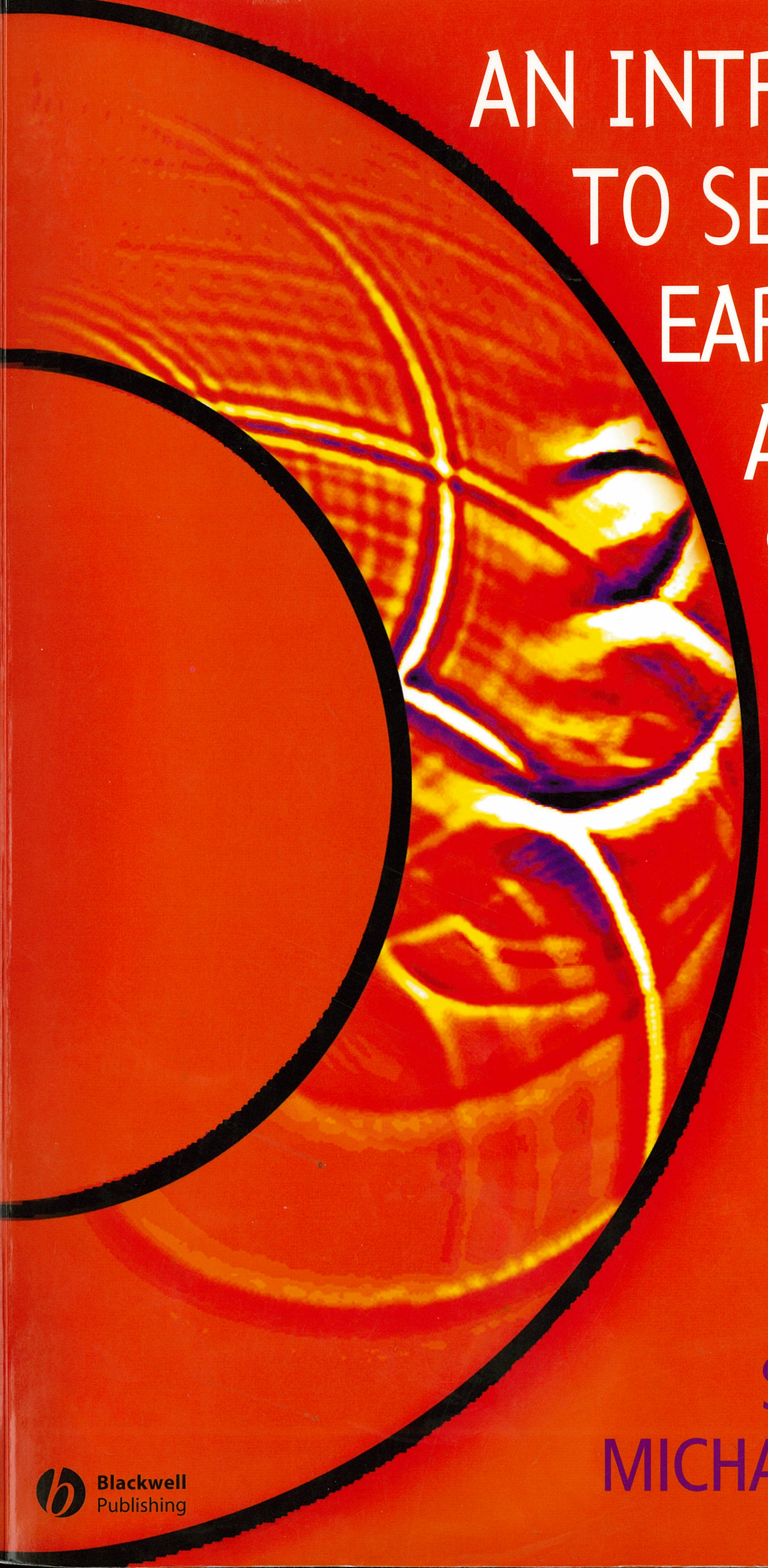


AN INTRODUCTION TO SEISMOLOGY, EARTHQUAKES, AND EARTH STRUCTURE

SETH STEIN and
MICHAEL WYSESSION

Blackwell
Publishing

5 Seismology and Plate Tectonics

The acceptance of continental drift has transformed the earth sciences from a group of rather unimaginative studies based on pedestrian interpretations of natural phenomena into a unified science that holds the promise of great intellectual and practical advances.

J. Tuzo Wilson, *Continents Adrift and Continental Aground*, 1976

5.1 Introduction

Two of the major advances in the earth sciences since the 1960s have been the growth of global seismology and the development of our understanding of global plate tectonics. The two are closely intertwined because seismological advances provided some of the crucial data that make plate tectonics the conceptual framework used to think about large-scale processes in the solid earth.

The theory of plate tectonics grew out of the earlier theory of continental drift, proposed in its modern form by Alfred Wegener in 1915. The idea that continents drifted apart was an old one, rooted in the remarkable fit of the coasts of South America and Africa. Still, without compelling evidence for motion between continents, the idea that such motions were physically impossible prevented most geologists from accepting Wegener's ideas. By the 1970s the story was very different. Geologists accepted continental drift in large part because paleomagnetic measurements, based on the geometry and history of the earth's magnetic field, showed that continents had in fact moved over millions of years. Combination of these observations with results from seismology and marine geology and geophysics led to the realization that all parts of the earth's outer shell, not just the continents, were moving.

Plate tectonics is conceptually simple: it treats the earth's outer shell as made up of about 15 rigid plates, about 100 km thick, which move relative to each other at speeds of a few cm per year.[1] The plates are rigid in the sense that little (ideally no) deformation occurs within them, so deformation occurs at their boundaries, giving rise to earthquakes, mountain building, volcanism, and other spectacular phenomena. These strong plates form the earth's *lithosphere*, and move over the weaker *asthenosphere* below. The lithosphere and asthenosphere are mechanical units defined by their strength and the way they deform. The lithosphere includes both the crust and part of the upper mantle.

Figure 5.1-1 shows the three basic types of plate boundaries. Warm mantle material upwells at *spreading centers*, also known as mid-ocean ridges, and then cools. Because the strength of rock decreases with temperature (Section 5.7.3), the cooling material forms strong plates of new oceanic lithosphere. The cooling oceanic lithosphere moves away from the ridges, and eventually reaches *subduction zones*, or trenches,[2] where it descends in *downgoing slabs* back into the mantle, reheating as it goes. The direction of the relative motion between two plates at a point on their common boundary determines the nature of the boundary. At spreading centers both plates move away from the boundary, whereas at subduction zones the subducting plate moves toward the boundary. At the third boundary type, *transform faults*, relative plate motion is parallel to the boundary.

As discussed in Section 3.8, seismology shows that the structure of the mantle and the core varies with depth, due to changes in temperature, pressure, mineralogy, and composition. Plate tectonics describes the behavior of the lithosphere, the strong outer shell of the mantle, which is the cold outer boundary layer of the thermal convection system involving the mantle and the core that removes heat from the earth's interior. Although much remains to be learned about this convective system, especially in the lower mantle and the core (Fig. 5.1-2), there is general agreement that at shallow depths the warm,

[1] This is about the speed at which fingernails grow.

[2] Boundaries are described either as mid-ocean ridges and trenches, emphasizing their morphology, or as spreading centers and subduction zones, emphasizing the plate motion there. The latter nomenclature is more precise, because there are elevated features in the ocean basins that are not spreading ridges, and spreading centers like the East African rift exist within continents.

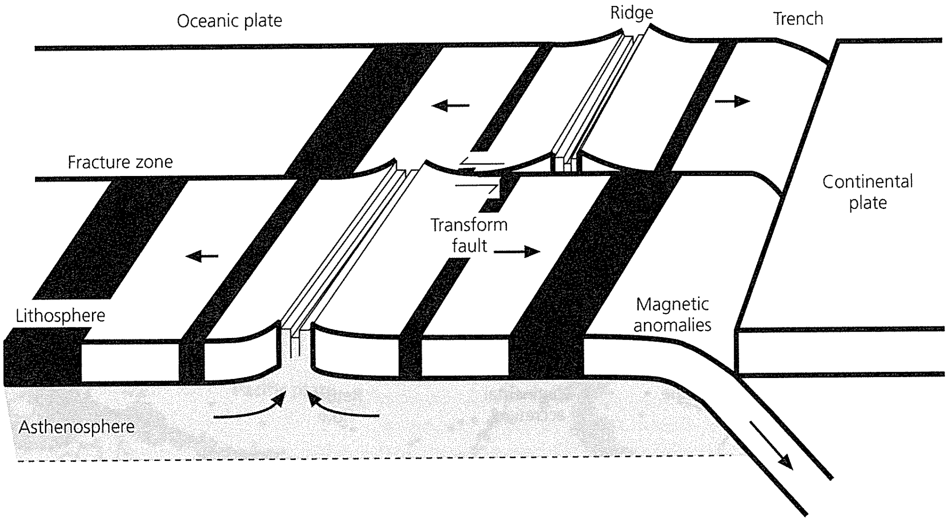


Fig. 5.1-1 Plate tectonics at its simplest. Oceanic lithosphere is formed at ridges and subducted at trenches. At transform faults, plate motion is parallel to the boundaries. Each boundary type has typical earthquakes.

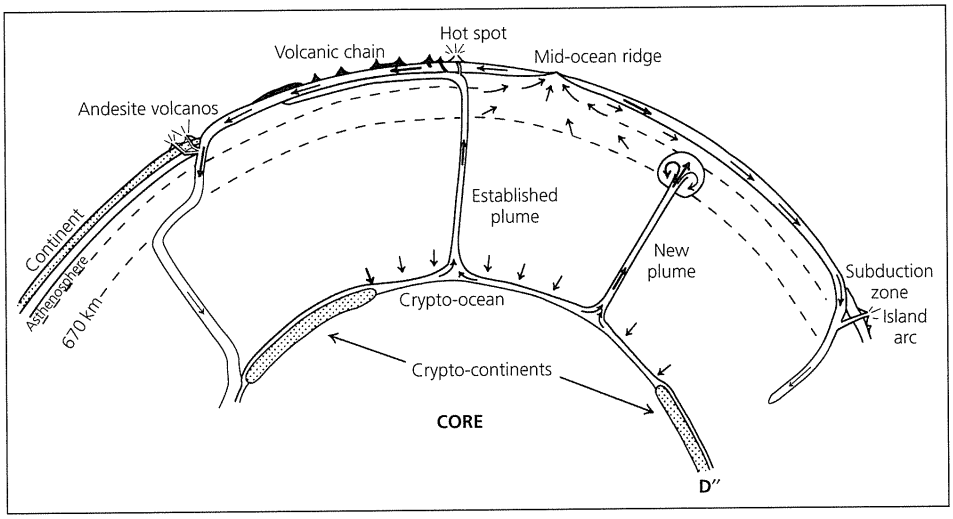


Fig. 5.1-2 Schematic diagram showing ideas about mantle convection. Ridges reflect upper mantle upwelling. Slabs penetrate into the lower mantle, causing heterogeneity there, and in some cases descend to the base of the mantle. Mantle (hot spot) plumes reflect lower mantle upwelling. Many features shown are controversial and subject to change without notice. (Modified from Stacey, 1992.)

and hence less dense, material rising below spreading centers forms upwelling limbs, whereas the relatively cold, and hence dense, subducting slabs form downwelling limbs. Although the lithosphere is a very thin layer compared to the rest of the mantle (100 km is 1/29 of the mantle's radius), it is where the greatest temperature change occurs, from about 1300° to 1400°C at a depth of 100 km to about 0°C at the surface. For this reason, the lithosphere is called a thermal boundary layer. Because of this temperature change, the lithosphere is much stronger than the underlying rock, and so is also a mechanical boundary layer. This strong boundary layer is thought to be a primary reason why plate tectonics is much more complicated than expected from simple convection models. Moreover, the lithosphere, which contains the crust, is also a chemical boundary layer distinct from the remainder of the mantle. Continental lithosphere is especially distinct: although individual plates can contain both oceanic and continental lithosphere, the latter is made of less dense rock than the former (recall the differences between granitic and basaltic rocks discussed in Section 3.2), and so does not subduct. The oceanic lithosphere is continuously subducted and reformed at ridges, and so never gets older than about 200 Myr. The continental lithosphere, however, can be billions of years old.

Put another way, plate tectonics is the primary surface manifestation of the heat engine whose nature and history govern the planet's thermal, mechanical, and chemical evolution.[3] Earth's heat engine is characterized by the balance between three modes of heat transfer from the interior: the plate tectonic cycle involving the cooling of oceanic lithosphere; mantle plumes, which are thought to be a secondary feature of mantle convection; and heat conduction through continents that are not subducted and hence do not participate directly in the oceanic plate tectonic cycle. Based on estimates from sea floor topography and heat flow, discussed shortly, terrestrial heat

[3] It has been said that heat is the geological lifeblood of planets.

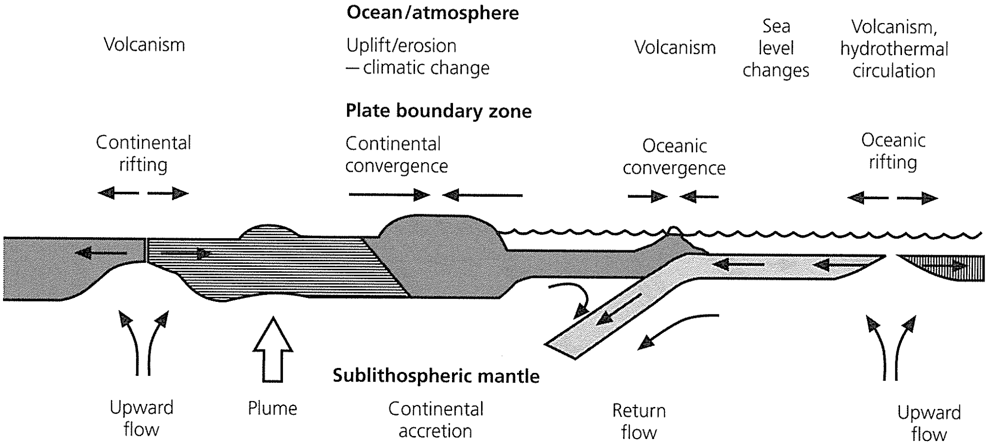


Fig. 5.1-3 Cartoon summarizing some of the primary modes of interaction between the solid earth's interior and the fluid ocean and atmosphere system. (Stein *et al.*, 1995. *Seafloor Hydrothermal Systems*, 425–45, copyright by the American Geophysical Union.)

loss seems to occur primarily (about 70%) via plate tectonics, with about 5% via hot spots (mantle plumes). By contrast, Earth's grossly similar sister planets, Mars and Venus, seem to function quite differently, because large-scale plate tectonics appears absent, at least at present.

Plate tectonics is also crucial for the evolution of Earth's ocean and atmosphere, because it involves many of the primary means (including volcanism, hydrothermal circulation through cooling oceanic lithosphere, and the cycle of uplift and erosion) by which the solid earth interacts with the ocean and the atmosphere (Fig. 5.1-3). The chemistry of the oceans and the atmosphere depends in large part on plate tectonic processes, and many long-term features of climate are influenced by mountains that are uplifted by plate convergence and the positions of continents that control ocean circulation. In fact, the presence of plate tectonics may explain how life evolved on earth (at mid-ocean ridge hot springs) and be crucial for its survival (the atmosphere is maintained by plate boundary volcanism, and plate tectonics raises the continents above sea level).

As a result, plate tectonics is heavily studied by earth scientists. Our goal in this chapter is to introduce some of the ways in which seismology contributes to these studies. Some sources for more general and more detailed treatments of these topics are listed at the end of the chapter.

Seismology plays several key roles in our studies of plate tectonics. The distribution of earthquakes provides strong evidence for the idea of essentially rigid plates, with deformation concentrated on their boundaries. Figure 5.1-4 shows maps of global seismicity covering the time period 1964–97. Such maps did not become available until the early 1960s, when the World Wide Standardized Seismographic Network (WWSSN) allowed accurate locations for earthquakes of magnitude 5 or greater anywhere in the world. The map shows several remarkable patterns.

The mid-ocean ridge system, where the oceanic lithosphere is created, is beautifully outlined by the earthquake locations. For example, the Mid-Atlantic ridge and East Pacific rise can be followed using epicenters for thousands of kilometers. The locations of the trenches, where oceanic lithosphere is subducted, are even more apparent in the lower panel showing earthquakes with focal depths greater than 100 km, because mid-ocean ridge earthquakes are shallow and thus do not appear.

It is especially impressive to plot the locations of earthquakes on cross-sections across trenches (Fig. 5.1-5). Inclined zones of seismicity delineate the subducting oceanic plates, which travel time and attenuation studies show to be colder and stronger than the surrounding mantle. These zones, identified before their plate tectonic significance became clear, are known as *Wadati–Benioff zones* after their discoverers.[4]

The *interplate* earthquakes both delineate plate boundaries and show the motion occurring there. We will see that the direction of faulting reflects the spreading at mid-ocean ridges and subduction at trenches. The earthquake locations and mechanisms also show that plate boundaries in continents are often complicated and diffuse, rather than the simple narrow boundaries assumed in the rigid plate model that are a good approximation to what we see in the oceans. For example, seismicity shows that the collision of the Indian and Eurasian plates creates a deformation zone which includes the Himalayas but extends far into China. Similarly, the northward motion of the Pacific plate with respect to North America creates a broad seismic zone, indicating that the plate boundary zone spans much of the western USA and Canada.

In addition, *intraplate* earthquakes occur within plate interiors, far from boundary zones. For example, Fig. 5.1-4 shows earthquakes in eastern Canada and central Australia. Such earthquakes are much rarer than plate boundary zone earthquakes, but are common enough to indicate that plate interiors are not perfectly rigid. In some cases these earthquakes are associated with intraplate volcanism, as in Hawaii. Intraplate earthquakes are studied to provide data about where and how the plate tectonic model does not fully describe tectonic processes.

[4] Kiyoo Wadati (1902–95) discovered the existence of deep seismicity and its geometry under Japan; Hugo Benioff (1899–1968), also known for important contributions to seismological instrumentation, discussed the global nature of deep earthquakes and their relation to surface features (Fig. 1.1-10).

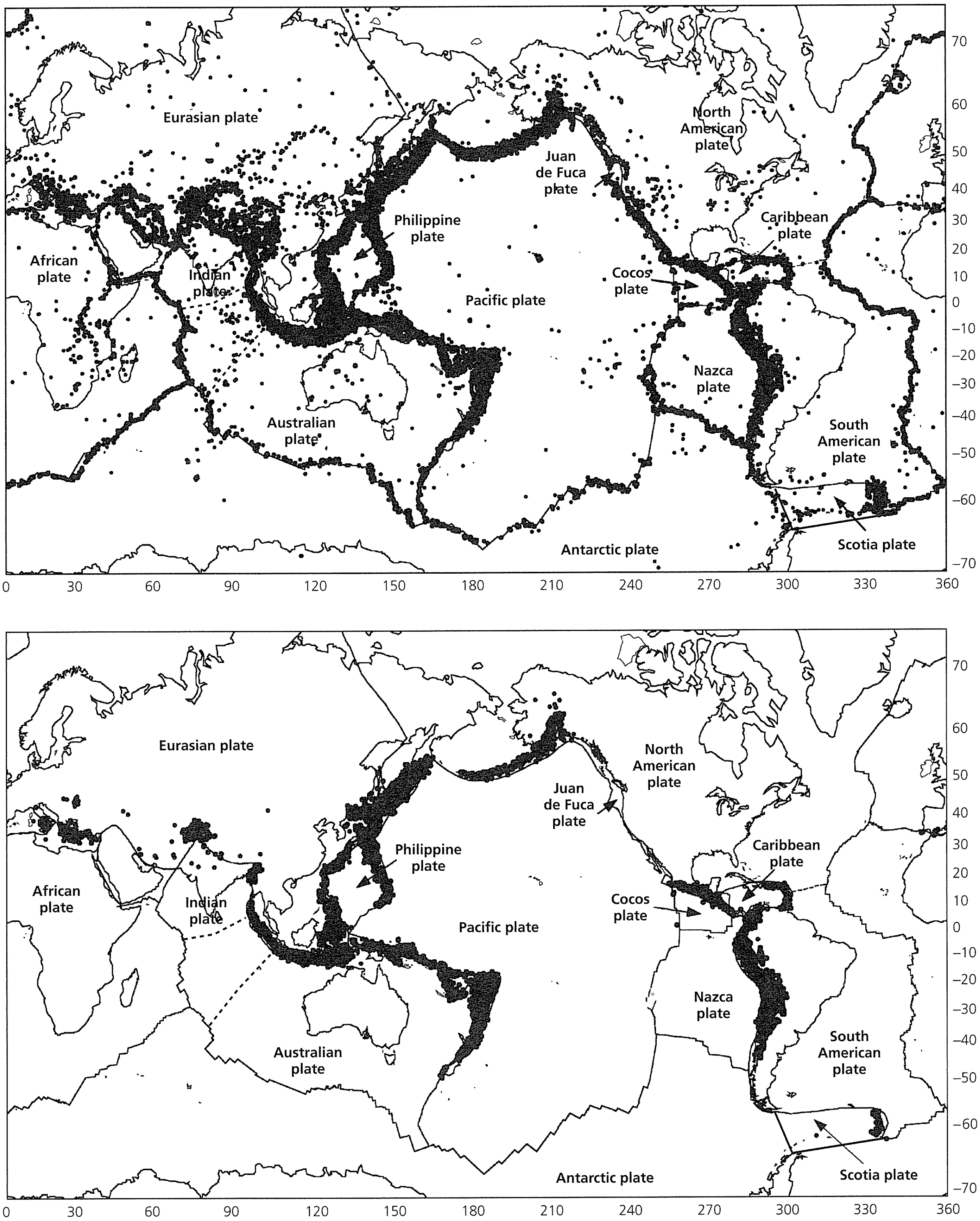


Fig. 5.1-4 Global seismicity (1964–97). *Top*: Earthquakes ($m_b \geq 5$, all depths) clearly delineate most plate boundaries, and show that some (e.g., India-Eurasia) are diffuse. Many intraplate earthquakes show internal plate deformation. *Bottom*: The locations of seismicity (of all magnitudes) below 100 km indicate the subduction zones.

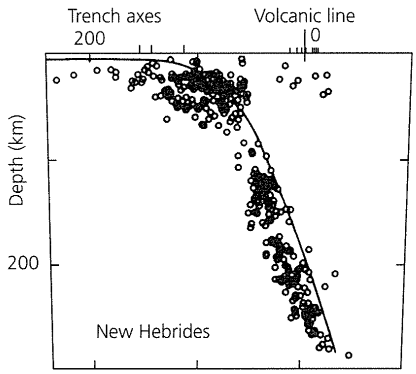


Fig. 5.1-5 Seismicity cross-section perpendicular to the New Hebrides trench showing the Wadati–Benioff zone. This dipping plane of earthquakes indicates the position of the subducting plate. (Isacks and Barazangi, 1977. *Island Arcs, Deep Sea Trenches and Back Arc Basins*, 99–114, copyright by the American Geophysical Union.)

In summary, seismology provides crucial information about both *plate kinematics*, the directions and rates of plate motions, and *plate dynamics*, the forces causing plate motions. As we will see, seismicity is one of the major tools used to identify and delineate plate boundary zones, and earthquake mechanisms are among the primary data used to determine the motion within plate boundary zones. The mechanisms also provide information about the stresses acting at plate boundaries and within plates, which, together with earthquake depths and seismic velocity structure, are important in developing ideas about the forces involved and the physical processes by which rocks deform and cause earthquakes. Conversely, plate motion data are used to draw inferences about the locations and times of future earthquakes and their societal risks. Thus it is often hard, and sometimes pointless, to decide where seismology ends and plate tectonics begins, or vice versa.

5.2 Plate kinematics

Understanding the distribution and types of earthquakes requires an understanding of the geometry of plate motions, or plate kinematics. In this section we sketch some basic results, of which we assume most readers have some knowledge. As full exploration of this topic is beyond our scope, readers are encouraged to delve into the suggested literature.

5.2.1 Relative plate motions

A basic principle of plate tectonics is that the relative motion between any two plates can be described as a rotation about an *Euler pole*[1] (Fig. 5.2-1). This condition controls the types of boundaries and the focal mechanisms of earthquakes resulting from relative motions, as discussed later. Specifically, at any point **r** along the boundary between plate i and plate j, with latitude λ and longitude μ, the *linear velocity* of plate j with respect to plate i is

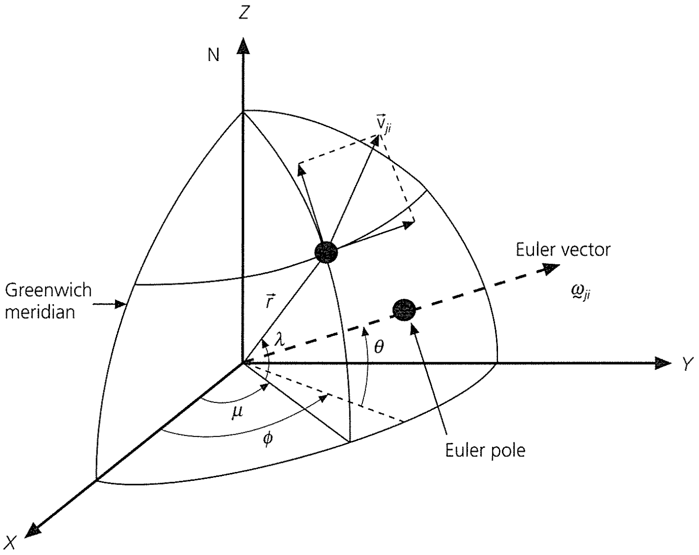


Fig. 5.2-1 Geometry of plate motions. Linear velocity at point **r** is given by $\mathbf{v}_{ji} = \boldsymbol{\omega}_{ji} \times \mathbf{r}$. The Euler pole is the intersection of the Euler vector with the earth's surface. Note that west longitudes and south latitudes are negative.

$$\mathbf{v}_{ji} = \boldsymbol{\omega}_{ji} \times \mathbf{r}. \tag{1}$$

This is the usual formulation for rigid body rotations in mechanics. **r** is the position vector to the point on the boundary, and $\boldsymbol{\omega}_{ji}$ is the angular velocity vector, or *Euler vector*. Both vectors are defined from an origin at the center of the earth.

The direction of relative motion at any point on the boundary is a small circle, a parallel of latitude *about the Euler pole* (not a geographic parallel about the North Pole!). For example, in Fig. 5.2-2 (*top*) the pole shown is for the motion of plate 2 with respect to plate 1. The convention used is that the first named plate ($j = 2$) moves counterclockwise (in a right-handed sense) about the pole with respect to the second named plate ($i = 1$). The segments of the boundary where relative motion is parallel to the boundary are transform faults. Thus transforms are small circles about the pole, and earthquakes occurring on them should have pure strike-slip mechanisms. Other segments have relative motion away from the boundary, and are thus spreading centers. Figure 5.2-2 (*bottom*) shows an alternative case. The pole here is for plate 1 ($j = 1$) with respect to plate 2 ($i = 2$), so plate 1 moves toward some segments of the boundary, which are subduction zones.

The magnitude, or rate, of relative motion increases with distance from the pole because

$$|\mathbf{v}_{ji}| = |\boldsymbol{\omega}_{ji}|\,|\mathbf{r}| \sin\gamma, \tag{2}$$

where γ is the angle between the Euler pole and the site (corresponding to a colatitude about the pole). All points on a plate

[1] This term comes from Euler's theorem, which states that the displacement of any rigid body (in this case, a plate) with one point (in this case, the center of the earth) fixed is a rotation about an axis.

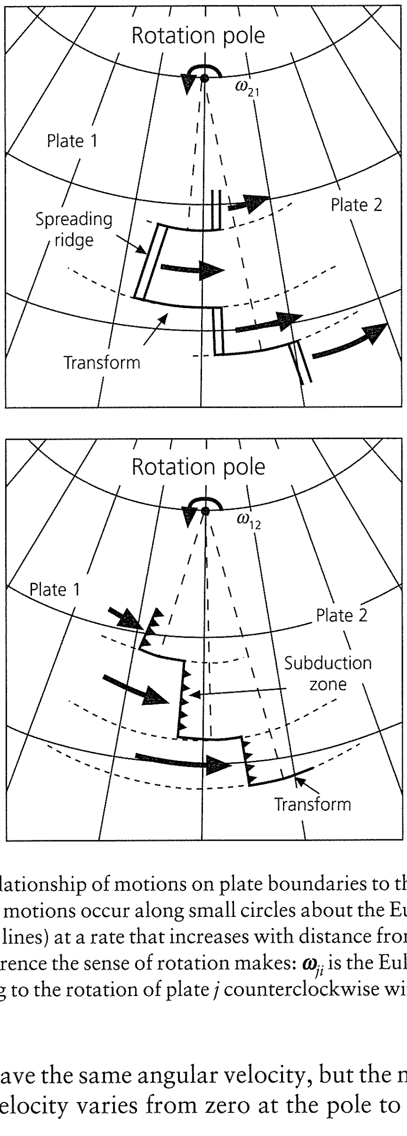


Fig. 5.2-2 Relationship of motions on plate boundaries to the Euler pole. Relative motions occur along small circles about the Euler pole (short dashed lines) at a rate that increases with distance from the pole. Note the difference the sense of rotation makes: $\boldsymbol{\omega}_{ji}$ is the Euler vector corresponding to the rotation of plate j counterclockwise with respect to i.

boundary have the same angular velocity, but the magnitude of the linear velocity varies from zero at the pole to a maximum 90° away.

The components of the vectors can be written in Cartesian (x, y, z) coordinates (Fig. 5.2-1). The position vector is

$$\mathbf{r} = (a \cos \lambda \cos \mu, a \cos \lambda \sin \mu, a \sin \lambda), \tag{3}$$

where a is the earth's radius. Similarly, if the Euler pole is at latitude θ and longitude ϕ, the Euler vector is written (neglecting the ij subscripts for simplicity) as

$$\boldsymbol{\omega} = (|\boldsymbol{\omega}| \cos \theta \cos \phi, |\boldsymbol{\omega}| \cos \theta \sin \phi, |\boldsymbol{\omega}| \sin \theta), \tag{4}$$

where the magnitude, $|\boldsymbol{\omega}|$, is the scalar angular velocity or rotation rate. To find the Cartesian components of the linear velocity $\mathbf{v}$, we evaluate the cross product (Eqn 1) using its definition (Eqn A.3.28), and find

$$\mathbf{v} = (v_x, v_y, v_z),$$

$$v_x = a|\boldsymbol{\omega}|(\cos \theta \sin \phi \sin \lambda - \sin \theta \cos \lambda \sin \mu)$$

$$v_y = a|\boldsymbol{\omega}|(\sin \theta \cos \lambda \cos \mu - \cos \theta \cos \phi \sin \lambda)$$

$$v_z = a|\boldsymbol{\omega}| \cos \theta \cos \lambda \sin (\mu - \phi). \tag{5}$$

At the point $\mathbf{r}$, the north–south and east–west unit vectors can be written in terms of their Cartesian components using Eqn A.7.4,

$$\hat{\mathbf{e}}^{\mathrm{NS}} = (-\sin \lambda \cos \mu, -\sin \lambda \sin \mu, \quad \cos \lambda),$$

$$\hat{\mathbf{e}}^{\mathrm{EW}} = (-\sin \mu, \cos \mu, \quad 0), \tag{6}$$

so we find the north–south and east–west components of $\mathbf{v}$ by taking dot products of its Cartesian components (Eqns 5) with the unit vectors (Eqns 6), and obtain

$$v^{\mathrm{NS}} = a|\boldsymbol{\omega}| \cos \theta \sin (\mu - \phi),$$

$$v^{\mathrm{EW}} = a|\boldsymbol{\omega}|[\sin \theta \cos \lambda - \cos \theta \sin \lambda \cos (\mu - \phi)]. \tag{7}$$

We can then find the rate and direction of plate motion,

$$\text{rate} = |\mathbf{v}| = \sqrt{(v^{\mathrm{NS}})^2 + (v^{\mathrm{EW}})^2}$$

$$\text{azimuth} = 90° - \tan^{-1} [(v^{\mathrm{NS}})/(v^{\mathrm{EW}})], \tag{8}$$

such that azimuth is measured in the usual convention, degrees clockwise from North.

In evaluating these expressions, it is important to be careful with dimensions. Although rotation rates are typically reported in degrees per million years, they should be converted to radians per year. The resulting linear velocity will have the same dimensions as Earth's radius. By serendipity, converting radius in km to mm and Myr to years cancel out, so only the degrees to radians ($\times \pi/180°$) conversion actually needs to be done to obtain a linear velocity in mm/yr. Plate motions are often quoted as mm/yr, because a year is a comfortable unit of time for humans and 1 mm/yr corresponds to 1 km/Myr, making it easy to visualize what seemingly slow plate motion accomplishes over geologic time.

To see how this works, consider Fig. 5.2-3, which shows the North America–Pacific boundary zone. The map is drawn in a projection about the Euler pole, so the expected relative motion is parallel to small circles like the one shown. By analogy to Fig. 5.2-2, this geometry predicts NW–SE-oriented spreading along ridge segments in the Gulf of California, which are rifting Baja California away from the rest of Mexico. Further north, the San Andreas fault system is essentially parallel to the relative motion, so is largely a transform fault. In Alaska, the eastern Aleutian arc is perpendicular to the plate motion, so the Pacific plate subducts beneath North America. Thus this plate boundary contains ridge, transform, and trench portions, depending on the geometry of the boundary.[2] In addition, the

[2] A good way to visualize the plate motion is to photocopy Fig. 5.2-3, cut along the boundary of the Pacific plate, and then photocopy the "Pacific" onto another piece of paper. Putting the "Pacific" beneath "North America" and rotating around a thumbtack through the pole shows the ridge, transform, and trench motions both forward and backward in time.

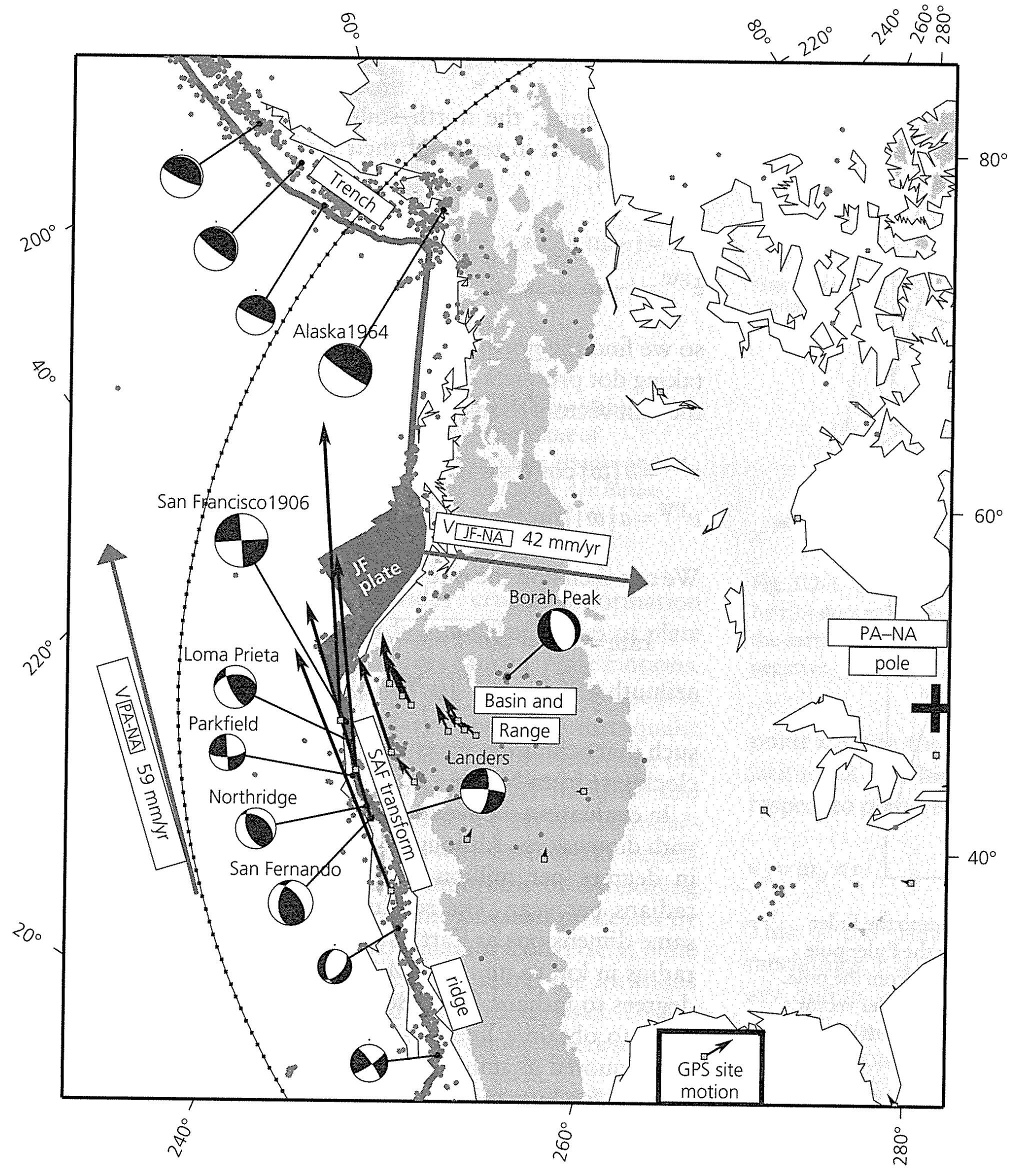


Fig. 5.2-3 Geometry and focal mechanisms for a portion of the North America–Pacific boundary zone that also includes the small Juan de Fuca (JF) plate. The map projection is about the Pacific–North America Euler pole, so the line with dots shows a small circle and thus the direction of plate motion. This small circle is further from the pole than the San Andreas fault, so the rate of motion on it is larger. The variation in the boundary type along its length from extension, to transform, to convergence, is shown by the focal mechanisms. The diffuse nature of the boundary zone is shown by seismicity (small dots), focal mechanisms, topography (elevation above 1000 m is shaded), and vectors showing the motion of GPS and VLBI sites (squares) (Bennett *et al.*, 1999) with respect to the stable interior of North America. The velocity scale is shown by the plate motion arrows; some site motion vectors are too small to be seen. (Stein and Klosko, 2002. From *The Encyclopedia of Physical Science and Technology*, ed. R. A. Meyers, copyright 2002 by Academic Press, reproduced by permission of the publisher.)

boundary zone contains the small Juan de Fuca plate, which subducts beneath the Pacific Northwest at the Cascadia subduction zone.

