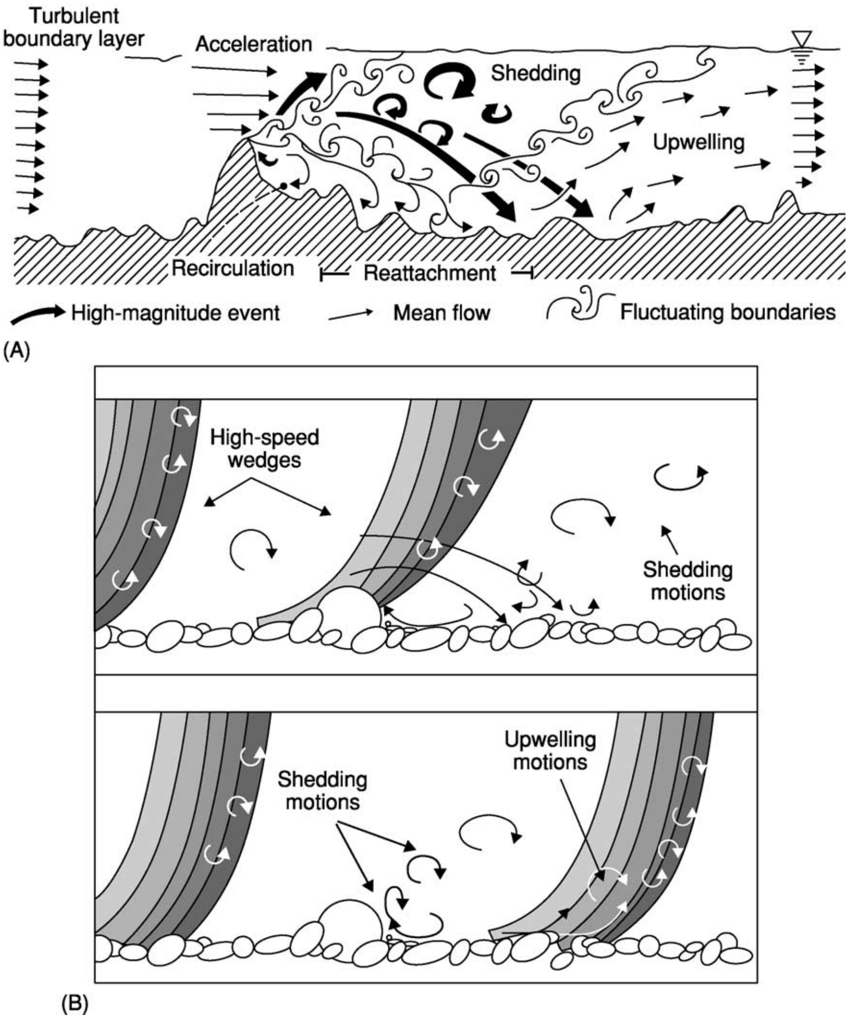


Figure 10 (A) Flow regions associated with the presence of a pebble cluster on the turbulent flow field. (B) Schematic diagram of the large-scale flow structures proposed to develop over a rough gravel surface, showing (top) how the passage of a large-scale high-speed flow structure expands the flow separation zone in the lee of the clast and (bottom) how this generates vortex shedding from the separation zone and upwelling in the region of flow reattachment. Both A and B are reproduced from Buffin-Belanger T and Roy AG (1998) Effects of a pebble cluster on the turbulent structure of a depth-limited flow in a gravel-bed river. *Geomorphology* 25: 249–267.

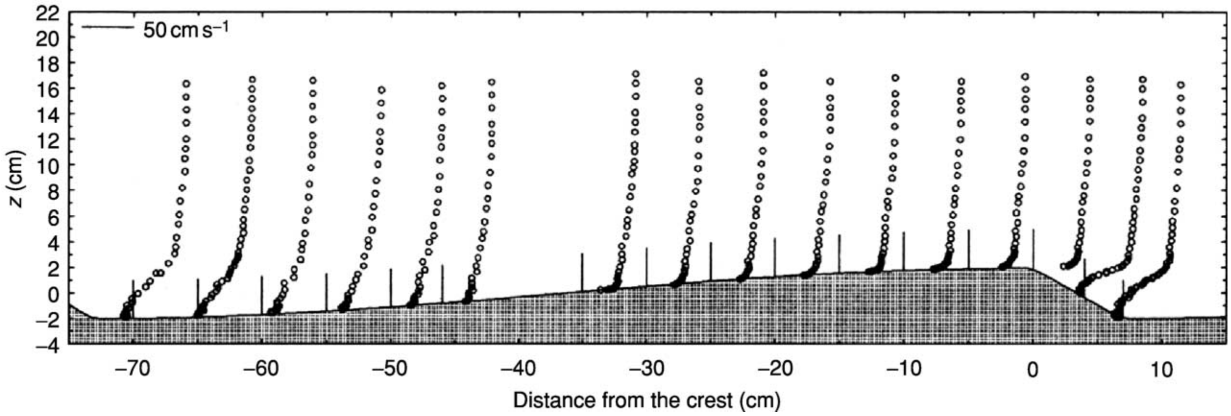


Figure 11 Velocity profiles measured over a dune bedform, showing flow separation in the dune leeside and boundary-layer recovery over the stoss side of the next downstream dune. Flow left to right. Reproduced from Nelson JM, McLean SR, and Wolfe SR (1993) Mean flow and turbulence fields over two-dimensional bedforms. *Water Resources Research* 29: 3935–3953, with permission from American Geophysical Union.

Other Factors Influencing Boundary Layer Structure

In many unidirectional aqueous flows, the precise nature of the mean and turbulent flow is influenced by a range of variables that can significantly alter the flow structure, bed shear stress, patterns of sediment transport, and, hence, development of bed morphology. Some of the most significant influences on the characteristics of unidirectional aqueous flows are described below.

The Nature of Bed Grain Roughness

Particle roughness significantly increases the potential for turbulent mixing near the bed and often results in an increase in the gradient of the near-bed velocity profile, with a concomitant increase in the bed shear stresses derived from the velocity gradient, Reynolds stress, or turbulent kinetic energy budget. Grain roughness may destroy the viscous sublayer and also increase the generation of turbulence near the bed, through either encouraging intensified bursting (quadrant 2 events) from between the grains (and thus larger-scale return quadrant 4 events) or generating regions of flow separation around individual grains or groups of particles, which may both create significant velocity gradients near the bed and generate large-scale coherent vortices, associated with flow separation both in front of and behind the particles, which can penetrate the entire flow depth (Figure 10).

The Presence and Type of Bedforms

Many bedforms such as ripples, dunes and larger-scale bar forms, create their own flow field through topographic, convective accelerations and decelerations of fluid that may significantly change the nature of a unidirectional flow. In addition, many bedforms are also associated with flow separation on their upstream stoss side or downstream lee side, which generates appreciable turbulence and a boundary layer that is recovering from flow separation downstream of the region in which the flow reattaches to the bed (Figure 11).

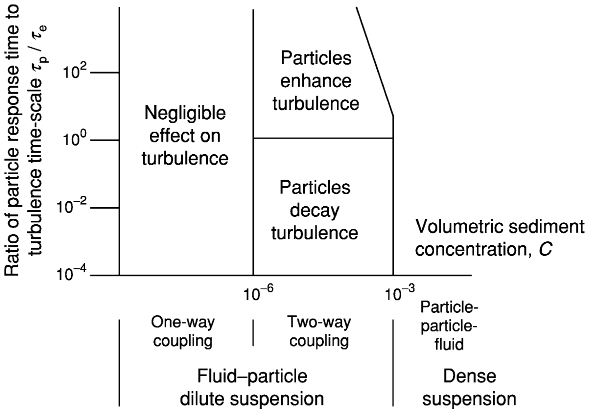

Ratio of particle response time to turbulence time-scale τ_p / τ_e
10^2
10^0
10^{-2}
10^{-4}
Negligible effect on turbulence
Particles enhance turbulence
Particles decay turbulence
Volumetric sediment concentration, C
10^{-6}
10^{-3}
One-way coupling
Two-way coupling
Particle-particle-fluid
Fluid–particle dilute suspension
Dense suspension

Figure 12 Schematic diagram of the attenuation or enhancement of turbulence due to the presence of sediment in a flow as a function of the volumetric sediment concentration, *C*, and the ratio of the particle response time to the turbulence time-scale, τ_p/τ_e (after Elghobashi S (1994) On predicting particle-laden turbulent flows. *Applied Scientific Research* 52: 309–329).

The Type and Quantity of Suspended Sediment

Many turbulent flows transport appreciable quantities of suspended sediment, with the suspended concentration in some flows reaching levels at which the flows become markedly non-Newtonian, such as in the Huanghe River in China, where concentrations of up to 1290 kg m^{-3} have been recorded. Suspension

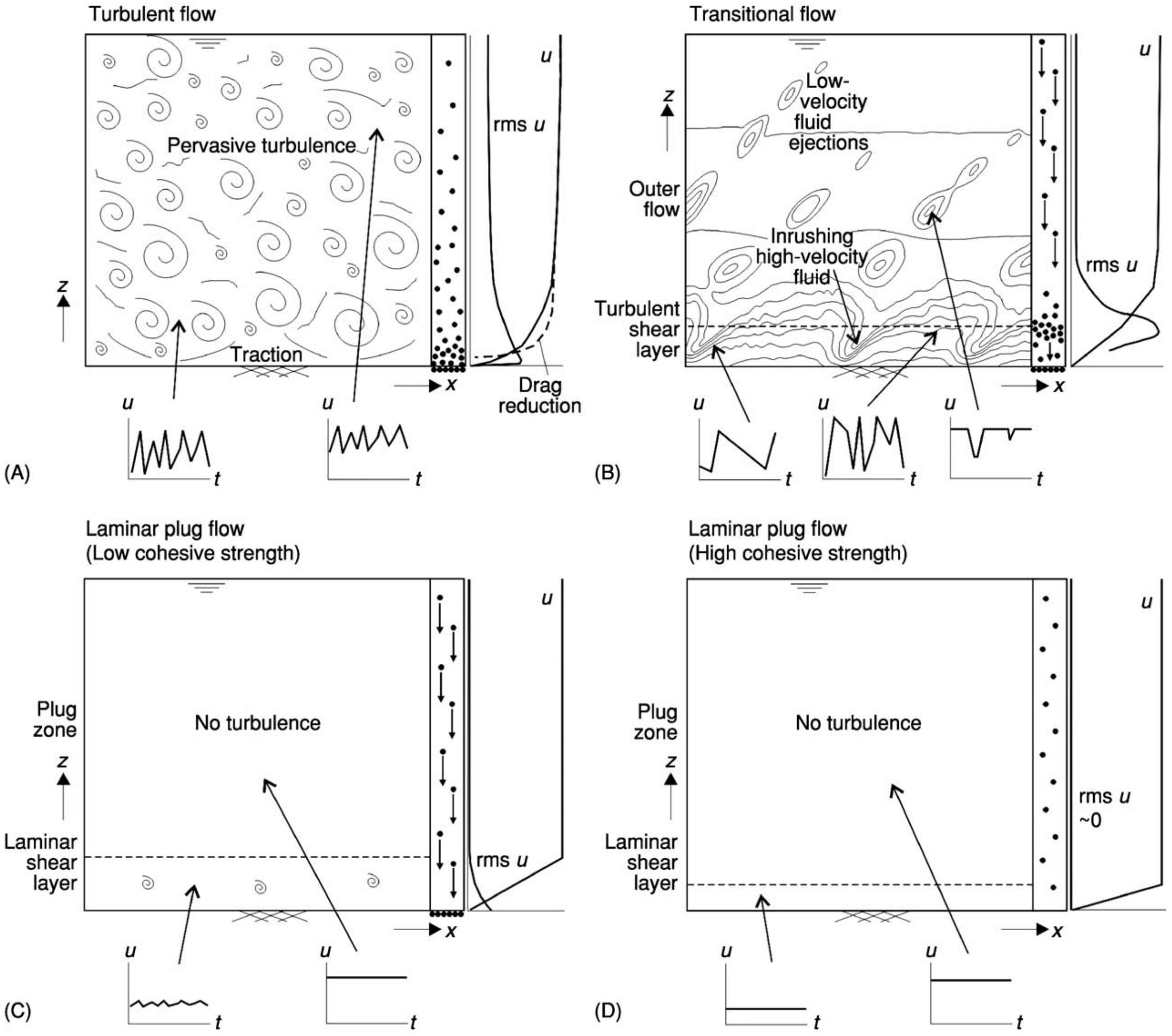


Figure 13 Schematic model for flows with increasing clay concentrations. Each model depicts the characteristic velocity profile and nature of flow, a vertical profile of the rms *u*, representative time series at several heights, and a representation of sediment settling. rms refers to the root mean square value of the downstream velocity component, or level of turbulence, within the flow, and indicates the principal regions of turbulence generation within each profile. (A) Turbulent flow with a logarithmic velocity profile and turbulence generation near the bed. At low clay concentrations, drag reduction may begin to occur. Coarser sediment is supported through turbulence, and sedimentary structures can develop. (B) Transitional flows with a developing shear layer, which separates a lower region of high velocity gradient from an upper layer of reduced shear. Sketch of flow depicts streamlines. Turbulence is strongest in the shear layer, along which Kelvin–Helmholtz instabilities are developed with a distinctive velocity signature (see inset sketches). Sediment entrained into the basal region is trapped, and parallel lamination may be produced by the variable shear stresses induced by the shear-layer instabilities. (C) Laminar plug flow without turbulence and with low cohesive strength. The cohesive strength of the flow is unable to support coarser sediment, which settles to the bed. (D) Laminar plug flow with high cohesive strength is able to support coarser sediment suspended within the flow. Reproduced from Baas JH and Best JL (2002) Turbulence modulation in clay-rich sediment-laden flows and some implications for sediment deposition. *Journal of Sedimentary Research* 72: 336–340. SEPM (Society for Sedimentary Geology).

of sediment requires turbulence within the flow, but a feedback is exerted where at some point the turbulence begins to be modified by the sediment in suspension. This complex feedback mechanism is poorly understood, with factors such as the concentration of sediment and ratios of grainsize: turbulent length and time scales thought to be important in causing either a decrease or increase in turbulence (turbulence attenuation and enhancement respectively) within the flow (Figure 12). Turbulence modulation is thus a key feedback mechanism within many unidirectional flows, acting to both enhance and suppress turbulence production, and is also known to change the downstream velocity profiles significantly with subsequent implications for sediment transport and sorting (Figure 13).

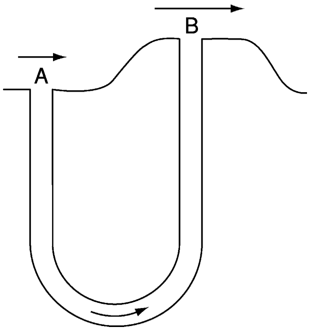


Figure 14 Flow acceleration above the raised limb of a burrow creates a lower pressure at B than at A, inducing flow within the tube (after Vogel S (1994) *Life in Moving Fluids*. Chichester: Princeton University Press).

The Porosity of the Bed Surface

Many studies of unidirectional flows have assumed that the bed is impermeable and that the subsurface flows exert little influence on the overlying boundary layer. However, this assumption is invalid, and the flow within porous beds, which comprise the surface of most sedimentary environments, can significantly alter the nature of the overlying boundary layer. This is especially true in the presence of bed morphology, which may generate differential velocities and fluid pressures around the topography. Such flow is seen, for instance, out of relict burrows that have a raised rim (Figure 14), where flow through the burrow and towards the area of raised topography results from lower fluid pressure associated with increased velocities at this raised opening. However, flow within porous beds may also lead to modification of the near-bed velocity profiles, and it has been suggested that the presence of a porous bed can decrease near-bed velocities and cause the velocity profile to deviate from a logarithmic form, with turbulence penetrating the top part of the porous bed.

See Also

Sedimentary Environments: Alluvial Fans, Alluvial Sediments and Settings; Deserts; Storms and Storm Deposits. **Sedimentary Processes:** Erosional Sedimentary Structures; Depositional Sedimentary Structures; Particle-Driven Subaqueous Gravity Processes; Deposition from Suspension; Fluxes and Budgets.

Further Reading

Ashworth PJ, Bennett SJ, Best JL, and McLelland SJ (1996) *Coherent Flow Structures in Open Channels*. Chichester: John Wiley and Sons.

Baas JH and Best JL (2002) Turbulence modulation in clay-rich sediment-laden flows and some implications for sediment deposition. *Journal of Sedimentary Research* 72: 336–340.

Bridge JS (2003) *Rivers and Floodplains*. Oxford: Blackwell Publishing.

Buffin-Belanger T and Roy AG (1998) Effects of a pebble cluster on the turbulent structure of a depth-limited flow in a gravel-bed river. *Geomorphology* 25: 249–267.

Cantwell B, Coles D, and Dimotakis P (1978) Structure and entrainment in the plane of symmetry of a turbulent spot. *Journal of Fluid Mechanics* 87: 641–672.

Elghobashi S (1994) On predicting particle-laden turbulent flows. *Applied Scientific Research* 52: 309–329.

Leeder MR (2000) *Sedimentology and Sedimentary Basins: From Turbulence to Tectonics*. Oxford: Blackwell Publishing.

López F and García MH (1999) Wall similarity in turbulent open channel flow. *Journal of Hydraulic Engineering* 125: 789–796.

Middleton GV and Southard JB (1984) *Mechanics of Sediment Movement*. SEPM Short Course Notes 3. Tulsa: Society of Economic Palaeontologists and Mineralogists.

Nelson JM, McLean SR, and Wolfe SR (1993) Mean flow and turbulence fields over two-dimensional bedforms. *Water Resources Research* 29: 3935–3953.

Nezu I and Nagkagawa H (1993) *Turbulence in Open-Channel Flows*. Balkema: International Association for Hydraulic Research.

Pope SB (2000) *Turbulent Flows*. Cambridge: Cambridge University Press.

Robinson SK (1991) Coherent motions in the turbulent boundary layer. *Annual Review of Fluid Mechanics* 23: 601–639.

Van Rijn LC (1990) *Principles of Fluid Flow and Surface Waves in Rivers, Estuaries, Seas and Oceans*. Oldemarkt: Aqua Publications.

Vogel S (1994) *Life in Moving Fluids*. Chichester: Princeton University Press.

Williams J (1996) Turbulent flow in rivers. In: Carling PA and Dawson M (eds.) *Advances in Fluvial Dynamics and Stratigraphy*, pp. 67–125. Chichester: Wiley and Sons.

URALS

See **EUROPE: The Urals**

URBAN GEOLOGY

A W Hatheway, Rolla, MO and Big Arm, MT, USA

Introduction

Most of the world's population lives in relatively crowded conditions in urban areas, affording them immediate contact with all forms of sustenance. At the same time, these teeming populations require huge imports of potable water, treatment of sanitary wastes, and export of solid, special, and hazardous wastes, along with a degree of infrastructure that impinges on and relies heavily on the constraints represented by the geological setting. Four major geological themes govern the application of geology to human life in cities and urban centres. First, although an abundance of regional geological information is applicable to urban development and life, the integration of most of such data is not readily discernible by the majority of the regional population. Second, urban life is concentrated such that there is 'loading' of the geological environment under various types of 'footprints' of engineered structures. Third, as the trend of importance of urban life for people continues to expand, 'megacities' become the centrepiece of new form of urban geology. Last, the concentration of urban populations in coastal regions interfaces with growing concerns over sea-level rise and global climate changes.

Cities historically have grown and developed around geological core areas where geological conditions were favorable to defense or security of construction. Although the form of these settings may seem geographical in nature, it is the underlying geology that has created such conditions. The following geological situations and the cities with which they are associated exemplify this:

- Natural, hard-rock sheltered seaports (Plymouth, England; Hong Kong; San Francisco, California; New York City).
- Confluences of major rivers (Pittsburgh, Pennsylvania).
- Mouths of major navigable rivers (Alexandria, Egypt; New Orleans, Louisiana; Para, Brazil).
- Heads of navigation of major rivers (Minneapolis–St. Paul, Minnesota).
- Confluences of rivers and pre-railroad trails (Paris; Rome).
- Defensive positions, underpinned by bedrock (Seoul; Rome).
- Sea-lane confluences for early trade (Singapore; Capetown, South Africa; Boston, Massachusetts).
- Confluences of pre-railroad trails (Kansas City, Kansas; Santa Fe, New Mexico; Edmonton, Canada).
- Mouths of mountain passes (Salt Lake City, Utah; Denver, Colorado; Reno, Nevada; Missoula, Montana).

Geoscience is a major potential contributor to maintenance of the health and welfare of cities and their populations. Successful implementation requires planning parameters for growth and redevelopment of the built environment and assessments of least-impact and least-cost alignments (e.g., the constantly needed improvements for rapid transit to move people around and in and out of cities). Geoscience plays an important role in the location, development, and delivery of potable water supplies and in the effective disposal of wastewater effluents. Previously used land must be characterized with respect to toxic contamination and its remediation, and risk from death, injury, and property losses stemming from geological hazards must be mitigated. Geological science is critical to understanding the potential for flooding, earthquakes, volcanic eruptions, ground collapse, mass movements, and seismic sea waves in densely populated areas. In the twenty-first century, yet another application of geoscience is in the consideration of how certain geological conditions might serve to enhance acts of terrorism.

Geological Influences on Urban Development

Modern engineered works in the urban environment represent relatively large and high-magnitude impacts on the substrate on which they are built. Geological site characterization of regional earth materials is thus an absolute necessity before laying the foundation for any large engineered structure, for economic reasons (i.e., to reduce construction costs) as well as for successful structure operation and maintenance and for basic environmental acceptability (Figure 1). Most of the high-impact concerns for site characterization are devoted to a few simple themes. The physical nature of foundation earth materials (soil, weak rock, or rock) must be identified as being able to support engineered works without unacceptable structural deformation or outright collapse. Geological anomalies that might compromise the integrity of building components or entire engineered work must be detected and delimited. Furthermore, 'bad ground' that would require premium foundations or difficult construction efforts and higher building costs must be recognized and presented to planners (Figure 1).

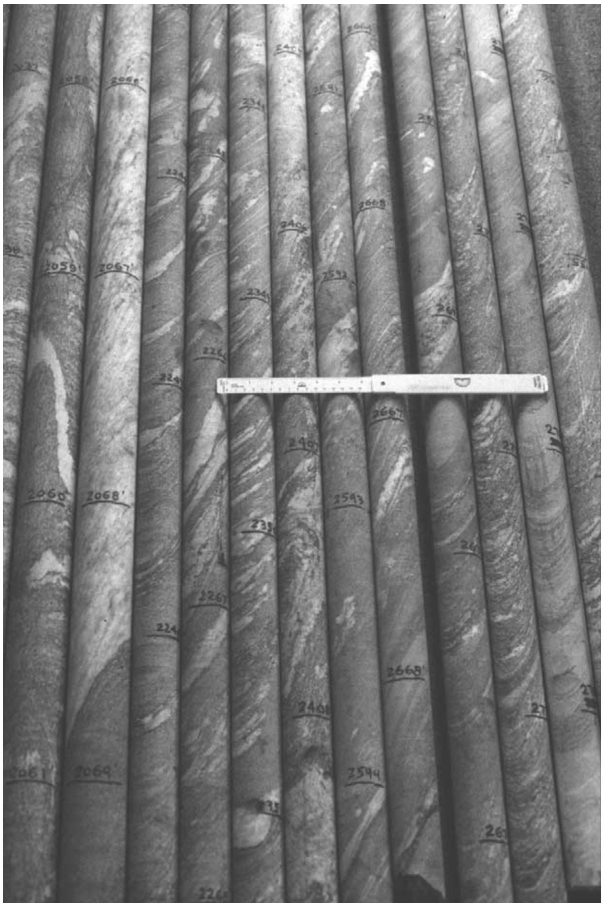


Figure 1 NX-size (2.125 inches) rock core recovered by oriented, triple-tube coring technique from more than 650 m below Kennedy International Airport, Long Island, New York City, at the site of a proposed liquefied natural gas storage cavern planned by the Brooklyn Union Gas Company (now Key Span Energy Co., Inc.). The rock is complexly mixed Fordham Gneiss and Manhattan Schist and these are select segments of 3-m lengths, all unbroken by natural discontinuities. The photograph shows the generally excellent foundation and deep cavern characteristics of New York City bedrock. Ruled portion of scale is 15 cm in length; depth marks on core are in metres. Photograph by the author.

Cities of The World Literature Series

For more than 20 years, the Association of Engineering Geologists (AEG) has fostered the incorporation of geoscience within urban development by publishing an international series of papers (*Cities of The World*) dedicated to the memory of the outstanding efforts of the late Canadian geological engineer, Robert F Legget (1904–94). The quarterly journal of the AEG (now a joint effort with The Geological Society of America), *Environmental and Engineering Geology*, seeks and publishes detailed accounts of urban geology, utilizing a standard format (Table 1; see also www.aegweb.org) that serves as a codification of the important elements of urban geology.

Problematic Conditions of Urban Construction

Construction in the built environment has to address a variety of geological 'constraints' that may be hidden from view during general observation of the existing ground surface. Without accurate geological information at the planning and design stages, new construction and urban renewal efforts will almost certainly encounter cost overruns, regulatory compliance infractions, and some type of construction failure. Three primary geological considerations (soil, groundwater, and geological dicontinuities) are integral to avoiding such difficulties. The nature and thickness of soil units that will bear the dead and live loads transferred downward from the intended construction must be established. Likewise, the presence, depth, and potential fluctuation of the groundwater surface below the intended engineered works must be evaluated and related to needs for dewatering without detriment to stability of existing nearby engineered works. The soil and groundwater profiles may pose potential excavation problems when deep basements are required to accommodate vehicle parking. The presence of and adverse geometrical orientation of geological discontinuities (bedding, joints, shear zones, and faults, to name a few) may

Table 1 Standard elements of urban geological considerations for *Cities of the World* journal series[a]

Chapter	*Section*
1. Background	1.1 Location 1.2 History of founding 1.3 Geological influences affecting founding
2. Geological setting	2.1 Brief on regional geology 2.2 Geology of the city 2.2.1 Basement rock 2.2.2 Surficial units (soils) 2.2.3 Stratigraphic chart
3. Geotechnical characteristics	3.1 General foundation-related geological units 3.2 Exploration methods 3.3 Typical foundation types in use 3.4 General laboratory test methods
4. Materials	4.1 Traditional types and uses 4.2 Sources and extraction methods 4.3 Regulations and zoning affecting extraction 4.4 Environmental impact of extraction
5. Geological constraints	5.1 Classification 5.2 Recurrence 5.3 Mitigation
6. Historic resource extraction	6.1 History 6.2 Classification of extracted ground; mines and fluids (water, oil, gas) 6.3 Areal extent 6.4 Constraints related to extracted ground 6.5 Mitigation of extracted-ground threats
7. Seismicity of the city	7.1 Historic record 7.2 Notable events 7.3 Generalized recurrence interval 7.4 Ground motion amplification factors 7.5 Seismic design provisions in force
8. Environmental concerns	8.1 Water supply 8.2 Wastewater treatment 8.3 Waste management (solid, special, and hazardous) 8.4 Remediation of uncontrolled wastes 8.5 Wetlands factor
9. Major engineered structures (tabulated)	Detail as appropriate
10. Use of underground space	
11. Summary	11.1 Conclusions 11.2 Predictions for the future
References	As appropriate
Illustrations (key illustrations of the geological situation of the city)	Frontispiece (color oblique emphasizing major geological features) Index map General geological planimetric map Stratigraphic column Geotechnical cross-section Seismicity plot Optional photographs

[a] Content and format recommendations of the Association of Engineering Geologists for papers submitted for publication in the series dealing with continuing development of the world's cities. Papers are published in the *Environmental and Engineering Geoscience Journal*.

affect the stability of basement excavations and the integrity of surrounding buildings and other structures. These aspects of engineering geology require the geological team to interface with urban historians, archaeologists, architects, urban planners, insurers, financiers, and others related to the design and construction processes (Tables 2 and 3).

Role of the Engineering Geologist

Geological information is critical to the siting, design, and construction of all engineered works. This is particularly important in the urban environment, where all physical aspects of construction are compounded in their effects by mandates in scheduling

Table 2 Geotechnical influences on urban construction

Geotechnical influence	*Emphasis*	*Key considerations*
Site geological knowledge reduces the risk of unknown ground conditions	Unknown, undetected, or undisclosed geological conditions can compromise scheduling, cost, and operational performance of engineered works	Capacity of foundation soils to support loads of the structure; stability of surrounding facilities; excessively 'weak' or 'strong' ground exacerbates the construction effort
Some earth materials have undesirable properties or characteristics	Avoidance of constraints or slowdowns to construction process	Detection in the site exploration process, notification of owner and design engineer; incorporation into the design and construction specifications
'Faces' across which geologic character of construction ground changes	When not anticipated by the contractor, can cause unwanted perturbations in schedule and in project cost; where spanned by one 'bay' (segment) of construction, may lead to unacceptable differential settlement of portions of the foundation	Avoidance requires adequate site characterization, funding by the owner, and judicious selection of geological and geotechnical support consultants
Pockets or zones of 'bad' ground	Three-dimensional bodies of degraded earth material not able to support design, construction, or operational efforts or roles of project	Generally related to geomorphic or tectonic considerations of origin and may have characteristics detrimental to construction or of performance of the facility
Near-surface groundwater	Always a problem	Generally interferes with construction, particularly in placing the foundation; may require dewatering, which may affect performance of surrounding existing foundations
Perched water or groundwater is generally detrimental to the construction process	Control and removal without impairment of construction or with performance of completed facility	When truly 'perched', drains into the construction excavation within hours and does not replenish
Nature of site preparation or construction 'spoil' (soil) or 'muck' (tunnel spoil)	Must be removed from the construction site and reused in some worthwhile manner acceptable to the community	Spoil and muck have geotechnical characteristics that must be heeded in considering their reuse

Table 3 Engineering geological contribution to urban geology

Contribution	*Purpose*	*Important considerations*
Stratigraphy	Define the nature and bounds of soil types and of geological formational units	Controls the suitability and relative cost of siting and dimensional design of virtually all projects
Engineering properties of the foundation soil or rock	Must be capable of bearing the combined live and dead loads of the engineered works to be constructed	Acceptability measured in terms of nil compressibility and sufficiently high shear strength to support the loads of the project
Geological structure of bedrock exposed in construction	Locate and define fault-displaced geological units and discontinuities that are of sufficient length of exposure to cross any one dimension of site excavation	Avoidance of adverse geological structures, premium foundation conditions, and expensive or 'bad' ground in terms of underground construction
Occurrence of groundwater	Protection from damage by human activities; avoidance of premium costs for foundations	In no way can groundwater be of beneficial consequence to the construction process
Surface-water hydrology	Using geological evidence to define the nature of flooding as an economic and human-welfare concern	Mainly involves interpretation of present geomorphic features that control the path, depth, and velocity of low-frequency/high-impact flood events
Avoidance of surrounding unstable ground	Protect from what may slip onto the site or move down-slope from the site	Mainly ground that is unstable under gravitational and slope-water conditions, along with rock falls
Avoidance of existing subsurface voids	Avoid presence of abandoned tunnels, mine shafts, mines in general, and natural voids such as karst caverns	All subsurface voids suffer from decreasing 'stand-up' time, which is the ability to span loads imposed at the ground surface, and from the surrounding overburden weight of the geological column of materials

Table 4 Problems of urban construction: naturally troublesome geological conditions

Condition	*Impact*	*Geological considerations*
Soft and/or otherwise compressible soils	When detected will lead to engineering selection of medium to deep foundations, in turn leading to premium foundation costs	Afflicted typically by clay–mineral-rich soils such as glacial, glaciomarine, and glaciolacustrine clay soils, as well as a variety of slack-water fluvial clays, silts, and 'muds'; learning to expect the presence of such conditions is based on regional and local geological knowledge
Certain marine silts and clayey silts are geotechnically 'quick' (unstable) by virtue of collapse-prone 'card-house' structure brought about by flocculated structure under saline depositional conditions	Can become unstable and can collapse under dynamic shock of earthquakes and a variety of man-made shocks such as pile driving, blasting, and dropping of large loads	Where bordering river and stream banks and other steep terrain; type formation is the Leda Clay of Quebec Province and the marine clay soils of Scandinavia where encountered on land
Clay ('mud') 'plugs' formed as still-water deposits of truncated river meanders	Creates unacceptable differential settlement of foundations, roads, and bridge abutments and piers	Endemic geotechnical problem in the Lower Mississippi River Valley of the United States
Organic soil of all types (notably peat)	Causes foundation deformation of all manner of engineered structures with all but the lightest bearing loads	Tends to have irregular but often oval to elliptical bounding contacts in the lateral sense; two type occurrences for this problem are the Newport Beach–Fountain Valley areas of the Los Angeles basin, especially at the Pacific Ocean mouth of the ancestral Santa Ana River and the southern metropolitan area of Edmonton, Alberta, Canada
'Highs' in bedrock surface	Escalates costs of site preparation for construction	Sites must be 'brought to grade' as a preparation for installation of foundation; rock can cost ten times or more to remove, compared to soils
Buried zones of degraded bedrock	Increased costs of tunnels for water supply, sewerage, and rapid transit	Often the most difficult to anticipate; some are due to fault movement and displacement, others are due to geochemical cation exchange by water
Buried valleys	Channels of unconsolidated, porous earth material ('soils'), with preferential movement of groundwater	May place earth materials of vastly different engineering properties in the line of linear projects such as roads and tunnels, representing unanticipated design/construction conditions
Near-surface groundwater	Calls for dewatering of basement excavations	Avoidance of disturbance of adjacent foundations and alteration of static ground surface, the latter leading to rot in older driven timber piles
Dumps of organic debris	Excessive settlement of foundations of successor buildings	Early residents routinely filled all topographic declivities with refuse, most of which is vertically compressible with time
Dumps of industrial (hazardous) wastes	Threats to health, construction workers, nearby residents, and occupants of the new facility	Mainly from low-level, chronic emissions of toxic vapors from semivolatile and volatile organic compounds
Floodplain development	Unwise development of land subject to poorly predictable low-frequency flooding	Geomorphic evidence of past flooding often crucial in defining the need for zoning against development
Proximity to seismically active faults and fault zones	The major seismotectonic zones are now well known throughout most of the world	Avoidance of high-density human occupancy along known, 'active' fault traces and in areas of 'bad ground' subject to loss of foundation stability in large-magnitude earthquakes
Proximity to major volcanoes	Ashfalls and lahars (volcanically induced mudflows)	In mild and humid climates, geochemical weathering rapidly produces volcanic soil, which attracts people because of cultivation advantages; geological evidence can predict rough levels of risk by time and location, for future exposure

and in costs of construction. The process by which geological input is discovered and provided to permitting officials, planners, owners, designers, and construction contractors is termed ‘site characterization’, and this task must include the discovery and description of man-made wastes (perhaps of archaeological/historical importance) left from previous site occupation. Site characterization as a concept generally arose in about 1900, but has taken on the particular meaning described here only since about 1980. It is a well-accepted practice in the field of engineered construction and environmental permitting.

Engineering geologists are generally tasked with planning and conducting site and waste characterizations, and the most important aspects related to the effort are experience, training, and professional competence of the geologist and the provision of adequate exploration funding by the builder/owner. In this respect, there has been a truly unfortunate trend towards unwarranted price-driven selection (‘bid-shopping’, or commoditization) of site-characterization consultants, by owners and some engineers, worldwide, since about 1980. Suffice it to say that the geologist who will perform the site characterization should be selected on the basis of qualifications and experience, rather than on the cost (budget) accorded to the effort.

Engineering Geological Site Characterization

Each construction site needs to be characterized for its geological, hydrological, and waste conditions, so that architects and design engineers have the necessary parameters when considering feasibility and design. In some cases, when geological constraints (the preferred term for ‘geological hazards’) are identified, the proposed construction may be shifted to a geologically ‘less expensive’ site in terms of building costs, operation, and maintenance of the proposed works. More often, however, for a variety of reasons, the owner of the planned construction, public or private, is committed to the site and the geological knowledge becomes essential to the success of the project.