Equation 8 lets us find how the motion varies. The predicted motion of the Pacific plate with respect to the North American plate at a point on the San Andreas fault (36°N, 239°E) has a rate of 46 mm/yr at an azimuth of N36°W. The predicted direction agrees reasonably well with the average trend of the San Andreas fault, N41°W. Thus, to first order, the San Andreas is a Pacific–North America transform plate boundary with right-lateral motion. However, there are some deviations from pure transform behavior. As we will see, the rate on the San Andreas fault is less than the total plate motion because some of the motion occurs elsewhere within the broad plate boundary zone. In addition, in some places the San Andreas trend differs enough from the plate motion direction that dip-slip faulting occurs. Hence we think of the San Andreas as the primary feature of the essentially strike-slip portion of the plate boundary zone.

Similarly, at a point on the Aleutian trench near the site of the great 1964 Alaska earthquake (Fig. 4.3-15) (62°N, 212°E), we predict Pacific motion of 53 mm/yr at N14°W with respect to North America. This motion is into the trench, which is a Pacific–North America subduction zone. It is worth noting that for a given convergent relative motion either plate can be subducting. However, the relative direction is important, so the plates cannot be interchanged: if N14°W were the direction of motion of North America with respect to the Pacific, the motion would be away from the boundary, which would then be a spreading center with the same rate. As for the San Andreas, the actual boundary zone shown by earthquakes and other deformation is wider and more complicated than the ideal.

Earthquake focal mechanisms within the boundary zone are consistent with the overall plate motions and illustrate some of their complexities. In the Gulf of California we see both strike-slip faulting along oceanic transforms and normal faulting on ridge segments. The San Andreas fault system, composed of the main fault and some others, has both pure strike-slip earthquakes (Parkfield) and earthquakes with some dip-slip motion (Northridge (Section 4.5.3), San Fernando, and Loma Prieta) when it deviates from pure transform behavior. The seismicity also shows that the plate boundary zone is quite broad. Although the San Andreas fault system is the locus of most of the plate motion (Fig. 4.5-13) and hence large earthquakes, seismicity extends as far eastward as the Rocky Mountains. For example, the Landers earthquake shows strike-slip motion east of the San Andreas, and the Borah Peak earthquake illustrates the extensional faulting that occurs in the Basin and Range. These focal mechanisms are consistent with the motions shown by space-based geodetic measurements, discussed shortly, and with geologic studies.

5.2.2 *Global plate motions*

The relative plate motions show how the plate boundary geometry is evolving and has evolved. The Juan de Fuca plate is subducting under North America faster than new lithosphere is being added to it by sea floor spreading at its boundary with the Pacific plate, so this plate was larger in the past and is shrinking. Rotating the Pacific plate backwards with respect to North America shows that 10 million years ago the Gulf of California had not yet begun to open by sea floor spreading. These changes are part of the evolution of the plate boundary in western North America, in which the large oceanic Farallon plate that used to be between the Pacific and North American plates began subducting under North America at about 40 Ma,[3] leaving the Juan de Fuca plate as a remnant and forming the San Andreas fault.

At this point you may be wondering how Euler poles are found. Until recently, this was done by combining three different types of data from different boundaries. The rates of spreading are found from sea floor magnetic anomalies, which form as the hot rock at ridges cools and acquires magnetization parallel to the earth's magnetic field. Because the history of reversals of the earth's magnetic field is known, the anomalies can be dated, so their distance from the ridge where they formed shows how fast the sea floor moved away from the ridge. The directions of motion are found from the orientations of transform faults and the slip vectors of earthquakes on transforms and at subduction zones. Euler vectors are found from the relative motion data, using geometrical conditions we have discussed. The process is easy to visualize. Because slip vectors and transform faults lie on small circles about the pole, the pole must lie on a great circle at right angles to them (Fig. 5.2-2). Similarly, the rate of plate motion increases with the sine of the distance from the pole (Eqn 2). These constraints make it possible to locate the poles. Determination of Euler vectors for all the plates can thus be treated as an overdetermined least squares problem whose solution (Section 7.5) gives a *global relative plate motion model*. Because these models use spreading rates determined from magnetic anomaly data that span several million years, they describe plate motions averaged over the past few million years.[4]

Table 5.2-1 gives such a model, known as NUVEL-1A,[5] which specifies the motions of plates (Fig. 5.2-4) with respect to North America. The vectors follow the convention that each named plate moves counterclockwise relative to North America. Although the table lists only Euler vectors with respect to North America, the motion of plates with respect to other plates is easily found using vector arithmetic. For example,

$$\omega_{ij} = -\omega_{ji}, \tag{9}$$

so we reverse the plate pair using the negative of the Euler vector. The pole for the new plate pair is the antipole, with latitude of opposite sign and longitude increased by 180°. The magnitude (rotation rate) stays the same. We can also reverse the plate pair by keeping the same pole and making the rotation rate negative (clockwise rather than counterclockwise). Although we usually use positive rotation rates, negative ones sometimes help us visualize the motion. For example, the table shows the Pacific–North America pole at about −49°N, 102°E, so the North America–Pacific pole is at about 49°N, (102 + 180 = 282)°E, which is in southeastern Canada. Thus, about this pole, North America rotates counterclockwise with respect to the Pacific, or the Pacific rotates clockwise with respect to North America, as shown in Fig. 5.2-3.

For other plate pairs we assume that the plates are rigid, so all motion occurs at their boundaries. We can then add Euler vectors,

$$\omega_{jk} = \omega_{ji} + \omega_{ik} \tag{10}$$

because the motion of plate *j* with respect to plate *k* equals the sum of the motion of plate *j* with respect to plate *i* and the motion of plate *i* with respect to plate *k*. Thus if we start with a set of vectors all with respect to one plate, e.g., *i*, we use

$$\omega_{jk} = \omega_{ji} - \omega_{ki} \tag{11}$$

to form any Euler vector needed. These operations are easily done using the Cartesian components (Eqn 4), as shown in this chapter's problems. We can also perform the analogous operations on linear velocity vectors at a specific site.

[3] "Ma" is often used to denote millions of years before the present.

[4] The most recent magnetic reversal occurred about 780,000 years ago, so any plate model based on paleomagnetic data must average at least over that interval.

[5] NUVEL-1 (Northwestern University VELocity) was developed as a new ("nouvelle") model (DeMets *et al.*, 1990). The multiyear development prompted the suggestion that "OLDVEL" might be a better name. Due to changes in the paleomagnetic time scale the model was revised to NUVEL-1A (DeMets *et al.*, 1994). This change caused a slight difference in the rates of relative motion, but not in the poles and hence directions of relative motion.

Table 5.2-1 Euler vectors with respect to North America (NA).

Plate	Pole latitude (°N)	Longitude (°E)	|ω| (°/Myr)
Pacific (PA)	−48.709	101.833	0.7486
Africa (AF)	78.807	38.279	0.2380
Antarctica (AN)	60.511	119.619	0.2540
Arabia (AR)	44.132	25.586	0.5688
Australia (AU)	29.112	49.006	0.7579
Caribbean (CA)	74.346	153.892	0.1031
Cocos (CO)	27.883	−120.679	1.3572
Eurasia (EU)	62.408	135.831	0.2137
India (IN)	43.281	29.570	0.5803
Nazca (NZ)	61.544	−109.781	0.6362
South America (SA)	−16.290	121.876	0.1465
Juan de Fuca (JF)	−22.417	67.203	0.8297
Philippine (PH)	−43.986	−19.814	0.8389
Rivera (RI)	22.821	−109.407	1.8032
Scotia (SC)	−43.459	123.120	0.0925
NNR*	2.429	93.965	0.2064

Source: After DeMets *et al.* 1994.
*No net rotation, defined in Section 5.2.4.

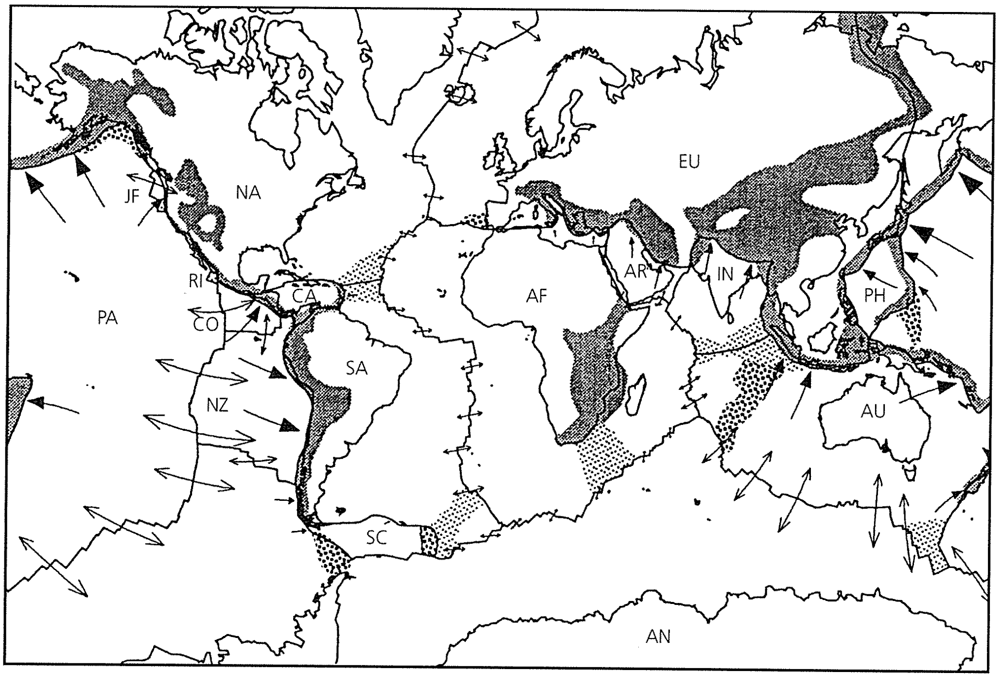


Fig. 5.2-4 Relative plate motions for the NUVEL-1 global plate motion model. Arrow lengths are proportional to the displacement if plates maintain their present relative velocity for 25 Myr. Divergence across mid-ocean ridges is shown by diverging arrows. Convergence is shown by single arrows on the underthrust plate. Plate boundaries are shown as diffuse zones implied by seismicity, topography, or other evidence of faulting. Fine stipple shows mainly subaerial regions where the deformation has been inferred from seismicity, topography, other evidence of faulting, or some combination of these. Medium stipple shows mainly submarine regions where the nonclosure of plate circuits indicates measurable deformation; in most cases these zones are also marked by earthquakes. Coarse stipple shows mainly submarine regions where the deformation is inferred mostly from the presence of earthquakes. The geometry of these zones, and in some cases their existence, is under investigation. (Gordon and Stein, 1992. *Science, 256*, 333–42, copyright 1992 American Association for the Advancement of Science.)

Such vector addition is important because we only have certain types of data for individual boundaries (Fig. 5.2-5). Although spreading centers provide rates from the magnetic anomalies and azimuths from both transform faults and slip vectors, only the direction of motion is directly known at subduction zones. As a result, convergence rates at subduction zones are estimated by global closure, combining data from all plate boundaries (Section 7.5). Thus the predicted rate at which

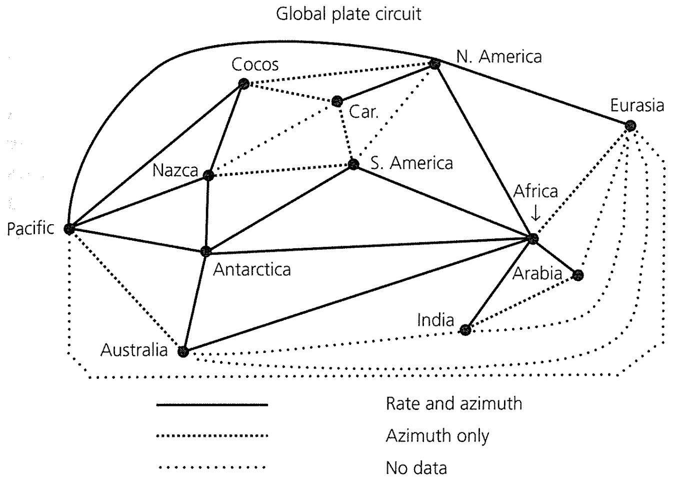


Fig. 5.2-5 Global plate circuit geometry for the NUVEL-1 plate motion model. Relative motion data are used on the boundaries indicated. (De Mets *et al.*, 1990. *Geophys. J. Int.*, *101*, 425–78.)

the Cocos plate subducts beneath North America, causing large earthquakes in Mexico, depends on the measured rates of Cocos–Pacific spreading on the East Pacific rise and Pacific–North America spreading in the Gulf of California. In some cases, such as relative motion between North and South America, no direct data were used because the boundary location and geometry are unclear, so the relative motion is inferred entirely from closure. Not surprisingly, the motions of plate pairs based on both rate and azimuth data appear to be better known.

Figure 5.2-4 shows the predicted relative motions at plate boundaries around the world. As shown for the Pacific–North America boundary in Fig. 5.2-3 and discussed in general terms in later sections, the predicted motions correspond to the earthquake mechanisms. Moreover, we can use the plate motions to make inferences about future earthquakes. For example, even though we do not have seismological observations of large earthquakes along the boundary between the Juan de Fuca and North American plates, the plate motions predict that such earthquakes could result from the subduction of the Juan de Fuca plate beneath North America. Evidence for this subduction is given by the presence of the Cascade volcanoes (such as Mount Saint Helens and Mount Rainer) and paleoseismic records (Section 1.2.5) that are interpreted as evidence of large past earthquakes.

Figure 5.2-4 also illustrates that boundaries between plates are often diffuse. Seismicity, active faulting, and elevated topography often indicate a broad zone of deformation between plate interiors. This effect is evident in continental lithosphere, such as the India–Eurasia collision zone in Asia or the Pacific–North America boundary zone in the western USA, but can also sometimes be seen in oceanic lithosphere, as in the Central Indian Ocean. Plate boundary zones cover about 15% of the earth's surface, and about 40% of the earth's population lives within them.

Earthquakes are among the best tools for investigating plate boundary zones and other deviations from plate rigidity. They provide one of the best indicators of the location of boundary zones, so new earthquakes often change our views. We also use plate motion data, many of which are earthquake slip vectors. For example, Fig. 5.2-4 shows zones of seismicity in the Central Indian Ocean (Section 5.5.2) as boundaries between distinct Indian and Australian plates, rather than as within a single Indo-Australian plate, because spreading rates along the Central Indian Ocean ridge are better fit by a two-plate model. A similar argument justifies the assumption of a small Rivera plate distinct from the Cocos plate. Another approach is to use the global plate circuit closures (Fig. 5.2-5). Recall that forming a Euler vector from two others (Eqn 10) assumes that all three plates are rigid. Hence this assumption can be used to test for deviations from rigidity. To do this, we form a *best-fitting vector* for a plate pair, using only data from that pair of plates' boundary, and a *closure fitting vector* from data elsewhere in the world. If the plates were rigid, the two vectors would be the same. However, a significant difference between the two indicates a deviation from rigidity, or another problem with the plate motion model. For example, such analysis shows systematic deviations along some subduction zones, suggesting that the slip vectors of the trench earthquakes do not exactly reflect plate motions because a sliver of forearc material in the overriding plate moves separately from the remainder of the overriding plate (Section 5.4.3).

A variant of this approach is to examine the Euler vectors for three plates that meet at a *triple junction*, compute best-fitting Euler vectors for each of the three plate pairs, and sum them. For rigid plates, Eqn 10 shows that the sum should be zero. However, when this was done for the junction in the Central Indian Ocean, assuming that it was where the African, Indo-Australian, and Antarctic plates met, the Euler vector sum differed significantly from zero, indicating deviations from plate rigidity. As plate motion data improve, it seems that what was treated as a three-plate system may include as many as six resolvable plates (Antarctica, distinct Nubia (West Africa) and Somalia (East Africa), India, Australia, and Capricorn (between India and Arabia)). Hence models of plate boundaries and motions improve with time (Fig. 1.1-9). For example, although the model in Fig. 5.2-4 has a single African plate, recent models seek to resolve the motion between Nubia and Somalia (Fig. 5.6-2).

5.2.3 Space-based geodesy

New plate motion data have become available in recent years due to the rapidly evolving techniques of space-based geodesy. Using space-based measurements to determine plate motions was suggested by Alfred Wegener when he proposed the theory of continental drift in 1915. Wegener realized that proving continents moved apart was a formidable challenge. Although geodesy — the science of measuring the shape of, and distances on, the earth — was well established, standard surveying

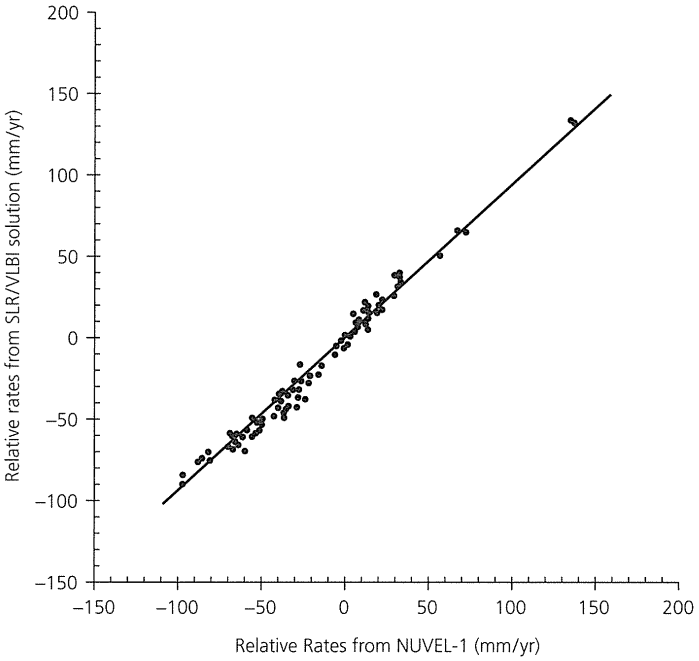


Fig. 5.2-6 Comparison of rates determined by space geodesy with those predicted by the NUVEL-1 global plate motion model. The space geodetic rates are determined from sites located away from plate boundaries to reduce the effects of deformation near the boundaries. The slope of the line is 0.94, indicating that plate motions over a decade are very similar to those predicted by a model averaging over 3 million years. (Robbins *et al.*, 1993. *Contributions of Space Geodesy to Geodynamics*, 21–36, copyright by the American Geophysical Union.)

methods offered no hope of measuring slow motions between continents far apart. Wegener thus decided to measure the distance between continents using astronomical observations.[6] However, because measuring continental drift called for measurement accuracies far greater than ever before to show small changes in positions over a few years, Wegener's attempts failed, and the idea of continental drift was largely rejected.

By the 1970s the story was very different. Geologists accepted continental drift, in large part because paleomagnetic measurements showed that continents had in fact moved over millions of years. It thus seemed natural to see if modern space-based technology could accomplish Wegener's dream of measuring continental motions over a few years. Three basic approaches were attempted. Each faced formidable technical challenges — and all succeeded. Hence, using the techniques discussed in Section 4.5.1, plate motions can now measured to a precision of a few mm/yr or better, using a few years of data from systems including Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR), and the Global Positioning System (GPS).

Space geodesy measures both the rate and the azimuth of the motions between sites, and can thus be used to compute relative plate motions. One of the most important results of space geodesy for seismology is that plate motions have remained generally steady over the past few million years. This is shown by the striking agreement between motions measured over a few years by space geodesy and the predictions of global plate motion models that average over the past three million years (Fig. 5.2-6). The general agreement is consistent with the idea that although motion at plate boundaries can be episodic, as in large earthquakes, the viscous asthenosphere damps out the transient motions (much like the damping element in a seismometer, Section 6.6) and causes steady motion between plate interiors. This steadiness implies that plate motion models can be used for comparison with earthquake data.

Space geodesy surmounts a major difficulty faced by models like NUVEL-1A: namely, that the data used (spreading rates, transform azimuths, and slip vectors) are at plate boundaries, so the model provides only the net motion across a boundary. By contrast, space geodesy can also measure the motion of sites within plate boundary zones. For example, Fig. 5.2-3 shows the motions of GPS and VLBI sites within the North America–Pacific boundary zone. Sites in eastern North America move so slowly — less than 2 mm/yr — with respect to each other that their motion vectors cannot be seen on this scale. These sites thus define a rigid reference frame for the stable interior of the North American plate. Sites west of the San Andreas fault move at essentially the rate and direction predicted for the Pacific plate by the global plate motion model. The site vectors show that most of the plate motion occurs along the San Andreas fault system, but significant motions occur for some distance eastward. The geodetic motions are consistent with the focal mechanisms and geological data. Thus, as discussed further in Section 5.6, the different data types are used together to study how the seismic and aseismic portions of the deformation vary in space and time in the diffuse deformation zones that characterize many plate boundaries. This is done both on large scales, as shown here, and for studies of smaller areas and individual earthquakes (Section 4.5).

Space geodesy is also used to study the relatively rare, but sometimes large, earthquakes within plates. Global plate motion models give no idea where or how often intraplate earthquakes should occur, beyond the trivial prediction that they should not occur because there is no deformation within ideal rigid plates. Space geodesy is being combined with earthquake locations, focal mechanisms, and other geological and geophysical data to investigate the motions and stresses within plates and how they give rise to intraplate earthquakes (Section 5.6.3).

5.2.4 *Absolute plate motions*

So far, we have discussed the relative motions between plates, which have traditionally been of greatest interest to seismologists because most earthquakes reflect these motions. However, in some applications it is important to consider *absolute* plate motions, those with respect to the deep mantle.

In general, both plates and plate boundaries move with respect to the deep mantle. To see this, assume that the African

[6] Using an extraterrestrial reference has a long history; in about 230 BC Eratosthenes found the Earth's size from observations of the sun's position at different sites, and navigators have found their positions by observing the sun and stars.

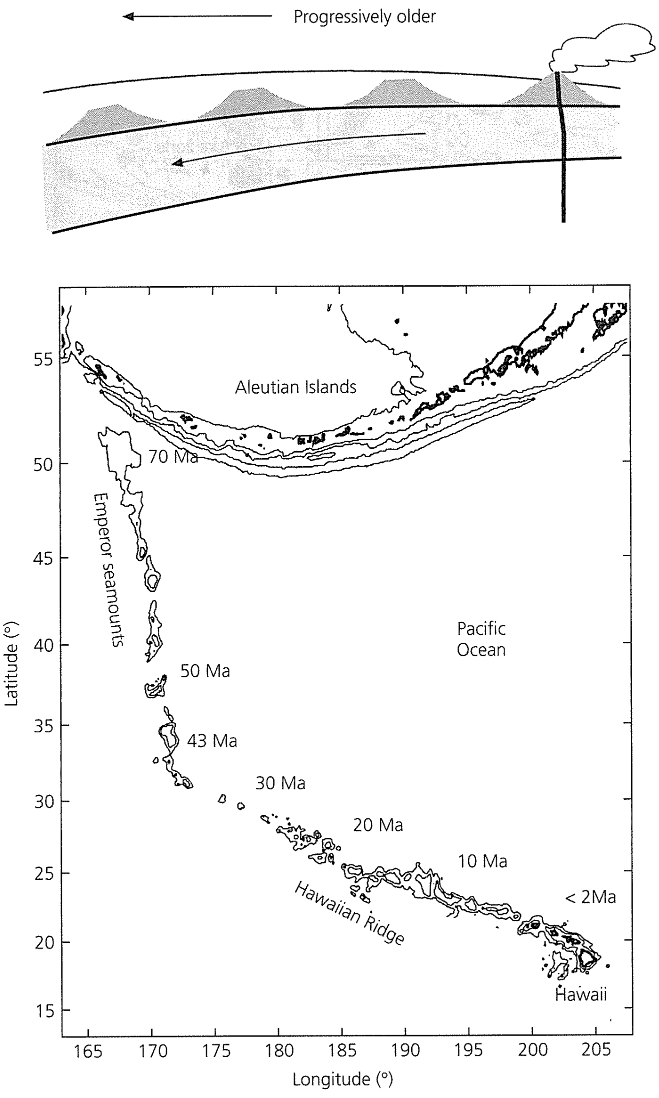


Fig. 5.2-7 *Top*: Illustration of the formation of a volcanic island chain by plate motion over a fixed hot spot. *Bottom*: Ages, in millions of years, of volcanoes in the Hawaiian–Emperor chain.

plate were not moving with respect to the deep mantle. In this case, as lithosphere was added to the plate by sea floor spreading at the Mid-Atlantic ridge (Fig. 5.2-4), both the ridge and the South American plate would move westward with respect to the mantle. Conversely, as the African plate lost area by subduction beneath the Eurasian plate in the Mediterranean, the trench would "roll backward," causing both it and Eurasia to move southward relative to the mantle. Such motions can have important consequences for processes at plate boundaries (e.g. Fig. 5.3-10).

Absolute plate motions cannot be measured directly. Hence we infer these motions in two ways. One uses the *hot spot* hypothesis, in which certain linear volcanic trends result from the motion of a plate over a hot spot, or fixed source of volcanism, which causes melting in the overriding plate (Fig. 5.2-7). If the overriding plate is oceanic, its motion causes a progression from active volcanism that builds the islands, to older islands, to underwater seamounts as the sea floor moves away from the hot spot, cools, and subsides. This process leaves a broad, shallow, topographic swell around the hot spot and a characteristic volcanic age progression away from it, as shown for the Hawaiian–Emperor seamount chain. The ages of volcanism range from present, on the currently active island of Hawaii, to a few million years on the other Hawaiian islands,[7] to about 28 Ma at Midway island, and about 70 Ma where the seamount chain vanishes into the Aleutian trench. Thus the direction and age of the volcanic chain give the motion of the plate with respect to the hot spot. For example, the bend in the Hawaiian–Emperor seamount chain has been interpreted as indicating that the Pacific plate changed direction about 40 million years ago. Hence using hot spot tracks beneath different plates, and assuming that the hot spots are fixed with respect to the deep mantle (or move relative to each other more slowly than plates), yields a hot spot reference frame.

It is often further assumed that hot spots result from plumes of hot material rising from great depth, perhaps even the core-mantle boundary (Fig. 5.1-2). The concepts of hot spots and plumes are attractive and widely used, but the relation between the persistent volcanism and possible deep mantle plumes remains a subject of active investigation because there are many deviations from what would be expected. Some hot spots move significantly, some chains show no clear age progression, evidence for plate motion changes associated with bends like that in Fig. 5.2-7 is weak, and oceanic heat flow data show little or no thermal anomalies at the swells. Seismological studies find low-velocity anomalies, but assessing their depth extent and relation to possible plumes is challenging. However, the hot spot reference frame is similar to one obtained by assuming there is no net rotation (NNR) of the lithosphere as a whole, and hence that the sum of the absolute motion of all plates weighted by their area is zero. Thus despite unresolved questions about the nature and existence of hot spots and plumes, NNR reference frames are often used to infer absolute motions.

To compute absolute motions, we recognize that motions in an absolute reference frame correspond to adding a rotation to all the plates. Thus we use the Euler vector formulation and treat the absolute reference frame as mathematically equivalent to another plate. We define Ω_i as the Euler vector of plate *i* in an absolute reference frame. For example, Table 5.2-1 gives the NNR Euler vector relative to the North American plate ($\boldsymbol{\omega}_{NNR-NA}$), so its negative ($\boldsymbol{\omega}_{NA-NNR}$) is the absolute Euler vector Ω_{NA} for North America in the NNR reference frame. The linear velocity at a point **r** is found by analogy to Eqn 1:

$$\mathbf{v}_i = \Omega_i \times \mathbf{r}. \qquad (12)$$

Thus we find the motion of North America with respect to the hot spot thought to be producing the volcanism and earthquakes in Yellowstone National Park (44°, −110°) to be

[7] This age progression was recognized by native Hawaiians, who attributed it to the order in which the volcano goddess Pele plucked the islands from the sea.

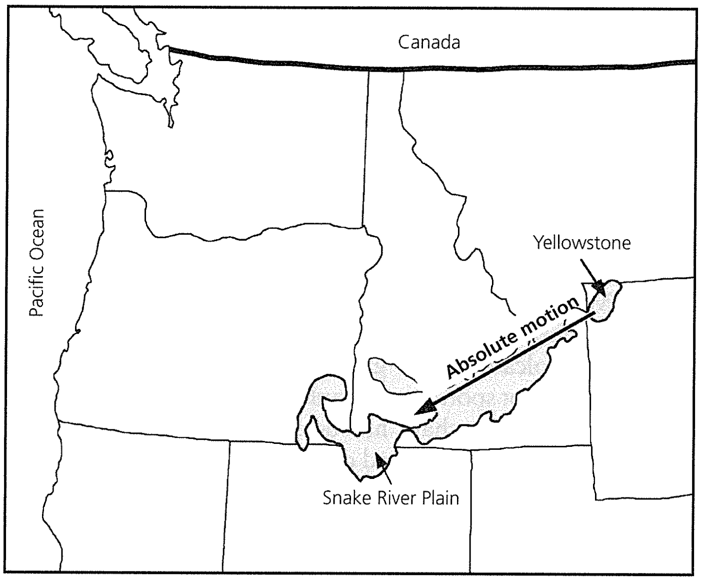


Fig. 5.2-8 Comparison of the predicted absolute motion of North America to the Snake River Plain basalts, which are thought to be the track of a hot spot now producing volcanism in Yellowstone National Park. (After Smith and Braile, 1994. *J. Volcan. Geotherm. Res., 61*, 121–87, with permission from Elsevier Science.)

18 mm/yr directed N239°E. This motion is along the trend connecting the present volcanism in Yellowstone to the Snake River Plain basalts (Fig. 5.2-8), which are thought to be its track, a continental analogy to the Hawaiian–Emperor seamount chain.

Relative and absolute Euler vectors are simply related because

$$\boldsymbol{\omega}_{ij} = \Omega_i - \Omega_j, \quad (13)$$

the relative Euler vector for two plates, is the difference between their absolute Euler vectors. Thus, if we know one plate's absolute motion, we can find all the others from the relative motions. For example, the absolute motion of the Pacific plate can be found from Table 5.2-1, which gives its vector relative to North America, using

$$\Omega_{PA} = \omega_{PA-NA} + \Omega_{NA}. \quad (14)$$

Absolute motions are important in several seismological applications. Seismology is used to study hot spots and their effects, including the resulting intraplate earthquakes like those associated with the volcanism in Hawaii. For example, Fig. 2.8-5 illustrated the use of surface wave dispersion to study the velocity structure under the Walvis ridge, which is thought to be the track produced by a hot spot under the Mid-Atlantic ridge. A second application involves seismic anisotropy in the mantle (Section 3.6), which is thought to reflect flow of olivine-rich material in a direction that is often consistent with the predicted absolute plate motions. Thus seismic anisotropy, seismic velocities, and absolute motions are being combined to model mantle flow.

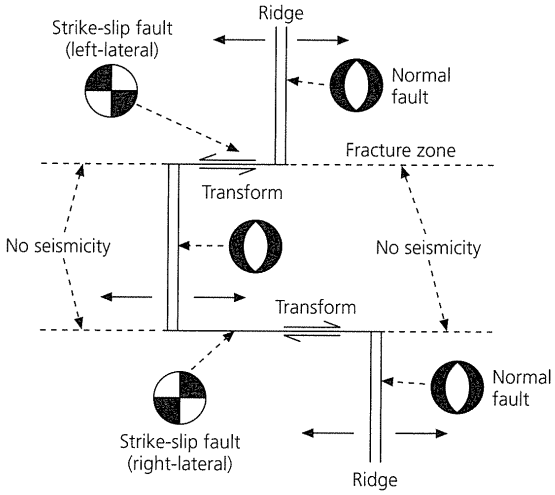


Fig. 5.3-1 Possible tectonic settings of earthquakes at an oceanic spreading center. Most events occur on the active segment of the transform and have strike-slip mechanisms consistent with transform faulting. On a slow-spreading ridge, like the Mid-Atlantic, normal fault earthquakes also occur.

5.3 Spreading centers

Because the lithosphere forms at spreading centers, we begin with an overview of such systems and the earthquakes within them. We will see that seismological observations both demonstrate and reflect the basic kinematic model for ridges and transforms. Moreover, they provide key evidence for the thermal-mechanical processes that control the formation and evolution of the oceanic lithosphere.

5.3.1 Geometry of ridges and transforms

Mid-ocean ridges are marked by earthquakes, which provide important information about the sea floor spreading process. Figure 5.3-1 is a schematic diagram of a portion of a spreading ridge offset by transform faults. Because new lithosphere forms at ridges and then moves away, transform faults are segments of the boundaries between plates, across which lithosphere moves in opposite directions. A given pair of plates can have either right- or left-lateral motion, depending on the direction in which a transform offsets the ridge; both reflect the same direction of relative plate motion. This motion across the transform is not what produced the offset of the ridge crest. In fact, in the usual situation such that spreading is approximately symmetric (equal rates on either side), the length of the transform will not change with time. This is a very different geometry from a transcurrent fault, where the offset between ridge segments is produced by motion on the fault and increases with time.

The focal mechanisms illustrate these ideas. Figure 5.3-2 (*top*) shows a portion of the Mid-Atlantic ridge composed of north–south-trending ridge segments that are offset by transform faults such as the Vema transform that trend approxim-

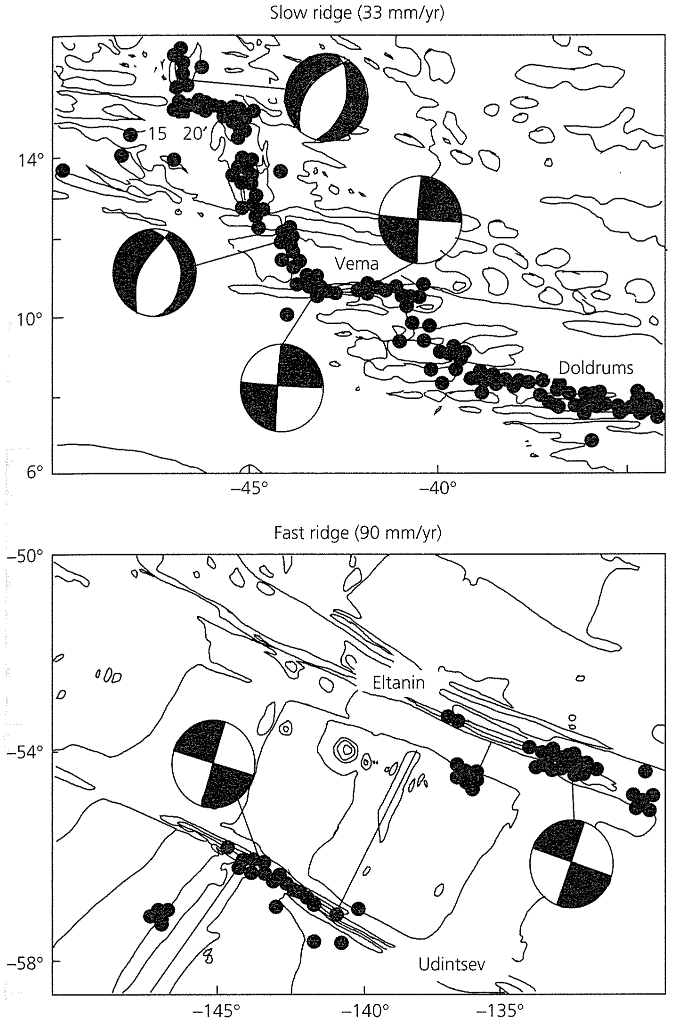


Fig. 5.3-2 Maps contrasting faulting on slow- and fast-spreading centers. *Top*: The slow Mid-Atlantic ridge has earthquakes on both the active transform and the ridge segments. Strike-slip faulting on a plane parallel to the transform azimuth is characteristic. On the ridge segments, normal faulting with nodal planes parallel to the ridge trend is seen. *Bottom*: The fast East Pacific rise has only strike-slip earthquakes on the transforms. (Stein and Woods, 1989.)

ately east–west. Both the ridge crest and the transforms are seismically active. The mechanisms show that the relative motion along the transform is right-lateral. Sea floor spreading must be occurring on the ridge segments to produce the observed relative motion. For this reason, earthquakes occur almost exclusively on the active segment of the transform fault between the two ridge segments, although an inactive extension known as a fracture zone extends to either side. Although no relative plate motion occurs on the fracture zone,[1] it is often marked by a topographic feature due to the contrast in lithospheric ages across it.