Engineering geologists generally begin site work with a walkover of the site, attempting to ‘peer’ below the ground surface to generate a working hypothesis and conceptual geological model of what is likely present and what may be hidden. The ultimate goal is to assess how geotechnical factors could impact the construction or performance of the engineered works. Tables 4 and 5 provide a summary of geological conditions that might be expected to impact project feasibility, design, and construction processes

Table 5 Problems of urban construction: societal pressures causing geological impacts

Condition	*Impact*	*Geological considerations*
Cities require off-street vehicle parking for a significant percentage of building occupants and visitors	Above-ground space is too valuable for parking vehicles, hence deep basements are required	Hardrock excavation costs are extreme; bedded sedimentary and jointed or foliated crystalline rock present slope stability problems for the basement excavation during construction; stability of adjacent buildings is a constant concern
Water must be supplied to the city	Need for conveyance projects and tunnels; importation of the water for the distribution system	Linear projects such as pipelines, aqueducts, highways, and railroads are particularly susceptible to earthquake ground motion
Transportation routes must not interfere with much of the existing ground-surface infrastructure	Reliance on tunnelling to achieve minimal grades between stations and minimal depth below ground for people movement	Places geological information at a premium for design and construction of stations; keeps transit tunnel alignment in sound rock, free of major ground-support considerations
Abandoned waste dumps of all types	Organic debris will naturally compress as much as 300% vertically (becoming as little as one-third of its preconstruction thickness, hence causing considerable settlement)	Use archival records, old topographic maps, historic aerial photographs, and confirming geophysical traverses
Previous quarrying, mining, and other mineral extraction activity	Many cities once relied on extraction of coal, building stone, gypsum, clay mineral, or mineral ore, leaving unstable ground	Present-day hazards include rotting support timber, collapse-prone workings, and movement of contaminated groundwater in workings
Excessive abstraction of groundwater	Causes rotting of historic timber foundation piles; these may also rise on termination of water abstraction	Saline intrusion in coastal areas, activation of non-seismic ‘growth’ faults in post-Cretaceous coastal embayments

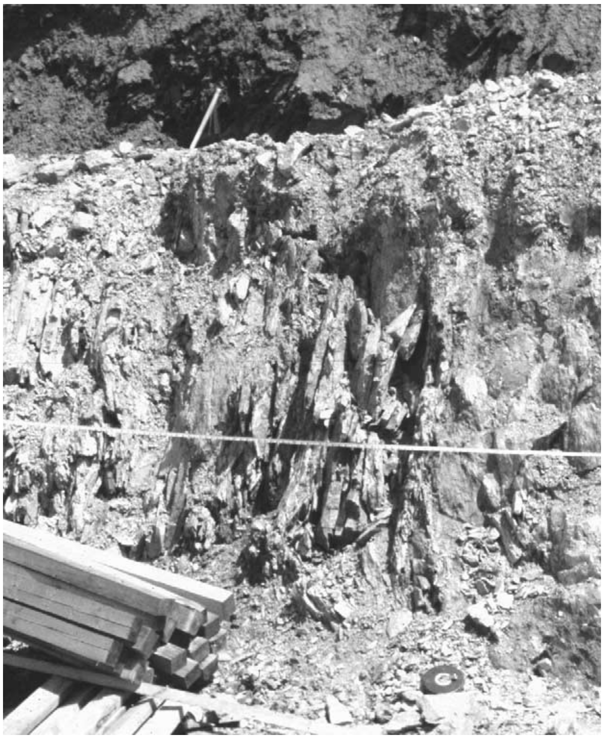


Figure 2 An example of a zone of 'bad ground' discovered and delimited in an engineering geological site characterization exploration. This zone represents poor engineering properties of foundation materials in terms of their ability to support structural loads that would be imposed by engineered works. Shown is a shear zone of ancient tectonic origin, as it perpendicularly enters the foundation excavation for the Lahey Medical Clinic, urban Boston, Massachusetts. Such features are important when they are planar to and cross one or more dimension of the foundation excavation, because they have the potential to induce instability of the surrounding ground and existing buildings. Timbers are 200 mm^2 in section; survey tape in midview. Photograph by the author.

and therefore must be incorporated into the site engineering geological characterization process.

Geological conditions decidedly influence geotechnical design measures to transfer dead and live loads of proposed structures into the foundation soil or rock (Figure 2). Transfer of load to geological horizons is integral to ensuring adequate structural support for engineered works. During the site characterization process, appropriate site exploration techniques must be incorporated to determine the engineering geological conditions likely to adversely impact the project.

Summary

The struggle to provide decent living conditions for humanity on earth is characterized to greatest degree by the interwoven impact of geology on the human habitat and the impact of humans on geologic conditions that serve to support that life. Cities are the focus of human activity and this activity takes place on and in the ground, which is itself only a complex of geologic conditions. Of all of the qualities of human life in the cities is its dynamism, all of which tends to obscure and often obliterate the delicate evidence of the nature of geologic conditions supporting the city. For this reason, we must turn to urban geologic observations and recording in order to compile a sequence of snap-shot views and vignettes of urban geology. When recorded, this information is invaluable in the continual struggle to accommodate the great demands.

See Also

Engineering Geology: Aspects of Earthquakes; Natural and Anthropogenic Geohazards; Made Ground; Site Classification; Ground Water Monitoring at Solid Waste Landfills. **Europe:** Holocene.

Further Reading

Association of Engineering Geologists (1982–2004) Cities of the world. *Environmental and Engineering Geoscience Journal* (annual series).

Baskerville CA (1992) *Bedrock and Engineering Geologic Maps of Bronx County and Parts of New York and Queens Counties, New York: USGS Miscellaneous Investigation Series MAPI-2003, two sheets, 1:24,000.* Washington, DC: US Geological Survey.

Culshaw MG (2005) Urban geoscience (the Glossop Lecture for 2004). *Quarterly Journal of Engineering Geology and Hydrogeology.*

Kaye CA Jr (1959) *Geology of the San Juan Metropolitan Area, Puerto Rico: US Geological Survey Professional Paper 417-A*. Washington, DC: US Geological Survey.

Kaye CA Jr (1976) *The Geology and Early History of the Boston Area of Massachusetts; A Bicentennial Approach: Geological Survey Bulletin 1476*. Washington, DC: US Government Printing Office.

Legget RF (1962) *Geology and Cities*. New York: McGraw Hill.

Legget RF and Hatheway AW (1988) *Geology and Engineering*. New York: McGraw-Hill.

McCall GJH, de Mulder EFJ, and Marker BR (1996) *Urban Geoscience*. Rotterdam: Balkema.

Schlocker J (1974) *Geology of the San Francisco North Quadrangle, California: US Geological Survey Professional Paper 782*. Washington, DC: US Geological Survey.

Schuberth CJ (1968) *The Geology of New York City and Environs*. New York: Natural History Press.

United States Geological Survey (1894–1955) *Folio series.* Select titles pertaining to cities; maps at 1:125,000 plus text and photographs. Washington, DC: US Geological Survey.

VENUS

See **SOLAR SYSTEM: Venus**

VOLCANOES

G J H McCall, Cirencester, Gloucester, UK

Introduction

Volcanoes are a major component of the Earth's present surface geology, both active and extinct; volcanic processes can be recognized throughout geological history and are important in generating certain types of mineral deposits (e.g., metallic sulphides, sulphur). They are studied in their relationship to both mantle (*see* **Earth:** Mantle) and crustal (*see* **Earth:** Crust) processes of igneous rock generation and plate tectonics. Volcanic eruptions comprise one of the most important of natural hazards (*see* **Engineering Geology:** Natural and Anthropogenic Geohazards) threatening populations living close to them. Volcanoes have been recognized on other bodies of the solar system – Venus (*see* **Solar System:** Venus), Mars (*see* **Solar System:** Mars) and Jupiter's satellite Io (*see* **Solar System:** Jupiter, Saturn and Their Moons) – the latter is the most volcanically active body in the solar system. Mars has the largest single volcano, Olympus Mons (440 km diameter; 24 km altitude). Venus, which we can only study from radar images on account of its dense volcanogenic CO_2-rich atmosphere, appears like Mars, to have no active volcanoes now, though there are many inactive ones.

Volcanoes and the Mantle

Liquid rock, poured out from volcanoes as lava (*see* **Lava**), makes up only a small portion of the planet, though a large part of the core is molten metal. The outer layer of the Earth, the lithosphere, is relatively cool, but the mantle below is so hot that rocks lose their cohesion. Indirect evidence obtained from seismology suggests that they move very slowly; and the theory of plate tectonics requires such movements to take the form of convection currents and localized upwellings (hot spots, mantle plumes) (*see* **Mantle Plumes and Hot Spots**). However, there is no agreement among geoscientists as to the depth of the lower boundary of these circulations. The hot, soft, solid material of the mantle, of peridotite and related compositions, only partially liquifies when the temperature exceeds the melting point of minerals in the rock. Because it is lighter than the solid rock above it, it will rise towards the surface, entering the lithosphere. It may pass through the crust directly and swiftly appear at the surface as a volcano, or it may be halted and form a large concentration of molten rock down in the crust, a magma chamber: from this it may later burst out to the surface as a residue, changed after some crystallization in the magma chamber. In narrow conduits to the surface, it may cool and crystallize, forming wall-like intrusions (dykes) or sheet-like intrusions (sills). Intrusive rocks formed by crystallization of the magma in the magma chamber, dykes, and sills, form the underworks of volcanoes, and may be all that is left after erosion: they frequently comprise ring complexes as on the Isle of Rum in Scotland. A quirk of erosion has left these underworks exposed, surrounded by Precambrian granite, in a hollow central to the remains of the lava and tuff pile in the 100-km wide Miocene Kisingiri volcano in western Kenya – Howel Williams called this the best preservation yet seen of a volcano complete with its upper- and under-works.

Distribution of Volcanoes

Volcanoes do not occur everywhere on the Earth's surface. The global distribution shown in Figure 1 reflects the present distribution of more than a dozen tectonic plates, rigid plates of the lithosphere, which move laterally across the Earth's surface at minute velocities in the order of centimetres per year. Of course, they cannot do this indefinitely without colliding with each other and this pattern is controlled by upwelling of magma on the lines of separation (the mid-ocean ridges) and either the diving down of the spreading oceanic plate under the one it spreads against (subduction) or, where continental parts of plates meet, collisions such as formed the Himalayas. In subduction there is both

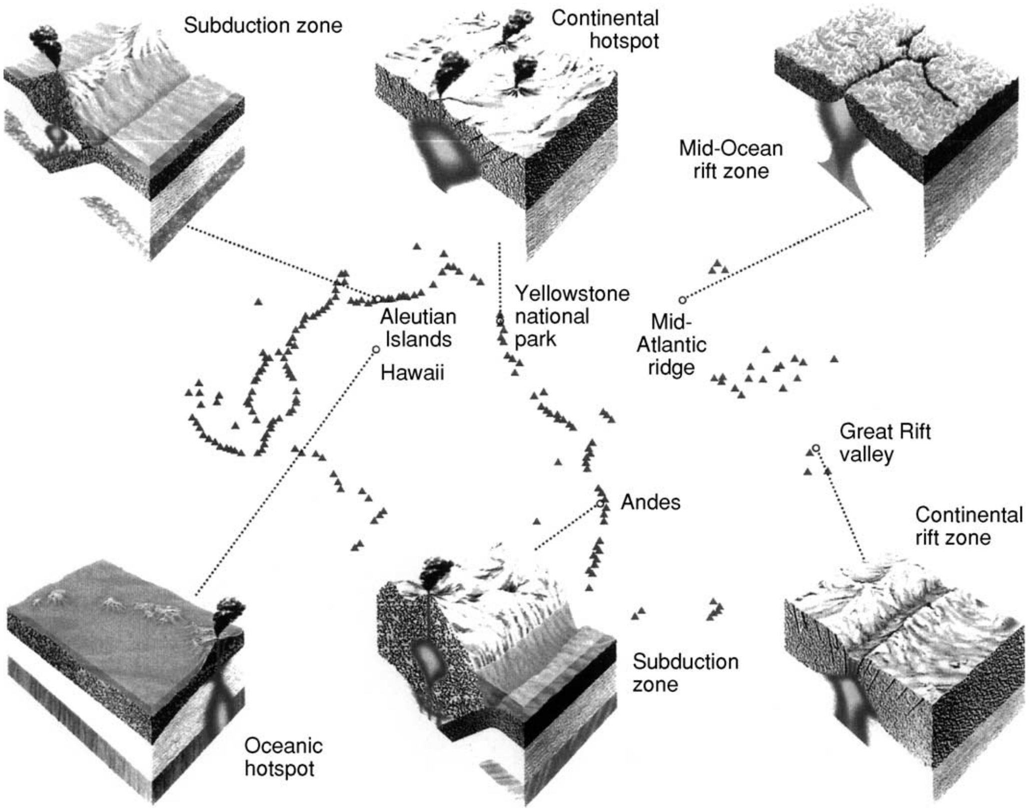


Figure 1 The global distribution of volcanoes and their geotectonic setting (after Scarth (1997)).

upwelling of magma from the mantle and melting in the lithosphere, after the descending oceanic plate has penetrated to some depth, and volcanic arcs are so formed, not at the surface boundary between the plates (the oceanic trench), but at some distance beyond it within the overriding plate.

As shown in Figure 1, most of the existing volcanoes of the Earth are concentrated in the arc zones; the ridge zones of upwelling produce much eruption deep under the oceans, but few subaerial volcanic piles occur on them: exceptions being the volcanoes on Iceland and Tristan da Cunha. Some volcanoes occur in the oceans unrelated to either type of plate tectonic zone, but are attributed to hot spots or plumes in the mantle: the Hawaiian Islands and the Galapagos are examples of these. Seamounts are stumps of such volcanoes. There are also volcanoes in rift zones within continents, for example Lengai, Tanzania, in the eastern rift and the volcanoes of the western rift (e.g., Nyiragongo), but such rift valleys are really aborted oceanic ridges, where the updoming and eruption has occurred, but the two flanks never spread and diverged away from the rift.

Classification of Volcanoes

There are two types of eruption:

Fissure Eruption

Concentrated in long, narrow fissures, the lava generally spreads to form extensive lava plateaus. If the eruption is explosive, plateaus of ash flows may form. Most fissure eruptions are, however, associated with fluid basalt lavas, which have given rise to the terms flood basalts and plateau basalts, but phonolites in Kenya also form extensive lava plateaus. Such eruptions have not been common in historical times, though the Laki fissure eruption in 1783 was of this type: however, extensive plateaus hundreds of kilometres wide of superimposed, quite thin fluid flows were formed in the geological past (e.g., the Deccan Plateau in India).

Central Eruption

Volcanic activity is more commonly concentrated in a central vent and gives rise to volcanic cones or, in the case of some viscous lavas, domes or necks with no summit crater. Single volcanoes, clusters, or chains are formed in this way. The chains may be associated with belts of extreme seismic disturbance (e.g., the Aleutian islands, Chile). Central volcanoes may have satellite cones superimposed on them: these are said to be nested (Figure 2) if they occur within the crater, parasitic or adventive if they occur on the outer slopes.

Central volcanoes of substantial dimensions may develop large circular cavities, calderas, focused on their summits: these may be circular, or of rather irregular annular shape; they may even be multiple, one inside or superimposed on the other. The Suswa caldera in Kenya is annular, a central plateau being preserved. Calderas may be formed either by subsidence, due to withdrawal of magma/lava support, or volcanic explosion. Calderas are evident on Mars and Io.

There are three types of central volcanic pile, all of which may develop calderas:

i. shield volcanoes: all lava flows
ii. ash cones: all pyroclastic
iii. composite or strato-volocanoes: alternating lava and pyroclastic (Figure 3).

The material that flows out as lava is termed effusive. Extrusive includes both lavas and pyroclastic rocks.

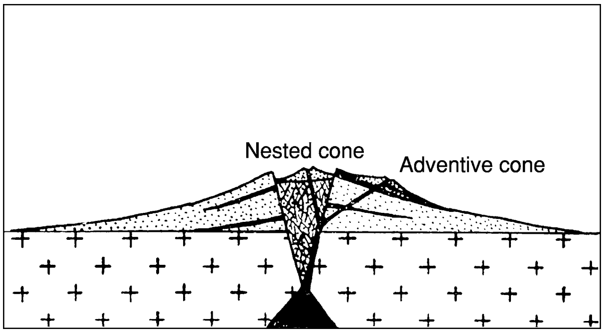


Figure 2 Diagram of a composite volcano: lava (black) and tephra (stippled) interbedded (from Green and Short (1971)).

There are seven types of volcano, classified according to the nature of the eruption. Four of the names stem from southern Italy, a field of classical study which can be termed the 'Birthplace of Vulcanology'; and which has lately been described by Guest *et al.* in modern terms. Iceland, Hawaii, and Martinique supply the remaining names. The list is given in Table 1.

This scheme is a very useful subdivision, but most volcanoes show some departure from these classical definitions derived from a handful of famous volcanoes. No other volcano, for example, matches Stromboli, where explosions occur every 10 minutes and activity has been continuous for 2500 years. Other types can be long dormant, Mt Pelée in Martinique, lay dormant for centuries, only, in 1902, to send off sudden blasts lasting a few minutes separated by weeks of relative inactivity while the plug in the vent built up into a spine (Figure 5).

A very useful table of explosivity of volcanoes was published in 1982 (Table 2):

Products of Volcanoes

Lavas

Magma erupted from volcanoes as lava consists of molten rock, crystals, and gas – carried as bubbles, mainly water, and carbon dioxide. The less siliceous lavas such as basalt flow freely and build up extensive shield volcanoes as in Hawaii. Other lava types may flow freely – a phonolite flow forming the Yatta Plateau in Kenya followed a valley for >300 km from its source. Less fluid flows such as trachyte

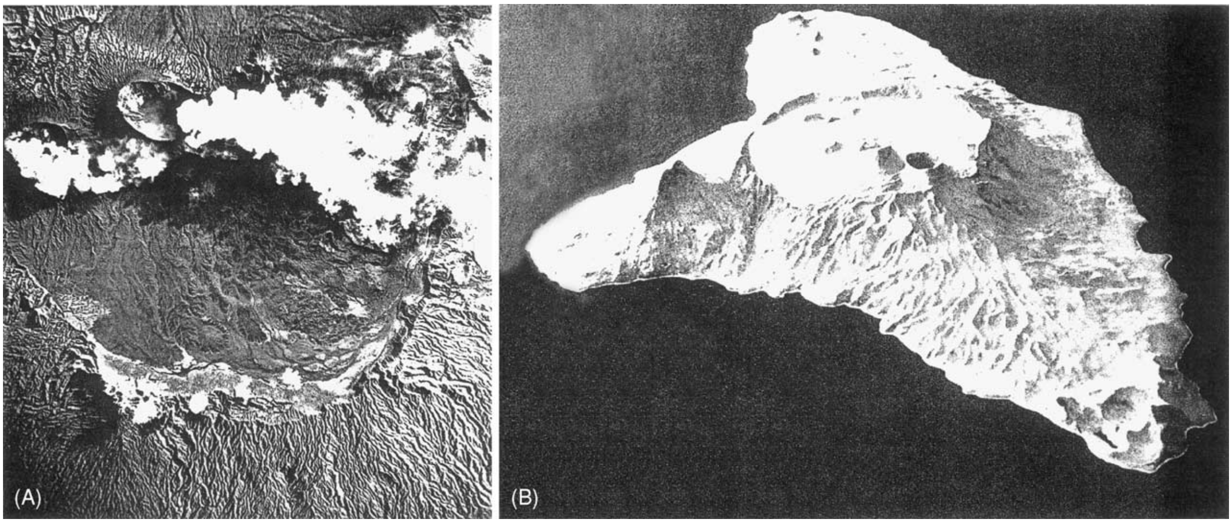


Figure 3 The 13 km diameter basaltic Ambrym volcano, Vanuatu. (A) Plaster model (by Jon Stephenson) (B) Air photo showing the caldera with two active craters nested within (Benbow and Marum).

Table 1 The seven types of volcano

Type	*Characteristics*
Icelandic	Fissure eruptions, releasing free-flowing basaltic magma; quiet, gas-poor; great volumes of lava flowing as sheets over large areas to build up plateaus
Hawaiian	Fissure, caldera, and pit-crater eruptions: mobile lavas with some gas; quiet to moderately active eruptions: occasional rapid emissions of gas charged lava produce fire-fountains; mainly basaltic; minor amounts of ash; builds up lava domes
Strombolian	Stratocones (composite) with summit craters: moderate, rhythmic to nearly continuous explosions, resulting from spasmodic gas escape; clots of lava ejected producing bombs and scoria; periodic more intense activity with lava outpourings; light coloured clouds, mostly steam reach only to moderate heights
Vulcanian	Stratocones: central vents; associated lavas more viscous; lavas crust over vent between eruptions, allowing gas buildup below surface; eruptions increase over long period of quiet until crust is broken up, clearing vent and ejecting bombs, pumice, and ash; lava flows from the top of flank after main explosive eruption; dark ash-laden clouds, convoluted and cauliflower-shaped, rise more or less vertically to moderate heights, depositing ash along the flanks of the volcano (note: other types, such as Hawaiin can produce similar effects when they suffer interference with groundwater, and phreatic eruption ensues, large steam clouds carrying fragmental material)
Vesuvian	More paroxysmal than the above two: extremely violent expulsion of gas-charged magma from stratocone vent; eruption occurs after long interval of quiescence or mild activity: vent tends to be emptied to considerable depth: lava ejects in explosive spray (glow above vent), with repeated clouds (cauliflower) that reach great heights and deposit ash
Plinian	More violent form of Vesuvian eruption (Figure 4): last major phase is uprush of gas that carries cloud vertically upward in a column for kilometres; narrow at base but expands outwards at upper elevations; cloud generally low in ash
Peléan	Results from high viscosity lava and delayed explosiveness, conduit of stratovolcano being usually blocked by dome or plug (Figure 5); gas (+ some lava) escaped from lateral (flank) openings or by destruction/uplift of plug; gas, ash, and blocks move with high velocity downslope in one or more blasts as nuees ardentes or glowing avalanches, producing directed deposits

After Guest *et al.* (2003).

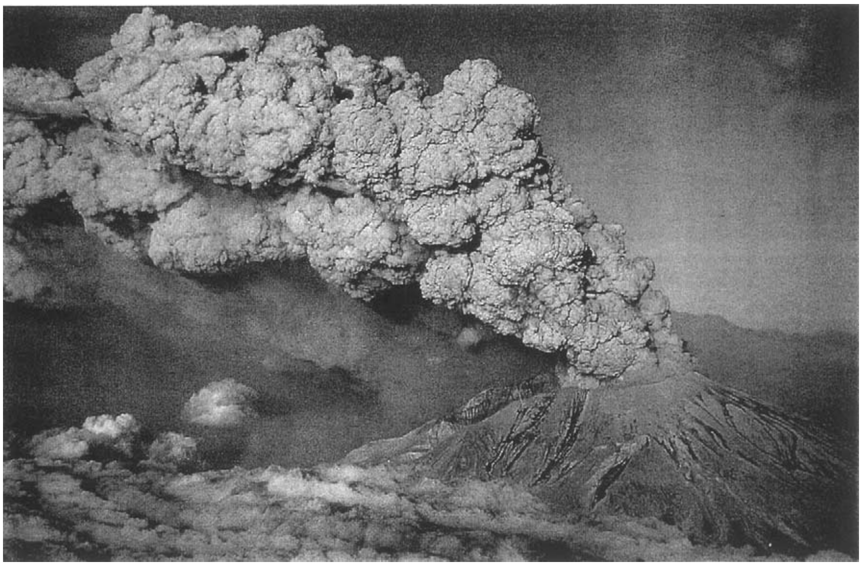


Figure 4 Plinian eruption of Mt St Helens, Oregon: ash, gas, and pulverized rock shooting out directionally (second eruption, July 1980) (from Pyle, 1998).

may terminate close to the vent (Figure 6). Even more siliceous lavas, such as rhyolite (equivalent in composition to granite), are sticky and form heaps close to the vent, or block the vent as plugs which may slowly extrude as spines. Lavas of this type are very dangerous as the plug finally breaks and there is a sudden explosive outburst which cannot be predicted. Both basalts and rhyolites tend to break up as they crust over, yielding slaggy flows. The solidified crust may be arrested and the molten lava flow underneath

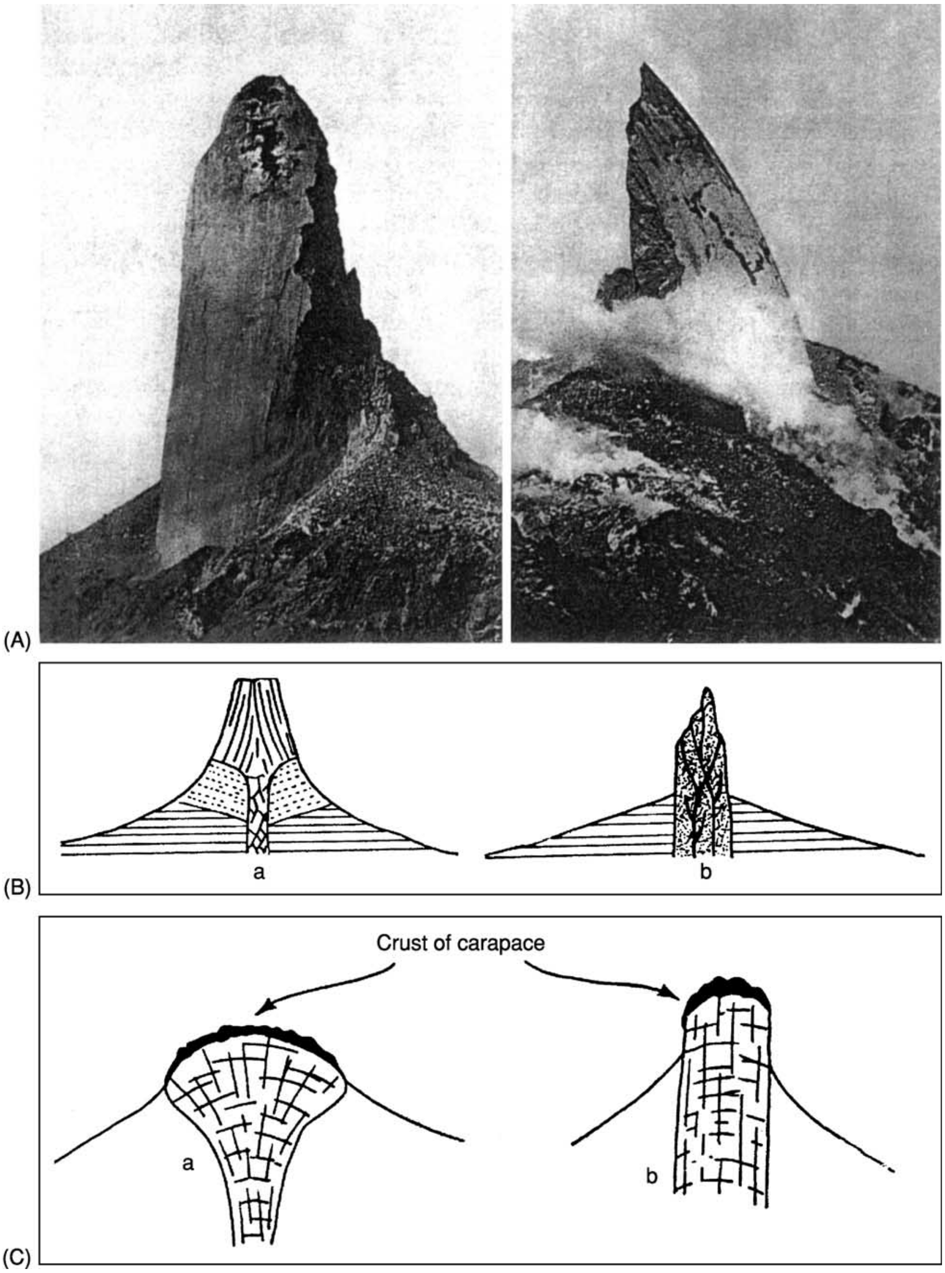


Figure 5 Mt Pelée, Martinique, (A) Two views of the rhyolite spine that plugged the vent prior to the catastrophe in 1902. (B) Types of lava domes: with spreading and no spreading. (C) Volcanic necks left behind by the above (from Green and Short (1971)).

it in a lava tube. There are four types of flow distinguished: pahoehoe (ropey), a'a' (slaggy); block lava; and pillow lavas (Figure 7). The latter form when the lava flows into water: sack-like glass covered bodies form, with concentric zones of vesicles or varioles (gas concentrations): they fissure radially and tend to break up easily, forming angular hyaloclastite breccias, containing pillow fragments. Fine fragmental glassy material formed in water or by deposition on wet surfaces is called peperite.

Some volcanoes erupt quite unusual peralkaline lavas; the Miocene 60 km diameter Kisingiri volcano in the Kavirondo Rift Valley, Kenya, erupted nephelinite and melilitite lavas; the magma chamber rocks below are nephelinitic ijolites and melilitic uncomprahgrites and turjaites, and the throat of the eroded volcano is a huge plug of carbonatite. Carbonatite has, in the past, been erupted as lavas from volcanoes elsewhere, but the only active emission known is of thin fluid flows of natrocarbonatite in

Table 2 Volcanic explosivity index

	Non-explosive	*Small*	*Moderate*	*Moderate large*	*Large*	*Very large*			
General description	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Volume of tephra (m^3)	10^4	10^6	10^7	10^8	10^9	10^{10}	10^{11}	10^{12}	
Cloud column height (km)*	<0.1	0.1–1	1–5	3–15	10–25	25			
Qualitative description	Gentle, effusive ←------	------------	Explosive------	------→	←------------	Cataclysmic, paroxysmal, colossal ------	------	------	------→
			←------------	------------	------------	Severe, violent, terrific ------	------	------	------→
Classification	←------------	Strombolian ------	------→ ←------	------------	------------	Plinian ------------	------→		
	Hawaiian←------	------------	------------	Vulcanian------	------→ ←------	Ultra-Plinian ------	------	------	------→
Total historic eruptions	487	623	3176	733	119	19	5	2	0
1975–1985 Eruptions	70	124	125	49	7	1	0	0	0

*For VEI 0–2, data are km above crater; for VEI 3–8. Data are in km above sealevel.

From Newhall and Self (1992).

the crater of Lengai volcano, Tanzania. The volcanoes Nyamuragira and Nyiragongo in the Western Rift Valley near Lake Kivu erupt potassic lavas rich in leucite.

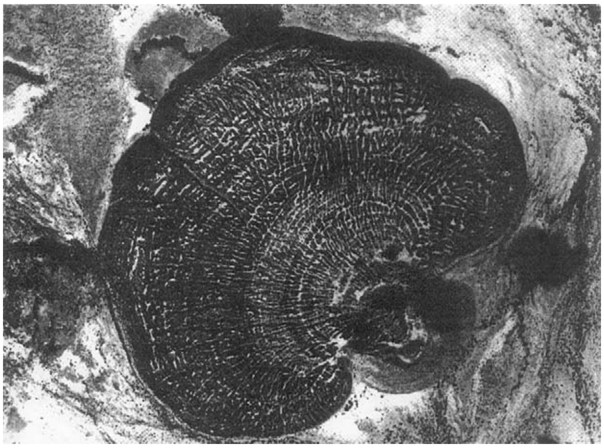


Figure 6 Trachyte flow extending only a few metres from the source vent in an ash cone (Silali volcano, Kenya) (from Green and Short (1971)).

Tephra

All clastic (fragmental) material issued from a volcano is covered by the term ‘tephra’. It is classified by size, as in Table 3:

Phreatic Eruptions

Interaction of rising magma and groundwater produces explosive eruptions, commonly with little material emission. Shallow depressions called ‘maars’ are formed this way and the extreme case is the crater with no material eruptive association except fragmented country rock (e.g., Hole-in-the-Ground, Oregon, Figure 9).

Volcanic Clouds

These may be of steam or ash rising to many kilometres and moving with the wind (Figure 10): the cloud from Stromboli is often carried as far as Greece. Ash clouds may be incandescent. An acid-bearing ash cloud from Klyuchevskaya, Kamchatka, rose to 20 000 m in 1994 and was carried 1000 km to the east by 240 km per hour winds, interfering with air travel.

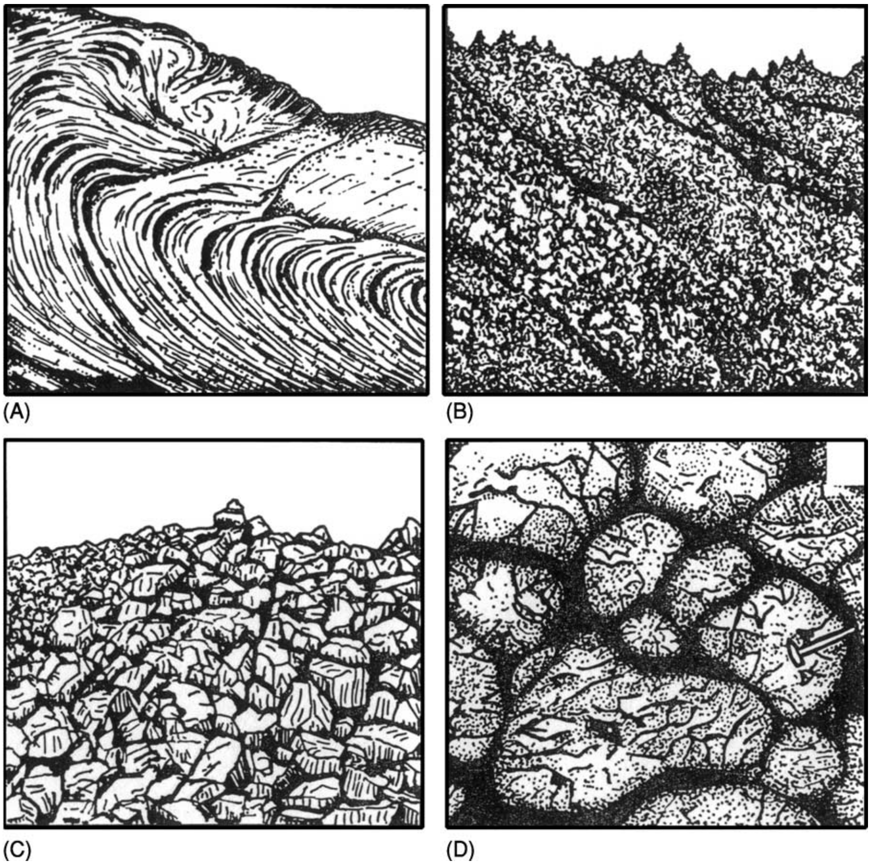


Figure 7 Four types of lava: (A) Pahoehoe; (B) A‘a’; (C) Block lava; (D) Pillow lava (from Green and Short (1971)).

Table 3 Tephrà classified by size

Name	*Size range*	*Remarks*
Dust	Less than 0.25 mm diameter	
Ash or sand	0.25–4 mm diameter	When lithified, the finer material is referred to as 'tuff' (Figure 8). The terms 'crystal' or 'lithic' are added according to the dominant component. Also the word 'pumice' may be added if pumice is dominant
Lapilli	0–32 mm diameter	
Bombs or blocks	More than 32 mm diameter	When lithified, these coarse deposits are referred to as 'agglomerates' or 'volcanic breccias'. The word 'scoria' is used for clasts with numerous open gas cavities, appearing like a sponge.
Bentonites		Bentonites are montmorillonitic clayey rocks formed by devitrification of volcanic glasses: palagonite is a yellow isotropic material formed by alteration of basaltic glass shards or hydration of them when hot

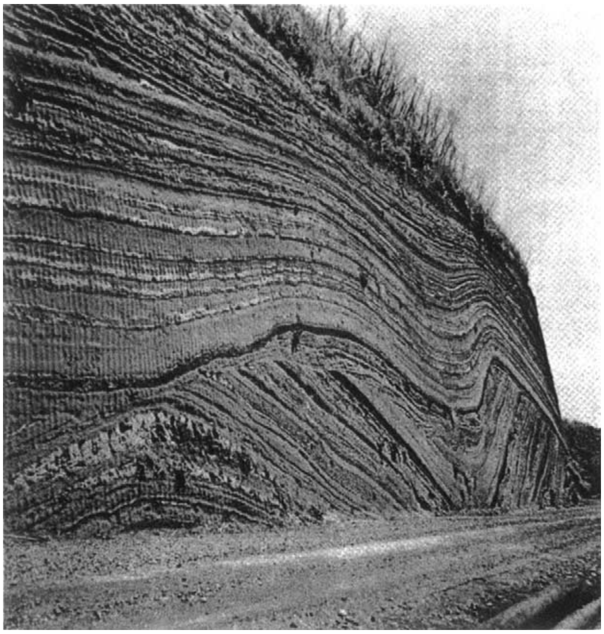


Figure 8 Bedded basaltic tuff of Pleistocene age, O Shima volcano, Japan (from Green and Short (1971)).