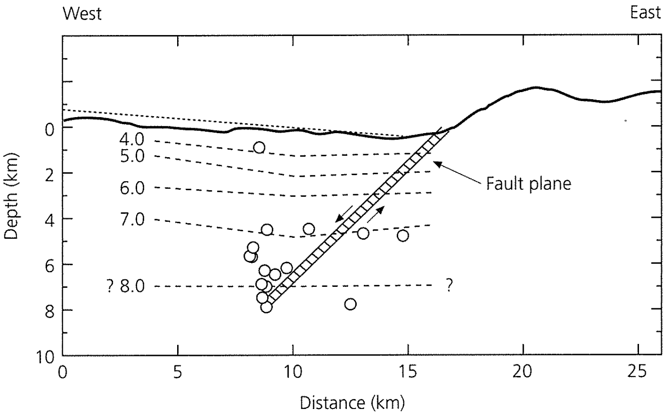


Fig. 5.3-3 Cross-section through the Mid-Atlantic ridge. The fault plane inferred from the focal mechanisms of large earthquakes is consistent with the locations of microearthquakes (dots) determined using ocean bottom seismometers. Dashed lines show P-wave velocity structure. (Toomey *et al.*, 1988. *J. Geophys. Res., 93*, 9093–112, copyright by the American Geophysical Union.)

Earthquakes also occur on the spreading segments. Their focal mechanisms show normal faulting, with nodal planes trending approximately along the ridge axis. These normal fault earthquakes are thought to be associated with the formation of the axial valley. For example, Fig. 5.3-3 shows a cross-section through the Mid-Atlantic ridge. The fault planes inferred from teleseismic focal mechanisms and the locations of microearthquakes determined using ocean bottom seismometers are consistent with normal faulting along the east side of the valley. Slip on this fault over 10,000 years would be enough to produce the observed geometry, including the eastward tilt of the valley floor.

The seismicity differs along the East Pacific rise. Here (Fig. 5.3-2, *bottom*) earthquakes occur on the transform faults with the expected strike-slip mechanisms, but few earthquakes occur on the ridge crest. This is probably because the East Pacific rise has an axial high, rather than the axial valley that occurs at the Mid-Atlantic ridge.[2] This difference appears to reflect the spreading rates: ridges spreading at less than about 60 mm/yr usually have axial valleys, whereas faster-spreading ridges have axial highs and thus do not have ridge crest normal faulting.

These examples show the spreading process at its simplest, but there can be complexities. Spreading can be asymmetric (one flank faster than the other) or oblique, such that the spreading is not perpendicular to the ridge axis. In addition, the geometry of a ridge system can change with time, as discussed in Section 5.3.3.

5.3.2 Evolution of the oceanic lithosphere

To understand the difference between fast- and slow-spreading ridges, and the nature of the earthquakes associated with them, it is important to understand the evolution of the oceanic

[1] Unfortunately, some transform faults named before this distinction became clear are known as "fracture zones" along their entire length.

[2] This is often shown incorrectly on older maps.

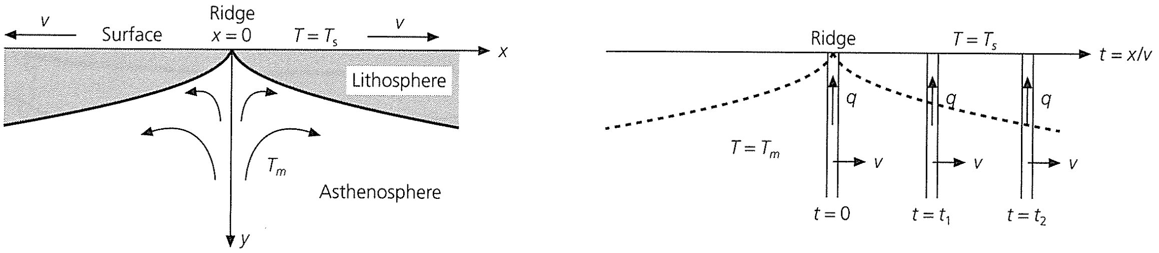


Fig. 5.3-4 Model for the cooling of an oceanic plate as it moves away from the ridge axis (*left*). Because a column moves away from the ridge faster than heat is conducted in the horizontal direction (*right*), the cooling in the vertical direction can be treated as a one-dimensional problem. (After Turcotte and Schubert, 1982.)

lithosphere. This process can be described using a simple, but powerful, model for the formation of the lithosphere by hot material at the ridge, which cools as the plate moves away.

In this model, material at the ridge at a mantle temperature T_m (1300–1400 °C) is brought to the ocean floor, which has a temperature T_s. The material then moves away at a velocity v, while its upper surface remains at T_s (Fig. 5.3-4). Because the plate moves away from the ridge faster than heat is conducted horizontally, we can consider only vertical heat conduction. Mathematically, this is the same as the cooling of a halfspace originally at temperature $T = T_m$, whose surface is suddenly cooled to T_s at time $t = 0$.

The temperature as a function of depth and time is given by the one-dimensional heat flow equation, which relates the temperature change with time in a piece of material to the rate at which heat is conducted out of it,

$$\frac{\partial T(z,t)}{\partial t} = \frac{k}{\rho C_p}\frac{\partial^2 T(z,t)}{\partial z^2} = \kappa\frac{\partial^2 T(z,t)}{\partial z^2}. \qquad (1)$$

κ, known as the *thermal diffusivity*, is a property of the material that measures the rate at which heat is conducted. It has units of distance squared divided by time, and is defined as $\kappa = k/\rho C_p$, where k is the thermal conductivity, ρ is the density, and C_p is the specific heat at constant pressure.

The well known solution to Eqn 1 is

$$T(z,t) = T_s + (T_m - T_s)\,\mathrm{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right), \qquad (2)$$

where

$$\mathrm{erf}\,(s) = \frac{2}{\sqrt{\pi}}\int_0^s e^{-\sigma^2}d\sigma \qquad (3)$$

is known as the error function. Figure 5.3-5 (*right*) shows how this function varies between erf (0) = 0 and erf (3) ≈ 1. Thus cooling starts at the surface and deepens with time (Fig. 5.3-5, *left*).

Assuming that any column of oceanic lithosphere cools this way, and that the sea floor temperature is $T_s = 0$ °C, then

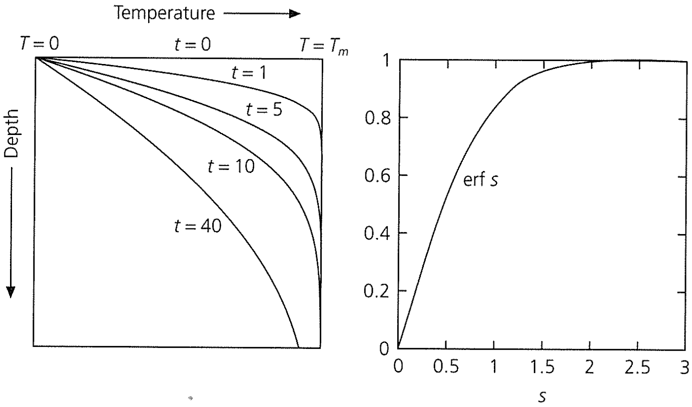


Fig. 5.3-5 *Left*: Cooling of a halfspace as described by the one-dimensional heat flow equation. The surface is cooled at time zero, and then the interior cools with time. *Right*: The error function, which controls the cooling solution shown.

$$T(z,t) = T_m\,\mathrm{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right) \qquad (4)$$

gives the temperature at a depth z for material of age t. The lithosphere moves away from the ridge at half the total spreading rate, so the age of the lithosphere is $t = x/v$, its distance from the ridge divided by the half-spreading rate v. Thus the temperature (Eqn 4) as a function of distance and depth is

$$T(x,z) = T_m\,\mathrm{erf}\left(\frac{z}{2\sqrt{\kappa x/v}}\right). \qquad (5)$$

It is useful to think of *isotherms*, lines of constant temperature, in the plate. An isotherm is a curve on which the argument of the error function is constant,

$$\frac{z_c}{2\sqrt{\kappa t}} = c, \quad \text{or} \quad z_c = 2c\sqrt{\kappa t}, \qquad (6)$$

so that the depth to a given temperature increases as the square root of the lithospheric age.

This is an example of a general feature of heat conduction problems: setting $c = 1$ and examining Fig. 5.3-5 for erf (1)

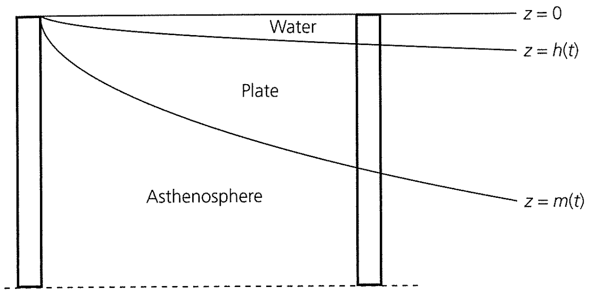


Fig. 5.3-6 The increase in ocean depth with lithospheric age due to the cooling of the lithosphere can be modeled using isostasy, the assumption that the mass in a vertical column is the same for all ages.

shows that most of the temperature change has propagated a distance $2\sqrt{\kappa t}$ in a time t. For example, after a lava flow erupts, it cools as the square root of time. Such square root of time behavior occurs for any process described by a diffusion equation, of which the heat equation is an example.

The concept that the lithosphere cools with time such that isotherms deepen with the square root of age has many observable consequences. The simplest is that ocean depth should vary with age, which makes sense, because spreading centers are ridges precisely because the ocean deepens on either side. To model this effect, we consider the mass in two columns, one at the ridge and one at age t, and invoke the idea of isostasy, which means that the masses in the two columns balance (Fig. 5.3-6).[3]

Assume that the lithosphere, defined by the $T = T_m$ isotherm, has thickness zero at the ridge and $z = m(t)$ at age t, where the water depth is $h(t)$. Similarly, we assume that the asthenosphere is at temperature T_m and has density ρ_m. However, the temperature and thus density in the cooling lithosphere vary, such that at the point (z, t) the temperature is $T(z, t)$ and the corresponding density is

$$\rho(z,t) \approx \rho_m + \frac{\partial \rho}{\partial T}[T(z,t) - T_m] = \rho_m + \rho'(z,t). \tag{7}$$

The change in density due to temperature, at constant pressure, is given by the coefficient of thermal expansion,

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P = -\frac{1}{\rho}\left(\frac{\partial \rho}{\partial T}\right)_P \tag{8}$$

(the minus sign is because $\partial\rho/\partial T$ is negative). Thus the density perturbation for the halfspace cooling model is

$$\rho'(z,t) = \alpha\rho_m[T_m - T(z,t)] = \alpha\rho_m T_m\left[1 - \operatorname{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right)\right]. \tag{9}$$

If the density of water is ρ_w, equal mass in the two columns requires that

$$\rho_m m(t) = \rho_w h(t) + \int_{h(t)}^{m(t)} [\rho_m + \rho'(z,t)]dz, \tag{10}$$

which gives the isostatic condition for ocean depth,

$$h(t) = \frac{1}{(\rho_m - \rho_w)} \int_{h(t)}^{m(t)} \rho'(z,t)dz. \tag{11}$$

Because temperature and density in the plate are defined for all values of z (the thickness of the plate is defined as some chosen isotherm), let $z' = z - h(t)$ and $m(t) \to \infty$. Then

$$h(t) = \frac{\alpha\rho_m T_m}{(\rho_m - \rho_w)} \int_0^\infty \left[1 - \operatorname{erf}\left(\frac{z'}{2\sqrt{\kappa t}}\right)\right] dz'. \tag{12}$$

To evaluate the integral, substitute $s = z'/2\sqrt{\kappa t}$ and integrate by parts (try it!) to show that

$$\int_0^\infty [1 - \operatorname{erf}(s)]\, ds = 1/\sqrt{\pi}. \tag{13}$$

Thus ocean depth should increase as the square root of plate age,

$$h(t) = 2\sqrt{\frac{\kappa t}{\pi}} \frac{\alpha\rho_m T_m}{(\rho_m - \rho_w)}. \tag{14}$$

The cooling of the lithosphere should also cause heat flow at the sea floor to vary with age. By Fourier's law of heat conduction, the heat flow at the sea floor is the product

$$q = k\frac{dT}{dz} \quad \text{at} \quad z = 0 \tag{15}$$

of the temperature gradient at the sea floor and the thermal conductivity k.[4] An easy approximation to see how heat flow varies with age is to consider the T_m isotherm as the base of the lithosphere, so that the thickness of the lithosphere increases

[3] Isostasy is the general idea that topography results from equal masses in different columns. Here we consider thermal isostasy, in which density changes produced by temperature variations cause topographic differences. Another common model, Airy isostasy, is used to explain the relation between crustal thickness variations and topography, such as crustal roots under mountains.

[4] Normally, this equation requires a minus sign because heat flows from hot objects to cold ones. Without this sign, hot objects would get hotter. There is none here because of our customary but inconsistent definitions: heat flow is measured upward whereas depth is measured downward.

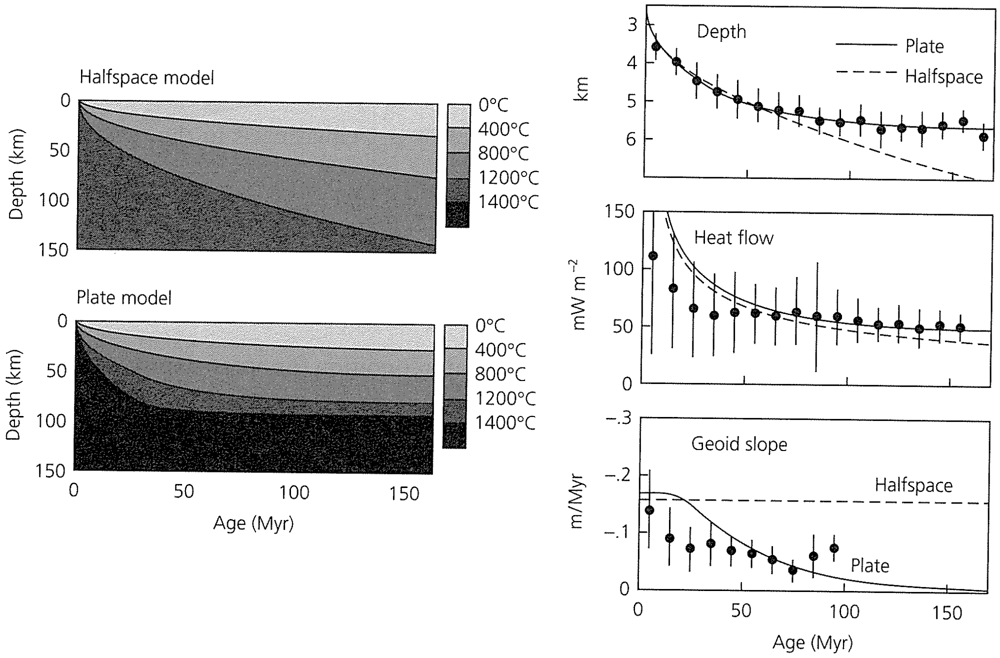


Fig. 5.3-7 Models and data for thermal evolution of the oceanic lithosphere. *Left*: Isotherms for thermal models. The lithosphere continues cooling for all ages in a halfspace model, but equilibrates for ~70 Ma lithosphere in a plate model with a 95 km-thick thermal lithosphere. The plate model shown has a higher basal temperature than the halfspace model. *Right*: Comparison of thermal model predictions to different data. All show a lithospheric cooling signal, and are better (but far from perfectly) fit by the predictions of a plate model (solid lines) than by those of a halfspace model (dashed lines). (Richardson *et al.*, 1995. *Geophys. Res. Lett.*, *22*, 1913–16, copyright by the American Geophysical Union.)

with the square root of age. Approximating the gradient at the surface by the average gradient through the lithosphere,

$$q(t) \approx k\frac{\Delta T}{\Delta z} \approx \frac{kT_m}{\sqrt{\kappa t}} \qquad (16)$$

predicts that the heat flow decreases as the square root of age. The same result can be obtained by differentiation of the temperature structure (Eqn 4) using

$$\frac{d}{dz}\mathrm{erf}(s) = \frac{d}{dz}\frac{2}{\sqrt{\pi}}\int_0^s e^{-\sigma^2}d\sigma = \frac{2}{\sqrt{\pi}}e^{-s^2}\frac{ds}{dz}, \qquad (17)$$

which gives

$$q(t) = k\left.\frac{dT}{dz}\right|_{z=0} = k\left.\frac{2T_m}{\sqrt{\pi}}e^{\frac{-z^2}{4\kappa t}}\frac{1}{2\sqrt{\kappa t}}\right|_{z=0} = \frac{kT_m}{\sqrt{\pi\kappa t}}. \qquad (18)$$

This model, which predicts that lithospheric thickness, heat flow, and ocean depth vary as the square root of age for all ages is called a halfspace model (Fig. 5.3-7, *upper left*). In it, the lithosphere is the upper layer of a halfspace that continues cooling for all time. (In reality, oceanic lithosphere never gets older than 200 million years old because it gets subducted.) The model does a good job of describing the average variation in ocean depth and heat flow with lithospheric age.

However, because ocean depth seems to "flatten" at about 70 Myr, we often use a modification called a plate model (Fig. 5.3-7, *lower left*), which assumes that the lithosphere evolves toward a finite plate thickness L with a fixed basal temperature T_m. In this model,

$$T(x,z) = T_m\left[\frac{z}{L} + \sum_{n=1}^{\infty} c_n \exp\left(-\frac{\beta_n x}{L}\right)\sin\left(\frac{n\pi z}{L}\right)\right], \qquad (19)$$

where $c_n = 2/(n\pi)$, $\beta_n = (R^2 + n^2\pi^2)^{1/2} - R$, $R = vL/(2\kappa)$. The constant R, known as the thermal Reynolds number, relates the rates at which heat is transported horizontally by plate motion and conducted vertically. In this model isotherms initially deepen as the square root of age, but eventually level out. The flattening reflects the fact that heat is being added from below, which the model approximates by having old lithosphere reach a steady-state thermal structure that is simply a linear geotherm (Fig. 5.3-8, *top*). As a result, the predicted sea floor depth and heat flow also behave for young ages like in the halfspace model, but evolve asymptotically toward constant values for old ages. Both have simple interpretations: the heat flow is proportional to the geotherm, and thus T_m/L, whereas the depth is proportional to the thermal subsidence and hence heat lost since the plate formed at the ridge, and thus the product T_mL. The model parameters can be estimated by an inverse problem, finding those that best fit a set of depth and heat flow data versus age (Fig. 5.3-8, *bottom*).

Comparison with data shows that the plate thermal model is a good, but not perfect, fit to the average data because processes other than this simple cooling are also occurring. For example, ocean depth is also affected by uplift associated with hot spots (Section 5.2.4). Water flow in the crust transports some of the heat for ages less than about 50 Ma, making the observed heat flow lower than the model's predictions, which assume that all heat is transferred by conduction. Some topographic effects, including the spectacular volcanic oceanic plateaus, result from crustal thickness variations. Because these and other effects vary from place to place, the data vary about their average values for a given age.

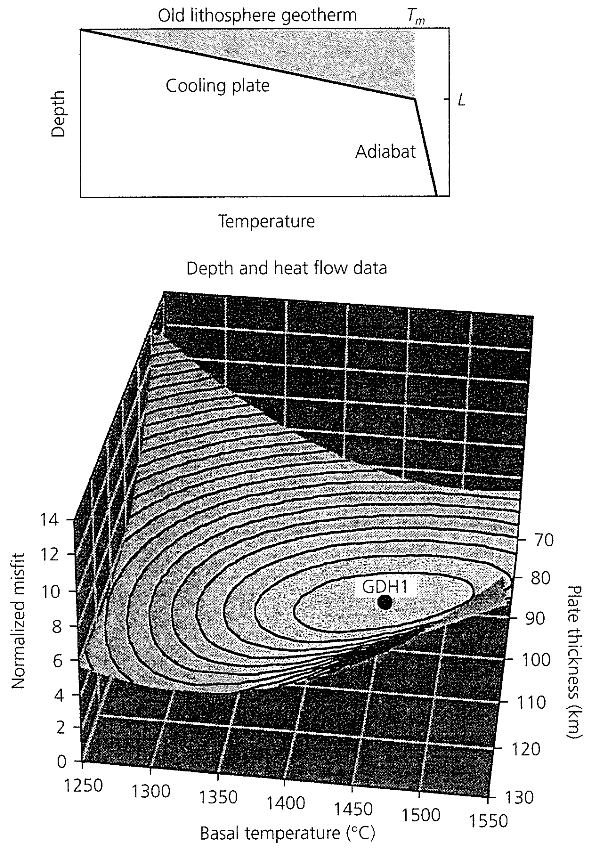


Fig. 5.3-8 *Top*: Asymptotic thermal structure for old lithosphere in a plate model. The sea floor subsidence from the ridge, and thus ocean depth, is proportional to the shaded area between the geotherm and $T = T_m$, whereas heat flow is proportional to the geotherm. A schematic adiabatic temperature gradient (Section 5.4.1) is shown beneath the plate. (Stein and Stein, 1992. Reproduced with permission from *Nature*.) *Bottom*: Fitting process used for thermal model parameters. The misfit to a set of depth and heat flow data has a minimum at the point labeled GDH1, a plate thermal thickness of 95 ± 15 km and basal temperature of 1450 ± 250°C. (Stein and Stein, 1996. *Subduction*, 1–17, copyright by the American Geophysical Union.)

Table 5.3-1 Constraints on thermal models $T(z, t)$.

Observable	Proportional to	Reflects
Young ocean depth	$\int T(z,t)dz$	$k^{1/2}\alpha T_m$
Old ocean depth	$\int T(z,t)dz$	$\alpha T_m L$
Old ocean heat flow	$\left.\dfrac{\partial T(z,t)}{\partial z}\right\vert_{z=0}$	kT_m/L
Geoid slope	$\dfrac{\partial}{\partial t}\int zT(z,t)dz$	$k\alpha T_m \exp(-kt/L^2)$

Source: Stein and Stein (1996).

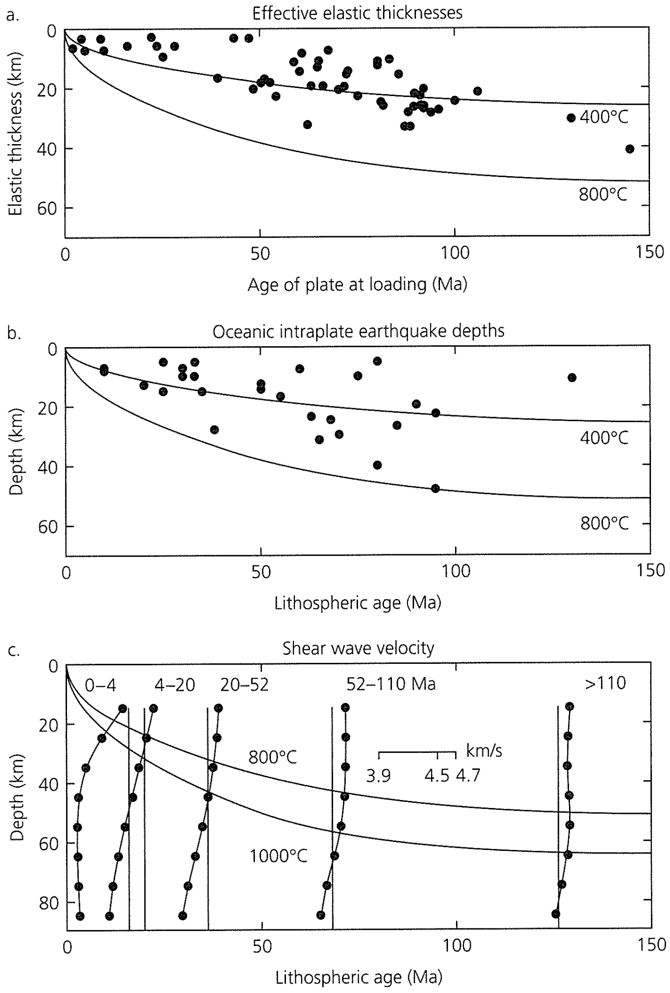


Fig. 5.3-9 Comparison of isotherms as functions of age for a plate model to three datasets whose variation with age is consistent with cooling of the lithosphere. The effective elastic thickness (a), deepest intraplate seismicity (b), and depth to the low-velocity zone, shown by velocity profiles at different ages (c), all increase with age. (After Stein and Stein, 1992. Reproduced with permission from *Nature*.)

We can view ocean depth, heat flow, and several other properties of the oceanic lithosphere as observable measures of the temperature in the cooling lithosphere. Because the observables depend on different combinations of parameters (Table 5.3-1), they can be used together to constrain individual parameters (a halfspace model corresponds to an infinitely thick plate). The depth depends on the integral of the temperature (Eqn 11), whereas the heat flow depends on its derivative at the sea floor (Eqn 15). Similarly, the slope of the geoid, a function of the gravity field depending on a weighted integral of the density, also varies with age in general agreement with the plate model's prediction (Fig. 5.3-7).

In addition, the elastic thickness of the lithosphere inferred from the deflection caused by loads such as seamounts (Fig. 5.3-9a), the maximum depth of intraplate earthquakes within the oceanic lithosphere (Fig. 5.3-9b), and the depth to the low-velocity zone determined from surface wave dispersion (Figs. 5.3-9c and 2.8-7), all increase with age. Hence the cooling of oceanic lithosphere causes the expected increase in strength and seismic velocity. Moreover, as discussed in Section 5.5, the resulting density increase is thought to provide a major force driving plate motions.

Because various properties vary with age, the oceanic lithosphere can be defined in various ways, so terms like "seismic lithosphere," "elastic lithosphere," and "thermal lithosphere" are often used. Interestingly, these thicknesses differ. It looks as if the deepest earthquakes are bounded by about 600–800 °C, such that hotter material cannot support seismic failure. The

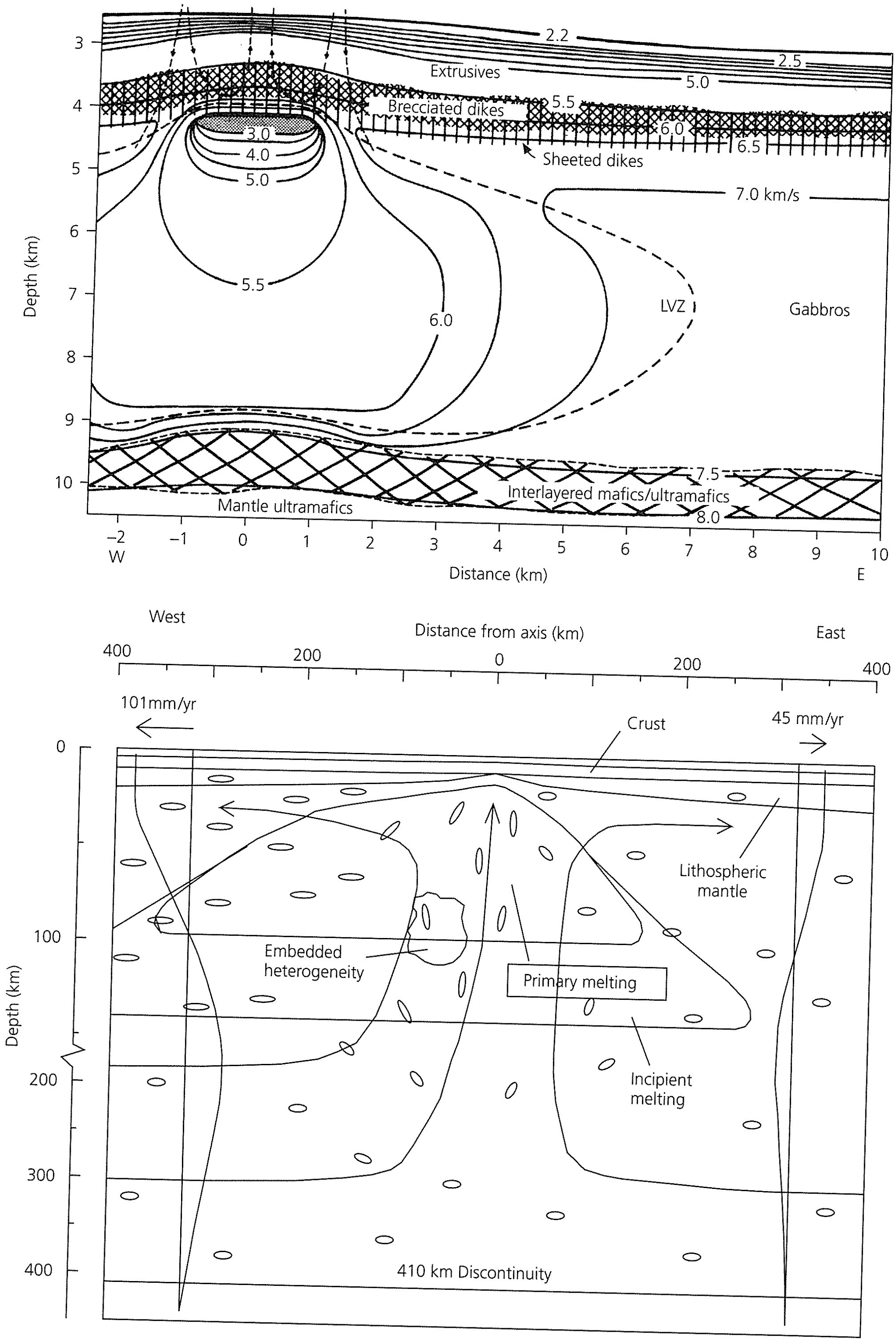


Fig. 5.3-10 *Top*: Geological interpretation of a multichannel seismic velocity study on the East Pacific rise. A low-velocity region under the axis is interpreted as a hot region of melting, capped by a magma lens. Dashed lines are possible paths of water circulation. (Vera *et al.*, 1990. *J. Geophys. Res., 95*, 15,529–56, copyright by the American Geophysical Union.) *Bottom*: Schematic cross-section across the East Pacific rise. The broad region of low velocities is interpreted as the primary melting region. Small ellipses are directions of preferred olivine alignment inferred from anisotropy. Lines with arrows indicate inferred mantle flow, causing the distortion shown of an initially vertical line. Absolute velocities of the two plates (Pacific on left, Nazca on right) are given by small horizontal arrows. (Forsyth *et al.*, 1998. *Science, 280*, 1215–18, copyright 1998 American Association for the Advancement of Science.)

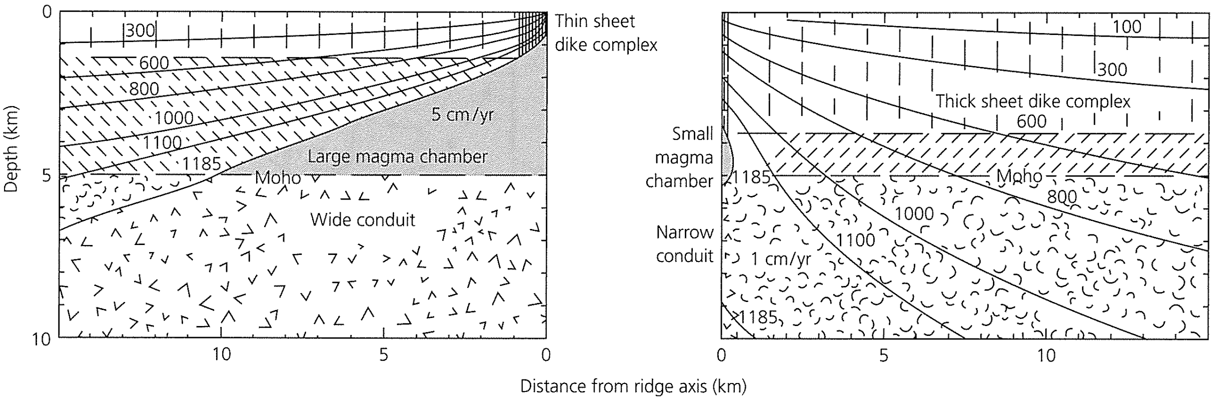


Fig. 5.3-11 Thermal and petrological model for the difference between fast-spreading (*left*) and slow-spreading (*right*) ridges. (Sleep and Rosendahl, 1979. *J. Geophys. Res.*, *84*, 6831–9, copyright by the American Geophysical Union.)

elastic thickness corresponds approximately to the 400 °C isotherm, whereas the low-velocity zone begins approximately below the 1000 °C isotherm (Fig. 5.3-9c). These differences, discussed in Section 5.7, likely result from rock being stronger for more rapid deformation. All of these thicknesses, however, only approximate what we would like to know but cannot directly measure: the depth of the base of the moving plate, which is likely to be a gradational rather than a distinct boundary.

5.3.3 Ridge and transform earthquakes and processes

Seismology makes important contributions to understanding the properties and behavior of spreading centers. Ocean bottom seismometers yield locations of microearthquakes and data for travel time and waveform studies. Larger earthquakes are also studied using teleseismic body and surface waves. The seismological results are being integrated with marine geophysical and petrological data to develop better models. For example, Fig. 5.3-10 (*top*) shows a geological interpretation of a multichannel seismic study (Section 3.3) that used air gun and explosive sources to image velocity structure under the East Pacific rise to a depth of about 10 km. A low-velocity region under the axis is interpreted as a hot melting region capped by a magma lens. Other studies using ocean bottom seismometers and distant earthquake sources map the structure to greater depth, including inferring flow directions under the ridge axis using anisotropy (Fig. 5.3-10, *bottom*). Such studies are finding interesting features of the spreading process. For example, the broad region of low velocity presumed to be the primary melting area extends further west than east of the axis. This asymmetry may occur because the westward absolute motion of the Pacific plate is much faster than the eastward absolute motion of the Nazca plate, causing the ridge to migrate westward relative to the deep mantle. Thus the spreading process, which depends on the relative plate motion (spreading rate), also seems affected by the absolute motion.

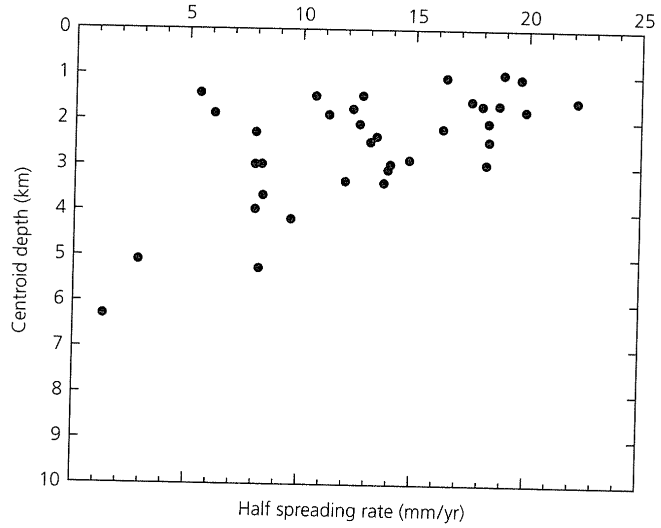
Some effects of the spreading rate are illustrated by a model shown in Fig. 5.3-11. At a given distance from the ridge, faster spreading produces younger lithosphere and isotherms closer to the surface than does slow spreading. If the region beneath the 1185 °C isotherm and above the Moho depth of 5 km is considered to be a magma chamber, a fast ridge has a larger magma chamber. Hence crust moving away from a fast-spreading ridge is more easily replaced than that moving away from a slow ridge. Thus, in contrast to the axial valley and normal faulting earthquakes on a slow ridge, a fast ridge has an axial high and an absence of earthquakes. Similarly, both the depths and the maximum seismic moments[5] of ridge crest normal faulting earthquakes decrease with spreading rate (Fig. 5.3-12). These observations are consistent with the fault area decreasing on faster-spreading and hotter ridges, because faulting requires that rock be below a limiting temperature, above which it flows (Section 5.7). The idea that the faulting depends on temperature is also implied by the increase in the maximum depth of oceanic intraplate earthquakes with age (Fig. 5.3-9b).

Transform fault earthquakes also depend on thermal structure. The temperatures along a transform fault should be essentially the average of the expected temperature on the two sides; coolest at the transform midpoint and hottest at either end (Fig. 5.3-13). As expected from the area available for faulting, the maximum seismic moment for transform earthquakes decreases with spreading rate (Fig. 5.3-14), consistent with the idea of faulting limited to a zone bounded by the isotherms.

An interesting question is how the seismic moments of transform earthquakes relate to the plate motion. The average slip rate from earthquakes can be inferred from the total seismic moment released on a transform, assuming that

$$\text{seismic slip rate} = \frac{\text{total seismic moment}}{(\text{fault area})(\text{rigidity})(\text{time period})}. \qquad (20)$$

[5] Recall (Section 4.6) that the seismic moment is the product of the rigidity, the slip in the earthquake, and the fault area.