Nuées Ardentes

These are glowing avalanches of unsorted tephra or incandscent ash flows. They are characteristic of volcanoes erupting siliceous magma (rhyolite or andesite), and occur when plugs rupture or a sector of the volcanic cone collapses. Another cause is loss of support from below of a lava column. The eruption of Vesuvius in AD 79 killed many by asphyxiation (Figure 11).

Lahars

These are volcanic debris flows and may be hot or cold. They may arise from interaction of hot pyroclastic flows and surges with ice and snow covering the volcanic cone summit, as at Nevado del Ruiz, Colombia, in 1985 (Figure 12); in this case the duration of the horrific event was small, but in the case of Pinatubo, Philippines, in 1993, pyroclastic flows continued to be reworked by heavy rains, forming lahars which continued for years.

Gases

Volcanic gases are commonly dominate by CO_2, which asphyxiates: such gases are invisible and concentrate in depressions in the land, as in Iceland: these concentrations are called 'mofettes': The Lake Nyos disaster in Cameroon was due to a sudden outburst of these gases from under the lake.

Volcanoes as a Major Natural Hazard

Volcanoes are distributed along plate boundaries and along rifts and fractures within plates, and they are thus fixed features. The hazards they pose are easier to predict and restrict geographically than those of earthquakes, but they pose a great hazard–particularly because the volcanic rocks form rich soils for agriculture and thus populations concentrate around them. Not only active volcanoes pose hazards: in 1943, Paricutin, Mexico, originated as a fissure opened up in a ploughed field and built up to a 2400 m high cone, from which 'a'a' lavas flowed down and buried buildings. Mt Pelée in Martinique was supposed, in 1902, be extinct and a nuée ardentes killed all but one of its 29 000 inhabitants (a prisoner in a gaol).

The toll on human life is an order less than from earthquakes. Since 1980 there have, however, been more than 30 000 deaths and some 1 million people have been detrimentally affected by eruptions.

In Table 4, some selected volcanic disasters, since the eighteenth century, are listed. Fatalities are not a

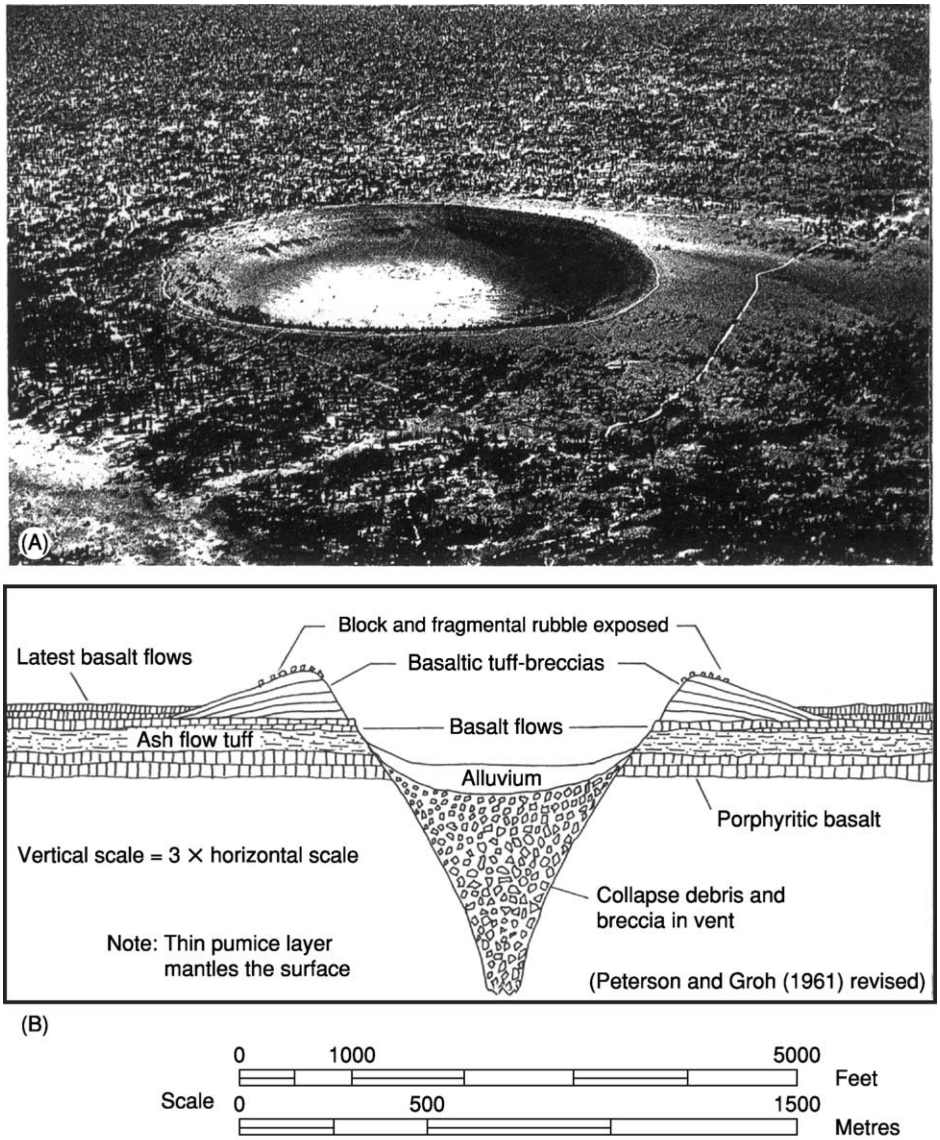

(A)
Block and fragmental rubble exposed
Latest basalt flows
Basaltic tuff-breccias
Basalt flows
Ash flow tuff
Alluvium
Porphyritic basalt
Vertical scale = 3 × horizontal scale
Collapse debris and breccia in vent
Note: Thin pumice layer mantles the surface
(Peterson and Groh (1961) revised)
(B)
Scale
0
1000
5000
Feet
0
500
1500
Metres

Figure 9 Hole-in-the-Ground, Oregon, a 1.6 km diameter Maar (from Green and Short (1971)).

satisfactory measure of disaster – for example, the 1997 collapse of the old crater wall surrounding the dome complex of Soufriere, Montserrat, cause only 19 fatalities, but the ongoing eruptions over several years caused the capital of the small island and more than half of the island's population to be evacuated indefinitely.

The principal volcanic hazards and effects are listed below:

Hazards

- Lava flows
- Ash falls
- Pyroclastic flows and surges
- Directed blasts and atmospheric shock waves
- Lahars and floods
- Landslides
- Volcanic gases
- Tsunamis
- Climate modification
- Crater lake emptying suddenly
- Ice/snow lava interaction

Effects

- Loss of land and buildings
- Disruption of social and economic infrastructure
- Famine
- Water pollution

- Disease
- Drowning
- Asphyxiation

Lava flows represent the principal hazard of basaltic volcanoes. On Etna, they stream down the flanks and threaten villages and coastal towns. The recent eruptions of Nyiragongo volcano above Goma in the Congo drove a lava flow through the middle of the town, destroying much property. They destroy buildings and usable land. Mitigation by diversion has been tried on Etna but is of limited success.

Pyroclastic flows may attain velocities of tens to hundreds of metres per second and temperatures of 300–800°C. They may generate secondary debris flows. Block and ash flows are slightly less deadly than pumice-rich ignimbrites. The worst of these flows and blasts destroy and kill everything in their path.

Debris avalanches, as at Mt St Helens in 1980 (Figure 4), travel rapidly and are equally lethal and destructive.

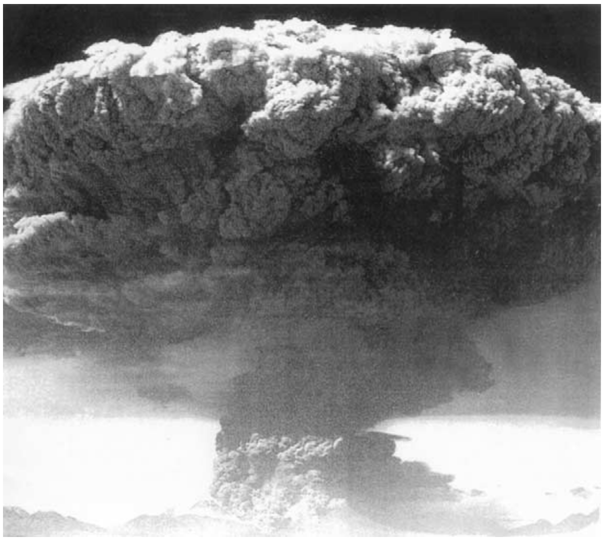


Figure 10 Ash cloud from a devastating eruption of Lascar volcano, Chile (from Pyle (1998)).

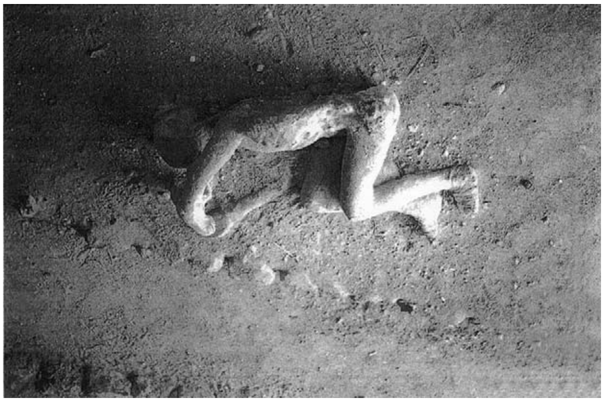


Figure 11 Cast of a child's body, asphyxiated by pyroclastic surge of Vesuvius, 79 AD (from Guest *et al.* (2003)).

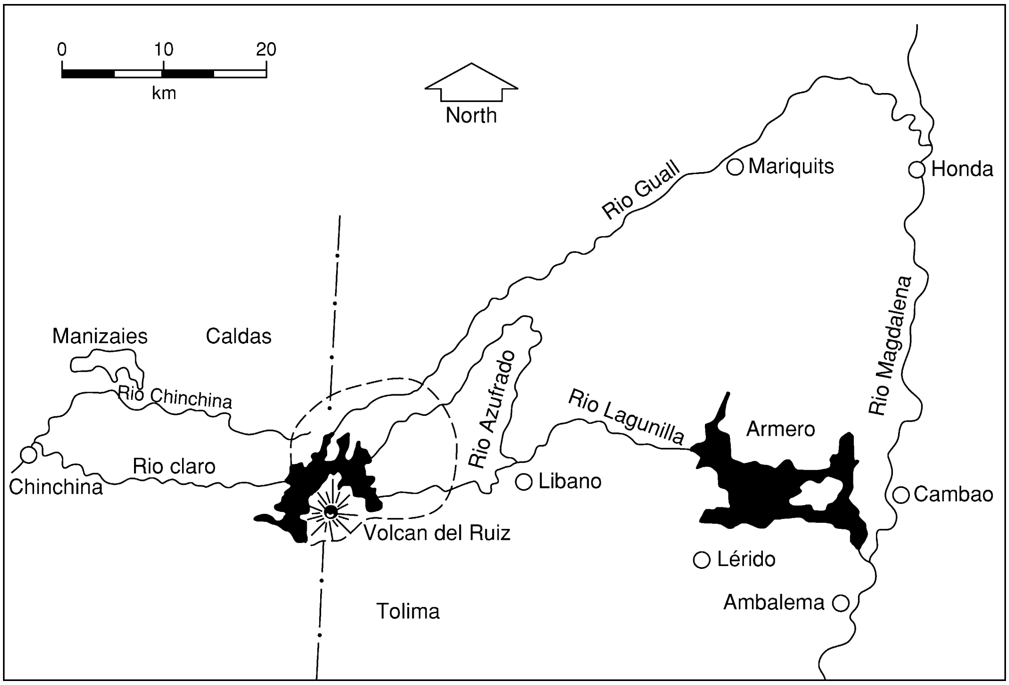


Figure 12 The pathway of the lahar from Nevado del Ruiz volcano, Colombia, to destroy Armero (from Hall, Geohazards Natural and Man Made, Chapman and Hall (1992)).

Table 4 Selected Volcanic disasters since the eighteenth century

Volcano (country)	*Year*	*Fatalities*	*Cause*
Laki, Iceland	1783	10 500	Famine
Unzen, Japan	1792	15 188	Tsunami
Tambora, Indonesia	1815	92 000	Mainly famine
Krakatau, Indonesia	1883	36 417	Tsunami
Mt Pelée, Martinique	1902	29 000	Pyroclastic flows
Kelut, Indonesia	1919	5110	Debris flows
Lamington, Papua-New Guinea	1951	2940	Pyroclastic flows
El Chichon, Mexico	1982	1700	Pyroclastic flows
Nevado del Ruiz, Colombia	1985	25 000	Debris flows
Lake Nyos, Cameroon	1986	1746	Volcanic gases (CO_2) – asphyxiation*
Pinatubo, Philippines	1991	500	Various
Soufriere, Montserrat	1997	19	Pyroclastic flows
Goma, Congo	2002	40	Lava flows

*This was not associated with a volcano: the gases came from a lake but were of volcanic origin.
After WJ Maguire.

Table 5 Volumes and non-out distances for volcanic debris

Volcano	*Volume (km^3)*	*Run out (km)*
Nevado di Colima	22–33	120
Socompa	17	35
Volcan di Colima	6–12	43
Shasta	26	50
Popocatapetl	28	33
Chimborazo	8.1	35
Mawenzi	7.1	60
Galunggung	2.9	25
Mt St Helens	2.5	24
Fuji	1.8	24
Shiveluch (1964)	1.5	12
Bandai-San (1888)	1.5	11
Egmont	0.35	31
Unzen (1792)	0.34	6.5
Asakusa	0.04	6.5

After WJ Maguire.

Lahars can be deadly, as at Nevado del Ruiz in 1985, where a flash flood was generated and spread out 40 km to destroy Armero completely (Figure 12).

A tsunami generated by eruption in the Kurile Is in 1737 swept up a fjord to a height of 65 m.

Tables 5 and 6 list volumes and run-out distances for volcanic debris avalanches and principal mitigating measures for these hazards. An interesting fact, as described by P Delos Reyes, is that animals give indications of forthcoming eruption by their unusual actions.

In Figure 13A and B, the distribution of hazards around a volcano is illustrated and the relative dispersal – the most dispersive fine ash carrying sulphuric acid droplets may extend over the entire circumference of the planet, high in the stratosphere and cause spectacular sunsets for years. In Figure 14, mitigation measures are illustrated.

Volcanoes listed for special study under UN IDNDR are named below (from WJ Maguire):

Decade volcanoes (UN sponsored)

- Colima (Mexico)
- Galeras (Colombia)
- Mauna Loa (USA)
- Merapi (Indonesia)
- Mount Rainier (USA)
- Nyirogongo (Zaire)
- Sakurajima (Japan)
- Santa Maria (Guatemala)
- Ta'al (Philippines)
- Ulawan (Papua New Guinea)
- Unzen (Japan)
- Vesuvius (Italy)

Laboratory volcanoes (EU sponsored)

- Etna (Sicily)
- Furnas (Azores)
- Krafla (Iceland)
- Piton de la Fournaise (Reunion Is)
- Teide (Tenerife)
- Santorini (Greece)

It is noteworthy that none are in Kamchatka, Siberia, often cited as the world's most active volcanic region: 20 volcanoes erupt there, 3–5 times per year.

Volcanoes and Earthquakes

Volcanic earthquakes provide a valuable warning of impending eruption, clusters of small seisms being detected on seismograph arrays. Major tectonic earthquakes are quite distinct from these seisms, but major earthquakes and eruptions can be interrelated:

Table 6 Principal mitigating measures for volcanic hazards

Hazard	*Principal mitigating measures*
Lava flows	Damming or diversion: flow front water cooling
Debris flows (lahars) and floods	Judicious siting of settlements; construction of elevated refuges, sediment dams, and baffles; dredging and levee construction; seismometer and trip-wire warning systems
Pyroclastic flows and surges	Judicious siting of settlements; pre-evacuation
Tephra	Evacuation of poorly constructed buildings; accumulation of accumulated ash, etc. from roofs; availability of face masks/protective headgear; appropriate medical care for respiratory problems and ingestion of glass microshards; contingency plans for power cut off, communications and transport disruption; availability of uncontaminated water supplies; measures to minimise crop and livestock damage; warnings to air traffic
Landslides and debris avalanches	Identification of collapse-prone areas; slope stability monitoring; pre-evacuation
Directed blasts and shock waves	Pre-evacuation
Volcanic gases	Gas monitoring; resettlement if a permanent problem: pre-evacuation if episodic and predictable: public safety guidelines and warning notices; construction of elevated refuges where appropriate
Tsunami ('tidal wave')	Identification of unstable slopes adjacent to water bodies; slope stability monitoring; pre-evacuation; establishment of a tsunami warning network both regionally and internationally

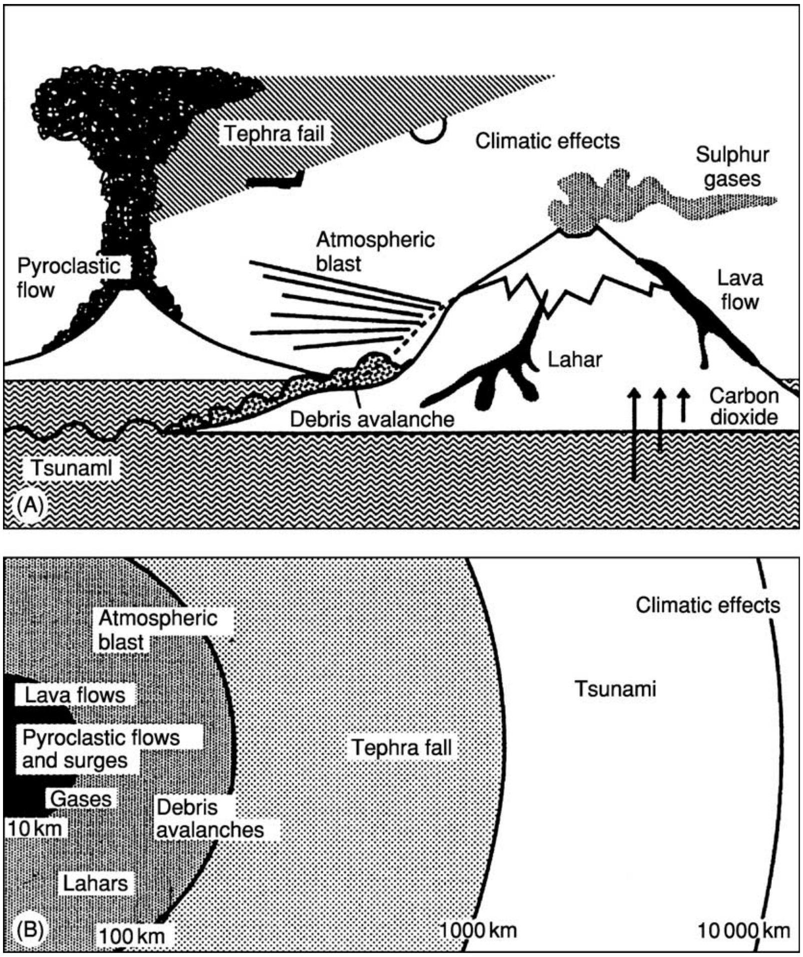


Figure 13 The destructive processes generated by volcanoes: (A) showing those confined to the immediate vicinity. (B) showing those dispersing further, even globally (from Maguire (1998)).

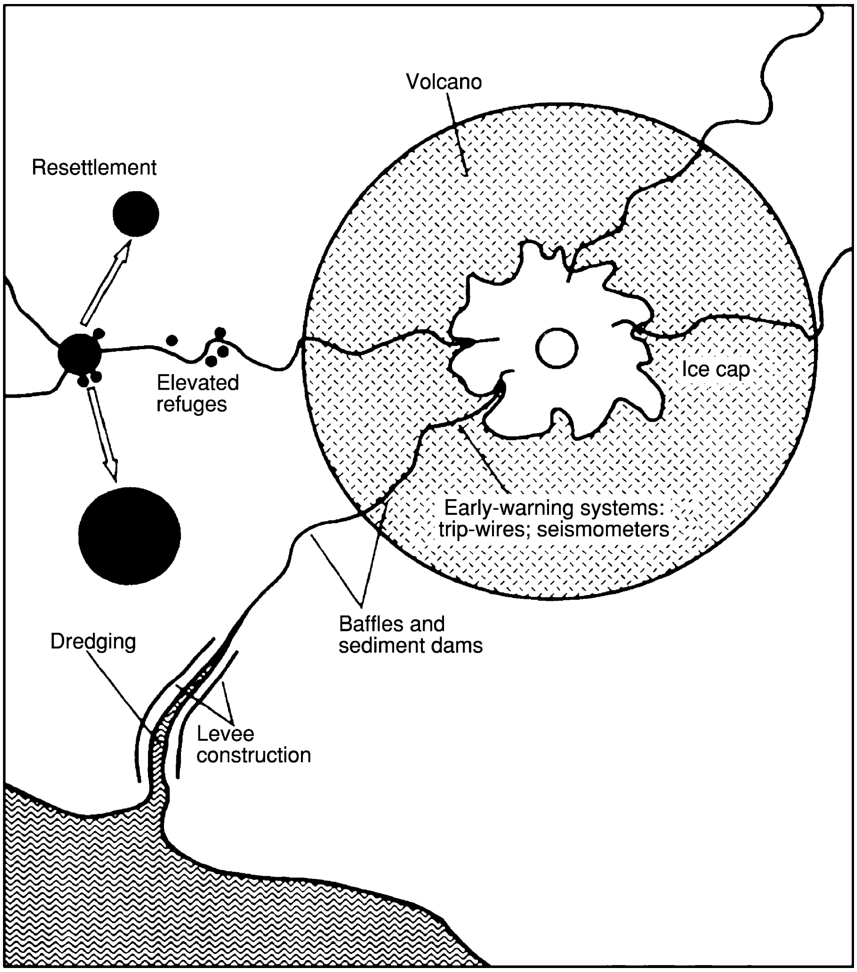

Volcano
Resettlement
Elevated refuges
Ice cap
Early-warning systems: trip-wires; seismometers
Baffles and sediment dams
Dredging
Levee construction

Figure 14 Mitigation measures possible in the case of lahars (from Maguire (1988)).

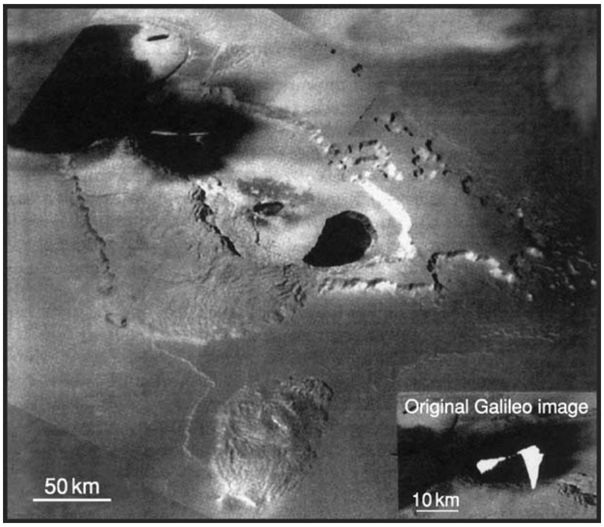


Figure 15 Tvashtar Catena, Io: a 160 km long caldera containing two nested smaller lava filled pits and an adventive volcano at the far end, associated with huge lava fountains and black lava effusion down the flanks of the host volcano. There is also a volcano of more viscous lava with a small summit crater in the foreground. Lighter areas are due to sulphur (NASA Galileo 1999 image).

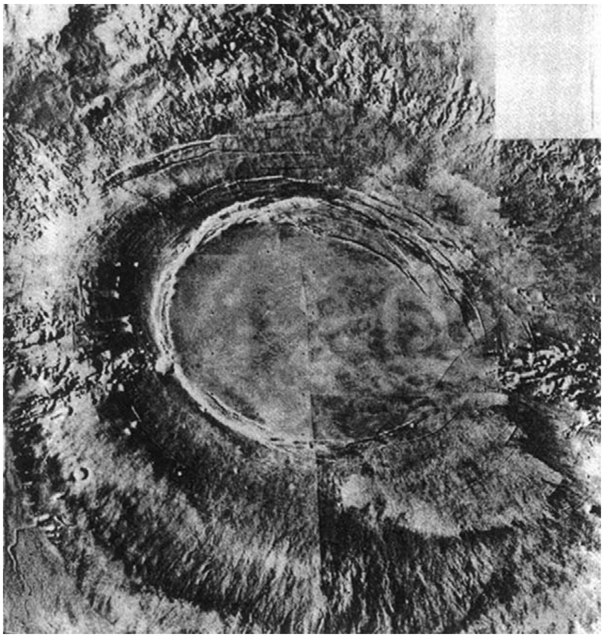


Figure 16 Arsia Mons volcano, Mars, with a 99 km wide caldera (NASA Viking Orbiter 1 image).

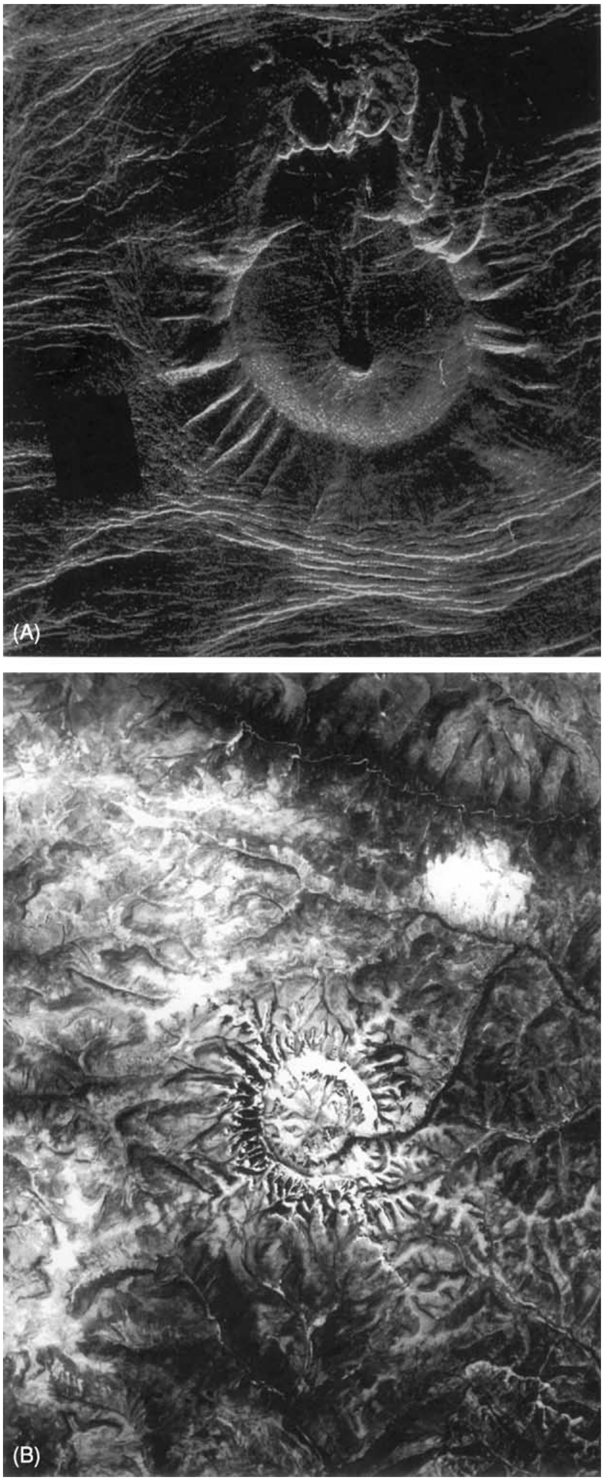


Figure 17 (A) False colour radar image of a volcanic crater on Venus (35 km diameter at summit). Note lava flow which has passed through a breach in the wall at top; also the eroded (?) lateral slope ridges and gullies which suggest Earth-like processes operating (NASA image). (B) The Gora Konder crater, diameter 22 km, Siberia: in remote terrain, the origin, volcanic or impact is unknown, illustrating the difficulty of differentiating impact and volcanism on physiography alone: the similarity of the flank ridges and gullies to a) is striking (NASA, Shuttle Discovery image).

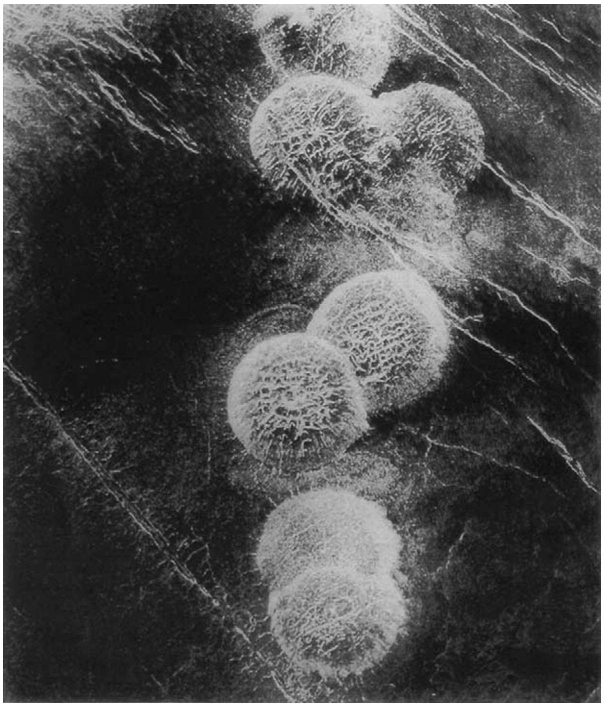


Figure 18 False colour radar image of seven eruptive domes, each 25 km in diameter, overlying one another, on Venus: possibly domes of viscous silicic lava? (NASA image).

a major earthquake in Chile in 1960 is recorded as activating several Andean volcanoes and a major earthquake in Sicily is recorded as having been preceded three days earlier by an eruption of Etna.

Extra-terrestrial volcanoes

Active volcanoes are known on Jupiter's satellite Io (Figure 15), and huge extinct volcanoes occur on Mars and Venus (Figures 16, 17 and 18). Venus's thick CO_2 rich atmosphere is believed to be volcanically sourced.

See Also

Earth: Mantle; Crust. **Engineering Geology:** Natural and Anthropogenic Geohazards. **Lava**. **Mantle Plumes and Hot Spots**. **Solar System:** Venus; Mars; Jupiter, Saturn and Their Moons.

Further Reading

Davies A and Bowler S (2001) Extra-terrestrial active volcanism. *Geoscientist* 11(8): 4–7.

Green J and Short NM (1971) *Volcanic Landforms and Surface Features – A Photographic Atlas and Glossary*. New York, Heidelberg, Berlin: Springer.

Guest JE, Cole PD, Duncan AM, and Chester DK (2003) *Volcanoes in Southern Italy*. London: Geological Society, Earth in View Series.

Maguire WJ (1998) Volcanic hazards and their mitigation. In: Maund JG and Eddleston M (eds.) *Geohazards in Engineering Geology*, pp. 79–95. London: Geological Society, Engineering Geology Special Publication 15.

McCall GJH (1956) Geology of the Gwasi Area. *Report No. 45, Geological Survey of Kenya*.

McCall GJH and Bristow CM (1965) An introductory account of Suswa volcano. *Bulletin Volcanologique* 28: 1–35.

McCall GJH, Laming DJC, and Scott SC (1992) *Geohazards: Natural and Man-made*. London: Chapman and Hall.

McCall GJH, Le Maitre RW, Malahoff A, Robinson GP, and Stephenson PJ (1970) The Geology and Geophysics of the Ambrym Caldera, New Hebrides. *Bulletin Volcanologique* XXXIV-3: 681–696.

Pyle D (1998) *Volcanoes*. London: Oceania.

Scarth A (1997) *Savage Earth*. London: HarperCollins.

Smith M, Dunkley PN, Deino A, Williams LAJ, and McCall GJH (1995) Geochronology, stratigraphy and structural geology of Silali volcano. *Journal of the Geological Society* 152: 293–310.

WEATHERING

W B Whalley and P A Warke, Queen's University Belfast, Belfast, UK

Introduction

Weathering can be defined as the irreversible structural and/or mineralogical breakdown of rock through the cumulative effects of chemical, physical, and biological processes operating at or near the Earth's surface (Table 1). However, this seemingly straightforward definition masks the complexity of rock weathering, in which interactions between the many different components of the weathering system give rise to an element of unpredictability that is characteristic of non-linear systems. The weathering behaviour of rock is a response to subaerial (Earth surface) conditions, where temperatures and pressures differ from those under which the minerals were formed. Consequently, adjustment to surface environments is manifest through rock breakdown, the rate of which is controlled by many factors: characteristics of the rock itself, the availability of weathering agents such as salt and moisture, biological agents such as lichens, and, especially, the microclimatic environment to which the rock is exposed.

Without weathering and, in particular, the breakdown of one mineral type into another, there would be no soils of any significance and little scope for widespread development of flora and fauna on land. Thus, long-term weathering is of paramount importance to the biosphere and is a crucial element of both long-term and short-term landscape development.

Nonequilibrium Conditions and the Lithological Cycle – General Significance

The lithological cycle provides a useful starting point when considering the role of weathering in landscape development. Erosion is generally preceded by a combination of weathering processes, which are usually crucial in the formation of silt, clays, and resultant solutes and in the release of residual components of crustal materials.

As rock approaches the Earth's surface, either through tectonic uplift or through erosion of the overburden, associated changes in pressure and/or temperature mean that it is no longer in a state of equilibrium. In this context, 'surface' should be taken to include the range of locations where the hydrosphere, atmosphere, and biosphere interact with the lithosphere; the maximum extension of these interactions is *ca.* 100 m (e.g. where there is tropical deep weathering), although it is normally much less than this. Differences in pressure and temperature at or near the Earth's surface give rise to important (intrinsic) aspects of rock breakdown that can create or reinforce positive-feedback conditions in weathering systems.

- Thermodynamically and chemically different conditions at the Earth's surface can destabilize minerals, thus increasing their susceptibility to subaerial weathering processes.
- Changing chemical and thermodynamic conditions may be accompanied by volume changes resulting from decreased overburden pressures, which lead to differential stresses that are realized as discontinuities at various scales, from joints and cracks to microcracks. Volume changes can also occur when one mineral is altered to another.

When the rock arrives at the Earth's surface, the interplay of hydrosphere, atmosphere, and biosphere provides more complexity, although, for the most part, we can reduce this to a small number of extrinsic factors, namely water (in all three phases), temperature (usually between −40°C and +40°C at or near the Earth's surface), and biotic activity (which depends on water and temperature). The interplay of these factors has had important consequences for long-term global climate change as well as landscape development, as discussed below.

Joints, Cracks, and Microcracks

One major result of the uplift of rocks to the Earth's surface is that, as pressure decreases, there is a volumetric expansion. The most significant way in which this manifests is through the creation of crack systems, from joints to microcracks. The intersection of joints, which can be many metres in length and depth, can substantially weaken a mass of rock. This gives rise to the concept of 'rock-mass strength' (as opposed to the 'intact strength' of small blocks). Therefore the 'strength' of the rock on a face in a quarry (rock mass) will differ from the crushing strength of the aggregate (intact strength). At a smaller scale, microcracks are plentiful in many rocks; they are usually a few micrometres wide and perhaps a few centimetres long.