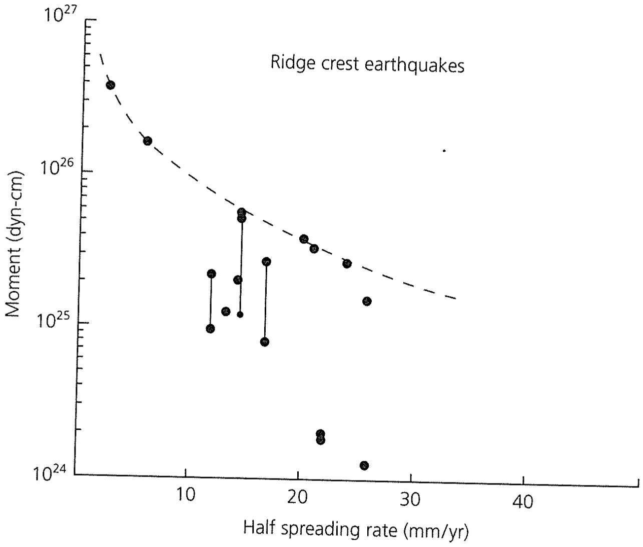


Fig. 5.3-12 *Left*: Shallowing of focal depth for ridge crest normal fault earthquakes with half-spreading rate. (After Huang and Solomon, 1988. *J. Geophys. Res.*, *93*, 13, 445–77, copyright by the American Geophysical Union.) *Right*: Corresponding decrease in maximum seismic moment. (After Solomon and Burr, 1979. *Tectonophysics*, *55*, 107–26, with permission from Elsevier Science.)

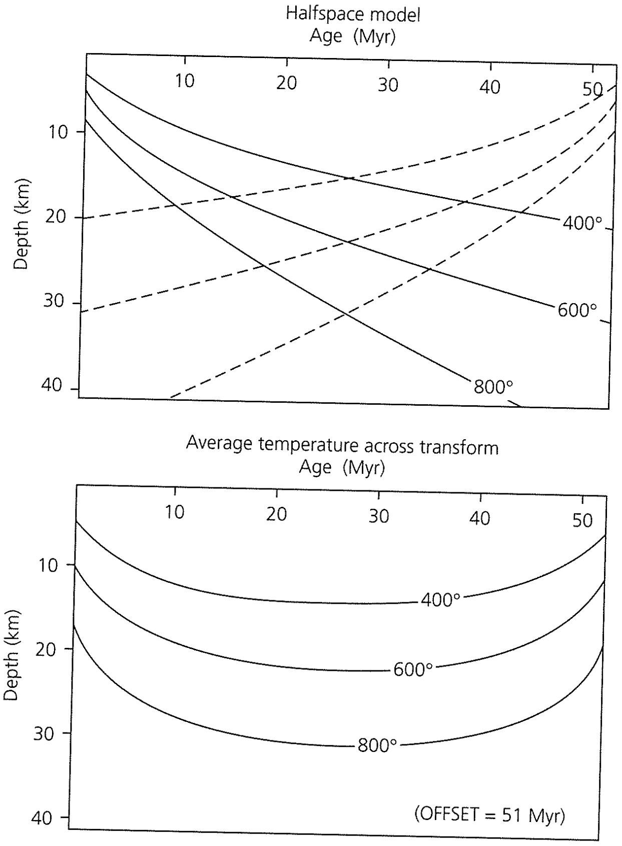


Fig. 5.3-13 Thermal model of the Romanche Transform.
Top: Temperatures on either flank predicted by the cooling halfspace model. *Bottom*: Average temperature distribution along the transform. (After Engeln *et al.*, 1986. *J. Geophys. Res.*, *91*, 548–77, copyright by the American Geophysical Union.)

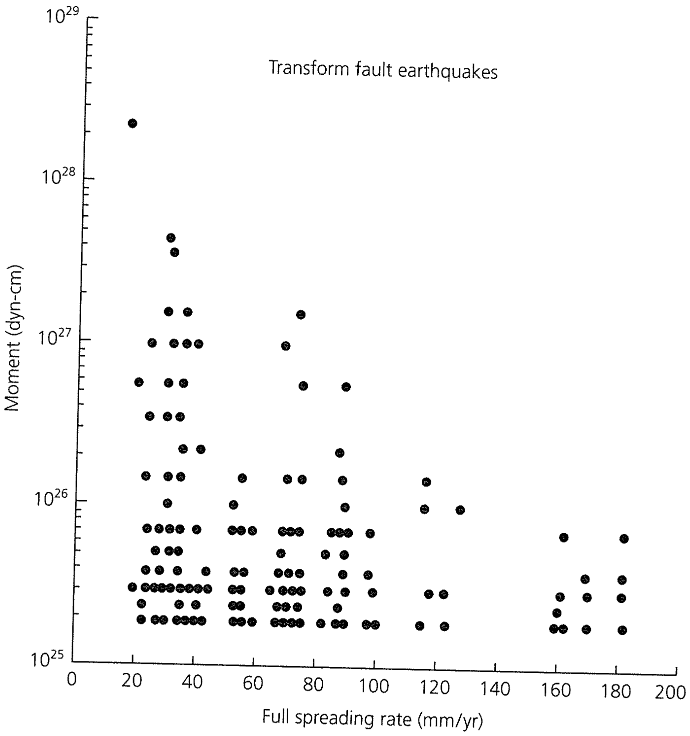


Fig. 5.3-14 Seismic moment versus spreading rate for oceanic transforms. The maximum moment decreases with spreading rate, as expected from thermal considerations. (After Solomon and Burr, 1979. *Tectonophysics*, *55*, 107–26, with permission of Elsevier Science.)

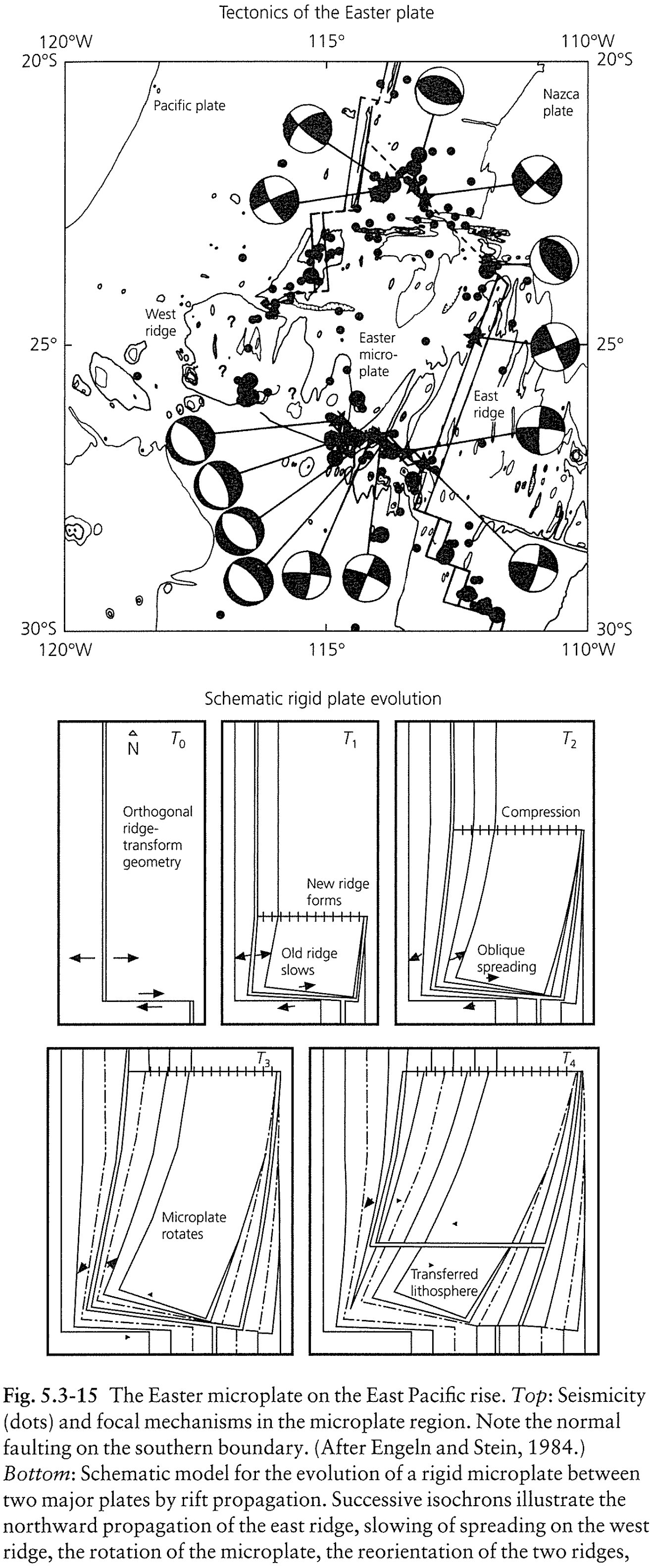


Fig. 5.3-15 The Easter microplate on the East Pacific rise. *Top*: Seismicity (dots) and focal mechanisms in the microplate region. Note the normal faulting on the southern boundary. (After Engeln and Stein, 1984.) *Bottom*: Schematic model for the evolution of a rigid microplate between two major plates by rift propagation. Successive isochrons illustrate the northward propagation of the east ridge, slowing of spreading on the west ridge, the rotation of the microplate, the reorientation of the two ridges, and the conversion of the initial transform into a slow and obliquely spreading ridge. (Engeln *et al.*, 1988. *J. Geophys. Res.*, *93*, 2839–56, copyright by the American Geophysical Union.)

Using this relation requires inferring the fault area, which depends on both the transform length and the depth to which faulting occurs. Assuming the area above the 600–700 °C isotherms fails seismically, the seismic slip rate for major Atlantic transforms is generally less than predicted by the plate motion. Thus, if the time period sampled is long enough to be representative — a major question — some of the plate motion occurs aseismically. The issue of how much slip occurs seismically remains unresolved, as we will see when we discuss subduction zones (Section 5.4.3) and intraplate deformation zones (Section 5.6.2).

In addition, seismology helps study how ridge-transform systems evolve. For example, the East Pacific rise near Easter Island contains two approximately parallel sections (Fig. 5.3-15, *top*). Earthquakes occur on these ridges, but not between them, suggesting that the area in between is an essentially rigid microplate. The normal fault earthquakes on the microplate's southern boundary are surprising because the East Pacific rise here is a very fast-spreading (15 cm/yr) ridge, which should not have normal fault earthquakes (Fig. 5.3-12). Magnetic anomalies show that the east ridge segment is propagating northward and taking over from the old (west) ridge segment. Figure 5.3-15 (*bottom*) shows a simplified model of this process. Because finite time is required for the new ridge to transfer spreading from the old ridge, both ridges are active at the same time, and the spreading rate on the new ridge is very slow at its northern tip and increases southward. As a result, the microplate rotates, causing compression (thrust faulting) and extension (normal faulting) at its north and south boundaries, respectively. Ultimately the old ridge will die, transferring lithosphere originally on the Nazca plate to the Pacific plate, and leaving inactive fossil ridges on the sea floor. Both V-shaped magnetic anomalies characteristic of ridge propagation and fossil ridges are widely found in the ocean basins, showing that this is a common way that ridges reorganize. Even for smaller (a few km) propagating ridge systems, studies of the associated earthquakes can yield useful information about the propagation process.

5.4 Subduction zones

We have seen that earthquakes at spreading centers, which at shallow depths are upwelling limbs of the mantle convection system, reflect the processes forming oceanic lithosphere there. In a similar way, earthquakes at subduction zones, downwelling limbs of the convection system, reflect the processes by which oceanic lithosphere reenters the mantle. Plate convergence takes different forms, depending on the plates involved. Figure 5.4-1 shows the basic model for a situation where oceanic lithosphere of one plate subducts beneath oceanic lithosphere of the overriding plate. Typically, a volcanic island arc forms, and sea floor spreading occurs behind the arc, forming a back-arc basin or marginal sea. Earthquakes occur both at the trench and to great depth, forming a dipping

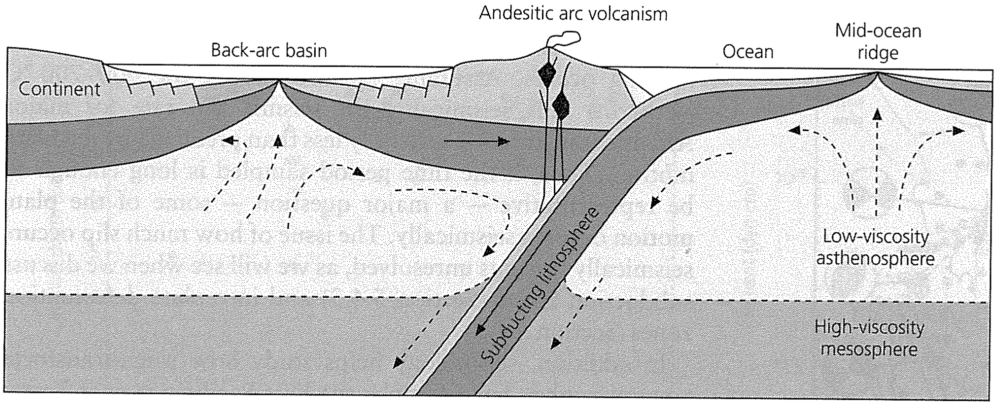


Fig. 5.4-1 Schematic diagram of processes associated with the subduction of one oceanic plate beneath another.

Wadati–Benioff zone. By contrast, when oceanic lithosphere subducts beneath a continent, a mountain chain like the Andes forms on the continent, and the oceanic lithosphere forms a Wadati–Benioff zone. Finally, because continental crust cannot subduct, convergence between two continental plates, as in the Himalayas, causes crustal thickening, mountain building, and shallow earthquakes but does not create a Wadati–Benioff zone.

Subduction zones have a wide variety of earthquakes with different focal mechanisms and depths. There are shallow (less than 70 km deep), intermediate (70–300 km deep), and deep (more than 300 km deep) focus earthquakes.[1] These earthquakes occur in different tectonic environments. The intermediate and deep earthquakes forming the Wadati–Benioff zone occur in the cold interiors of downgoing slabs. The shallow earthquakes are associated with the interaction between the two plates. The largest and most common of these shallow earthquakes occur at the interface between the plates, and release the plate motion that has been locked at the plate interface. In addition, shallow earthquakes can occur within both the overriding and the subducting plates. Figure 5.4-2 shows some features of seismicity observed in subduction zones. Not all features have been observed at all places. For example, the dips and shapes of subduction zones vary substantially. Some show double planes of intermediate or deep seismicity, whereas others do not.

In discussing subduction zones, we follow an approach similar to that used in the last section for ridges. We introduce thermal models for subduction, then use them to gain insight into earthquake and seismic velocity observations. We will see that seismological observations, thermal models, and calculations of the behavior of materials at high temperature and pressure are combined to investigate these complicated regions. In general, the seismological observations are fairly clear, but they can be interpreted in terms of a variety of models. As a result, subduction zone studies remain active, fruitful, and exciting.

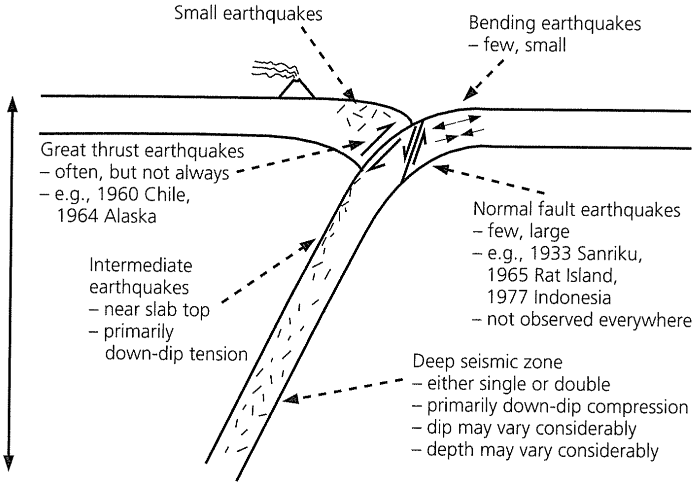


Fig. 5.4-2 Composite subduction zone showing some earthquake types. Not all are observed at all subduction zones.

5.4.1 Thermal models of subduction

The essence of subduction is the penetration and slow heating of a cold slab of lithosphere as it descends into the warmer mantle. As we will see, slabs subduct rapidly compared to the time needed for heat conducted from the surrounding mantle to warm them up. Thus they remain colder, denser, and mechanically stronger than the surrounding mantle. Consequently, slabs transmit seismic waves faster and with less attenuation than the surrounding mantle, making it possible to map slabs and to show that deep earthquakes occur within them. Moreover, the negative thermal buoyancy of cold slabs appears to be the primary force driving plate motions and provides a major source of stress within them that causes deep earthquakes.

To explore the thermal evolution of slabs, we use two approaches. First, we discuss a simplified analytic thermal model that allows insights into the physics. We then discuss numerical models that incorporate additional effects in the hope of providing a more realistic description. We highlight some significant points, and more complete information can be found in the references.

[1] Slightly different definitions have been used for these depth ranges; for example, 325 km has also been used as the upper limit for deep earthquakes.

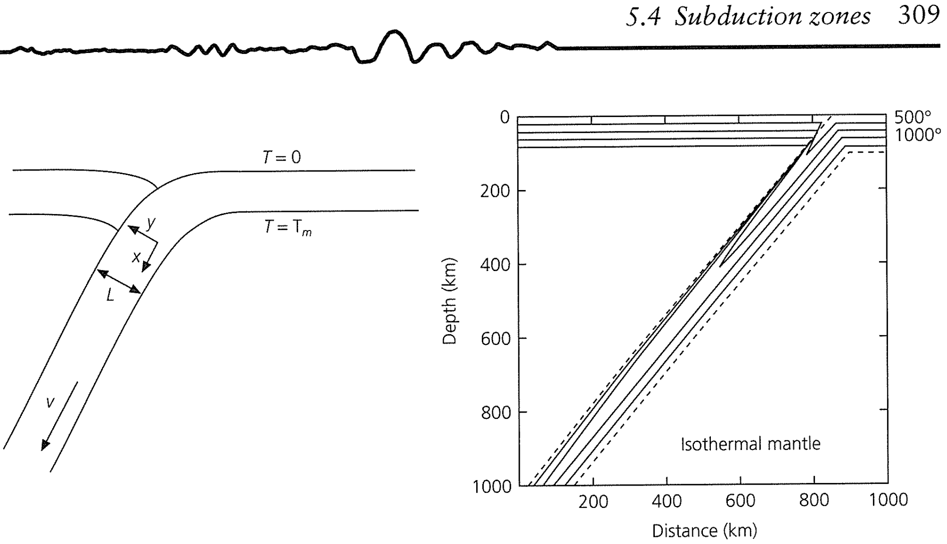


Fig. 5.4-3 An analytic model for temperatures in a subducting plate. *Left*: model geometry. *Right*: Results, showing the cold slab heating up as it descends through the hotter surrounding mantle.

The analytic model (Fig. 5.4-3) considers a semi-infinite slab of thickness L subducting at rate v. The surrounding mantle is at temperature T_m, and the plate enters the trench with a linear temperature gradient from $T = 0$ at its top to T_m at its base. We define the x axis down the dip of the slab, and the y axis across the slab. The evolution of the region is given by a slightly more complicated version of the heat equation (Eqn 5.3.1) used to model the cooling of the lithosphere as it moves away from the ridge. This version,

$$\rho C_p\left(\frac{\partial T}{\partial t} + v\nabla T\right) = \nabla \cdot (k\nabla T) + \varepsilon, \qquad (1)$$

describes the evolution of the temperature field, $T(x, y, t)$, as a function of time and the two space coordinates. In addition to the heat conduction term $\nabla \cdot (k\nabla T)$, Eqn 1 includes a $v\nabla T$ term describing the transfer (or advection) of heat by movement of material, and the ε term representing additional sources or sinks of heat such as radioactivity and phase changes. This form allows key parameters such as the density ρ, specific heat C_p, thermal conductivity k, and heat sources or sinks ε to vary with position. For a simple analytic solution, we assume that the problem is steady state ($\partial T/\partial t = 0$) and neglect heat sources and sinks ($\varepsilon = 0$). We further assume that the physical properties of the material (ρ, C_p, k, and hence the thermal diffusivity $\kappa = k/\rho C_p$) are independent of position.

With these simplifications, Eqn 1 becomes

$$\rho C_p v\frac{\partial T}{\partial x} = k\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right), \qquad (2)$$

which has a series solution

$$T(x, y) = T_m[1 + 2\sum_{n=1}^{\infty} c_n \exp(-\beta_n x/L) \sin(n\pi y/L)], \qquad (3)$$

with

$$c_n = (-1)^n/(n\pi), \quad \beta_n = (R^2 + n^2\pi^2)^{1/2} - R, \quad R = vL/(2\kappa).$$

R, the dimensionless thermal Reynolds number, is the ratio of the rate at which cold material is subducted to that at which it heats up by conduction. This solution resembles the temperature field in the plate model of cooling lithosphere (Eqn 5.3.19), because both models describe the thermal evolution of a plate of finite thickness with temperature boundary conditions at the top, bottom, and one end. In the previous case the plate cools, whereas in this case it heats up.

To find how far along the slab a given isotherm penetrates, we approximate the series by its first term and use the fact that $R \gg \pi$, so

$$T(x, y) \approx T_m[1 - (2/\pi) \exp(-\pi^2 x/(2RL)) \sin(\pi y/L)]. \qquad (4)$$

Solving for the point where $\partial T/\partial y = 0$ yields $y = L/2$, the middle of the slab. In fact, taking additional terms shows that this point is actually closer to the colder top (Fig. 5.4-3). Using the first-term approximation, a temperature T_0 goes furthest into the subduction zone at

$$T_0(x_0, L/2) = T_m[1 - (2/\pi) \exp(-\pi^2 x_0/(2RL))], \qquad (5)$$

and reaches a maximum down-dip distance

$$x_0 = -vL^2/(\pi^2\kappa) \ln[\pi(T_m - T_0)/(2T_m)]. \qquad (6)$$

To convert this distance to depth in the mantle, we multiply by $\sin\delta$, where δ is the slab dip. This correction converts the subduction rate v to the slab's vertical descent rate $v \sin\delta$. Thus an isotherm's maximum depth is proportional to the subduction rate and the square of the plate thickness, so faster subduction or a thicker slab allows material to go deeper before heating up. If we assume that the square of the plate thickness is proportional to its age, the maximum depth to an isotherm in the downgoing slab is proportional to the vertical descent rate times the age, t, of the subducting lithosphere.

This idea can be tested by assuming, as we did for spreading center earthquakes, that the maximum depth of earthquakes is temperature-controlled, so earthquakes should cease once material reaches a temperature that is too high. To compare

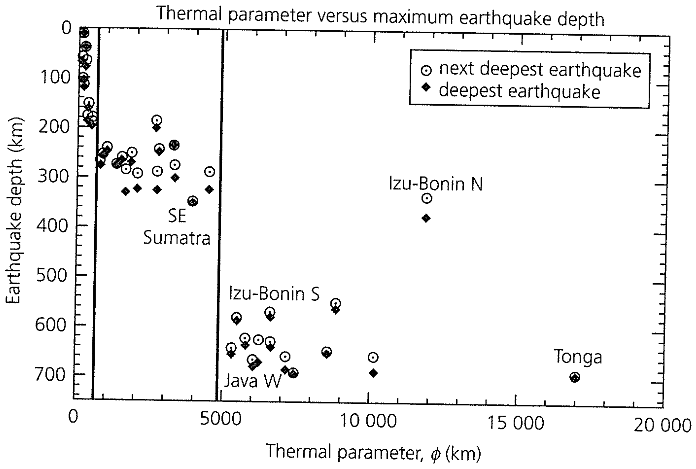


Fig. 5.4-4 Maximum earthquake depths for different subduction zones as a function of thermal parameter, the product of vertical descent rate and lithospheric age. If earthquakes are limited by temperature, this observation is consistent with the simple thermal model's prediction that the maximum depth to an isotherm should vary with the thermal parameter. (After Kirby *et al.*, 1996b. *Rev. Geophys.*, *34*, 261–306, copyright by the American Geophysical Union.)

various subduction zones, we examine the maximum depth of earthquakes as a function of their *thermal parameter*

$$\phi = tv \sin \delta. \tag{7}$$

Figure 5.4-4 shows that the maximum depth increases with thermal parameter, and deep earthquakes below 300 km occur only for slabs with a thermal parameter greater than about 5000 km.

However, the fact that the earthquakes stop does not mean that the slab has equilibrated with the surrounding mantle. Figure 5.4-5 shows the predicted minimum temperature within a slab as a function of time since subduction, assuming it maintains its simple planar geometry and does not buckle or thicken. The coldest portion reaches only about half the mantle temperature in about 10 Myr, which is about the time required for the slab to reach 660 km. Thus the restriction of seismicity to depths shallower than 660 km does not indicate that the slab is no longer a discrete thermal and mechanical entity. From a thermal standpoint, there is no reason for slabs not to penetrate into the lower mantle, an issue we discuss shortly. If a slab descended through the lower mantle at the same rate (in fact, it would probably slow down due to the more viscous lower mantle), it would retain a significant thermal anomaly at the core–mantle boundary, consistent with some models of that region (Section 3.8.4).[2]

The thermal model can be improved with simple modifications. Although we assumed that the slab subducts into an isothermal mantle, temperature should increase with depth, as the material is compressed due to increasing pressure from the overlying rock. Because the mantle below the lithosphere is thought to be convecting, it is often assumed that self-compression occurs adiabatically, such that material moving vertically neither loses nor gains heat. In this case, equilibrium thermodynamics requires that the effects of temperature and pressure changes exactly offset each other,

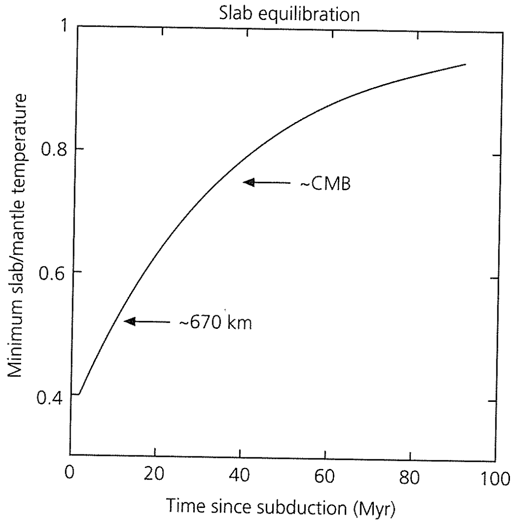


Fig. 5.4-5 Minimum temperature within a slab as a fraction of the mantle temperature, as a function of the time since subduction, computed using the analytic thermal model (Fig. 5.4-3). The coldest portion reaches half the mantle temperature in about 10 Myr, by which time a typical slab is approximately at 670 km depth, and 80% of it in 40 Myr, by which time a slab that continued descending at the same rate would reach the core–mantle boundary. Slabs can thus remain thermally distinct for long periods of time. (Stein and Stein, 1996. *Subduction*, 1–17, copyright by the American Geophysical Union.)

$$dS = \frac{C_p}{T} dT - \frac{\alpha}{\rho} dP = 0, \tag{8}$$

so that the entropy S does not change. This condition gives the adiabatic temperature gradient, or *adiabat*, as

$$\left(\frac{dT}{dP}\right)_s = \frac{\alpha}{\rho C_p} T, \tag{9}$$

where α is the coefficient of thermal expansion. Because pressure increases with depth as $dP/dz = \rho g$, temperature increases with depth as

$$\left(\frac{dT}{dz}\right)_s = \frac{\alpha g}{C_p} T. \tag{10}$$

We can thus correct the temperatures for the isothermal mantle case to include adiabatic heating. Using the entropies requires using absolute (Kelvin) temperatures, equal to the Celsius temperature plus 273.15°. Thus if the absolute temperature at

[2] The oceanic lithosphere takes about 70 Myr to cool to equilibrium with the mantle below, and so takes about half that time to heat up again from both sides after it subducts.

depth z_0, the base of the plate, is T_0^K, we integrate Eqn 10 to find the absolute temperature at depth z,

$$T^K(z) = T_0^K \exp\,[(\alpha g/C_p)(z - z_0)]. \qquad (11)$$

Another possibly important effect is that of heat sources and sinks. For example, the olivine to spinel transition, which gives rise to the 410 km discontinuity outside the slab, should release heat as it occurs in the slab. Heat might also be generated by friction at the top of the downgoing slab. The heat produced is the product of the subduction rate and the shear stress on the slab interface. The magnitude of this effect is difficult to estimate. It should not be significant unless the shear stress is greater than a few kilobars. As discussed later (Section 5.7.5), the stress on faults is unknown. A further complexity results from the fact that the viscosity of the mantle, which controls the stress, decreases exponentially with temperature. Thus, if frictional heating raises the temperature at the slab interface, viscosity, and hence stress, would decrease, tending to counteract the effect.

To address these complexities, we use numerical models to solve the heat equation at every point in the slab. These models allow parameters such as density to vary with position. In addition, heat sources and sinks such as radioactive heating, phase changes, and frictional heating can be incorporated. The results of such calculations are similar to those of the analytic model and are used to explore how temperatures should vary between subduction zones. For example, Fig. 5.4-6 compares models for a relatively younger and slower-subducting slab (thermal parameter about 2500 km), approximating the Aleutian arc, and an older, faster-subducting slab (thermal parameter approximately 17,000 km), approximating the Tonga arc. As expected, the slab with the higher thermal parameter warms up more slowly, and is thus colder. This prediction is consistent with the observation that Tonga has deep earthquakes, whereas the Aleutians do not (Fig. 5.4-4).

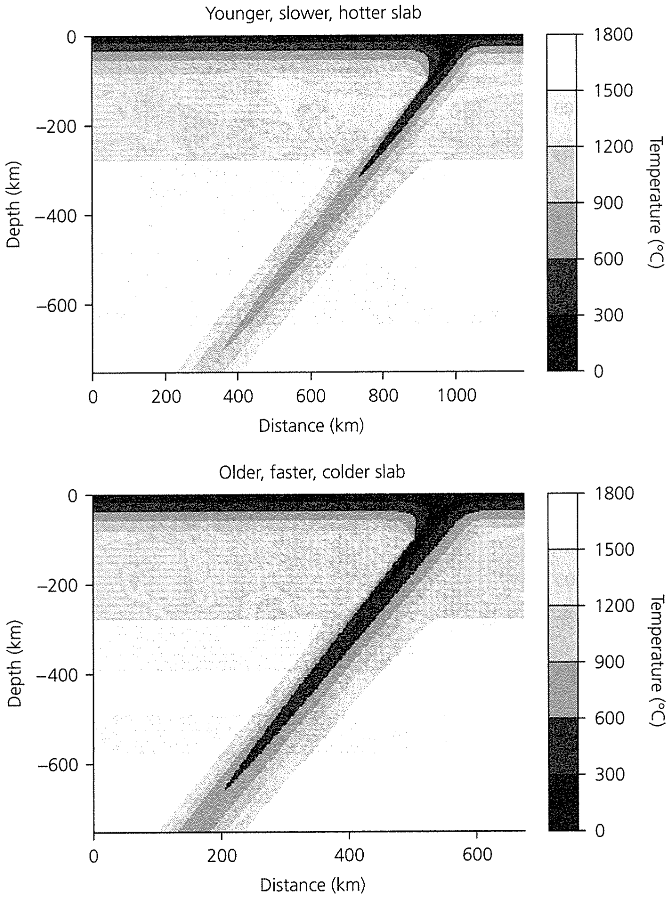


Fig. 5.4-6 Comparison of thermal structure for a relatively younger, slower-subducting slab (50 Myr-old lithosphere subducting at 70 mm/yr; thermal parameter about 2500 km), which approximates the Aleutian arc, and an older, faster-subducting slab (140 Myr-old lithosphere subducting at 140 mm/yr; thermal parameter about 17,000 km) which approximates the Tonga arc. (Stein and Stein, 1996. *Subduction*, 1–17, copyright by the American Geophysical Union.)

Although we can compute such thermal models, a question is whether they make sense. We test them using two seismological datasets: earthquake locations and seismic velocities. Travel time tomography (Section 7.3) across subduction zones shows high-velocity slabs (Fig. 5.4-7). These results are compared to the velocities predicted using a thermal model of the subducting slab and laboratory values for the variation in velocity with temperature. The model predicts coldest temperatures in the slab interior where the earthquakes occur. Because the tomographic inversion finds the velocity within rectangular cells, the model is converted to that grid and then "blurred" because the seismic rays do not uniformly sample the slab. As shown by the hit count, the number of rays sampling each cell, most rays go down the high-velocity slab, yielding a somewhat distorted image. The fact that this image and the tomographic result are similar suggests that the model is a reasonable description of the actual slab. A similar conclusion emerges from the observation that the tomographic result also resembles parts of the model image that are artifacts, velocity anomalies that are not present in the original model. These artifacts, generally of low amplitude, cause the slab to appear to broaden, shallow in dip, or flatten out. Hence, although slab thermal models are simplifications of complicated real slabs, and many key parameters are not well known, it seems likely that the models are reasonable approximations (perhaps accurate to a few hundred degrees) to the temperatures within actual slabs.

Seismology provides other tools to study the contrast between the cold, rigid, downgoing plate and the hotter, less rigid material around it. Figure 3.7-20 showed that a cold slab transmits seismic energy with less attenuation than its surroundings. Figure 5.4-8 shows some of the earliest data for this effect: seismograms from a deep earthquake are contrasted at stations NIU, to which waves travel through the downgoing plate; and VUN, to which waves arrive through the surrounding mantle. The VUN record shows much more long-period energy, especially for S waves, than that at NIU. Thus the

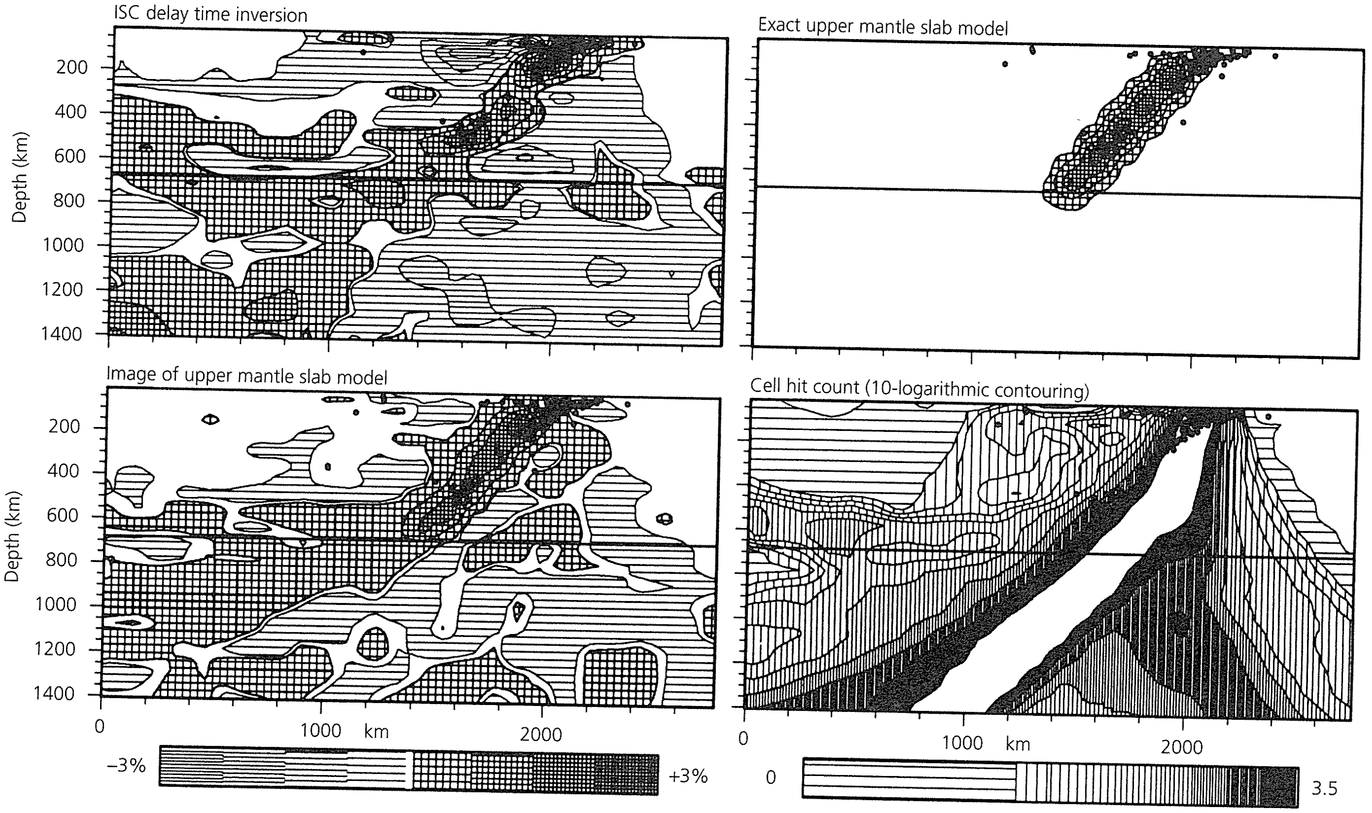


Fig. 5.4-7 Comparison of a seismic tomographic image of a subducting slab, indicated by the velocity anomaly and earthquake hypocenters (dots) (*upper left*) to the image (*lower left*) predicted for a slab thermal model. The seismic velocity anomaly predicted by the thermal model (*upper right*) is imaged by a simulated tomographic study using the same seismic ray path sampling as the data. The hit count (*lower right*) shows the number of rays sampling each cell used in the inversion. As a result of ray geometry and noise, the slab model gives a somewhat distorted image (*lower left*), showing how the model would appear in such a tomographic study. The similarity of the image of the model and the tomographic result suggests that the model generally describes the major features of the actual slab. Left scale bar gives velocity perturbations in percent, with positive values representing fast material. Right scale bar is for hit count, showing values as logarithm to base 10; the white region in the hit count plot is densely sampled and off scale. (Spakman *et al.*, 1989. *Geophys. Res. Lett.*, *16*, 1097–110, copyright by the American Geophysical Union.)

high-frequency components were more absorbed on the path to VUN due to higher attenuation (lower Q) than on the more rigid slab path to NIU. In addition, the sharp contrast in seismic velocity at the top of slabs can be detected using reflected and converted seismic waves (Fig. 2.6-15).