Table 1 Examples of weathering processes and mechanisms

Weathering process	*Weathering mechanisms* (*Main mechanisms outlined, not a full discussion of hypotheses*)
Sometimes referred to as physical or mechanical weathering	
Salt weathering (haloclastis)	**Salt crystallization** In pores and microfractures salt crystallization can result in the creation of expansive stresses in excess of the tensile strength of the rock. Repeated exposure to the stress effects of salt crystallization can result in the disruption of intergranular bonds and a reduction in structural coherence
	Crystalline phase change Changes from dehydrated to hydrated states through the absorption of atmospheric humidity result in volumetric expansion of salt crystals in pores and microfractures. Typically, take-up of moisture by an anhydrous salt forms a crystallographically different mineral
	Thermal expansion and contraction Interstitial salt crystals exhibit coefficients of thermal expansion that are often greater than those of the rock minerals that surround them
	Mobilization of silica Under highly alkaline conditions disruption of aluminosilicate minerals can occur together with the dissolution of quartz and silica cement
Frost weathering (macrogelivation)	**Freeze–thaw** Repeated freezing results in volumetric expansion of water in pore spaces and fractures. This can also enhance the disruptive hydration effects of swelling clay minerals
	Hydrofracture Moisture freezes in a microfracture sealing off the surface end, and any unfrozen water trapped in the substrate may be forced under pressure, through volumetric expansion of the ice, towards the tip of the microfracture and thus extend it
Thermal weathering (insolation weathering, thermoclastics)	**Thermal fatigue** The effects of insolation weathering arise from differential volumetric expansion of individual mineral grains and/or surface and near-surface rock layers in response to repeated (diurnal) heating and cooling
	Thermal shock is a rapid increase in rock surface temperature (typically associated with bush fires). Thermal gradients develop quickly, giving rise to tensile stresses between expanding heated surface rock layers and cooler substrate material
Chemical weathering	**Solution of minerals** occurs as a result of exposure to water, its effectiveness is influenced by contact time, the pH of the water, and the solubility characteristics of the elements of the individual minerals
	Carbonation is the reaction of minerals with 'carbonic acid' (CO_2 dissolved in water) and is a particularly important reaction in limestone weathering
	Hydrolysis is the chemical reaction between hydrogen ions in water and the ions in any mineral structure
	Oxidation is when electrons are lost from an atom and usually describes the reaction with oxygen to form oxides, e.g. the conversion of ferrous to ferric iron (such as the weathering of olivine) with an associated volume increase
	Reduction is when an atom gains an electron, usually under anaerobic conditions (typically, gleys in soils)
	Hydration is an exothermic reaction involving the addition of water to a mineral. It is an important weathering characteristic of many clay minerals, which undergo considerable volumetric expansion when water is incorporated into their crystal lattices, e.g. swelling clay minerals such as bentonite from weathered volcanic ash
Biological weathering	**Chemical dissolution of minerals** through the action of organic and inorganic acids
	Chelation is the removal of metallic ions by chelating agents of organic origin
	Plucking is the dislodgement of rock or mineral fragments through the contraction of lichen thalli and fungal hyphae on drying
	Boring into rock by biota (e.g. snails, sea urchins)
	Fracturing by root penetration and exploitation of joints and cracks.

This table gives only the main mechanisms considered to be involved in rock weathering.

The surface area of rock components (say intact rock blocks) is the area upon which physical and chemical weathering processes act. This surface area is extended substantially by the presence of microcracks, joints, and cracks because, as long as water can reach these surfaces, weathering reactions can act upon them. Such physical and chemical action can further extend microcrack and joint systems by increasing the tensile stresses at crack tips. The interpenetration of discontinuities at any of these scales aids weathering in general.

Minerals and Rates of Weathering

Minerals in rocks can be divided into those that are relatively resistant to weathering (such as quartz), those that are rather vulnerable to weathering (such as olivine), and those that have been produced by weathering. Figure 1 shows the Goldich weathering series (which is the inverse of the Bowen reaction series, in that the minerals that crystallize at the lowest temperatures in the melt are also the most stable in the conditions found at the Earth's surface).

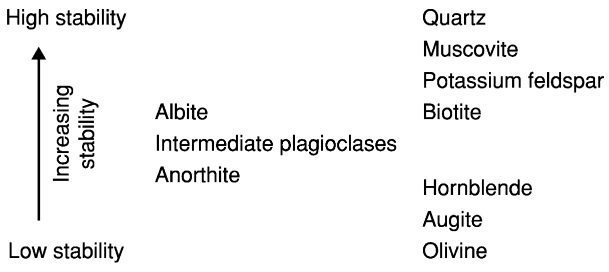


Figure 1 Weathering and the stability of some rock-forming minerals. The continuous plagioclase series is given in a separate column.

With the exception of thermoclastis, all Earth-surface weathering (physical, chemical, and biological) involves water, which enables chemical reactions, facilitates the mobilization and ingress of salt into rock fabrics, and, in its solid phase (ice), contributes directly to rock breakdown. Chemical weathering is particularly complex in that the efficacy of reactions depends on moisture availability, temperature, and the nature of the mineral assemblage. Mineral reactions are therefore controlled by the mineralogical nature of the reactants and the acidity (pH) and reduction–oxidation potential (Eh or redox potential) of the water. The dissociation of water into H^+ (protons or hydrogen ions) and OH^- (hydroxyl radicals) is important in weathering reactions, especially in hydrolysis, where one or other of these ions replaces ions in a mineral structure.

Weathering products are of three main types: layer silicates in the clay minerals, silica in solution (usually given the formula H_4SiO_4 – 'silicic acid'), and alkali-metal ions in solution.

The example of orthoclase weathering to muscovite can be expressed as:

$$3KAlSi_3O_8 + 12H_2O + 2H^+ \rightarrow KAl_3Si_3O_{10}(OH)_2 + 6H_4SiO_4 + 2K^+.$$

Muscovite can then degrade into kaolinite:

$$2KAl_3Si_3O_{10}(OH)_2 + 3H_2O + 2H^+ \rightarrow 3Al_2Si_2O_5(OH)_4 + 2K^+$$

(It should be noted that there are several ways of expressing these reactions and that these two examples are rather generalized because of the complex nature of feldspars and clay minerals.)

The significance of water-soluble products, which may be removed by agents such as groundwater, is clearly seen as the solute loads of rivers. Additionally, the relict nature of weathering products not removed in solution can be seen in saprolites and duricrusts, which may be legacies of previous weathering stages.

In terms of rock weathering, one of the most dramatic examples of the role of chemical weathering in landscape development is provided by 'karst' landforms, where the reactions between rainwater, atmospheric gases, and calcareous rocks (typically limestones) can over time create distinctive and dramatic surface features (Figure 2). Rainwater dissolves atmospheric CO_2 to give bicarbonate ions, HCO_3^-, in solution as a weak acid, carbonic acid, as a result of a reversible reaction:

$$H_2O(l) + CO_2(g)\{ReversReact\}H^+(aq) + HCO_3^-(aq)$$

Calcium carbonate is an important constituent of chalks and limestones and is soluble in carbonic acid:

$$CaCO_3 + H_2CO_3\{ReversReact\}Ca^{++} + 2HCO_3^-$$

These reactions also create subsurface weathering structures, such as cave systems, and features such as stalagmites, which are formed when the reaction is reversed and calcite is deposited.

The rate laws of chemical reactions show how rates depend on the concentrations of the reactants. Although this is significant in reactions that take place over very short periods of time (usually less than a few hours), it can be ignored for our purposes. More significant is the application of Arrehenius' equation to chemical reactions. In particular, increases in temperature speed up reactions. A general rule is that an increase in temperature of 10 K doubles the 'weathering rate'. Clearly, there are direct links with the extrinsic factor of climatic temperature. Assuming a constant (mean annual) temperature at a location, we might reasonably expect, at least over relatively short periods of time such as a few thousand years, the rate of weathering to be constant. When a distinctive weathering product is produced, this can be used as a marker or even as a somewhat primitive chronometer. Rock varnish is a good example of the latter. Weathering horizons in soils buried by subsequent deposition (palaeosols) also provide distinctive marker horizons.

Weathering in Surface Landscape Interpretation

Weathering studies have tended to be somewhat overshadowed by other geomorphological disciplines, and yet weathering is extremely significant because it is the principal means of sediment production for erosion, transport, and deposition by aeolian and fluvial processes. There has been a tendency to investigate more visually prominent weathering features, such as tafoni (shallow rounded cavities in rocks produced by weathering), but it is now widely accepted that such features are exceptional and not truly

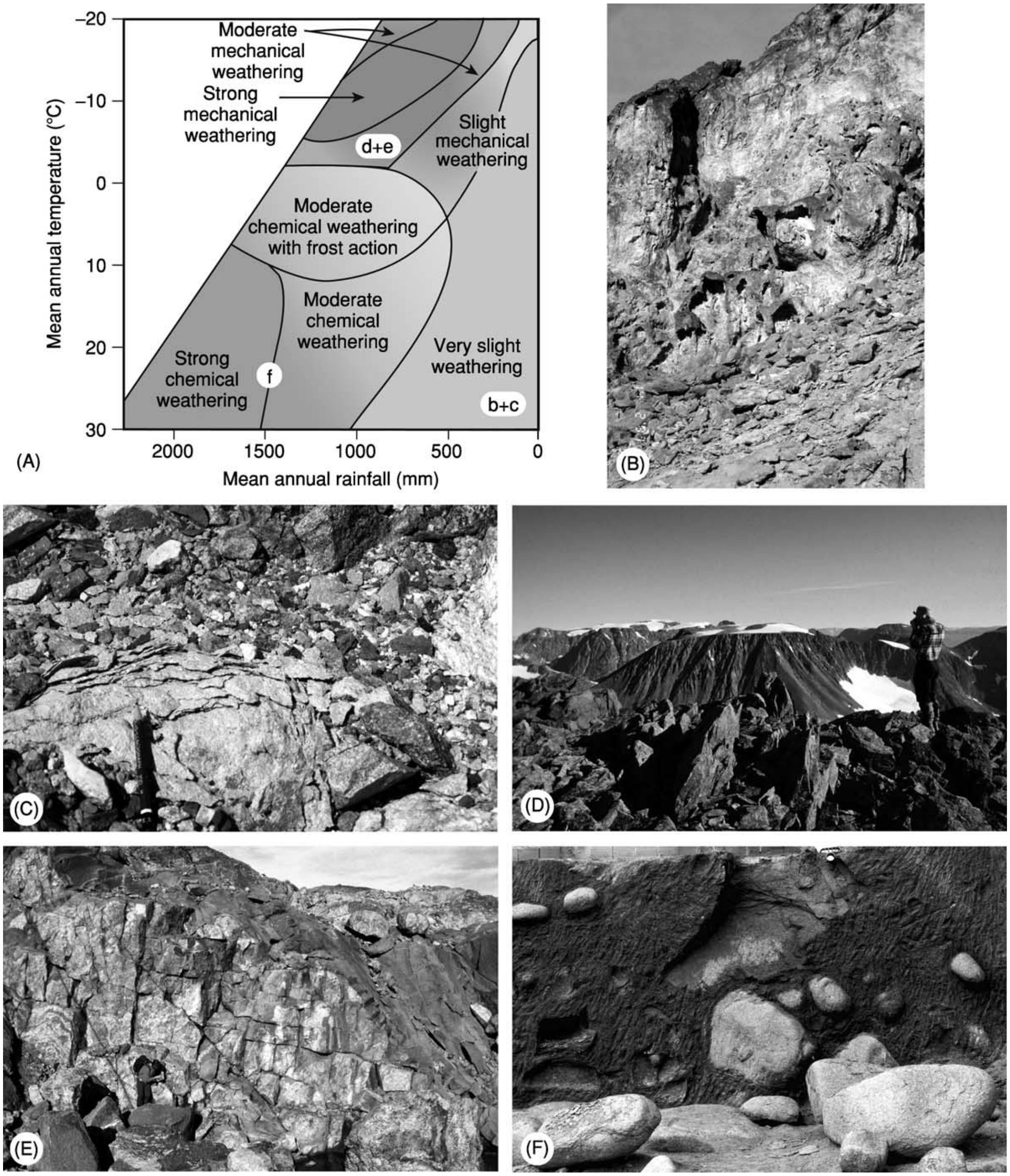


Figure 2 (A) Peltier's zonal classification of weathering, based on average annual temperature and moisture availability. Images shown in parts (B)–(F) are marked on the corresponding weathering zones. (B) Present-day weathering of a sandstone outcrop in Death Valley, California (a hyperarid hot desert); breaching of the casehardened outer layers of stone and subsequent development of basal tafoni is contributing to undercutting and periodic collapse and retreat of the rock face. Present-day weathering processes comprise a complex assemblage of physical, chemical, and biological mechanisms, which are of both spatially and temporally variable effectiveness (cliff height is approximately 4 m). (C) Also in Death Valley, the active disintegration of clasts occurs under present-day conditions primarily but not solely because of the effect of groundwater that is rich in a complex assemblage of salts, which facilitates salt-weathering mechanisms and the preferential chemical weathering of silicates owing to the strongly alkaline conditions. Owing to the often prolonged absence of surface water, debris disintegrates *in situ*, leading to a local accumulation of sediment, which is periodically transported away from the site during flash-flood events. (D) High-latitude regions, such as this site in the Lyngen Alps, northern Norway (70°N), tend to be associated with physical weathering by frost action because of the present-day climatic conditions, which are dominated by low average annual temperatures. However, this inference is not wholly accurate as the gabbroic rock of the region has been subject to long-term chemical weathering (perhaps extending as far back as the Tertiary), leading to an accumulation of clay minerals in a weathering mantle on the plateaux that can be up to 2 m deep. Thus, the present-day environment belies the long-term activity and changing environmental conditions, especially temperature. This inheritance is also

representative of most landscapes, which are dominated by extensive debris mantles (saprolites) and where weathering phenomena tend to be more mundane. Increasing understanding of the factors controlling weathering phenomena in a wide variety of climatic regions, in the latter half of the twentieth century, has highlighted several major conceptual issues that underpin much of contemporary rock-weathering research – issues that have significant implications for the understanding of landscape development. These issues include: climate, rock weathering, and classification; magnitude and frequency of rock weathering; feedback mechanisms; equifinality (form convergence) – the problem of linking process and form; and inheritance effects.

Climate, Rock Weathering, and Classification

The traditional classification of weathering tends to be climatically zonal, reflecting temperature characteristics and precipitation, with the implication that maximum weathering occurs in the humid tropics and minimum weathering occurs in hot and cold desert regions. However, better understanding of the conditions under which specific weathering processes operate demonstrates that the efficacy of these processes cannot necessarily be clearly associated with specific climatic zones, primarily because of the temporal and spatial variability of conditions at the rock–air interface (Figure 3). For example, chemical weathering is relevant to rock breakdown in arid environments because the spatial variability of microenvironments can create atypical 'pockets' of activity. Many early schemes neglected the temporal aspects of weathering systems, whereby, if sufficient time is allowed, many weathering features hitherto ascribed to specific climatic zones can develop in regions with quite different climatic parameters. For example, the development of karstic (limestone) weathering phenomena associated with moisture availability proceeds slowly in many present-day hot arid environments but rapidly in colder conditions (because of the greater solubility of carbon dioxide in water at low temperatures).

Further complexity is introduced by the temporal variability of conditions at the rock–air interface, whereby long-term (tens to thousands of years) and/or short-term (diurnal and seasonal) environmental changes can alter the nature and extent of rock weathering at a specific location. This is particularly important in hot and cold deserts, where small changes in environmental parameters may have significant effects on weathering activity. Furthermore, viewing weathering as a consequence of the existing climate neglects the effects of inheritance (see below), which are important when considering the landscape as a whole and as a continuum. It is now accepted that, rather than trying to ascribe particular weathering processes to specific climatic zones, it is better to view weathering features as azonal phenomena *per se*.

Magnitude and Frequency in Rock Weathering

Sudden high-magnitude increases in the stress burden on rock through the operation of weathering processes may result in rapid or catastrophic breakdown of the rock fabric. Breakdown might not have occurred, or might have occurred less dramatically, if the stress had been applied more gradually. For example, the freezing rate and the number of oscillations across the freezing point of water in rock (which may be several degrees below 0°C) have been shown to be important in determining the efficacy of frost weathering. The rapid freezing of moisture within the rock fabric reduces losses through evaporation and 'cryosuction' (converting liquid water to solid), increasing the energy available for shattering. This in turn is related to the threshold concept, whereby, if the critical threshold of material strength is greater than the stress applied, no apparent change will occur. The absence of obvious visual damage in a rock can, however, be misleading, as it implies that the material is unaltered, even though microscopic external and internal changes may have occurred, the accumulation of which may eventually reduce cohesive strength and lead to 'fatigue' failure. Similarly, the effects of extreme heating and high rates of surface temperature change (thermal shock) are clearly demonstrated in the natural environment during bush fires, where widespread splitting and spalling of natural rock outcrops occurs as critical thresholds of strength are exceeded over very short periods of time (minutes). Depending on the nature of the

seen in the concordance of plateau summits, which are thought to be Mesozoic in age and probably the result of long-term *in situ* chemical weathering producing an 'etch-plain'. The effects of glacial activity and periglacial weathering in this environment are far less important than would be implied by a superficial view regarding only present-day processes. (E) The gabbroic cliffs (Øksfordsjøkel plateau, northern Norway) show clear evidence of jointing and removal by small rockfalls as well as glacial activity, illustrating the importance of weathering and crack or joint frequency in the wider context of erosive agents. Unlike in (D), the deeply weathered plateau blockfield material has been removed, and the weathering seen is mainly 'cryogenic'; nevertheless, chemical weathering continues to act on mafic minerals, producing a weathering rind. (F) Corestones remain in a deeply weathered Brazilian saprolite as a consequence of long-term chemical weathering processes. Without a good stratigraphical marker, the age of these deposits and time taken to form such materials is indeterminate (profile section depth is approximately 2 m).

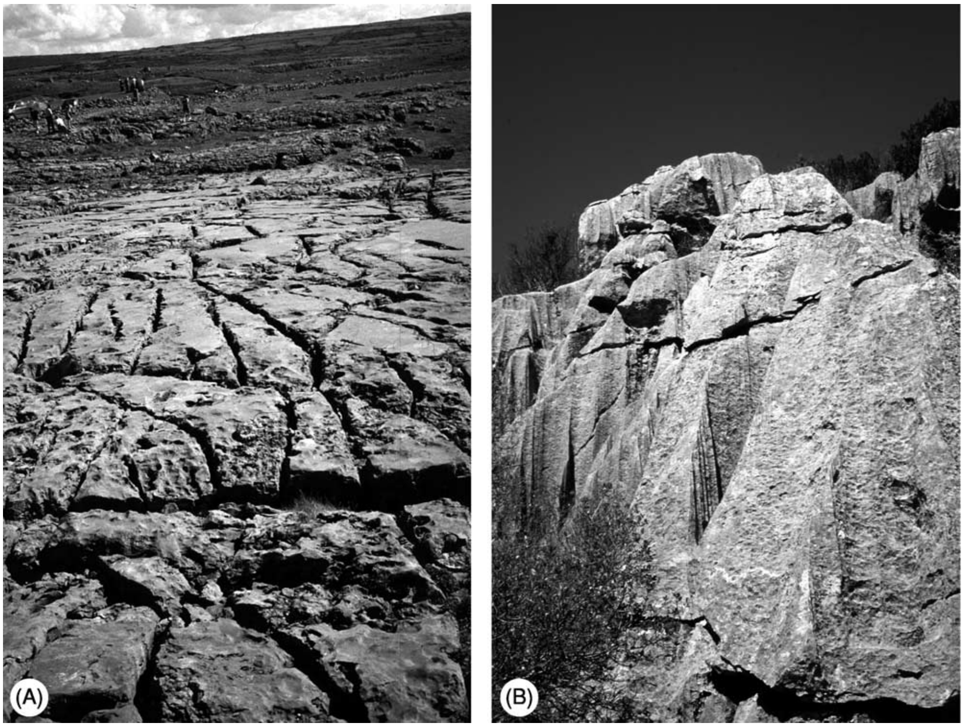


Figure 3 (A) Limestone pavement ('clints' and 'grykes') developed in the horizontally bedded Carboniferous limestone of the Burren, County Clare, Ireland. This limestone has a low porosity, resulting in the chemical dissolution effect of rainwater being concentrated along vertical joints, which form naturally occurring lines of weakness. (B) In contrast, the Jurassic limestone outcrops of the Sierra Norte mountains in Mallorca have been contorted and tilted by orogenesis, with the resultant flow of rainwater over surfaces preferentially weathering joints to create a different limestone landscape of pinnacles and towers on which a hierarchy of smaller solutional weathering features have formed.

environmental stress, prolonged and continuous stressing of rock may be less destructive than – or not so obviously destructive as – repeated high-magnitude short-term events.

Although the magnitudes and frequencies of stress events are important, material characteristics must also be considered. For example, the impermeability of some rock types may leave them relatively unaffected by high-magnitude weathering events, primarily because weathering agents such as moisture and salt cannot penetrate the rock fabric because of a lack of microcracks. Other rock types, exposed to the same environmental conditions, may be more susceptible because their structural and mineralogical characteristics facilitate the ingress of exploitative weathering agents. Consequently, in a landscape where different lithologies are present, it is probable that the nature and extent of the weathering response will vary.

Feedback Mechanisms

The weathering system may be best characterized as a nonlinear system primarily because of the variety of pathways and the unpredictability of outcomes. The unpredictable nature of rock-weathering pathways reflects the operation of positive- and negative-feedback conditions, with the former resulting in an overall change in system state through the acceleration of weathering activity while the latter maintains the system status quo or acts to retard weathering. The trigger for a change in feedback conditions can be seemingly insignificant but may result in considerable system destabilization. For example, a change in groundwater conditions can be sufficient to initiate widespread salt weathering of both natural rocks and manmade building material. Similarly, a change in microenvironmental parameters such as atmospheric humidity may activate salts that have hitherto been present but 'dormant' as regards weathering of rock.

Feedback conditions may help to explain some anomalous rock weathering responses. Part of a rock outcrop, for example, may exhibit well-developed weathering phenomena while an adjacent part of the same outcrop with the same exposure conditions shows no obvious evidence of deterioration. Examples of this are shown in Figure 4. Figure 4A shows a combination of basal tafoni with smaller hollows (alveoli) forming above; in Figure 4B there are no basal hollows but there are well-developed

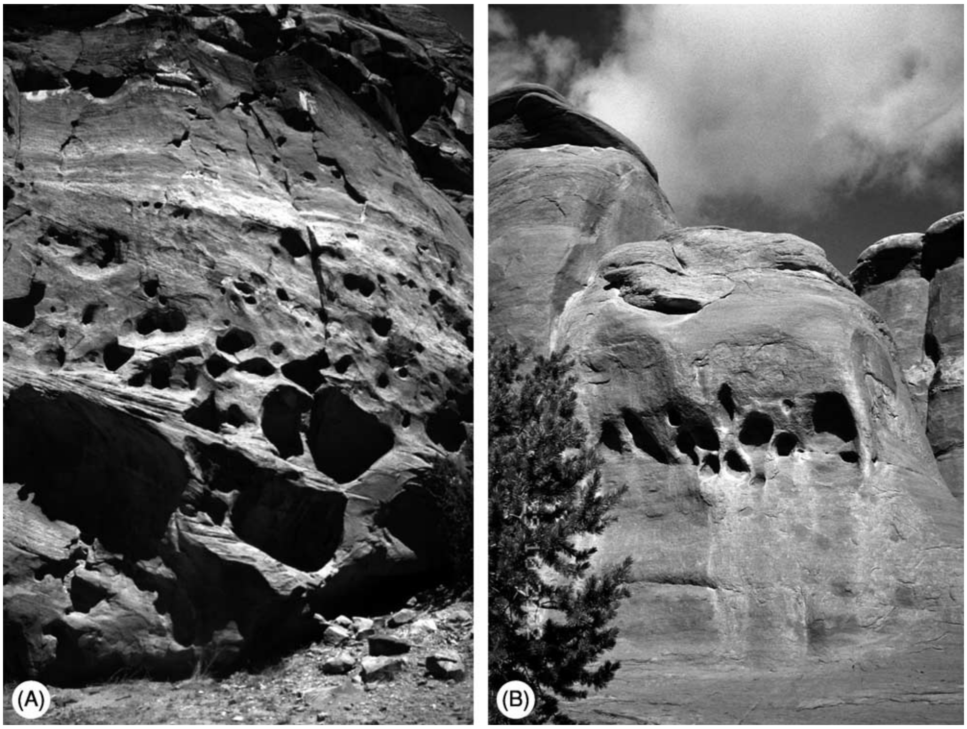


Figure 4 (A) Sandstone outcrop at Capitol Reef, Utah, showing well-developed basal tafoni and alveolar weathering features some 2–3 metres above ground level. (B) Sandstone monolith in Arches National Park, Utah, with well-developed alveoli formed some 2–3 m above ground level.

alveoli perched several metres above ground level. In both examples alveoli have formed adjacent to seemingly intact rock that has been exposed to the same environmental conditions and has the same general lithological characteristics. Positive-feedback will have influenced decay pathways initially through the creation of conditions conducive to the initiation of these hollows and subsequently through their progressive development. Initiation of the alveoli may have occurred as a result of structural or compositional anomalies within the sandstone, which were preferentially weathered because they allowed moisture or salt to penetrate. Hollow development may have been slow until increasing size allowed the establishment of microenvironmental conditions that were conducive to the more intense action of various weathering processes not effective on the outcrop surface. The weathering significance of such features lies not only in their development but also in the reasons why the adjacent rock remains relatively unaffected.

Equifinality – The Problem of Linking Process and Form

Another complicating factor affecting rock weathering is the issue of equifinality. This arises when different weathering processes produce similar weathering forms, thus preventing the identification of simple correlations between process, climate, and form. For example, angular shattered debris is common in both hot desert and high-latitude or mountainous regions. In the former, rock typically weathers *in situ* with little or no fluvial abrasion in an environment characterized by large diurnal temperature fluctuations and limited moisture availability. In the latter, moisture is normally more abundant and air temperatures frequently fall below 0°C. In these different climatic environments different groups of weathering processes contrive to produce similar debris forms. The similarity of clasts from these two environments is a consequence of the elongation of crack tips, typically by salt crystals in hot deserts and by ice in cold deserts. However, it should be noted that salt weathering is not exclusive to hot deserts: the sculptured cavernous weathering features (tafoni) observed in the Dry Valleys of Antarctica are attributed to the action of salt weathering rather than freeze–thaw processes.

At the landform scale, a classic example is that of ‘U’-shaped valleys, which were once (and still are by some) viewed as being characteristic of glaciated landscapes. However, these features can also be found in the subtropics, where intense chemical weathering at the water table contributes to ‘basal sapping’ (weathering plus erosion of products) and the maintenance of steep-sided valleys. In glaciated regions vertical-sided 'U' shapes can be a consequence of rock sheeting ('exfoliation'; as seen in the granites of Yosemite) and massive block failure. Both can be considered to be large-scale weathering phenomena involving overburden removal by long-term denudation.

Inheritance Effects

Many weathering studies have sought to explain patterns and rates of breakdown in terms of the prevailing environmental conditions and have sometimes relied upon the inference of process from 'form' alone at various scales, from landscape through to individual mineral grains. Too often, weathering processes that are immediately associated with observed rock weathering forms are judged to be solely responsible for their formation. In fact, this 'guilt by association' obscures the role of previous events and conditions – the weathering history of the material.

Exposure to various environmental conditions through either long-term climatic change or the spatial relocation of clasts gives a rock a complex weathering history. Rock outcrops and debris in many present-day landscapes therefore carry within their fabric an inheritance of structural and mineralogical weaknesses incurred under former conditions, which influence their response to present-day processes.

In addition to depending on the prevailing environmental conditions, weathering rates depend on the physical and mineralogical characteristics of the rock, and any factor that alters these characteristics will therefore influence subsequent rates of decay. Such changes may result from present-day weathering processes. However, many changes may be cumulative, each reflecting a set of environmental conditions and/or processes that acted in the past. Indeed, the significance of inheritance effects lies in the fact that they are frequently unrelated to present-day environmental conditions but were incurred under former conditions. For example, in arid environments many weathering features owe their initiation and development to the greater availability of water in the past. In Death Valley, now a hyperarid desert, the influence of water on landscape development is clearly evident, with extensive alluvial fans and, at a smaller scale, corroded cobbles, which can be found along the ancient strandlines that mark lake high stands from the early Holocene. These cobbles exhibit well-developed hollows covered by a manganese- and iron-rich rock varnish, indicating that the processes responsible for their development are no longer active under the present-day hyperarid conditions. Such weathering features created under former conditions are preserved within the present-day landscape and while some may be 'inactive' others may be undergoing some limited modification.

Weathering, Near-Surface Diagenesis and the Geological Record

Weathering of a variety of rocks, and sediments in particular, provides useful evidence of past climatic conditions. The weathering products can be substantial or present just as relict traces. Soils in the geological record (palaeosols) are found in a wide variety of sediments (including loess and river gravels) and in extrusive igneous rocks (typically Tertiary basalts). Substantial deposits related to weathering are exemplified by 'duricrusts'. These are hard durable layers or concretions found exposed (and even actively forming) on the surface of the Earth, typically in semiarid areas. The best known are laterite (Fe and Al rich), calcrete ($CaCO_3$), ferricrete (Fe^{3+}), gypcrete ($CaSO_4$), and silcrete (SiO_2). The main formative process is the upward flow of water with evaporation at the surface. This tends to be a self-limiting process, as the thickening duricrust eventually limits its own development by restricting evaporation.

Silcretes are orthoquartzites formed from fluids containing high concentrations of silica in solution (silicic acid), and silcrete remnants form the triliths at Stonehenge. The deposit from which these were derived (by weathering and removal of the surrounding uncemented sand) has been denuded but *in situ* silcretes have been found in the Paris basin, which forms the southern side of the Tertiary basin of southern England.

Weathering Related to Engineering and Economic Geology

As weathering and weathering-derived products are near-surface phenomena, they have important implications for engineering and economic geology. Frequently these implications are problematical; for example, deep weathering in relation to dams, other containment structures, foundations etc. In tropical areas, deep piling techniques may have to be used for high-rise buildings or substantial thicknesses of deep weathering may have to be removed to reach rock-head. Weathering processes influence many aspects of economic geology, not only by making ores (or traces of ore deposits) visible but also by producing valuable deposits in their own right. Of the former, perhaps the weathering of kimberlite to leave diamonds in alluvial deposits is the most famous.

It is important to remember that the weathering processes that affect rock in the natural environment also operate in urban environments, where their effectiveness can be enhanced by atmospheric pollutants, particularly through the effects of 'acid deposition', which results from increased concentrations of atmospheric SO_2, SO_3, NO_2, and NO (sometimes referred to colloquially as SOx and NOx) dissolved in rainwater. Although carbonates suffer most from such attacks, especially but not uniquely in industrial areas, other rock types used for building facades also suffer. Feldspars in certain granites are susceptible, as are calcareous sandstones. The main reactions are the result of atmospheric CO_2 dissolved in water (carbonic acid) and acid deposition (most commonly as aqueous H_2SO_4) on calcium carbonate, which is a common constituent of many rock types.

$$\mathrm{CaCO_3(s) + H_2SO_4(aq) \rightarrow Ca^{2+}(aq) + SO_4^{2-}(aq) + H_2O(l) + CO_2(g)}$$

In this reaction (which is also believed to occur in some natural karstic systems in addition to the usual CO_2 reaction), rainwater removes the dissolved, although sparingly soluble, calcium sulphate. Where there is no running water, or where there is air containing water droplets of low pH (occult precipitation), hydrated calcium sulphate (gypsum; $CaSO_4 \bullet 2H_2O$) can form. Where 'soot' particles are also scavenged from the local atmosphere and precipitated with the gypsum, 'black crusts' may form. There is still debate as to whether such crusts protect the underlying stone or exacerbate decay.

Biological Influences on Weathering

The traditional view of biological factors in rock weathering relates to tree roots penetrating pre-existing cracks in rocks, forcing mechanical breakdown. This is probably insignificant in comparison with the effects of micro-organisms on (and very near) rock surfaces and, in particular, bacterial action and root gas–water exchanges in soils. The metabolism of plants (and bacteria) produces organic acids (e.g. humic acid and fulvic acid), which can lower the pH of soil water and directly attack some minerals such as calcite. In addition, roots can increase gaseous CO_2 thus giving rise to high concentrations of HCO_3^- and increased acidity. The earliest soils (which were poor in metabolically important elements) were probably altered by the action of prokaryotic and early eukaryotic organisms, which broke down minerals and enriched soils with iron and phosphorous by chelation mechanisms.

The very presence of root structures and decaying vegetation not only increases CO_2 concentrations but also helps to retain moisture, allowing chemical weathering reactions to take place around soil particles. It has been shown that in modern streams the rate of chemical weathering is about seven times higher in a forested area than in surrounding bare areas.

Long-Term Changes in Weathering – Some Complicating Factors

There are several ways in which changes in weathering rates have had a marked effect on the geological record. The weathering of calcium and magnesium silicate minerals removes CO_2 from the atmosphere. This may be represented by

$$\mathrm{CO_2 + CaSiO_3\{Revers\ React\}CaCO_3 + SiO_2}$$

The concentration of CO_2, moderated by weathering, may influence climate. A computer simulation by Berner estimated levels of CO_2 in the Phanerozoic and showed a marked drop in the later Palaeozoic. This has been interpreted as the result of an increase in weathering during this period. The following aspects are significant.

In the Devonian land plants became more diverse and colonized previously barren land (especially upland). This extension of forested areas resulted in an increase in weathering. Atmospheric CO_2 continued to fall during the Carboniferous, even after the spread of forests, and this is attributed to burial of those forests in swampy areas (CO_2 sequestration). This total decline in atmospheric CO_2 is assumed to have weakened the effect of greenhouse warming and thus induced, or at least exacerbated, an extension of the southern hemisphere ice-sheets at about this time. An increase in levels of atmospheric CO_2 in the Mesozoic is attributed in part to minimal amounts of orogenesis (which increases weathering and removes CO_2 from the atmosphere) but mainly to increased emission during the metamorphism of calcareous oozes in subduction zones. In the Cenozoic a slow decline in the CO_2 concentration is attributed to orogenesis, in particular that of the Himalaya–Karakorum chains. This is an area of investigation that has been prompted by the work of Ruddiman and Raymo and is still an area of considerable research. A further factor is the position of land masses relative to the tropics: higher temperatures and precipitation increase weathering, and a lack of vegetation in more extreme conditions may reduce weathering (thus decreasing CO_2 sequestration). Clearly, the area of continental masses exposed to prevailing climatic belts also affects the total amount of weathering and is part of the self-regulating system of global climate. For example, increases in precipitation as a result of maritime influences increase terrestrial vegetation, weathering, and CO_2 capture, reducing the greenhouse effect and leading to decreased temperatures (and evaporation rates). Overall, temperature and precipitation affect the amounts and the extent of weathering. Thus, in moderating atmospheric CO_2, weathering on the Earth's surface exerts substantial negative feedback on global climatic change.

Concluding Comments

Weathering processes rarely, if ever, operate in isolation. Consequently, the weathering forms that we see are probably products of the cumulative and sequential effects of a variety of physical, chemical, and biological processes. Recent investigations focusing on the mechanisms of weathering rather than on the resultant forms have led to a wider and deeper understanding of the importance of weathering in topics ranging from building-stone conservation to long-term controls of climate and evolutionary processes.

See Also

Building Stone. **Clay Minerals**. **Geomorphology**. **Geotechnical Engineering**. **Mining Geology:** Hydrothermal Ores. **Sedimentary Processes:** Karst and Palaeokarst. **Soils:** Modern; Palaeosols.

Further Reading

Berner RA (1991) A model for atmospheric CO_2 over Phanerozoic time. *American Journal of Science* 291: 339–375.

Bland W and Rolls D (1998) *Weathering*. London: Arnold.

Drever JI and Clow DW (1995) Weathering rates in catchments. In: White AF and Brantley SL (eds.) *Chemical Weathering Rates of Silicate Minerals*, pp. 463–483. Reviews in Mineralogy. Washington DC: Mineralogical Society of America.

Goudie AS and Viles HA (1997) *Salt Weathering Hazards*. Chichester: Wiley.

Lsaga AC, Soler JM, Ganor J, Burch TE, and Nagy KL (1994) Chemical weathering rate laws and global geochemical cycles. *Geochimica et Cosmochimica Acta* 58: 2361–2386.

Ollier CD (1984) *Weathering*. London: Longman.

Peltier L (1950) The geographic cycle in periglacial regions as it is related to climatic geomorphology. *Annals of the Association of American Geographers* 40: 214–236.

Phillips JD (1999) *Earth Surface Systems*. Oxford: Blackwell.

Retallack GJ (2001) *Soils of the Past*. Oxford: Blackwell.

Ruddiman WF and Raymo ME (1988) Northern Hemisphere climate regimes during the past 3 Ma: possible tectonic connections. *Philosophical Transactions of the Royal Society, Series B* 318: 411–430.

Thomas MF (1994) *Geomorphology in the Tropics: A Study of Weathering and Denudation in Low Latitudes*. Chichester: Wiley.

Twidale CR (1982) *Granite Landforms*. Amsterdam: Elsevier.

White WB (1988) *Geomorphology and Hydrology of Karst Terrains*. New York: Oxford University Press.

Yatsu E (1988) *The Nature of Weathering*. Tokyo: Sozosha.