5.4.2 *Earthquakes in subducting slabs*

The deep and intermediate earthquakes forming the Wadati–Benioff zone extend in some places to depths of almost 700 km (Fig. 5.4-9). These are the deepest earthquakes that occur: away from subduction zones, earthquakes below about 40 km are rare. The Wadati–Benioff zone earthquakes illustrate that material cold enough to fail seismically (rather than flow) is being subducted, and give our best information about the geometry and mechanics of slabs.

The number of earthquakes as a function of depth illustrates why we distinguish intermediate and deep earthquakes; seismicity decreases to a minimum near about 300 km, and then increases again. Deep earthquakes, those below about 300 km, are thus generally treated as distinct from intermediate earthquakes. Deep earthquakes peak at about 600 km, and then decline to a minimum before 700 km. The focal mechanisms also vary with depth; those shallower than 300 km show generally down-dip tension, whereas those below 300 km show generally down-dip compression (Fig. 5.4-10).

Various explanations for this distribution of earthquakes and focal mechanisms are under consideration. One is that near the surface the slab is extended by its own weight, whereas at depth it encounters stronger lower mantle material, causing down-dip compression. Another possible factor may be mineral phase changes that occur at different depths in the cold slab than in the surrounding mantle.

It is generally assumed that the most crucial effect is the negative buoyancy (sinking) of the cold and dense slabs. The thermal model gives the force driving the subduction due to the integrated negative buoyancy of a slab resulting from the density contrast between it and the warmer and less dense material at the same depth outside. Because the slab does not have a discrete lower end in the analytic model, the net force is

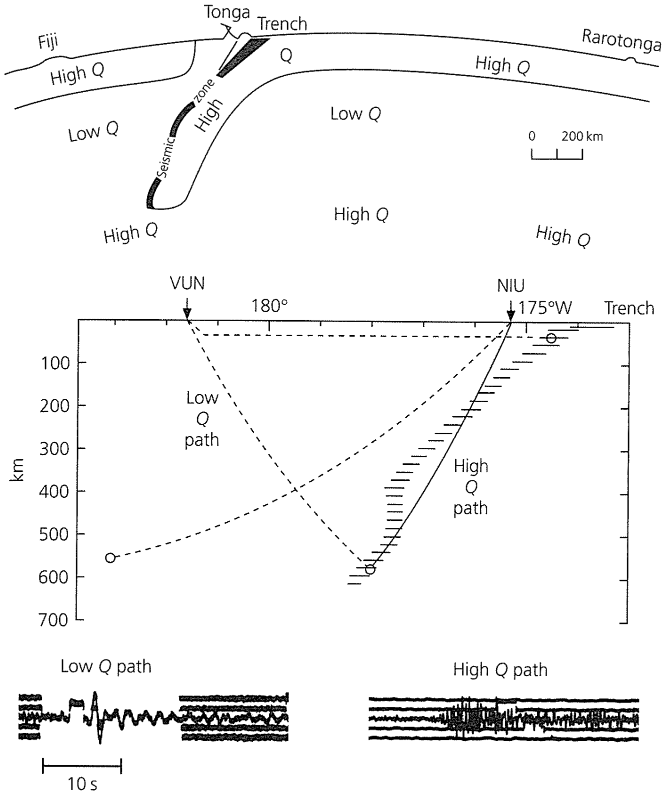


Fig. 5.4-8 Seismological observations showing the difference between the cold slab and hotter ambient mantle. Comparison of the seismograms at NIU and VUN shows that high frequencies are transmitted better by the slab, so the slab is a less attenuating, or higher Q path. (Oliver and Isacks, 1967. *J. Geophys. Res., 72*, 4259–75, copyright by the American Geophysical Union.)

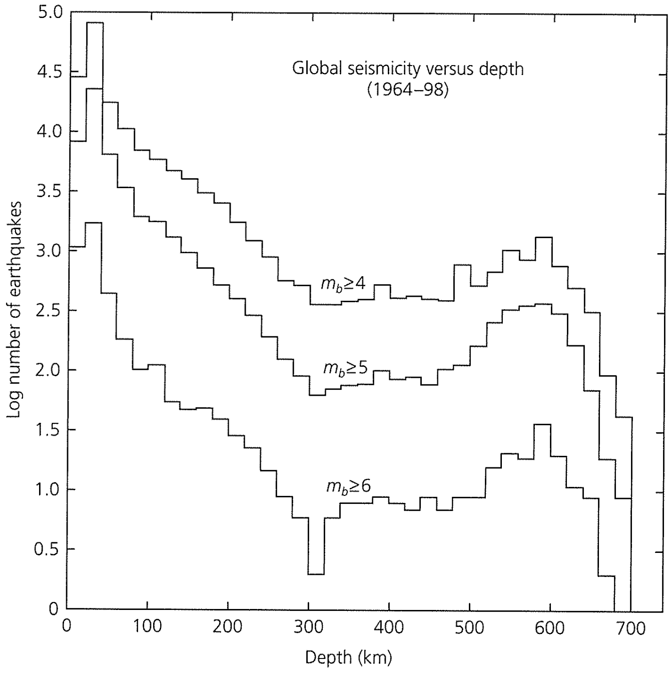


Fig. 5.4-9 Distribution of seismicity with depth.

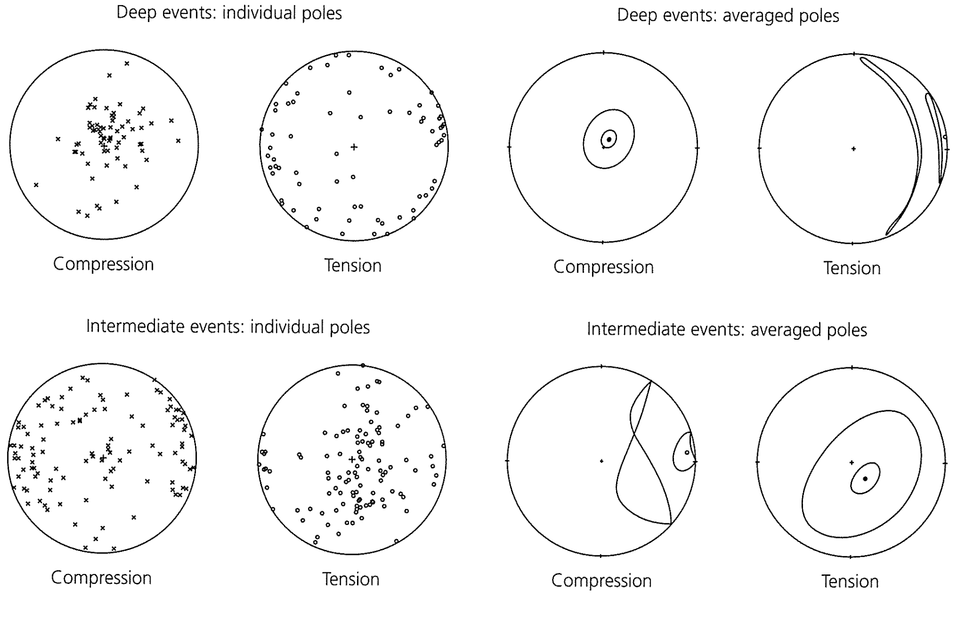


Fig. 5.4-10 Stress orientations inferred from focal mechanisms of subduction zone earthquakes. The P and T axes are rotated so that the down-dip direction is at the center of each plot, and their distributions are contoured. *Top*: Events below 300 km are dominated by down-dip compression. *Bottom*: Events from 70–300 km are dominated by down-dip tension. (After Vassiliou, 1984. *Earth Planet. Sci. Lett., 69*, 195–201, with permission from Elsevier Science.)

$$F = \int_0^L \int_0^\infty g[\rho(x,y) - \rho_m]dx\,dy. \tag{12}$$

If material outside the slab is at temperature T_m and density ρ_m, material in the slab at the point (x, y) has density

$$\rho(x,y) \approx \rho_m + \frac{\partial \rho}{\partial T}[T(x,y) - T_m] = \rho_m + \rho'(x,y). \tag{13}$$

As for the cooling plate (Eqn 5.3.9), the density perturbation is

$$\rho'(x,y) = \alpha\rho_m[T_m - T(x,y)], \tag{14}$$

so for the analytic temperature model (Eqn 3) the integral over the slab yields a force

$$F = \frac{g\alpha\rho_m T_m v L^3}{24\kappa}. \tag{15}$$

This force, known as "slab pull," is the plate driving force due to subduction. Specifically, it is the negative buoyancy associated with a cold downgoing limb of the convection pattern. Its significance for stresses in the downgoing plate and for driving plate motions depends on its size relative to the resisting forces at the subduction zone. There are several such forces. As the slab sinks into the viscous mantle, the material displaced causes a force depending on the viscosity of the mantle and the subduction rate. The slab is also subject to drag forces on its sides and to resistance at the interface between the overriding and downgoing plates, which is often manifested as earthquakes.

To gain insight into the relative size of the negative buoyancy ("slab pull") and resistive forces, we consider the stress in the downgoing slab and the resulting focal mechanisms. Figure 5.4-11 shows a simple analogy, the stress due to the weight of a vertical column of length L of material with density ρ. Using the equilibrium equation (Eqn 2.3.49), we equate the stress gradient to the body force,

$$\frac{\partial \sigma_{zz}(z)}{\partial z} = -\rho g, \tag{16}$$

so the stress as a function of depth is found by integration,

$$\sigma_{zz}(z) = -\rho g z + C, \tag{17}$$

where C is a constant of integration. To determine C, and thus the stress in the column, the boundary conditions must be known.

First, suppose the stress is zero at the top, $z = 0$. In this case $C = 0$ and

$$\sigma_{zz}(z) = -\rho g z, \tag{18}$$

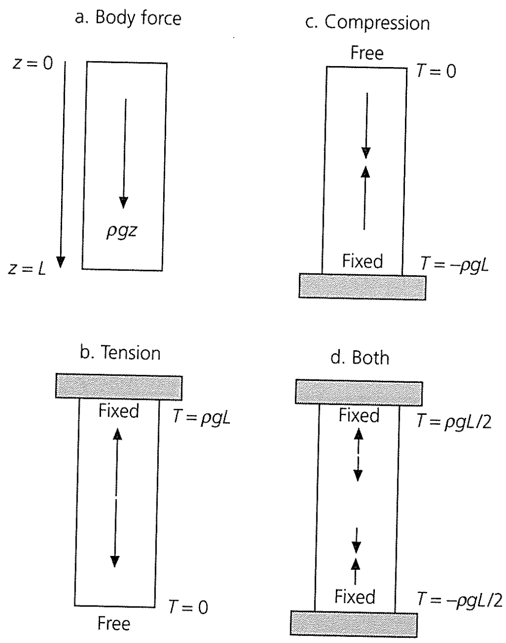


Fig. 5.4-11 Stress within a vertical column of material under its own weight, a simple analogy to stress within a downgoing slab. For the same body force, different stress distributions result from different boundary conditions. If the load is supported at the bottom, the column is under compression; if the support is at the top, the column is under tension. A combination of the two produces a transition.

which is negative, corresponding to compression everywhere. The forces required at the top and the bottom to maintain equilibrium are given by the relation between the traction, stresses, and outward normal vector on a surface (Eqn 2.3.8),

$$T_z = \sigma_{zz} n_z. \tag{19}$$

At the top $T_z(0) = 0$, whereas at the bottom a force

$$T_z(L) = -\rho g L \tag{20}$$

holds the column up. This situation is like a column of material sitting on the earth's surface, under compression everywhere.

Alternatively, suppose the stress is zero at the bottom. In this case the constant is chosen so that

$$\sigma_{zz}(z) = \rho g(L - z) \tag{21}$$

and the column is in extension (σ_{zz} positive) everywhere. The force at the bottom is zero, and the force at the top,

$$T_z(0) = \rho g L, \tag{22}$$

supports the column, because n_z points in the $-z$ direction. This situation corresponds to the material hanging under its own weight.

If the column is supported equally at both ends, the forces at either end are equal, so we find the stress from the condition

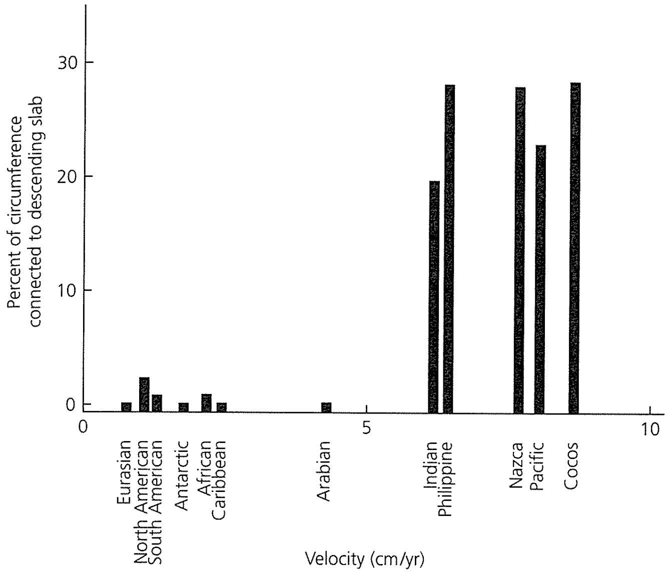


Fig. 5.4-12 The absolute velocity of lithospheric plates increases with the fraction of the plate's boundary formed by subducting slabs, suggesting that slabs provide a major driving force for plate motions. (Forsyth and Uyeda, 1975.)

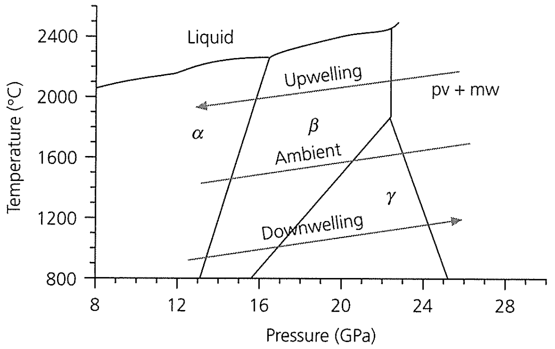


Fig. 5.4-13 Phase diagram for transitions in olivine with increasing depth. The phase boundaries as functions of temperature and pressure are known as Clapeyron curves. The downwelling and upwelling lines contrast conditions in slabs and plumes, respectively, to those in the ambient mantle. A reaction with a positive slope, such as the olivine (α phase) to spinel (β phase) change thought to give rise to the 410 km discontinuity outside the slab, is displaced upward (to lower pressure) within the cold slab. By contrast, the γ spinel to perovskite plus magnesiowustite (pv + mw) transition has a negative slope, so the 660 km discontinuity should be deeper in slabs than outside. (After Bina and Liu, 1995. *Geophys. Res. Lett.*, 22, 2565–8, copyright by the American Geophysical Union.)

$$T_z(0) = -T_z(L), \tag{23}$$

which gives

$$\sigma_{zz}(z) = \rho g(L/2 - z). \tag{24}$$

Thus the column is in extension in its upper half, $z < L/2$, and in compression below this point.

The stress in the column shows how the body force due to gravity is balanced by forces on the boundaries. By analogy, if the downgoing slab were in tension, the negative buoyancy force must exceed the resistive forces at the subduction zone, and the slab would be "pulling" on and supported by the remainder of the plate outside the subduction zone. In fact, most earthquakes in the deeper portions of the slab show down-dip compression, whereas the intermediate earthquakes show down-dip tension (Fig. 5.4-10). This situation is like the column supported at both ends.

These ideas about the forces within subduction zones are consistent with two important pieces of data. First, the average absolute velocity of plates increases with the fraction of their area attached to downgoing slabs (Fig. 5.4-12), suggesting that slabs are a major determinant of plate velocities. Second, as discussed in Section 5.5.2, earthquakes in old oceanic lithosphere have thrust mechanisms, demonstrating deviatoric compression. Thus the net effect of the subduction zone on the remainder of the plate is not a "pull," so the term "slab pull" is misleading. Instead, as implied by the slab stress models, the "slab pull" force is balanced by local resistive forces, a combination of the effects of the viscous mantle and the interface between plates. This situation is like an object dropped in a viscous fluid, which is accelerated by its negative buoyancy until it reaches a terminal velocity determined by its density and shape and the viscosity and density of the fluid.

An interesting possible complication is that slabs are not just thermally different from their surroundings; they are probably also mineralogically different. Slabs extend through the mantle transition zone, where mineral phase changes are thought to occur (Section 3.8). However, because a downgoing slab is colder than material at that depth elsewhere, phase changes within the slab are displaced relative to their normal depth. The displacement can be calculated using the thermodynamic relation, known as the Clapeyron equation, for the boundary between two phases as a function of pressure and temperature. If ΔH and ΔV are the heat and volume changes resulting from the phase change, then a change dT in temperature moves the phase change by a pressure dP given by the Clapeyron slope (the reciprocal of Eqn 9),

$$\gamma = \frac{dP}{dT} = \frac{\Delta H}{T\Delta V}. \tag{25}$$

For example, the 410 km discontinuity is attributed to the phase change with increased pressure from olivine to a denser spinel structure (the β phase, wadsleyite) described by a phase diagram like that in Fig. 5.4-13. Because the spinel phase is denser, ΔV is less than zero. This reaction is exothermic (gives off heat), so ΔH is also negative, causing a positive Clapeyron slope. If we know the depth (pressure) and temperature at which a phase change occurs in the mantle, the Clapeyron equation gives its position in the slab. The slab is colder than

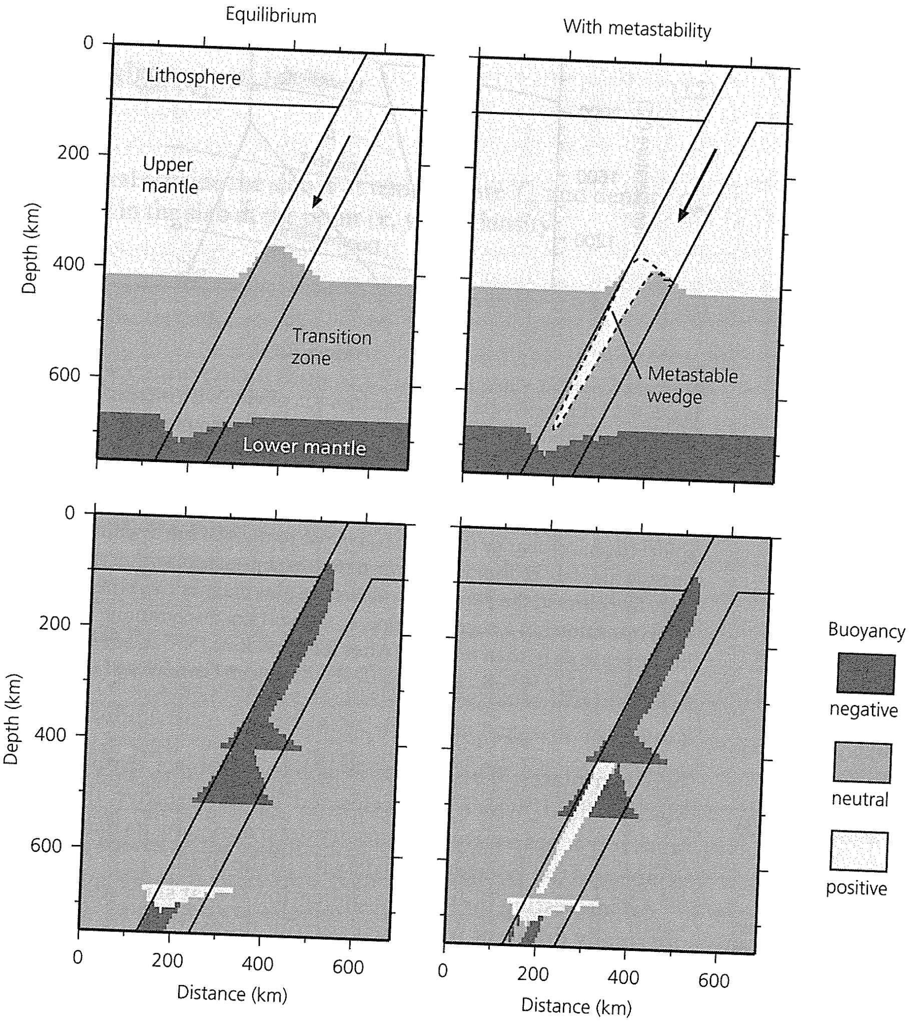


Fig. 5.4-14 Predicted mineral phase boundaries and resulting buoyancy forces in a downgoing slab without (*left panels*) and with (*right panels*) a metastable olivine wedge. Assuming equilibrium mineralogy the cold slab has negative thermal buoyancy, negative compositional buoyancy associated with the elevated 410 km discontinuity, and positive compositional buoyancy associated with the depressed 660 km discontinuity. A metastable wedge gives positive compositional buoyancy and hence decreases the force driving subduction. Negative buoyancy favors subduction, whereas positive buoyancy opposes it. (Stein and Rubie, 1999. *Science, 286*, 909–10, copyright 1999 American Association for the Advancement of Science.)

the ambient mantle ($dT < 0$), so this phase change occurs at a lower pressure ($dP < 0$), corresponding to a shallower depth. Converting the pressure change to depth, the vertical displacement of this phase change is

$$\frac{dz}{dT} = \frac{\gamma}{\rho g}. \qquad (26)$$

By contrast, the ringwoodite (γ spinel phase) to perovskite plus magnesiowustite transition, thought to give rise to the 660 km discontinuity, is endothermic (absorbs heat), so ΔH is positive. Because this is a transformation to denser phases (ΔV less than zero), the Clapeyron slope is negative, and the 660 km discontinuity should be deeper in slabs than outside. These opposite effects – upward deflection of the 410 km and downward deflection of the 660 km discontinuities (Fig. 5.4-14) – have been observed in travel time studies. An interesting way to think about these is to note that the negative buoyancy associated with the elevated 410 km discontinuity helps the subduction, whereas the positive buoyancy associated with the depressed 660 km discontinuity opposes the subduction. The reverse effect should not occur at the 660 km discontinuity for upcoming plumes, however, because the phase diagram shows that at these higher temperatures the Clapeyron curve for the perovskite plus magnesiowustite transition is vertical, so the transition is not displaced (Fig. 5.4-13).

The position of the olivine–spinel phase change may be further affected. The Clapeyron slope predicts what happens if a phase change occurs at equilibrium. However, the phase change actually occurs by a process in which grains of the high-pressure phase nucleate on the boundaries between grains of the lower-pressure phase and then grow with time (Fig. 5.4-15). Studies of mineral nucleation and growth rates suggest that in the coldest slabs the phase transformation cannot keep pace with the rate of subduction, causing a wedge of olivine in the cold slab core to persist metastably[3] to greater depths (Fig. 5.4-14).

[3] *Metastability* describes the situation where a mineral phase survives outside its equilibrium stability field in temperature–pressure space. Such metastable persistence is expected because the relatively colder temperatures in slabs should inhibit reaction rates. This effect explains why diamonds, which are unstable at the low pressures of earth's surface, survive metastably rather than transform to graphite. The situation in slabs is similar to that of supercooled water, which persists as a liquid at temperatures below its equilibrium freezing point.

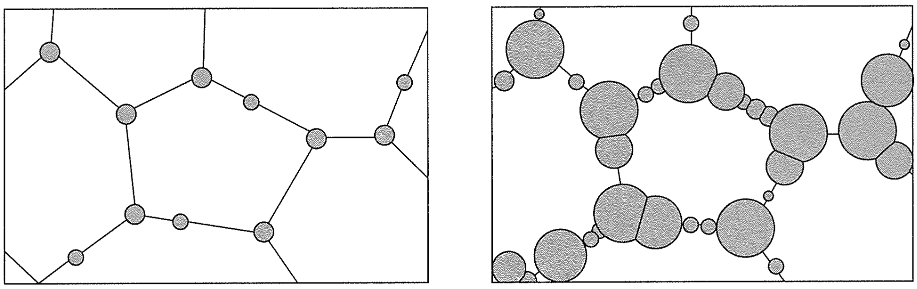


Fig. 5.4-15 Diagram showing the early stages of a phase transformation. Grains of the new phase (shaded) nucleate on grain boundaries and grow by consuming the original phase until none remains. (Kirby *et al.*, 1996b. *Rev. Geophys.*, *34*, 261–306, copyright by the American Geophysical Union.)

The deflections of the phase boundaries have several possible consequences. First, phase changes affect the thermal structure of the slab due to the heat of the phase change. Thus the exothermic olivine–spinel change should add heat to slabs. This effect is simulated in thermal models by increasing the temperature at the phase change. Second, the phase boundaries are probably important for the buoyancy and stresses within slabs. We have already discussed the idea that the cold slabs are denser than their surroundings, causing negative thermal buoyancy, which favors sinking. The phase boundaries cause additional mineralogical buoyancy. For example, if the olivine–spinel boundary is uplifted in the slab, the presence of slab material denser than at that depth outside causes additional negative buoyancy. However, if a wedge of metastable olivine exists, it would be less dense than material at that depth outside and produce positive buoyancy (Fig. 5.4-14) in addition to that caused by the downward deflection of the 660 km discontinuity. Although the net buoyancy must be negative because slabs subduct, the details of the buoyancy can be important. For example, metastable olivine may help regulate subduction rates. Faster subduction would cause a larger wedge of low-density metastable olivine, reducing the driving force and slowing the slab.

A third possibility is that a phase change causes deep earthquakes. Although this idea is a natural consequence of the observation that deep earthquakes occur at transition zone depths, it was not given serious consideration for a long time because deep earthquake focal mechanisms show slip on a fault, rather than isotropic implosions (Section 4.4.6). However, laboratory studies now suggest that an instability

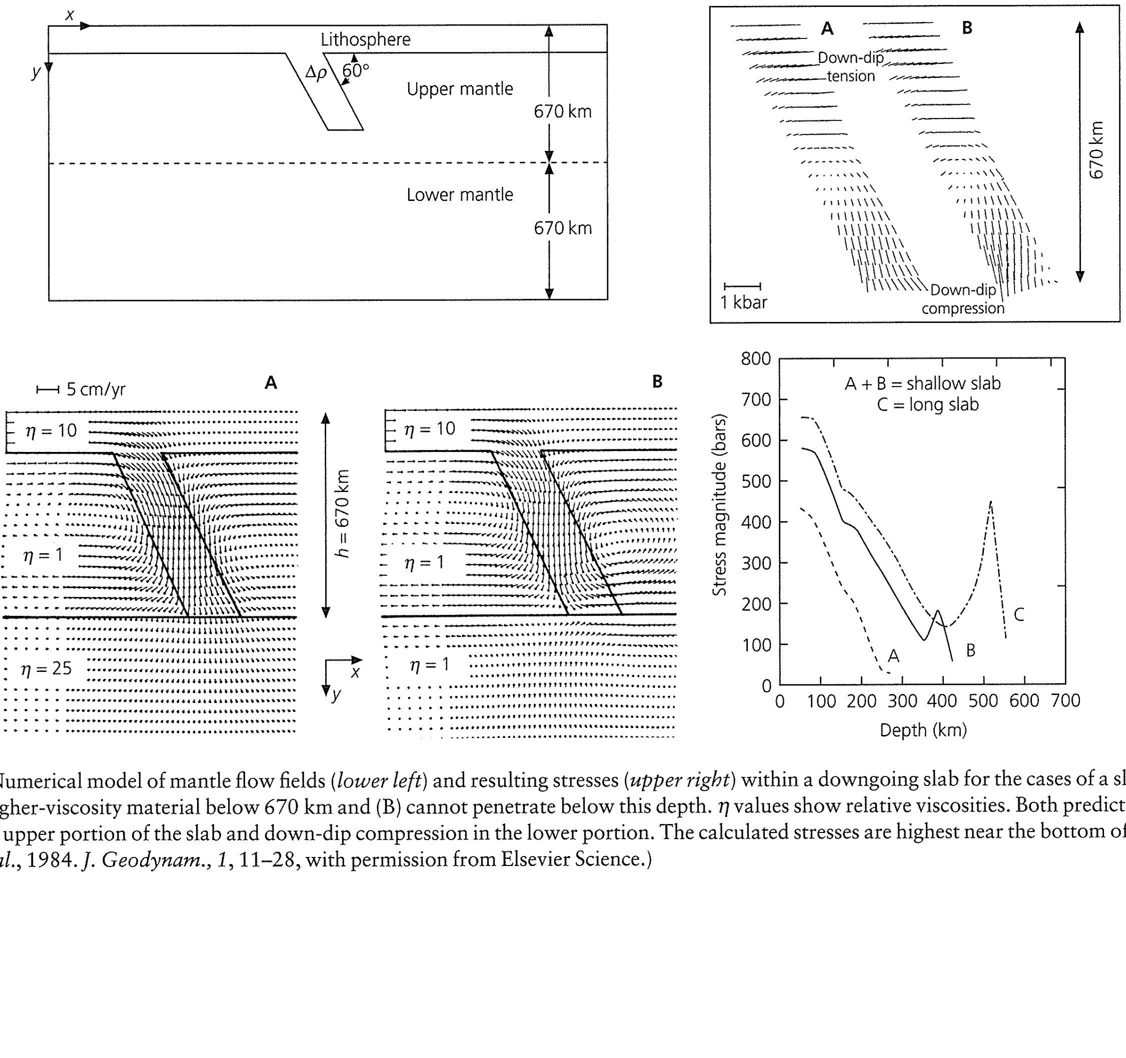


Fig. 5.4-16 Numerical model of mantle flow fields (*lower left*) and resulting stresses (*upper right*) within a downgoing slab for the cases of a slab that (A) encounters higher-viscosity material below 670 km and (B) cannot penetrate below this depth. η values show relative viscosities. Both predict down-dip tension in the upper portion of the slab and down-dip compression in the lower portion. The calculated stresses are highest near the bottom of the slab. (Vassiliou *et al.*, 1984. *J. Geodynam.*, *1*, 11–28, with permission from Elsevier Science.)

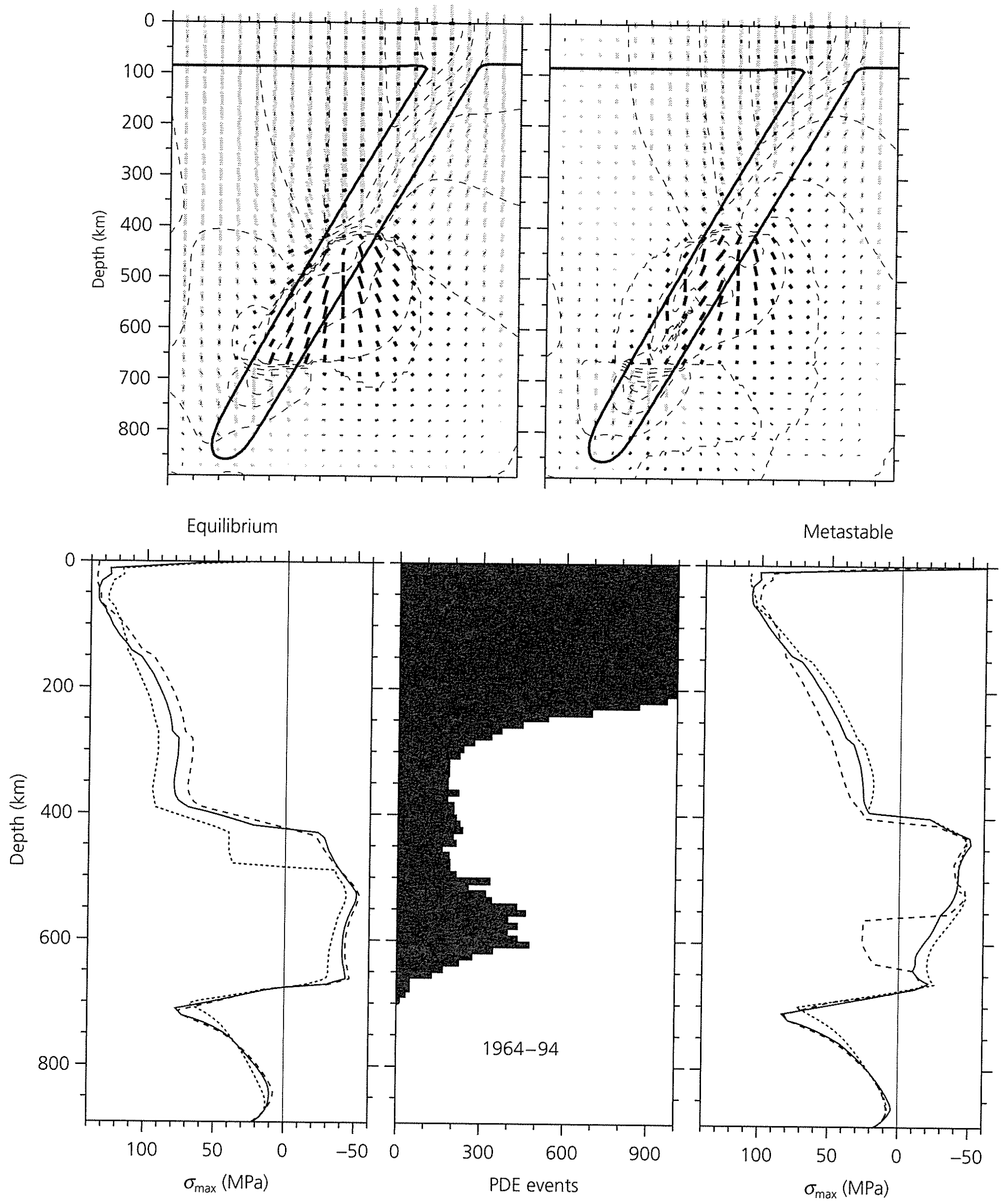


Fig. 5.4-17 Numerical models of stresses within a downgoing slab assuming the density distribution corresponding to equilibrium mineralogy (*left panels*) and with metastable olivine (*right panels*). Upper panels show stress orientations, and lower panels show stress magnitudes, with compression as negative, compared to the distribution of seismicity (*lower center*). (Bina, 1997. *Geophys. Res. Lett., 24,* 3301–4, copyright by the American Geophysical Union.)

called transformational faulting can cause slip along thin shear zones where metastable olivine transforms to denser spinel. Such faulting can occur for the exothermic olivine to spinel transition, but not for the endothermic spinel to perovskite plus magnesiowustite transition, so deep earthquakes would occur only in the transition zone. Because the metastable wedge's lower boundaries are essentially isotherms, this model offers a physical mechanism for the observation (Fig. 5.4-4) that the depth of earthquakes increases with thermal parameter. This idea is attractive, but to date seismological studies show no evidence for a metastable wedge, and large deep earthquakes occur on fault planes that appear to extend beyond the boundaries of the expected metastable wedge. If such wedges exist, earthquakes may nucleate by transformational faulting, but then propagate outside the wedge via another failure mechanism.

Together these ideas offer several possible explanations for features of slab earthquakes. One key feature is the depth variation in seismicity and focal mechanisms. The first explanation is that the depth distribution and stresses are largely due to the negative thermal buoyancy of slabs and their encountering either a region of much higher viscosity or a barrier to their motion at the 660 km discontinuity. Numerical models (Fig. 5.4-16) predict stress orientations similar to those implied by the focal mechanisms. Moreover, the magnitude of the stress varies with depth in a fashion similar to the depth distribution of seismicity — a minimum at 300–410 km and an increase from 500 to 700 km. Alternatively, numerical models including the buoyancy effects of the phase changes (Fig. 5.4-14) also predict a similar variation in stress magnitude and orientation with depth (Fig. 5.4-17), without invoking a barrier or higher viscosity in the lower mantle. Thus, in such

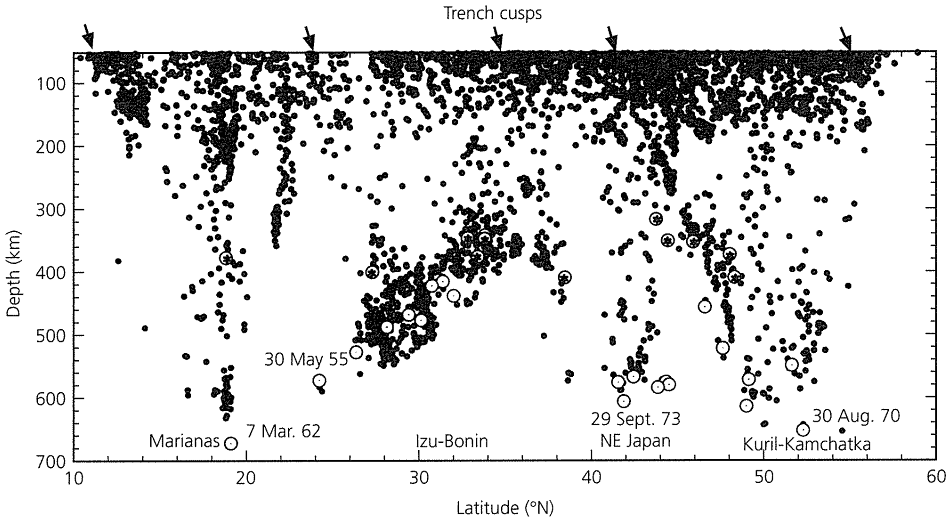


Fig. 5.4-18 North–south cross-section showing seismicity of subduction zones of the Northwest Pacific. Seismicity shallows near the cusps where arcs meet, making individual Wadati–Benioff zones tongue-shaped. Large deep earthquakes (M_0 greater than 10^{26} dyn-cm), shown by open circles, tend to be at the edges or bottoms of deep seismicity, or isolated from the main Wadati–Benioff zones. (Kirby *et al.*, 1996b. *Rev. Geophys.*, *34*, 261–306, copyright by the American Geophysical Union.)

models, deep earthquakes need not be physically different from intermediate ones, because the minimum in seismicity reflects a stress minimum.

A second key issue is how deep earthquakes can occur at all. As discussed in Section, 5.7, the strength of rock that must be exceeded for fracture increases with pressure. The pressures deep in a subducting slab should be high enough to prevent fracture. One possibility is that the slabs become hot enough that water released by decomposition of hydrous minerals lubricates (reduces the effective stress on) faults. Another possibility, mentioned earlier, is transformational faulting in metastable olivine. It is also possible that the earthquakes occur by very rapid creep, possibly associated with weakening due to unusually small spinel grains formed in the coldest slabs.

The different explanations offered by these models all have attractive features and may be true in part. However, although such simple models based on idealized slabs explain some gross features of deep earthquakes, none fully explains the complexity of deep earthquakes. As shown by Fig. 5.4-18, a cross-section along the subduction zones of the Northwest Pacific, deep seismicity is "patchy" and variable. For example, it shallows dramatically at the cusps between the Marianas, Izu-Bonin, NE Japan, and Kuril-Kamchatka arcs. Moreover, the largest earthquakes occur at the edges of the regions of deep seismicity, as especially evident at the northern edge of the Izu-Bonin seismicity. These sites may reflect tears in the downgoing lithosphere at the junctions between arcs, where hot mantle material penetrates slabs. A further complexity is that some deep earthquakes occur in unusual locations off the down-dip extension of the main Wadati–Benioff zones and have focal mechanisms differing from those of the deepest earthquakes in the main zone (Fig. 5.4-19). Some other deep earthquakes are isolated from actively subducting slabs. Such unusual earthquakes may occur in slab fragments where metastable olivine survives, and thus have mechanisms related to local stresses rather than those expected for continuous slabs.

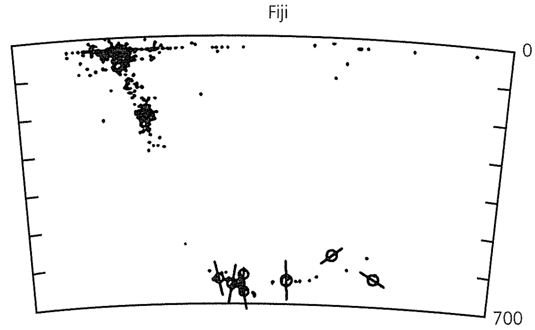


Fig. 5.4-19 Seismicity cross-section for the Fiji subduction zone, showing "outlier" deep earthquakes. Lines through symbols show P axes, which often differ from those for the main Wadati–Benioff zone. (Lundgren and Giardini, 1994. *J. Geophys. Res.*, *99*, 15, 833–42, copyright by the American Geophysical Union.)

Another interesting observation from precise earthquake locations in some subduction zones (Fig. 5.4-20) shows that the Wadati–Benioff zone is made up of two distinct planes, separated by 30–40 km. The upper plane seems to coincide with the conversion plane for *ScSp* (Fig. 2.6-15), a sharp velocity contrast that is presumably near the slab top. Focal mechanisms suggest that the upper plane is in down-dip compression and the lower one in down-dip extension. A variety of models have been proposed. One is that the double plane results from "unbending" of the slab – the release of the bending stresses produced when the slab began to subduct. Another model is that the slab "sags" under its own weight, because at depth it runs into a more viscous mesosphere, while at intermediate depths it encounters a less viscous asthenosphere. Explaining the phenomenon is complicated by the observation that only some subduction zones have double zones.

The nature of deep earthquakes, especially the mechanism restricting them to the transition zone, has implications for mantle flow. The simplest explanation for the cessation of deep seismicity is that slabs cannot penetrate the lower mantle. However, as shown in Fig. 5.4-21, tomographic studies (Chapter 7) indicate that although some slabs are deflected at 660 km, they eventually penetrate deeper. Hence models in which earthquakes stop either because the stress is not high enough or because the phase changes causing them no longer occur seem more likely. The issue is important because heat and mass transfer between the upper and lower mantles have major implications for the dynamics and evolution of the earth (Section 3.8). At present, most models favor some degree of communication between the two (Fig. 5.1-2). Slabs are sometimes deflected at the 660 km discontinuity, where they warm further, lose any buoyant metastable wedge, and then penetrate into the lower mantle. Thus the slab geometry we see likely reflects a complex set of effects. To cite another, some flat-lying slabs at the 660 km discontinuity may be caused by the trench "rolling backward" in the absolute (mantle) reference frame.

There has also been considerable discussion about the nature of intermediate depth earthquakes. Figure 5.4-22 shows a

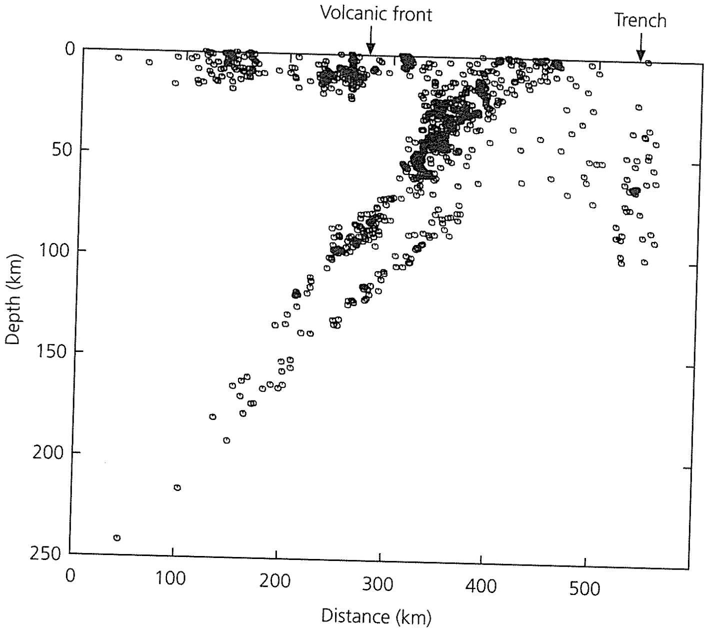


Fig. 5.4-20 Double seismic zone beneath Tohoku, Japan. (Hasegawa *et al.*, 1978. *Tectonophysics*, *47*, 43–58, with permission from Elsevier Science.)

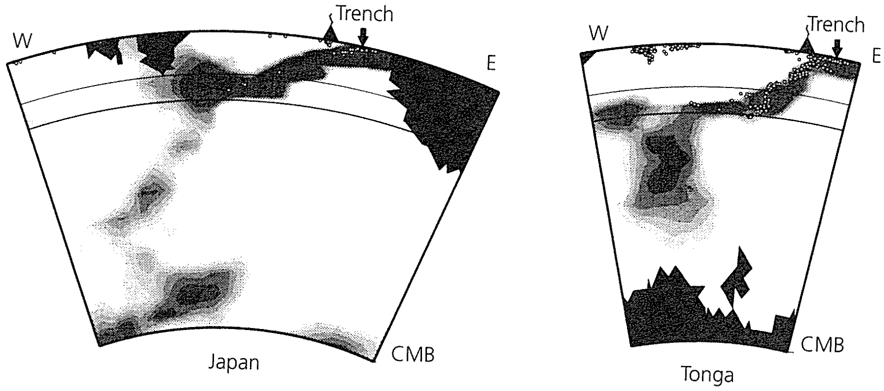


Fig. 5.4-21 Tomographic images across Pacific subduction zones with deep earthquakes. Horizontal lines are at 410 and 660 km depth. White dots are earthquake hypocenters. The Wadati–Benioff zone seismicity generally coincides with the high-velocity anomaly (dark regions) due to the cold subducting slab. Slabs are deflected at the base of the transition zone before penetrating into the lower mantle. (van der Hilst *et al.*, 1998. *The Core–Mantle Boundary Region*, 5–20, copyright by the American Geophysical Union.)

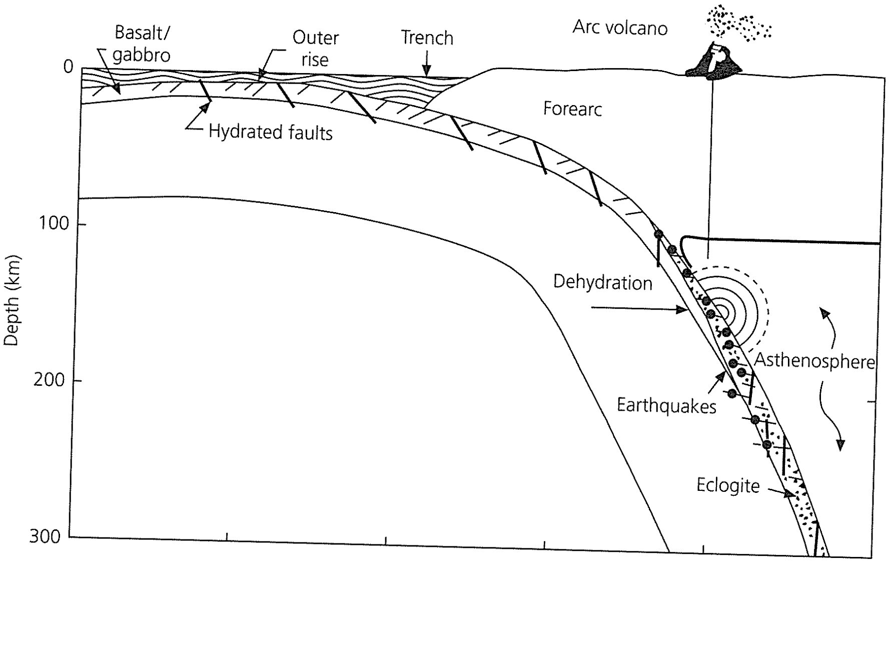


Fig. 5.4-22 Schematic model for intermediate depth earthquakes. Earthquakes are assumed to occur in subducting crust and be associated with the dehydration of mineral phases and the gabbro to eclogite transition. (Kirby *et al.*, 1996a. *Subduction*, 195–214, copyright by the American Geophysical Union.)

schematic model in which the earthquakes are presumed to occur in subducting oceanic crust, rather than throughout the subducting mantle that makes up most of the slabs, because detailed location studies show that the earthquakes are close to the top of the subducting slabs. The crust should undergo two important mineralogical transitions as it subducts. Hydrous (water-bearing) minerals formed at fractures and faults should warm up and dehydrate. Eventually, the gabbro transforms to eclogite, a rock of the same chemical composition composed of denser minerals.[4] Under equilibrium conditions, eclogite should form by the time slab material reaches about 70 km depth. However, travel time studies in some slabs find a low-velocity waveguide interpreted as subducting crust extending to deeper depths. Hence it has been suggested that the eclogite-forming reaction is slowed in cold downgoing slabs, allowing gabbro to persist metastably. Once dehydration occurs, the freed water weakens the faults, favoring earthquakes and promoting the eclogite-forming reactions. In this model the intermediate earthquakes occur by slip on faults, but the phase changes favor faulting. The extensional focal mechanisms may also reflect the phase change, which would produce extension in the subducting crust. Support for this model comes from the fact that the intermediate earthquakes occur below the island arc volcanoes, which are thought to result when water released from the subducting slab causes partial melting in the overlying asthenosphere.

The fact that various explanations are under discussion illustrates the difficulty in understanding the complex thermal structure, mineralogy, rheology, and geometry of real slabs. We can think of the deep subduction process as a chemical reactor that brings cold shallow minerals into the temperature and pressure conditions of the mantle transition zone, where these phases are no longer thermodynamically stable (Fig. 5.4-23). Because we have no direct way of studying what is happening and what comes out, we seek to understand this system by studying earthquakes that somehow reflect what is happening. This is a major challenge, and we have a long way to go.

5.4.3 *Interplate trench earthquakes*

Much of what is known about the geometry and mechanics of the interaction between plates at subduction zones comes from the distribution and focal mechanisms of shallow earthquakes at the interface between the plates. These include the largest earthquakes that occur, as illustrated by Fig. 5.4-24, showing the largest earthquakes (surface wave magnitude greater than 8.0) during 1904–76. Among these are the two largest earthquakes ever recorded seismologically: the 1960 Chilean (M_0 2 × 10^{30} dyn-cm, M_s 8.3) and 1964 Alaska (M_0 5 × 10^{29} dyn-cm, M_s 8.4) earthquakes. Figure 5.4-25 shows the geometry of the Chilean earthquake: 21 meters of slip occurred on a fault 800 km long along strike, and 200 km wide down-dip. The mechanism shows thrusting of the South American plate over the subducting oceanic lithosphere of the Nazca plate. The aftershock zone was 800 km long, and the surface deformation was dramatic, reaching 6 meters of uplift in places. Thrust earthquakes of this type, although smaller, make up most of the large, shallow events at subduction zones. Such *interplate* earthquakes release the plate motion that has been locked at the plate interface. As we saw in Section 4.6.1, these can be much bigger than the largest earthquakes at transform fault boundaries like the San Andreas. For example, even the 1906 San Francisco earthquake was tiny (100 times smaller seismic moment) compared to the 1964 Alaska earthquake, although both occurred along different segments of the same plate boundary. The difference reflects the fact that faulting occurs only when rock is cooler than a limiting temperature. Thus a vertically dipping transform like the San Andreas has a much shorter cold down-dip extent than the shallow-dipping thrust interfaces (sometimes called megathrusts) at subduction zones.

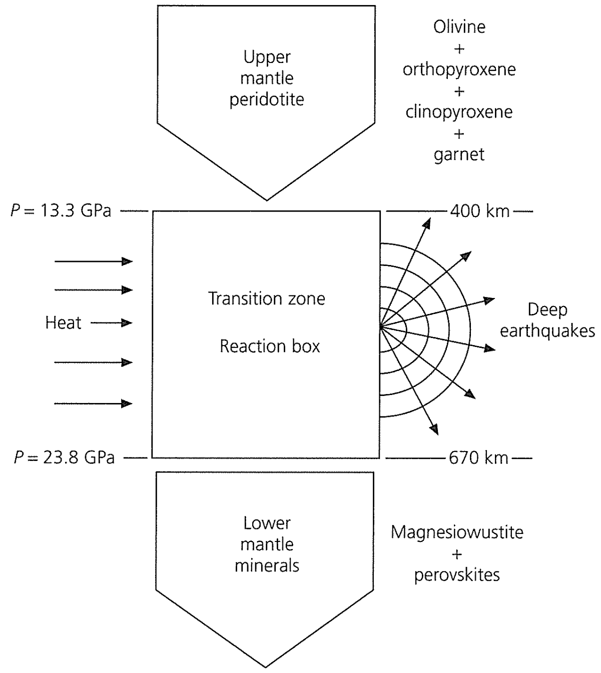


Fig. 5.4-23 Cartoon of subducting slabs in the transition zone as a chemical reactor. (Kirby *et al.*, 1996b. *Rev. Geophys.*, *34*, 261–306, copyright by the American Geophysical Union.)

Major thrust earthquakes at the interface between subducting and overriding plates directly indicate the nature of subduction. In most cases, their focal mechanisms show slip toward the trench, approximately in the convergence direction predicted by global plate motion models or space-based geodesy (Section 5.2) (Fig. 5.2-3). However, in some cases when the plate motion is oblique to the trench, a forearc sliver moves separately from the overriding plate (Fig. 5.4-26). This effect,

[4] Most of the oceanic crust consists of gabbro, the intrusive version of the extrusive basalt seen at mid-ocean ridges (Section 3.2.5). With increasing pressure, gabbro becomes eclogite as feldspar and pyroxene transform to garnet.

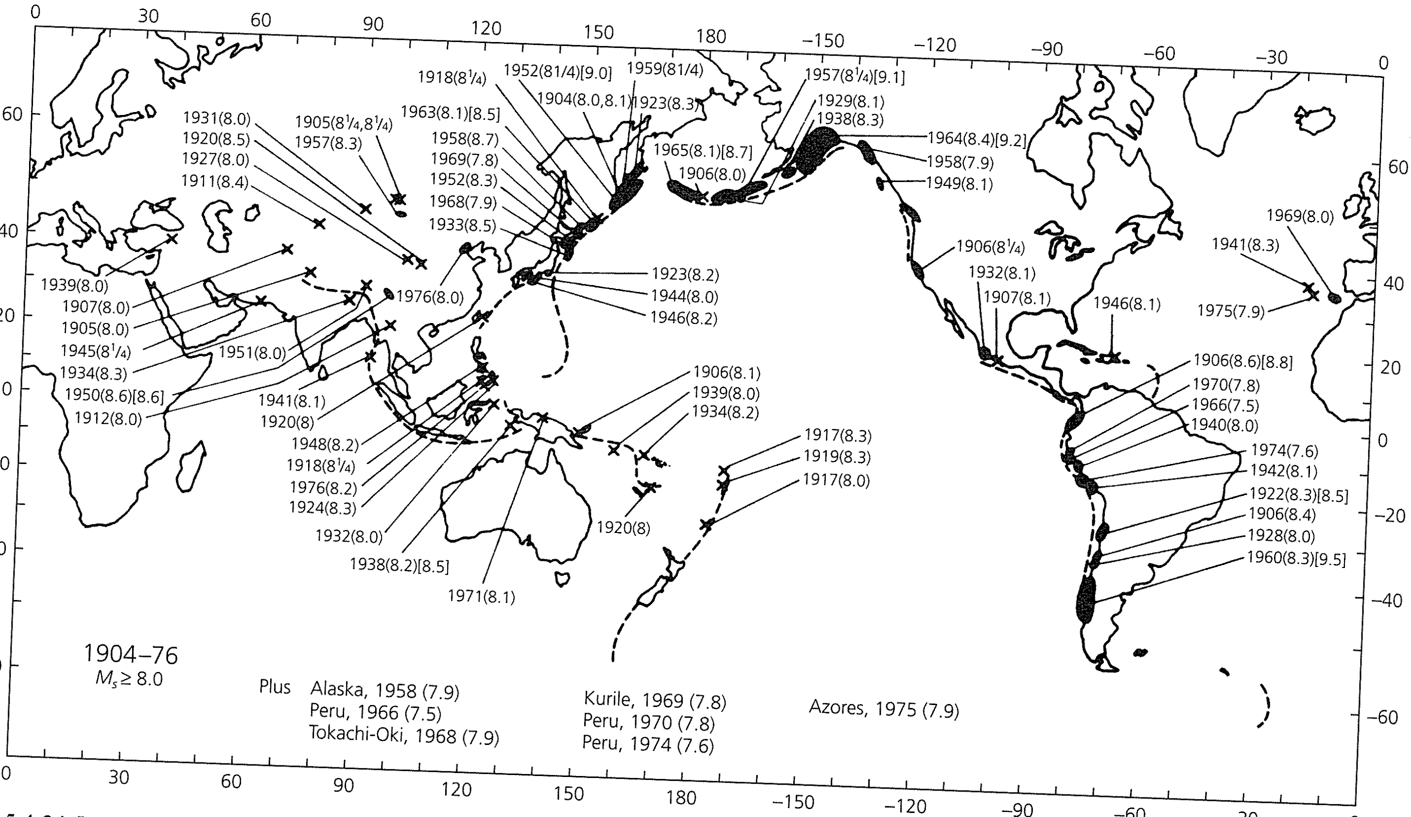


Fig. 5.4-24 Location of the largest earthquakes between 1904 and 1976. M_s values are in parentheses and M_w values in square brackets. Most are at subduction zones and result from thrust faulting at the interface between the two plates. (Kanamori, 1978. Reproduced with permission from *Nature*.)

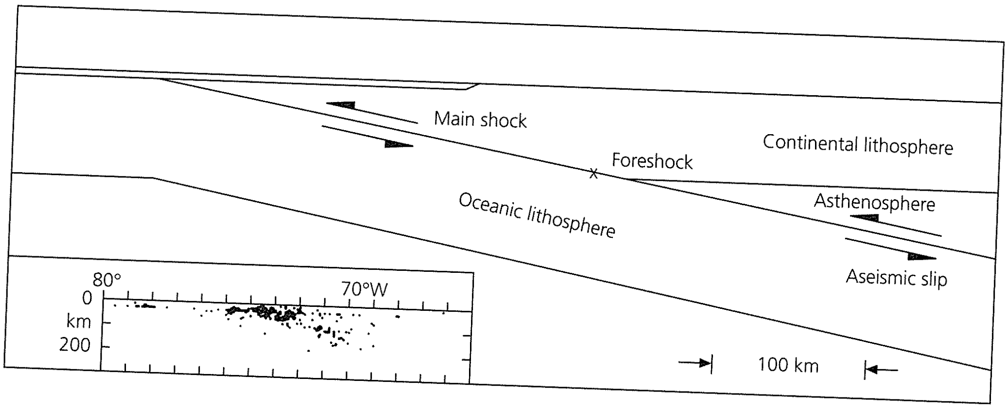


Fig. 5.4-25 Fault geometry and aftershock distribution (*insert*) for the 1960 Chilean earthquake. (Kanamori and Cipar, 1974. *Phys. Earth Planet. Inter.*, 9, 128–36, with permission from Elsevier Science.)

called *slip partitioning*, makes earthquake slip vectors at the trench trend between the trench-normal direction and the predicted convergence direction, and causes strike-slip motion between the forearc and the stable interior of the overriding plate. This effect can be seen in plate motion studies and with GPS data, and can cause misclosure of plate circuits. In the limiting case of pure slip partitioning, pure thrust faulting would occur at the trench, and all the oblique motion would be accommodated by trench-parallel strike-slip.

How the thrust earthquakes release the accumulated plate motion is both interesting scientifically and important for assessing earthquake hazards. In many subduction zones, thrust earthquakes have characteristic patterns in space and time. For example, large earthquakes have occurred in the Nankai trough area of southern Japan approximately every 125 years since 1498 with similar fault areas (Fig. 5.4-27).[5] In some cases the entire region seems to have slipped at once; in others, slip was divided into several events over a few years.

[5] Due to its location where between four and six plates (North America, Pacific, Philippine, Eurasia, and perhaps Okhotsk and Amuria) interact, Japan has a high level of seismicity, which was originally attributed to the motion of the namazu, a giant underground catfish. As a result, Japan has an outstanding tradition of seismology and some of the best data in the world for studying subduction-related earthquakes.

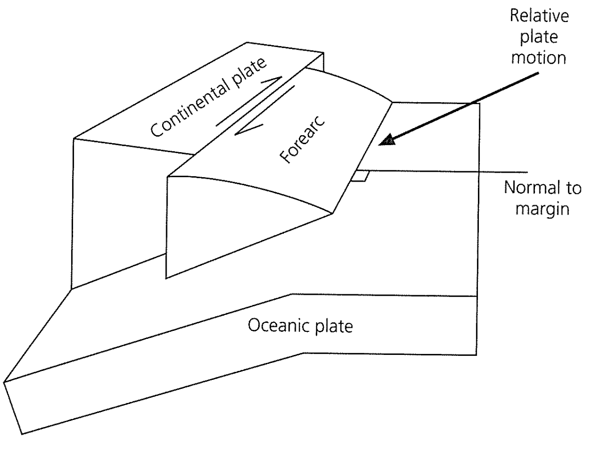


Fig. 5.4-26 Schematic illustration of forearc sliver motion when convergence is oblique. (Courtesy of D. Davis.)

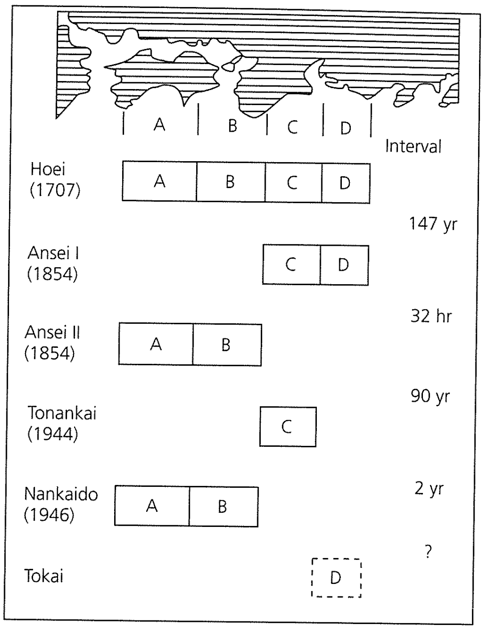


Fig. 5.4-27 Time sequence of large subduction zone earthquakes along the Nankai trough, suggesting both some space and time periodicity and some variability. (Ando, 1975. *Tectonophysics, 27,* 119–40, with permission from Elsevier Science.)

Given such repeatability, it seems likely that a segment of a subduction zone that has not slipped for some time constitutes a seismic gap and is due for an earthquake. For example, the Tokai area (segment D) may be such as case and is the focus of extensive earthquake prediction studies. However, despite the intuitive appeal of the gap idea, efforts to predict the location of future earthquakes using it have not generally been successful (Sections 1.2.5, 4.7.3).

One difficulty is that not all of the plate motion occurs seismically. Figure 5.4-28 shows that during 1952–73 a large segment of the Kuril trench slipped in a series of six major earthquakes with similar thrust fault mechanism. Seismic moment studies show that the average slip was 2–3 meters. Since the previous major earthquake sequence in the area occurred about 100 years earlier, the average seismic slip rate is 2–3 cm/yr, about one-third of the plate motion predicted from relative motion models. The remaining two-thirds of the slip occurs aseismically, as postseismic or interseismic motion. Similar studies around the world find that the fraction of plate motion that occurs as seismic slip, sometimes called the *seismic coupling* factor, is generally much less than 1, implying that much of the plate motion occurs aseismically if the time interval sampled is adequate.

The Chilean subduction zone shows the other extreme. The seismic slip rate, estimated from the slip in the great 1960 earthquake and historical records indicating that major earthquakes occurred about every 130 years during the past 400 years, exceeds the convergence rate predicted by plate motion models (Fig. 5.4-29). Because the convergence rate is an upper bound on the seismic slip rate, the two estimates are inconsistent. One possibility is that the seismic slip is overestimated: either the earlier earthquakes were significantly smaller than the 1960 event or their frequency in the past 400 years is higher than the long-term average.

More generally, these examples illustrate the difficulty in inferring seismic slip from historical seismicity, owing to problems including the variability of earthquakes on a given plate boundary, the issue of whether the time sample is long enough, and the difficulty in estimating source parameters for earthquakes that pre-dated instrumental seismology. Given the uncertainties in estimating the slip in an earthquake even with seismological data (Section 4.6), doing so without such data is particularly challenging. An alternative approach to estimating plate coupling, discussed in Sections 4.5.4 and 5.6.2, uses GPS geodesy to measure the deflection of the overriding plate, which will be released in future large earthquakes. This deflection depends on the mechanical coupling at the interface, so directly measures what we infer indirectly from the earthquake history. However, the GPS data sample only the present earthquake cycle, which may not be representative of long-term behavior.

Perhaps for similar reasons, efforts to interpret the seismic slip fraction in terms of the physical processes of subduction have not yet been successful. Although the term "seismic coupling" implies a relation between the seismic slip fraction with properties such as the mechanical coupling between the subducting and overriding lithospheres, this has been hard to establish. This relation was originally posed in terms of two end members: coupled Chilean-type zones with large earthquakes and uncoupled Mariana-style zones with largely aseismic subduction. The largest subduction zone earthquakes appear to occur where young lithosphere subducts rapidly (Fig. 5.4-30, *top*), where we might expect the minimum "slab pull" effects and hence the strongest coupling. However,

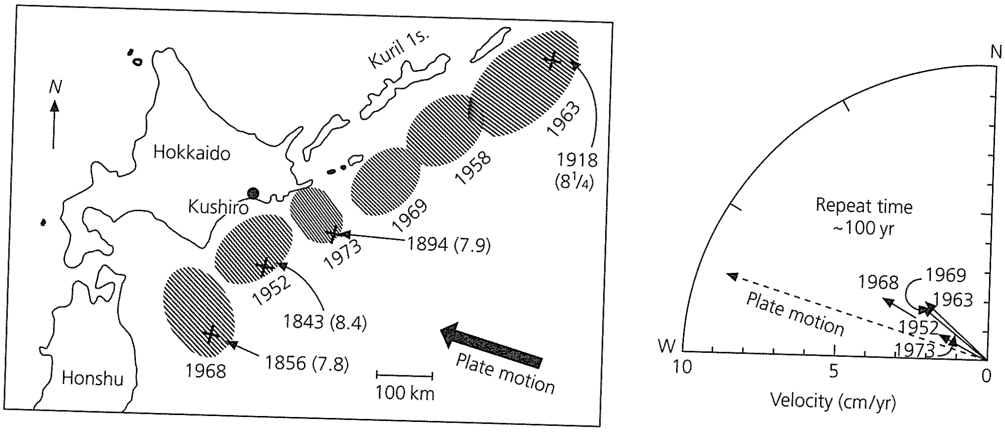


Fig. 5.4-28 Rupture areas for a sequence of large subduction zone earthquakes along the Kuril trench. Different segments of the boundary slip seismically over time. Arrows show the direction and rate of seismic slip and plate motion. If such sequences occur about every 100 years and this time sample is representative, the seismic slip is only about one-third of the plate motion. (Kanamori, 1977b. *Island Arcs, Deep Sea Trenches and Back Arc Basins*, 163–74, copyright by the American Geophysical Union.)

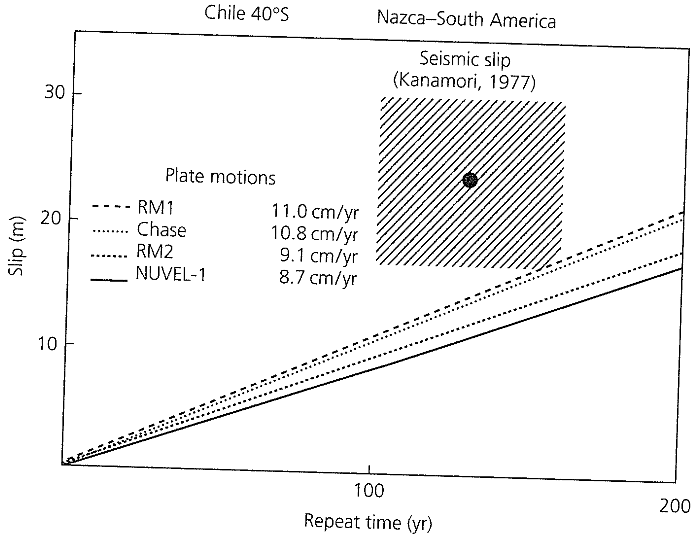


Fig. 5.4-29 Comparison of seismic slip rate and plate motions for the area of the great 1960 Chilean earthquake. Shaded region gives slip rate estimated from slip in the 1960 event and recurrence of large trench earthquakes in the last 400 years. The estimated slip rate exceeds that predicted by any of the four plate motion models shown. (Stein *et al.*, 1986. *Geophys. Res. Lett.*, *13*, 713–16, copyright by the American Geophysical Union.)

efforts to correlate the seismic slip fraction with subduction zone properties such as convergence rate or plate age find no clear pattern (Fig. 5.4-30, *bottom*). It has also been suggested that seismic coupling may be lowest for sedimented trenches and where normal stress on the plate interface is low, although these plausible ideas have yet to be demonstrated. Thus, although seismic coupling can be defined from the seismic slip fraction, its relation to the mechanics of plate coupling is still unclear. It appears that most subduction zones have significant components of aseismic slip, as do oceanic transforms and many continental plate boundaries (Section 5.6.2). Hence, even given the considerable uncertainties in such estimates, it appears common for a significant fraction of plate motion to occur aseismically.

The difficulty in estimating seismic coupling and understanding the process of aseismic plate motion has consequences for estimating the recurrence of earthquakes on a plate boundary and the seismic gap concept. It may be difficult to distinguish between gaps and areas where much of the slip is aseismic. For example, we would not want to say both that areas with recent major seismicity have high seismic hazard and that areas with little recent seismicity are gaps with high seismic hazard.[6] Moreover, as discussed in Sections 1.2 and 4.7.3, the process of earthquake faulting may be sufficiently random that it is hard to use the plate motion rate and seismic history to usefully predict how long it will be until the next large earthquake.

Although most shallow subduction zone seismicity is at the plate interface, some earthquakes occur within either plate. Some appear to result from flexural bending of the downgoing plate as it enters the trench (Fig. 5.4-31). Focal depth studies show a pattern of normal faulting in the upper part of the plate to a depth of 25 km, and thrusting in its lower part, between 40 and 50 km. These observations constrain the position of the neutral surface dividing the mechanically strong lithosphere (Section 5.7.4) into upper extensional and lower compressional zones. In some cases the normal fault earthquakes are so large that they may be "decoupling" events due to "slab pull" that rupture the entire downgoing plate (Fig. 5.4-32). Aftershock distributions and studies of the rupture process indicate that faulting extended through a major portion, and perhaps all, of the lithosphere. Rupture through the entire lithosphere favors the decoupling model. If only a portion of the lithosphere breaks, the interpretation is more complicated. Rupture may have been restricted to one side of the neutral surface (in the flexural model) or reflect the material below being too hot and weak for seismic rupture. In the latter case, the entire lithosphere could have failed, with the deeper rupture being aseismic.

[6] The observation that more recent grizzly bear attacks have occurred in Montana than in Illinois might indicate either a perilous "gap" in Illinois or a greater intrinsic hazard in Montana.

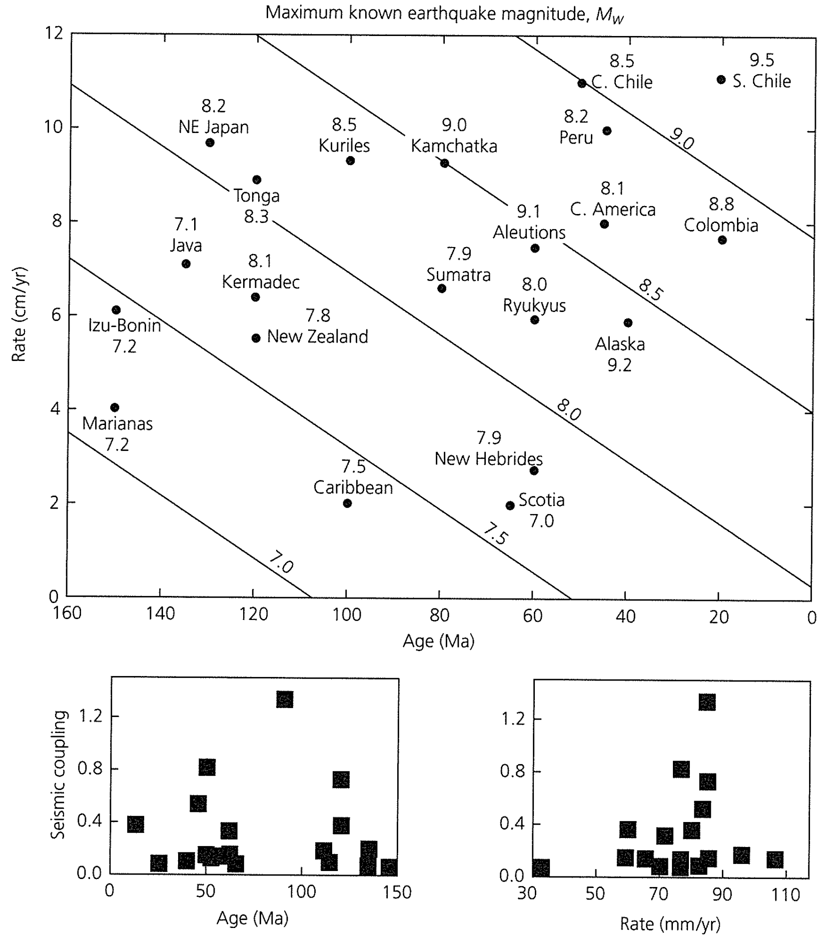


Fig. 5.4-30 *Top*: Variation in the magnitude (M_w) of the largest known subduction thrust fault earthquake between subduction zones as a function of the convergence rate and age of the subducting lithosphere. (Ruff and Kanamori, 1980. *Phys. Earth Planet. Inter.*, *23*, 240–52, with permission from Elsevier Science.) *Bottom*: Seismic coupling fraction estimated from historical seismicity at various subduction zones. Although most subduction zones show considerable aseismic slip, there is no obvious correlation with either age of the subducting lithosphere (*left*) or subduction rate (*right*). (Pacheco *et al.*, 1993. *J. Geophys. Res.*, *98*, 14, 133–59, copyright by the American Geophysical Union.)

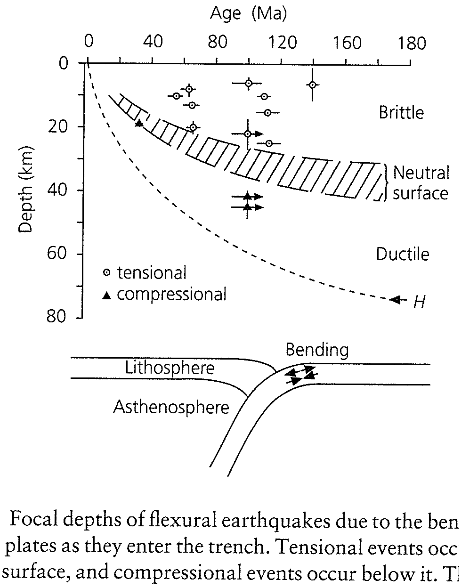


Fig. 5.4-31 Focal depths of flexural earthquakes due to the bending of subducting plates as they enter the trench. Tensional events occur above the neutral surface, and compressional events occur below it. The plate mechanical thickness, *H*, increases with age, as expected from thermal models. (After Bodine *et al.*, 1981. *J. Geophys. Res.*, *86*, 3695–707, copyright by the American Geophysical Union.)

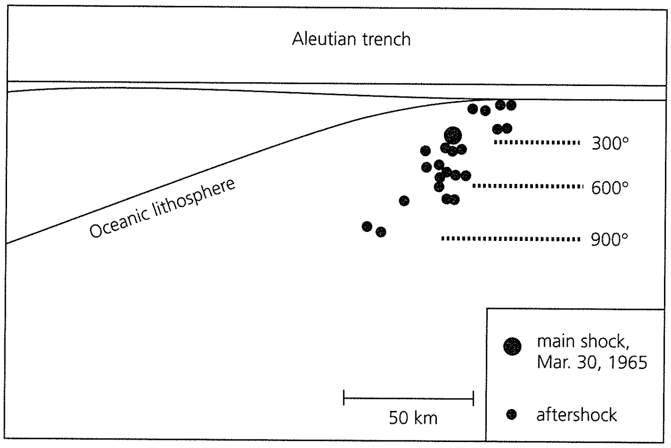


Fig. 5.4-32 Large normal faulting earthquakes at trenches, such as the 1965 M_s 7.5 Rat Island earthquake, may be due to flexure or failure of the lithosphere under its own weight. The extent of aftershocks, which appear not to cut the entire lithosphere, may reflect the extent of rupture or be a temperature effect. (Wiens and Stein, 1985. *Tectonophysics, 116*, 143–62, with permission from Elsevier Science.)

5.5 Oceanic intraplate earthquakes and tectonics

The vast majority of earthquakes — especially when measured in terms of seismic moment release — occur on plate boundaries and reflect the relative plate motions there. However, *intraplate* earthquakes, those within plates, also provide important tectonic information. We discuss intraplate earthquakes that occur in oceanic lithosphere in this section, and then discuss their counterparts in continental lithosphere in the next.

5.5.1 Locations of oceanic intraplate seismicity

Figure 5.5-1 illustrates the distribution of earthquakes in the Atlantic Ocean, excluding those along the Mid-Atlantic ridge. Although these earthquakes are rarer than those along the ridges and transforms making up the Mid-Atlantic ridge plate boundaries, there are enough to justify interest. They nicely illustrate that plates deviate from the ideal case of perfect rigidity without internal deformation, such that all motion occurs at narrow boundaries. Instead, as noted in Section 5.2, real plates are complicated entities that have both internal deformation and diffuse boundary zones.

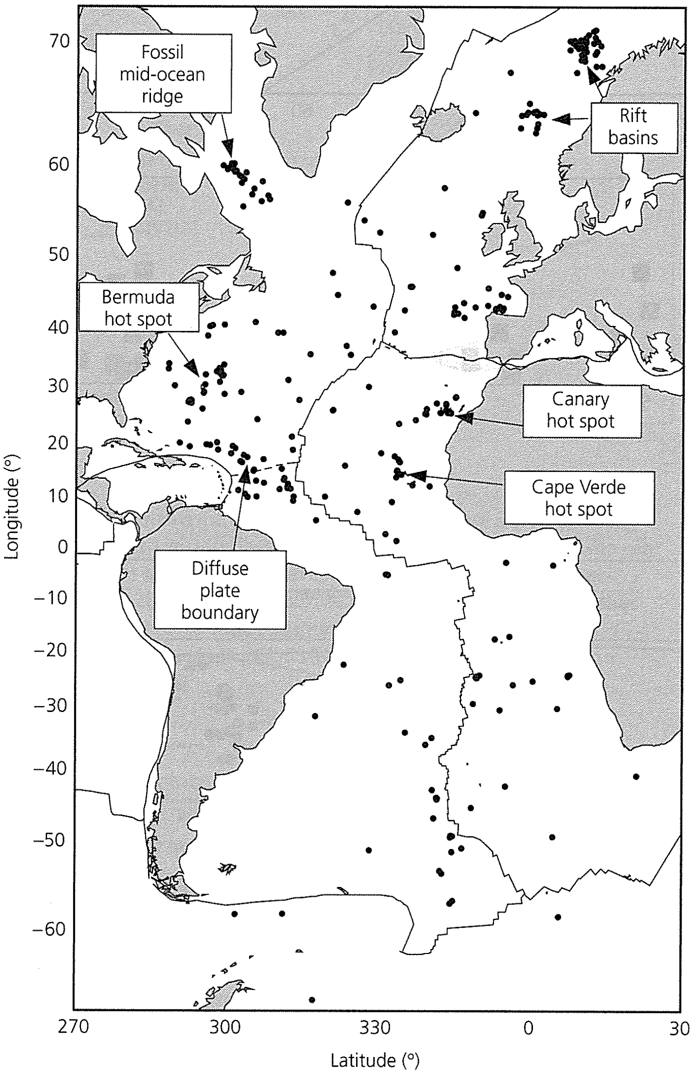


Fig. 5.5-1 Distribution of earthquakes in the Atlantic Ocean other than those on ridge and transform segments of the Mid-Atlantic ridge system. (Wysession *et al.*, 1995. © Seismological Society of America. All rights reserved.)

One way to think about these earthquakes is to consider a hierarchy, from slow-moving plate boundaries, to recognizable weak structures, and then to apparently isolated earthquakes. For example, the Atlantic portion of the boundary between the Eurasian and African plates, which stretches from Gibraltar to the Azores, is poorly defined by topography and seismicity compared to the Mid-Atlantic ridge. However, the focal mechanisms (Fig. 5.5-2, *top*) show a transition from extension at the Terceira Rift near the Azores, to strike-slip along a segment that includes the mapped Gloria transform fault, to compression near Gibraltar, and then into the Mediterranean. This transition reflects the fact that the Euler pole is close enough that the relative motions are small and change rapidly with distance (Fig. 5.5-2, *bottom*). For example, near the triple junction the NUVEL-1A model (Table 5.2-1) predicts 4 mm/yr of extension resulting from the small difference between Eurasia–North America (23 mm/yr at N97°E) and Africa–North America (20 mm/yr at N104°E) spreading across the Mid-Atlantic ridge. Even in the western Mediterranean, the motions are too slow to generate a well-developed subduction zone like those of the Pacific, but instead cause a broad convergent zone indicated by large earthquakes like the 1980 M_s 7.3 El Asnam, Algeria, earthquake.

Even slower motion appears to be why sea floor topography shows no clear evidence for the boundary between the North American and South America plates shown by the dashed line in Fig. 5.5-1, despite a diffuse zone of seismicity in this area. This zone is considered to be a plate boundary, based on detailed studies of plate motions. These studies invert plate motion data (spreading rates, transform fault directions, and earthquake slip vectors; Section 5.2.2) to find Euler vectors under two different assumptions: either there is a single American plate, or there are two. The Euler vectors derived by assuming there are two plates fit the data better, which would be expected, because a model with more parameters always fits data better. However, statistical tests (Section 7.5.2) show that the fit to the data improves more than expected purely by chance due to the additional parameters, implying that the two plates are distinct.

The North America–South America Euler vector that results from inverting the data is not well constrained, because it is not derived directly from data recording the motion between North America and South America, but is estimated from closure of the plate circuit (Fig. 5.2-5). Thus the estimate of motion results from the difference between North America–Africa and South America–Africa motions, which are quite similar (if they were not, the data would clearly show two distinct American plates). The predicted motion along the North

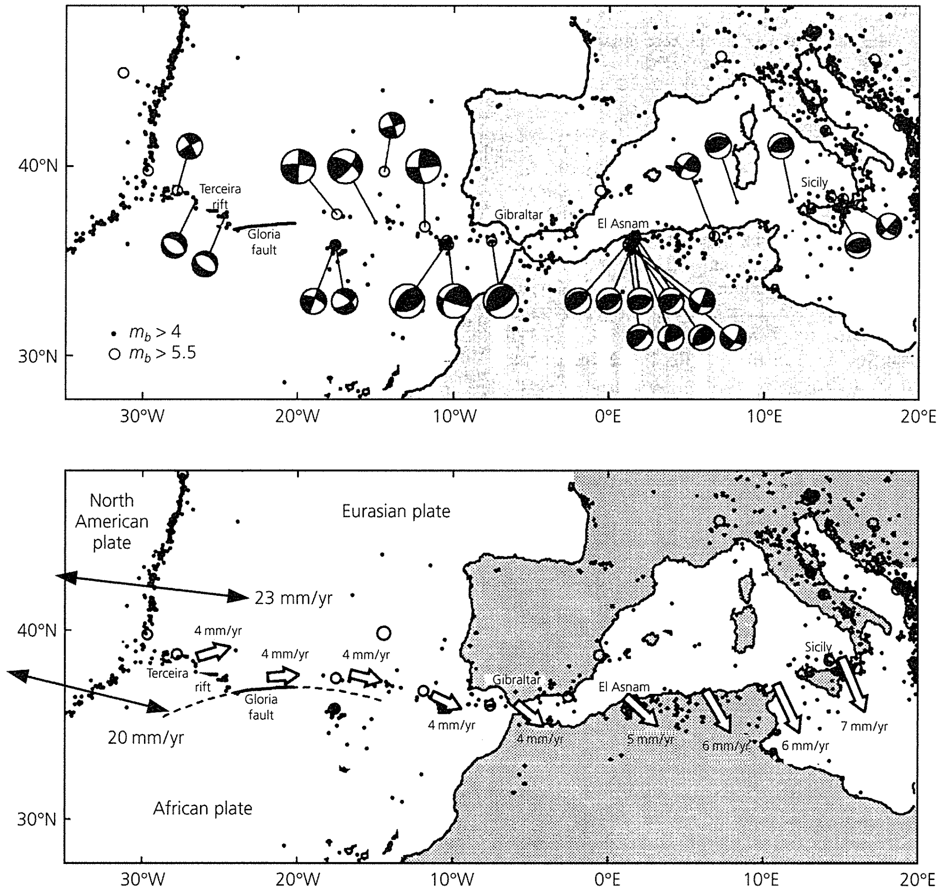


Fig. 5.5-2 *Top*: Focal mechanisms along the western section of the Eurasia–Africa plate boundary. Note the transition from extension near the Azores, to strike-slip (the Gloria fault is a transform), to compression near Gibraltar and into the Mediterranean. *Bottom*: Motions with respect to Africa along the boundary predicted by an Euler pole slightly south of the mapped area, near 20°N, 20°W. The dashed line is a small circle about this pole. (Argus *et al.*, 1989. *J. Geophys. Res.*, *94*, 5585–5602, copyright by the American Geophysical Union.)

America–South America boundary is only about 1 mm/yr — much slower than the approximately 20 mm/yr along the Mid-Atlantic ridge. The North America–South America boundary is thus considered a diffuse, slow-moving boundary zone, although its location and motion are not well constrained. Another reason for treating this as a boundary zone is that paleomagnetic reconstructions find that over the past 70 Myr the two plates have moved relative to each other as the Atlantic Ocean opened.

In general, 1–2 mm/yr is an approximate lower limit for plate boundary deformation. Regions with motions faster than this are generally viewed as plate boundaries, and slower deformation is generally treated as intraplate. However, there is no generally accepted criterion, and evidence from seismicity and topography is also considered. Put another way, in many cases one can regard a region as either a slow-moving plate boundary zone or a zone of intraplate deformation, and "intraplate" earthquakes are often just ones not on an obvious plate boundary.

The Atlantic example (Fig. 5.2-1) shows that in addition to the North America–South America boundary zone, some intraplate seismicity is concentrated in other areas associated with tectonic features. For example, seismicity between Greenland and North America is likely related to the former spreading ridge that opened this part of the Atlantic (the Labrador Sea). Although this spreading stopped about 43 Myr ago, the fossil ridge appears to remain a weak zone along which intraplate stresses cause some motion. Intraplate seismicity is often associated with such fossil structures. Concentrations of seismicity are also associated with the Bermuda (32°N, 65°W), Cape Verde (17°N, 25°W), and Canary (26°N, 17°W) hot spots. Focal mechanism studies are consistent with the earthquakes reflecting heating of the lithosphere by the hot spots.

Hawaii, the most impressive hot spot trace in the oceans (Fig. 5.2-7),[1] provides the best example of intraplate earthquakes associated with hot spot processes (Fig. 5.5-3). Small earthquakes are associated with magma upwelling in the rift zones. Larger earthquakes, which occur on a time scale of tens of years, reflect sliding of the volcanic edifice on subhorizontal faults that are thought to be a layer of weak sediments at the top of the old oceanic crust on which the volcanic island formed. These earthquakes can be quite large — the 1975

[1] Numerical models that infer the amount of upwelling mantle material from how elevated the sea floor is relative to the normal depth–age curves estimate that Hawaii has a buoyancy flux 5–10 times greater than that of Bermuda (Sleep, 1990).

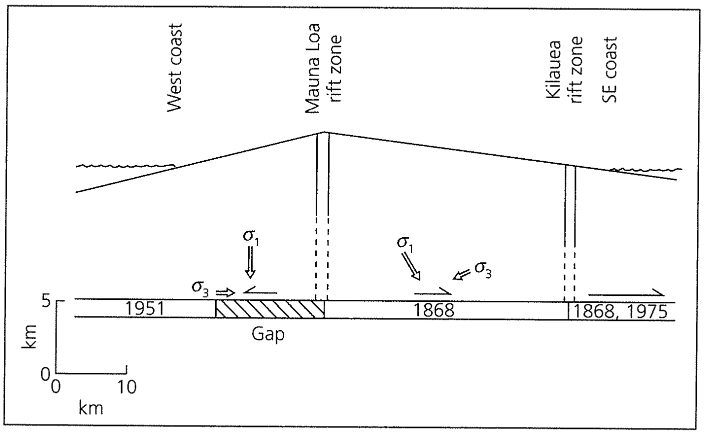


Fig. 5.5-3 Schematic model for large intraplate earthquakes below the island of Hawaii. Small earthquakes are associated with magma upwelling in the rift zones. Larger earthquakes, at dates shown, reflect sliding of the volcanic edifice on subhorizontal faults. The portion of the basal fault that has not ruptured in historic time may be a seismic gap. (Wyss and Koyanagi, 1992. © Seismological Society of America. All rights reserved.)

Kalapana earthquake had M_s 7.2, caused a tsunami that killed two campers on the seashore, and did considerable property damage. The earthquake was followed by a small volcanic eruption near the summit of Kilauea, perhaps because the ground shaking triggered an eruption of shallow magma. Curiously, some earthquakes occur to considerable depths under Hawaii, including a magnitude 6.2 earthquake at 48 km depth.

Although many oceanic intraplate earthquakes are associated with tectonic features, some appear to occur far from plate boundaries, hot spots, or major bathymetric features. Thus the stresses generated by plate driving forces and other sources, including mantle flow near hot spots, appear to reactivate weak zones in the plate resulting from small-scale structure acquired during the lithosphere's evolution.[2]

These earthquakes can be dramatic. For example, the enormous (M_w 8.2) intraplate earthquake that occurred near the Balleny Islands in an oceanic part of the Antarctic plate (63°S, 149°E) in March 1998 was the largest earthquake that had occurred on earth for several years. The fault inferred from waveform modeling (Section 4.3) followed no observable lineaments and cut straight across existing fracture zones. Moreover, in the previous hundred years, no other earthquakes had been located in this region. It is not clear what caused the earthquake or whether this area has any special properties or stress acting there. Although the earthquake occurred south of a puzzling hypothesized deformation zone in the extreme southeast corner of the Australian plate (Fig. 5.2-4), its fault plane solution is inconsistent with its being on the boundaries of a microplate. It is thus unclear whether this area is now any more prone to future earthquakes than other areas, and what the recurrence time of such earthquakes might be. Similar issues arise in considering the intraplate seismicity and associated seismic hazard in the more structurally complex continents.

Oceanic intraplate seismicity often occurs in swarms. Regions without previously known seismicity sometimes become active for several years, with hundreds of teleseismically located earthquakes.[3] The seismicity then dies out, and seems not to recur. For example, during 1981–3, an intraplate earthquake swarm occurred near the Gilbert Islands in Micronesia. A total of 225 earthquakes were detected, mostly over a 15 month period, with 87 above m_b 5. No major tectonic features are known in this area, and a ship survey found no bathymetric anomalies. Before and after the swarm, no other earthquakes have been recorded in this region. The swarms thus differ from plate boundary seismicity, which occurs on features that remain active for long periods even if there are intervening quiet intervals. Moreover, the intraplate swarms often appear not to have a single well-developed fault, and no event is significantly bigger than the others. By contrast, plate boundary earthquakes usually have one or two main ruptures and many aftershocks, perhaps reflecting local adjustments to the stress field after the mainshock has ruptured the entire fault.

These swarms raise an interesting issue. We can assume that these areas are analogous to plate boundaries in having special, if not yet understood, tectonic significance. If so, they are likely to be the sites of future swarms. Alternatively, perhaps all areas of oceanic lithosphere are equally susceptible to such swarms. In this case, over time, swarms will occur in many places, and future swarms are no more likely in one place than another. We will see that similar issues surface in trying to estimate seismic hazards due to intraplate earthquakes within continents.

5.5.2 *Forces and stresses in the oceanic lithosphere*

In addition to using oceanic intraplate seismicity to investigate the specific processes acting at individual sites, we study the seismicity to learn about plate-wide processes. For example, Fig. 5.5-4 shows the variation of mechanism type with lithospheric age. Most of the oceanic lithosphere seems to be in horizontal deviatoric compression, as shown by thrust and strike-slip mechanisms. This compression is in approximately the spreading direction, and is thought to be related to "ridge push": the plate driving force due to lithospheric cooling and subsidence. The major exceptions are the extensional events occurring in the central Indian Ocean. Although originally regarded as intraplate, these earthquakes now appear to be in a diffuse plate boundary zone (Section 5.2.2). In the model shown, the focal mechanisms (Fig. 5.5-5) reflect counterclockwise rotation of Australia with respect to India, causing normal fault earthquakes in the young lithosphere near the Euler pole

[2] This situation is analogous to timbers creaking as a wooden boat rocks in the waves.

[3] There may be many more smaller earthquakes associated with these swarms, but because the swarms often occur in remote regions, only the larger events are detected.

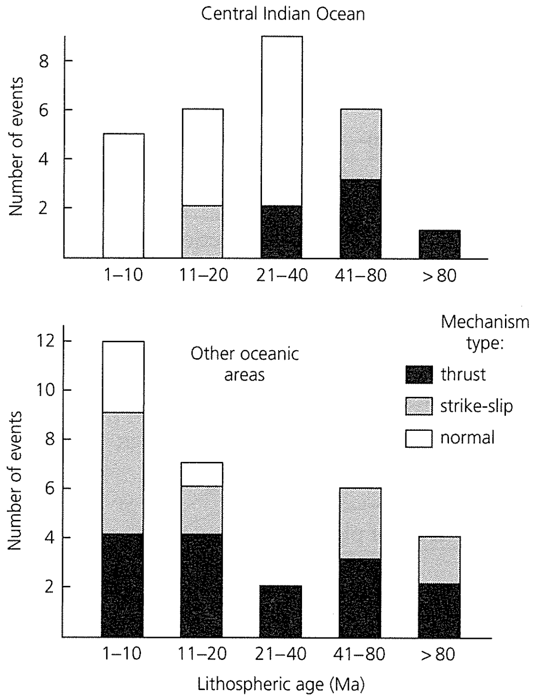


Fig. 5.5-4 Focal mechanism type as a function of lithospheric age for oceanic intraplate earthquakes. Older oceanic lithosphere is in compression, whereas younger lithosphere has both extensional and compressional mechanisms. Extensional events are located primarily in the central Indian Ocean. (Wiens and Stein, 1984. *J. Geophys. Res.*, *89*, 11, 442–64, copyright by the American Geophysical Union.)

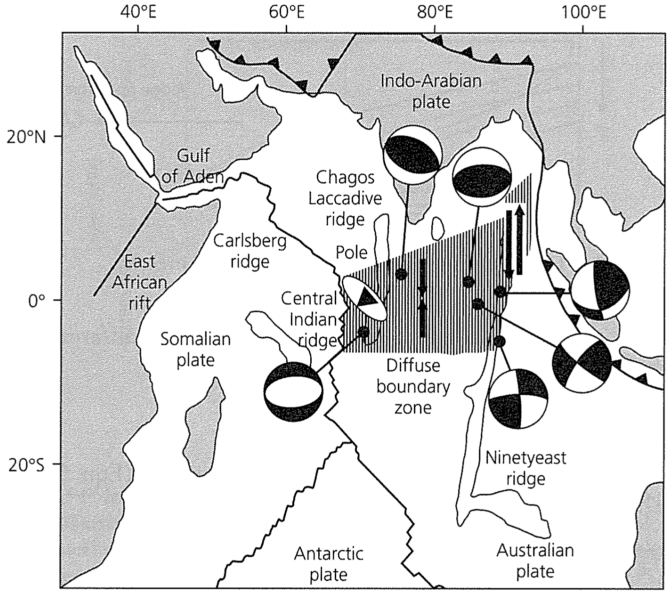


Fig. 5.5-5 Schematic map of earthquake mechanisms in the central Indian Ocean, shown here as a diffuse boundary zone (shaded) between the Indian and Australian plates. Later studies have refined the location and geometry of the boundary zone (Fig. 5.2-4) and pole (triangle) (Wiens *et al.*, 1985. *Geophys. Res. Lett.*, *12*, 429–32, copyright by the American Geophysical Union.)

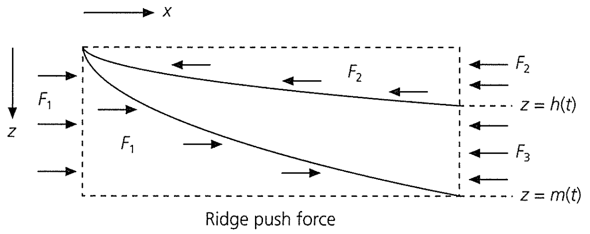


Fig. 5.5-6 Derivation of the "ridge push" force.

and thrust and strike-slip earthquakes to the east. These earthquakes reach magnitude 7 on the Ninetyeast ridge.[4]

The general trend of compressive mechanisms in the oceanic plates is consistent with the plate driving force due to the cooling of the oceanic lithosphere. Consider a plate, defined as the area above the $m(t)$ isotherm, out to age t, where the water depth is $h(t)$ (Fig. 5.5-6). The plate is cooler, and thus denser, than material below. The thermal model we used for ocean depth and heat flow also predicts the resulting force.

The total horizontal force on the base of the lithosphere, F_1, equals the integrated horizontal pressure force of the asthenosphere at the ridge, because the material is in hydrostatic equilibrium:

$$F_1 = \int_0^{m(t)} \rho_m gz dz = \rho_m g(m(t))^2/2. \qquad (1)$$

Similarly, F_2, the horizontal force due to water pressure on the plate, equals the integrated horizontal pressure force of the water,

$$F_2 = \int_0^{h(t)} \rho_w gz dz = \rho_w g(h(t))^2/2. \qquad (2)$$

F_3 is the remaining horizontal force due to lithospheric pressure $P(z, t)$,

$$F_3 = \int_{h(t)}^{m(t)} P(z, t) gz dz, \qquad (3)$$

where the pressure depends on the density perturbation due to lithospheric cooling (Eqn 5.3.7),

$$P(z, t) = \rho_w gh(t) + g \int_{h(t)}^{z} [\rho_m + \rho'(z', t)] dz'. \qquad (4)$$

[4] Although hot spot tracks like the Ninetyeast and Chagos-Laccadive ridges have been termed "aseismic" ridges, to distinguish them from spreading ridges, these two are more seismically active in terms of moment release than many spreading ridges.

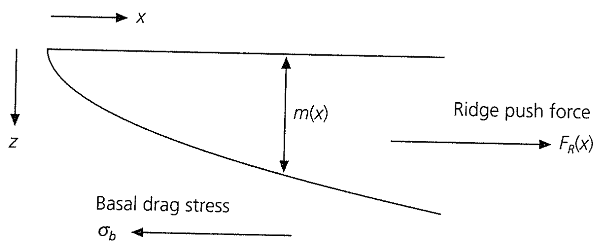


Fig. 5.5-7 Geometry for a simple model of intraplate stresses.

If the plate is not accelerating, the force difference is balanced by a net horizontal force

$$F_R = F_1 - F_2 - F_3. \tag{5}$$

For the cooling halfspace temperature structure (Eqn 5.3.2), this force is

$$F_R = g\alpha\rho_m T_m \kappa t, \tag{6}$$

whereas for a plate model it approaches a constant value for old lithosphere. The convention of calling this force "ridge push" is confusing because it is zero at the ridge and increases linearly with plate age. It results not from force at the ridge but from the total force due to the density anomaly within the cooling plate out to any given age.

The expression for the "ridge push" force is similar to that for the "slab pull" force (Eqn 5.4.15) because both are thermal buoyancy forces due to the density contrast resulting from the temperature difference between the plate and its surroundings. The two depend in the same way on the $g\alpha\rho_m T_m$ term that describes the force due to the density contrast, but differently on κ because faster cooling increases ridge push whereas faster heating decreases slab pull. Although it is useful to think of the forces separately, both are net buoyancy forces due to the mantle convection system of which the plates are a part.[5]

To discuss the stresses within the oceanic lithosphere, we compare the ridge push force to the other forces applied at the boundaries of the plate. These include forces at the plate base and forces at the subduction zone. As for the downgoing slab, earthquake focal mechanisms constrain the relative size of the forces. Here, we use the observation (Fig. 5.5-4) that stress in the spreading direction is typically compressive at all ages.

Consider a simple model of stress in the oceanic lithosphere, using the geometry of Fig. 5.5-7. Using the stress equilibrium equation (Eqn 2.3.49) in the spreading (x) direction, we relate the deviatoric stresses to the body force $f(x, z)$, which is the contribution to ridge push from the material at (x, z),

$$\frac{\partial \sigma_{xx}(x, z)}{\partial x} + \frac{\partial \sigma_{xz}(x, z)}{\partial z} + f(x, z) = 0. \tag{7}$$

Integrating first with respect to x and then with respect to z from $z = 0$ to the base of the lithosphere $m(x)$ yields the force balance

$$\bar{\sigma}_{xx}(x) = \frac{\sigma_b x - F_R(x)}{m(x)} + \sigma_r. \tag{8}$$

Here the stress in the spreading direction is given by its vertical average $\bar{\sigma}_{xx}(x)$; $\sigma_r = \bar{\sigma}_{xx}(0)$ characterizes the strength of the ridge; the drag force at the base of the plate is given by the basal shear stress σ_b; and $F_R(x)$ is the net ridge push force

$$F_R(x) = \int_0^{m(t)} \int_0^{x} f(x, z) dx dz. \tag{9}$$

Written in terms of plate age, t,

$$\bar{\sigma}_{xx}(t) = \frac{\sigma_b vt - F_R(t)}{m(t)} + \sigma_r, \tag{10}$$

where v is a half spreading rate, assumed constant. A useful form for comparing different plates comes from the usual assumption that the basal drag force equals the product of absolute velocity u and drag coefficient C ($\sigma_b = Cu$),

$$\bar{\sigma}_{xx}(t) = \frac{Cuvt - F_R(t)}{m(t)} + \sigma_r \tag{11}$$

Thus a drag depending on absolute velocity is applied over an area proportional to the spreading rate. For simplicity, we assume that $v = u$, spreading rate equals absolute velocity (the ridge is fixed with respect to the mantle), so the net drag force is proportional to velocity squared.

A subduction zone would provide a boundary condition on the oldest lithosphere. For example, if focal mechanisms in the lithosphere near trenches were extensional, an extensional condition could be imposed. Because such mechanisms are not seen, it is often assumed that the negative buoyancy of slabs (slab pull) is balanced by local resistive forces (Section 5.4.2). Thus, although the ridge push force is probably smaller than the slab pull forces, the thrust fault mechanisms suggest that it is more crucial for determining stress in oceanic lithosphere.

Although this stress model is schematic and does not describe any individual plate, it lets us use focal mechanism observations to estimate several important quantities. Figure 5.5-8 shows the predicted intraplate stress as a function of plate age and drag coefficient. For zero drag the stress is purely compressive ($\bar{\sigma}_{xx} < 0$) and varies as $\sqrt{t}$, because the force increases linearly with age, whereas the plate thickens as its square root. For larger drag coefficients, $\bar{\sigma}_{xx}$ follows $\sqrt{t}$ curves corresponding to less and less compression, until the lithosphere is in extension for all ages. All lithospheric plates

[5] Verhoogen (1980) offers the analogy that rain occurs because of the negative buoyancy of the drops relative to the surrounding air, as part of the process by which solar heat evaporates water which rises as vapor due to positive buoyancy and is transported by wind to the point where it cools, condenses into drops, and then falls.

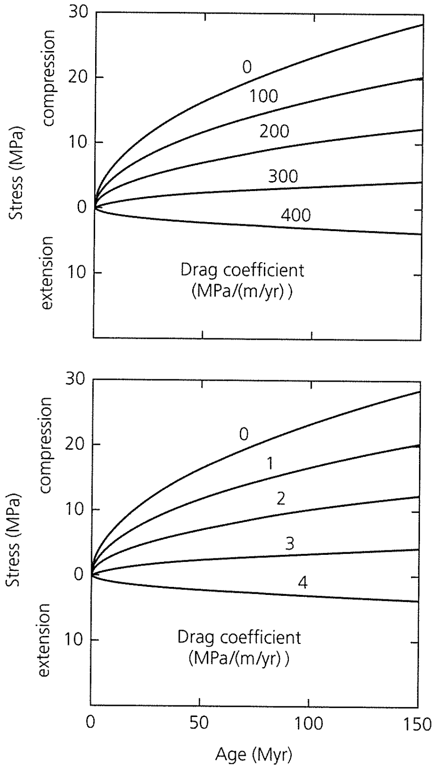


Fig. 5.5-8 Intraplate stress in the spreading direction as a function of lithospheric age and assumed basal drag coefficient for slow-moving (1 cm/yr, *top*) and fast-moving (10 cm/yr, *bottom*) plates. The compressional stresses in old oceanic lithosphere place an upper bound on the drag coefficient of 4 MPa/(m/yr). (Wiens and Stein, 1985. *Tectonophysics, 116*, 143–62, with permission from Elsevier Science.)

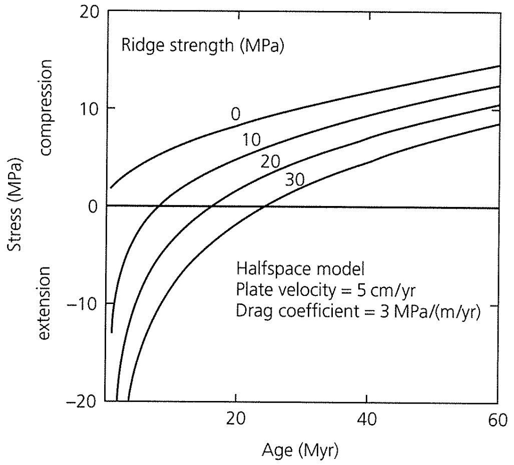


Fig. 5.5-9 Intraplate stress in the spreading direction as a function of lithospheric age computed for several values of ridge strength. The age of the transition from ridge-normal extension to compression increases with the strength of the ridge. (Wiens and Stein, 1984. *J. Geophys. Res., 89*, 11, 442–64, copyright by the American Geophysical Union.)

appear to be in compression, so a rapidly moving plate (such as the Pacific, which moves at about 10 cm/yr) constrains the drag coefficient to less than about 4 MPa/(m/yr). Similar results emerge for a cooling plate model.

This model assumes a zero stress boundary condition at the ridge axis, so the axis has no tensile strength. The predicted stress in young lithosphere, especially the location of a possible transition from compression to extension in the direction of spreading, would be sensitive to the strength of the ridge (Fig. 5.5-9). Models with substantial strength at the axis predict a wide band of extension in the spreading direction. Since such a zone of normal-faulting earthquakes is not observed, the axis seems weak.

Although this simple model describes only a hypothetical average plate, more sophisticated models use realistic plate geometries to calculate the stresses expected from ridge push, slab pull, and basal drag forces. These models' predictions can be compared to earthquake focal mechanisms and other data for specific areas. For example, Fig. 5.5-10 shows stresses predicted for the Indian Ocean region. Although the model was calculated assuming a single Indo-Australian plate, it predicts stresses in the region now considered a diffuse boundary zone (Fig. 5.5-5) that are generally consistent with the focal mechanisms and the folding seen in gravity and seismic reflection data.

5.5.3 *Constraints on mantle viscosity*

The last section's analysis relating earthquake mechanisms to drag at the base of the lithosphere also gives insight into the viscosity of the mantle. The viscosity,[6] the proportionality constant between shear stress and the strain rate (or velocity gradient), controls how the mantle flows in response to applied stress, and is thus crucial for mantle convection. If the drag on the base of a plate is due to motion over the viscous mantle, compressive earthquake mechanisms in old lithosphere constrain the viscosity.

Consider a simple two-dimensional geometry where mass flux due to the moving plate is balanced by a return flow at depth (Fig. 5.5-11, *top*). The drag coefficient is proportional to the viscosity and inversely proportional to the flow depth. Figure 5.5-12 shows that the basal drag constraint from the focal mechanism data, $C \leq 4$ MPa/(m/yr), requires an average mantle viscosity less than 2×10^{20} poise if flow occurs to a depth of 700 km in the upper mantle, or 10^{21} poise if flow occurs in the entire mantle. These values are lower than the $1\text{–}5 \times 10^{22}$ poise typically estimated from glacial rebound, earth rotation, and satellite orbits.

This discrepancy can be reconciled by assuming that the plate is underlain by a thin, low-viscosity asthenosphere (Fig. 5.5-11, *bottom*). The low-viscosity layer, in which only a fraction of the return flow occurs, decouples the plates from the underlying

[6] Viscosity, defined in Section 5.7, is given in cgs units as poise (dyn-s/cm^2) or in SI units as Pascal-seconds (1 poise = 0.1 Pa-s).

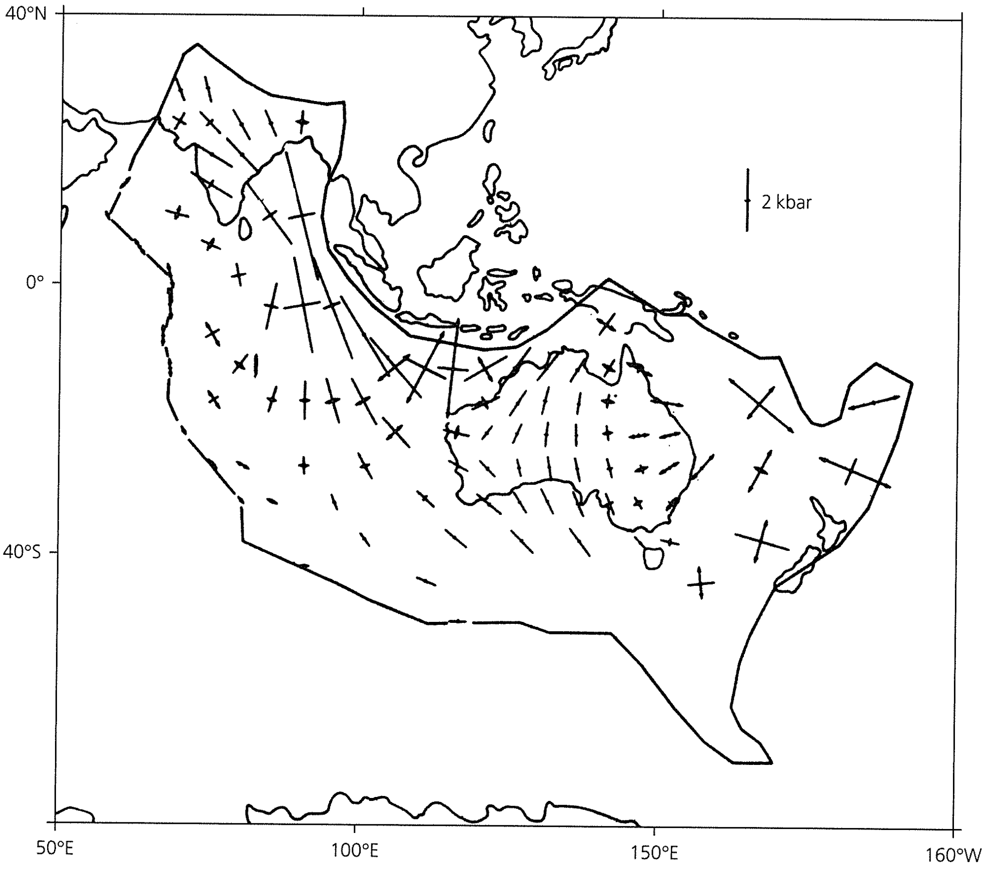


Fig. 5.5-10 Intraplate stress predicted by a force model for the Indo-Australian plate The bars show the principal horizontal deviatoric stresses, with arrowheads marking tension. The location and orientation of the highest stresses, such as the transition between compression and tension, are generally consistent with earthquake mechanisms in the region now regarded as a diffuse plate boundary (Fig. 5.5-5). (Cloetingh and Wortel, 1985. *Geophys. Res. Lett.*, *12*, 77–80, copyright by the American Geophysical Union.)

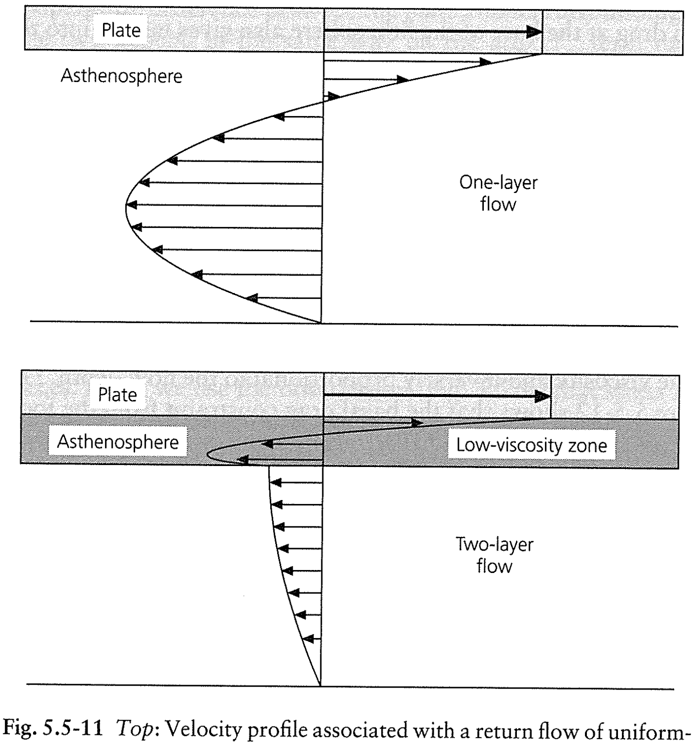


Fig. 5.5-11 *Top*: Velocity profile associated with a return flow of uniform-viscosity asthenosphere that balances the mass flux due to plate motions. *Bottom*: Velocity profile associated with a return flow of two layers of different viscosity. The upper, low-viscosity layer decouples the plates from the underlying mantle. (McKenzie and Richter, 1978.)

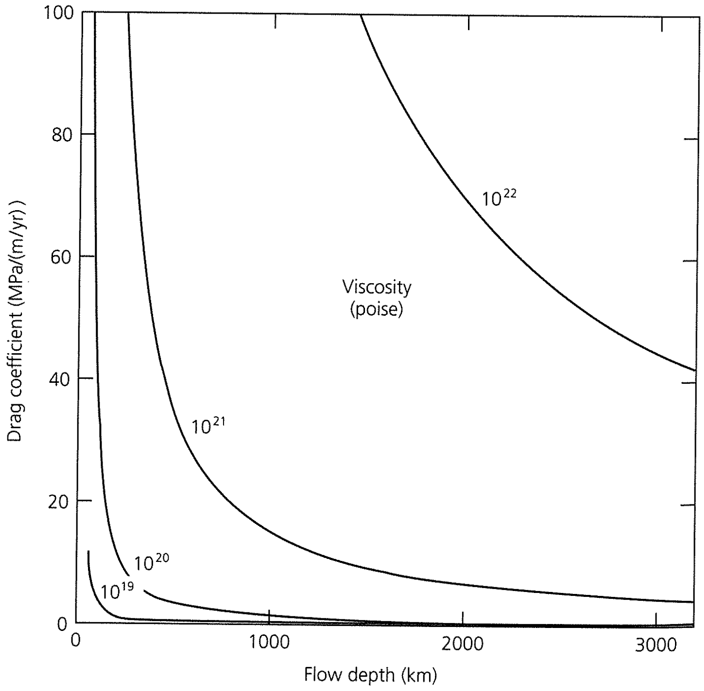


Fig. 5.5-12 Basal drag coefficients as a function of the mantle viscosity and flow depth, assuming single-layer flow. (Wiens and Stein, 1985. *Tectonophysics*, *116*, 143–62, with permission from Elsevier Science.)

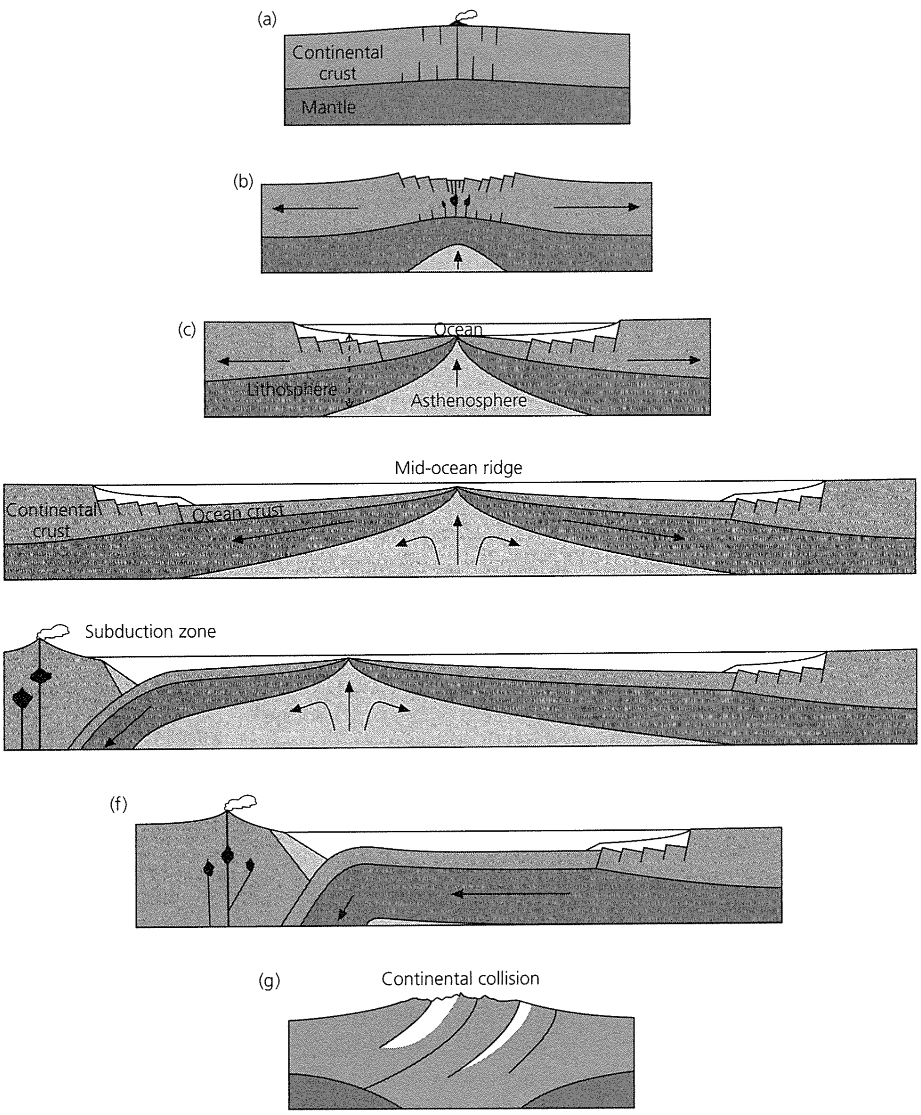


Fig. 5.6-1 Schematic illustration of the Wilson cycle, the fundamental geological process controlling the evolution of the continents. (a)–(b): A continent rifts, such that the crust stretches, faults, and subsides. (c): Sea floor spreading begins, forming a new ocean basin. (d): The ocean widens and is flanked by sedimented passive margins. (e): Subduction of oceanic lithosphere begins on one of the passive margins, closing the ocean basin (f) and starting continental mountain building. (g): The ocean basin is destroyed by a continental collision, which completes the mountain building process. At some later time, continental rifting begins again.

mantle. Viscosity values that satisfy the focal mechanisms are consistent with constraints from gravity and glacial isostasy, and such decoupling is consistent with the lack of correlation between oceanic plate area and absolute velocity (Fig. 5.4-12).

5.6 Continental earthquakes and tectonics

Although the basic relationships between plate boundaries, plate interiors, and earthquakes apply to continental as well as oceanic lithosphere, the continents are more complicated. The continental crust is much thicker, less dense, and has different mechanical properties from the oceanic crust. As a result, plate boundaries in continental lithosphere are generally broader and more complicated than in the oceanic lithosphere (Fig. 5.2-4).

Studies of continental plate boundaries, which rely heavily on seismology, provide important insights into the fundamental geological processes controlling the evolution of the continents. The basic process, known as the *Wilson cycle*,[1] is illustrated in Fig. 5.6-1. A continental region undergoes extension, such that the crust is stretched, faulted, and subsides, yielding a rift valley like the present East African rift. Because the uppermost mantle participates in the stretching, hotter mantle material upwells, causing partial melting and basaltic volcanism. Sometimes the extension stops after only a few tens of kilometers, leaving a failed or fossil rift such as the 1.2 billion-year-old mid-continent rift in the central USA. In other cases the extension continues, so the continental rift evolves into an oceanic spreading center (identifiable from sea floor magnetic anomalies), which forms a new ocean basin like

[1] Named after J. Tuzo Wilson (1908–93), whose key role in developing plate tectonic theory included introducing the ideas of transform faults, hot spots, and that the Atlantic had closed and then reopened.

the Gulf of Aden or the Red Sea. With time, the ocean widens and deepens due to thermal subsidence of oceanic lithosphere (Section 5.3.2), and thick sediments accumulate on the continental margins, such as those on either side of the Atlantic. These margins are not plate boundaries – the oceanic and continental crust on the two sides are on the same plate – and are called *passive margins*, to distinguish them from active continental margins, which are plate boundaries. Subduction often begins along one of the passive margins, and the ocean basin closes, such that magmatism and mountain building occur, as along the west coast of South America today. Continental collision like that currently in the Himalayas occurs eventually, and the mountain building process reaches its climax. If the continental materials on either side cease to move relative to each other, this process leaves a mountain belt within the interior of a single plate. At some future time, however, a new rifting phase can begin, often near the site of the earlier rifting, and a new ocean will start to grow. Thus the Appalachian Mountains record a continental collision that closed an earlier Atlantic Ocean about 270 million years ago, and remain despite the opening of the present Atlantic Ocean during the past 200 Myr.

As a result, continental and oceanic crust have very different life cycles. Because the relatively less dense continental crust is not subducted, the continents have accreted over a much longer time than the 200-million-year age of the oldest oceanic crust. Hence the continents preserve a complex set of geologic structures, many of which can be sites of deformation, including earthquake faulting. Thus both plate boundary and intraplate deformation zones within continents are more complex than their oceanic counterparts.

Earth scientists seek to understand the continental evolution process for both intellectual and practical reasons. The process is fundamental to how the planet works, but also provides information about geologic hazards (earthquakes, volcanism, uplift, and erosion) and mineral resources. In addition, the large mountain belts have major impacts on earth's climate. Seismology contributes to these studies by providing data about earthquakes and velocity structure in regions where different parts of the evolutionary cycle occur today or occurred in the past. These data are combined with other geophysical and geological data to form an integrated picture of the complicated continental evolution processes. Hence, although the processes are not fully understood, important progress continues to be made.

5.6.1 Continental plate boundary zones

As for oceanic boundaries, we seek to first describe the motion (kinematics) within boundary zones, and then to combine the kinematics with other data to investigate their mechanics (dynamics). One example is the East African rift (Fig. 5.6-2), a spreading center between the Nubian (West Africa) and Somalian (East Africa) plates. The extension rate is so slow, less than 10 mm/yr, that it is hard to resolve in plate motion models, and the two plates are often treated as one (Fig. 5.2-4). However, the rift topography, normal faulting, and seismicity distribution show the presence of an extensional boundary zone broader, more diffuse, and more complex than at a mid-ocean ridge. For example, the seismicity ends in southern Africa and has no clear connection to the southwest Indian ridge, where the plate boundary must go. A recent estimate is that the northern East Africa rift opens at about 6 mm/yr, whereas the southern part opens at about half that, because the Euler pole is to the south. Some of the complexity of such continental extensional zones results from the fact that, unlike a mid-ocean ridge, the lithosphere starts off with reasonable thickness and then is stretched and thinned in the extending zone. The rifting process can eventually progress far enough that a new oceanic spreading center forms. This has already occurred in the Gulf of Aden and the Red Sea, which are newly formed (and hence narrow) oceans separating the Arabian plate from Somalia and Nubia at rates of about 22 and 16 mm/yr, respectively. Whether the East African rift will evolve this far is still unclear, because the geologic record shows many rifts that, although active for some time, failed to develop into oceanic spreading centers and simply died. As we will see, these fossil rifts can be loci for intraplate earthquakes.

The earthquakes also indicate that the thermal and mechanical structure of continental rifts is more complicated than on mid-ocean ridges. Normal-faulting earthquakes extend to depths of 25–30 km, considerably deeper than at mid-ocean ridges. Hence the lower crust appears to be surprisingly stronger and colder than might be expected in an active rift.

Continental transforms are also more complicated than their oceanic counterparts. As we saw in Section 5.2, the transform portion of the Pacific–North America plate boundary in western North America is an active seismic zone hundreds of kilometers wide (Fig. 5.2-3), in contrast to widths of less than 10 km for oceanic transforms. Thus the focal mechanisms show primarily strike-slip motion on the San Andreas fault itself and demonstrate complexities including thrust faulting for events like the 1971 San Fernando and 1994 Northridge earthquakes and normal faulting due to the regional extension in the Basin and Range province. The earthquakes and space-geodetic data show that although most of the motion occurs along the San Andreas (Fig. 4.5-13) and nearby faults, a reasonable fraction of the motion occurs elsewhere (Figs. 5.6-3 and 5.2-3). The boundary zone is further complicated by volcanism in areas including the Long Valley caldera in eastern California and the Yellowstone hot spot, which also have associated seismicity. Hence, we think of a boundary zone in which the overall steady motion between the plate interiors is distributed in both space and time (Fig. 5.6-4). Although much of the motion occurs in occasional large earthquakes or steady creep on the main boundary segment, some deformation occurs elsewhere in the zone.

The breadth of continental plate boundary zones has important implications for seismic hazards within them. Because ground shaking decays rapidly with distance (Fig. 1.2-5), nearby smaller earthquakes within a boundary zone, but not

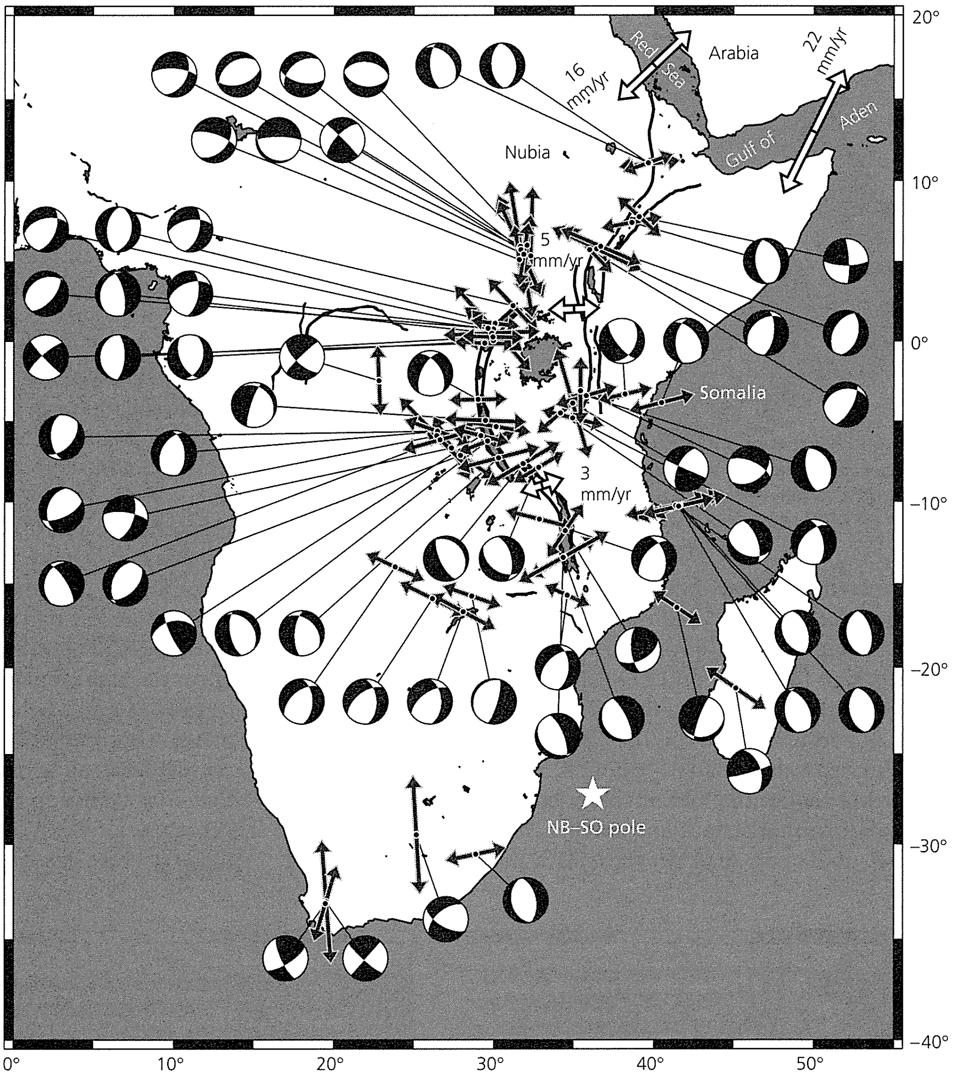


Fig. 5.6-2 Seismicity and focal mechanisms (*T* axes shown by black arrows) for the East African rift system, with relative plate motions (white arrows) from Chu and Gordon (1998, 1999).

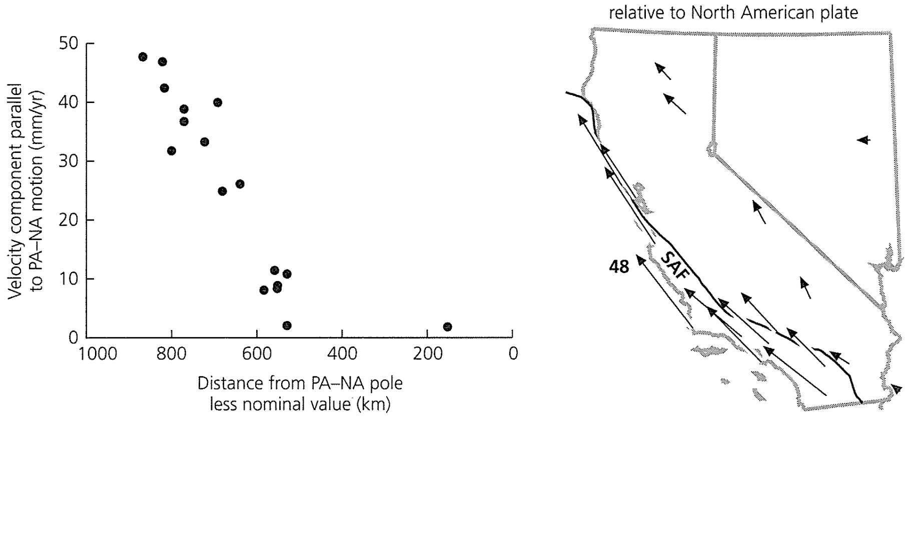


Fig. 5.6-3 Variation in motion of space-geodetic sites across part of the Pacific–North America boundary zone. *Right*: Horizontal velocities of sites in California, Nevada, and Arizona relative to stable North America. The velocity of the southwesternmost site nearly equals the predicted 48 mm/yr velocity of the Pacific plate relative to the North American plate. *Left*: Component of motion tangent to small circles centered on the Pacific–North America Euler pole versus angular distance from that pole. Velocities increase with distance from the Euler pole, with a discontinuity due to the approximately 35 mm/yr of time-averaged slip across the San Andreas fault. (Gordon and Stein, 1992. *Science*, *256*, 333–42, copyright 1992 American Association for the Advancement of Science.)

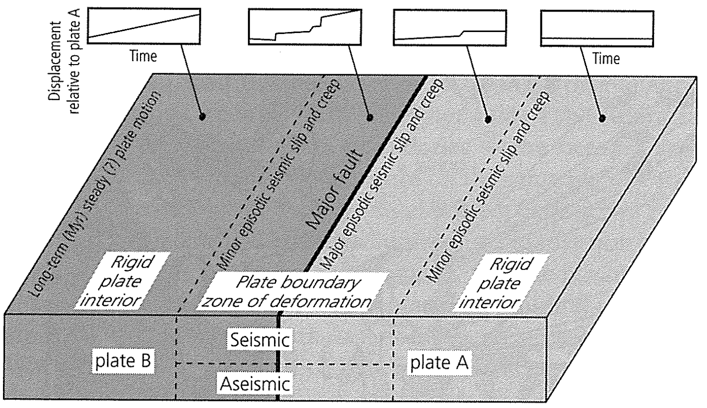

Displacement relative to plate A
Time
Time
Long-term (Myr) steady (?) plate motion
Minor episodic seismic slip and creep
Major fault
Major episodic seismic slip and creep
Minor episodic seismic slip and creep
Rigid plate interior
Plate boundary zone of deformation
Rigid plate interior
Seismic
Aseismic
plate B
plate A
Plate boundary zone slip distribution

Fig. 5.6-4 Schematic illustration of the distribution of motion in space and time for a strike-slip boundary zone between two major plates. (Stein, 1993. *Contributions of Space Geodesy to Geodynamics*, 5–20, copyright by the American Geophysical Union.)

on the main boundary fault, can be more damaging than larger but more distant ones on the main fault. Hence the Los Angeles area is vulnerable to both nearby earthquakes like the 1994 Northridge (M_w 6.7) or 1971 San Fernando (M_s 6.6) earthquakes and larger ones on the more distant San Andreas Fault, such as a recurrence of the 1857 Fort Tejon earthquake which is estimated to have had M_w about 8. Similarly, the earthquake hazard in the Seattle area involves both great earthquakes at the subduction interface and smaller, but closer, earthquakes in the subducting Juan de Fuca plate (like the 2001 M_w 6.7 Nisqually earthquake) or at shallow depth in the North American plate.

Of the three boundary types, continental convergence zones may be the most complicated compared to their oceanic counterparts. One primary difference is that because continental crust is much less dense than the upper mantle, it is not subducted, and a Wadati–Benioff zone is not formed. As a result, continental convergence zones in general do not have intermediate and deep focus earthquakes. However, the plate boundary tectonics occur over a broader and more complex region than in the oceanic case.

A spectacular example is the collision between the Indian and Eurasian plates. This area is the present type example of mountain building by continental collision, which has produced a boundary zone extending thousands of km northward from the nominal plate boundary at the Himalayan front (Fig. 5.6-5). The total plate convergence is taken up in several ways. About half of the convergence occurs across the locked Himalayan frontal faults such as the Main Central Thrust (Fig. 5.6-6), and gives rise to large destructive earthquakes. These faults are part of the interface associated with the underthrusting Indian continental crust, which thickens the crust under the high Himalayas. However, the earthquakes also show normal faulting behind the convergent zone, in the Tibetan plateau, presumably because the uplifted and thickened crust spreads under its own weight. GPS data (Fig. 5.6-5) show that this extension is part of a large-scale process of crustal "escape," or "extrusion," in which large fragments of continental crust are displaced eastward by the collision along

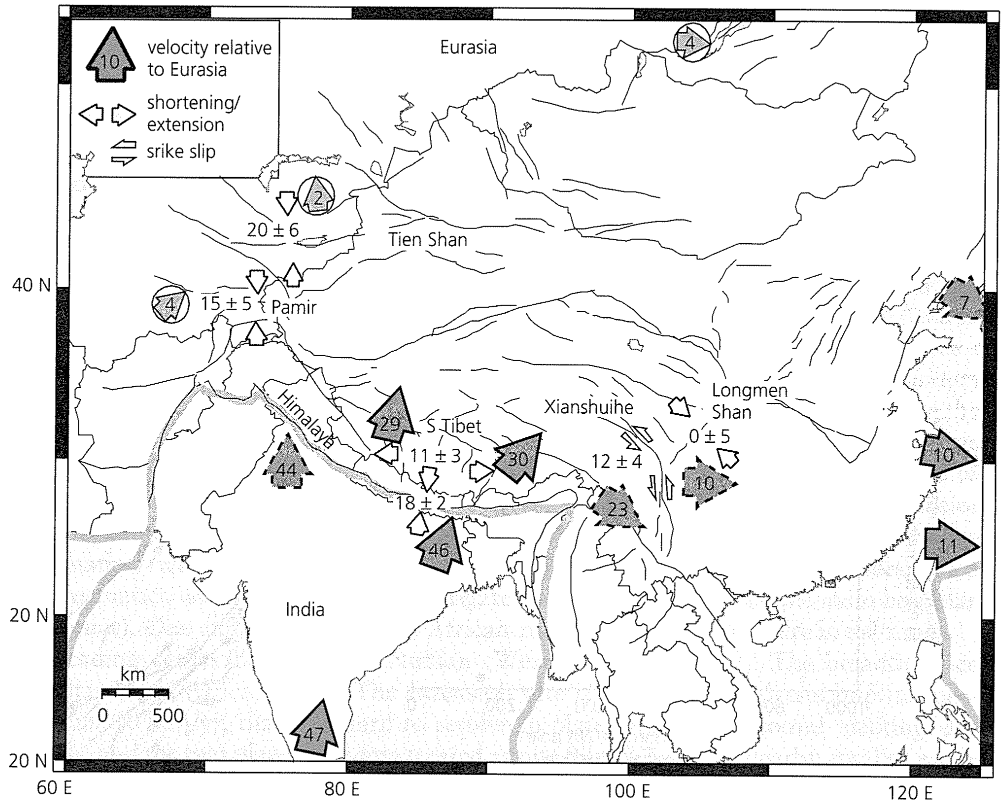


Fig. 5.6-5 Summary of crustal motions determined using space geodesy in the India–Eurasia plate collision zone. Large arrows indicate velocities relative to Eurasia. Arrows in circles show velocities with no significant motion with respect to Eurasia. Small arrows show local relative deformation. (Larson *et al.*, 1999. *J. Geophys. Res.*, *104*, 1077–94, copyright by the American Geophysical Union.)

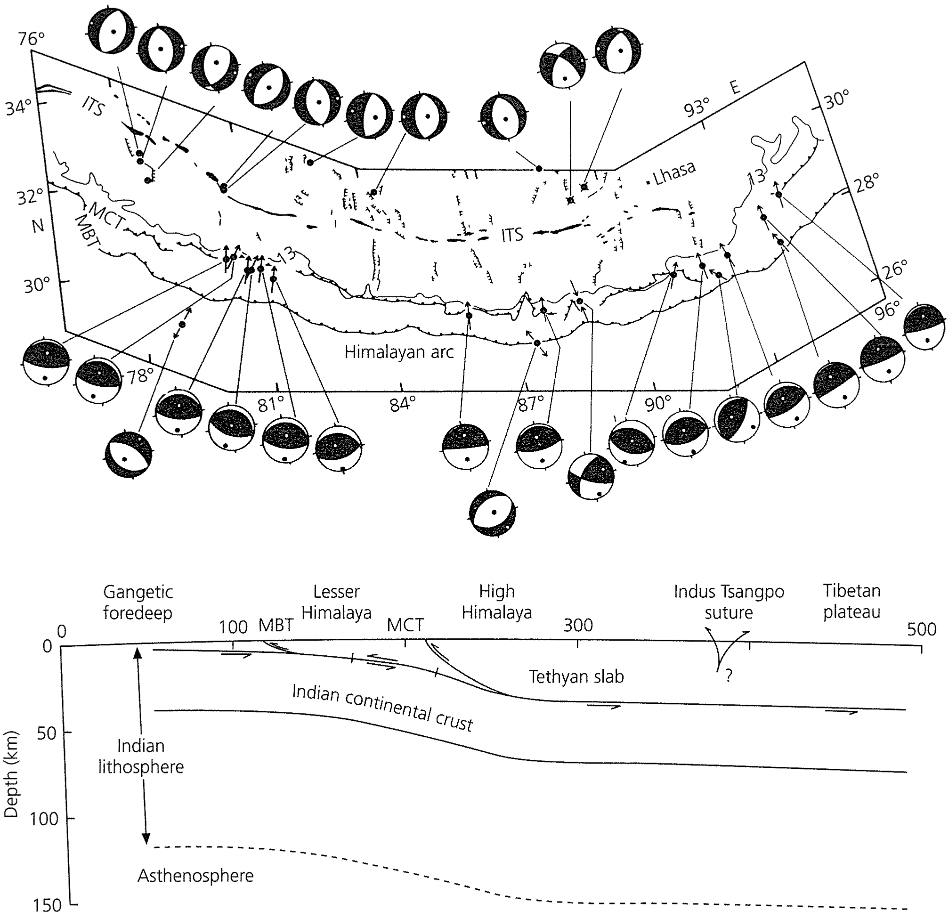


Fig. 5.6-6 Focal mechanisms and tectonic interpretation for the Himalayan continental convergence zone. MCT and MBT are the Main Central and Main Boundary thrust faults. (After Ni and Barazangi, 1984. *J. Geophys. Res., 89*, 1147–64, copyright by the American Geophysical Union.)

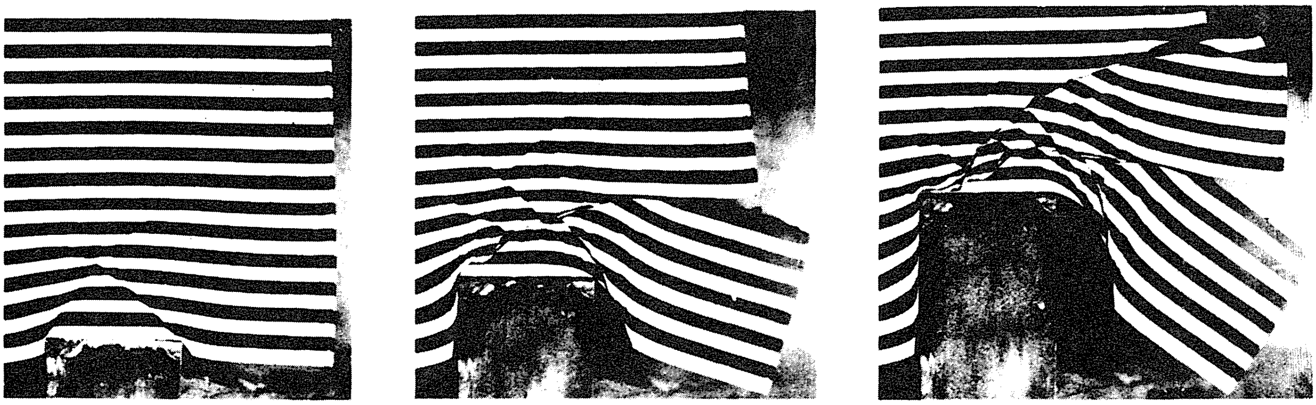
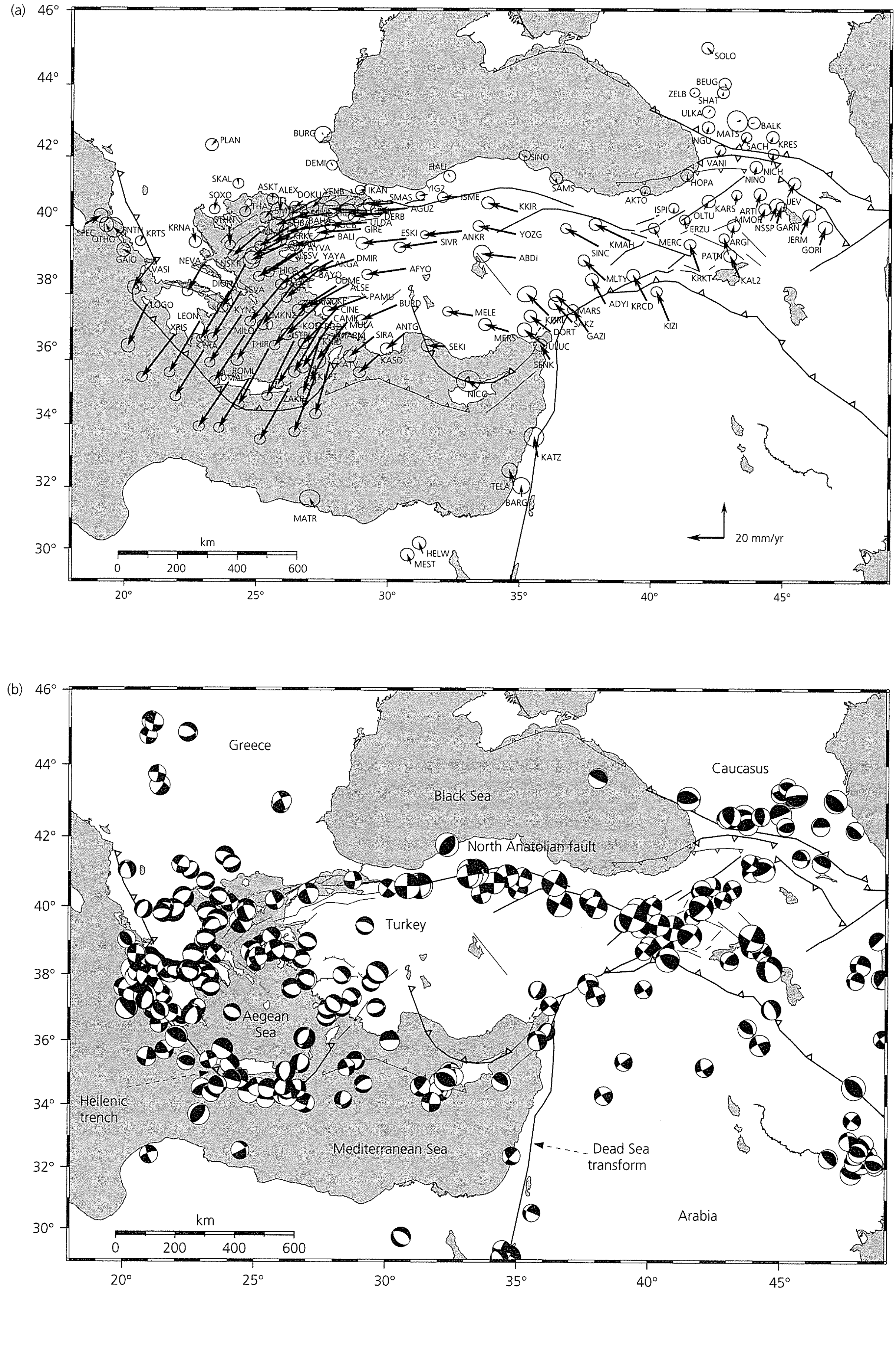


Fig. 5.6-7 Demonstration of the deformation of Asia, modeled by a striped block of plasticine, as the result of a collision with a rigid block simulating the Indian subcontinent. The plasticine is constrained on the left side, so the impact forces blocks to be extruded to the right, analogous to the eastward motion of blocks in Indochina and China. (Tapponnier *et al.*, 1982. *Geology, 10*, 611–16, with permission of the publisher, the Geological Society of America, Boulder, Co. © 1982 Geological Society of America.)

major strike-slip faults. This extrusion has been modeled assuming that India acts as a rigid block indenting a semi-infinite plastic medium (Asia), giving rise to a complicated faulting and slip pattern (Fig. 5.6-7). The extent of the collision is illustrated by GPS data and focal mechanisms showing that the Tien Shan intracontinental mountain belt, 1000–2000 km north of the Himalayas, accommodates almost half the net plate convergence in the western part of the zone.

In addition to providing data about a collision region's kinematics, seismological studies provide insight into its mech-


(a)
46°
44°
42°
40°
38°
36°
34°
32°
30°
20°
25°
30°
35°
40°
45°
SOLO
BEUG
ZELB
SHAT
ULKA
BALK
MATS
KRES
SACH
NGU
BURG
PLAN
DEMI
SINO
HALI
VANI
NICH
NINO
HOPA
SAMS
AKTO
SKAL
SOXO
ASKT
ALEX
DOKU
YENB
IKAN
YIG2
SMAS
ISME
AGUZ
DERB
ULDA
GIRE
KKIR
YOZG
ESKI
SIVR
ANKR
ABDI
BALI
AYVA
YAYA
DMIR
AFYO
KMAH
SINC
MLTY
ADYI
KRCD
KIZI
MERC
ERZU
OLTU
KARS
ARTI
MMOR
NSSP
GARN
JERM
GORI
IJEV
ARGI
PATN
KRKT
KAL2
ISPI
SPEC
OTHO
PNTN
KRTS
KRNA
GAIO
NEVA
NSKR
VASI
LOGO
LEON
XRIS
KYRA
MILO
THIR
KYNS
MKNZ
CINE
CAMK
MULA
SIRA
ANTG
KASO
KATV
ROML
OMAL
ZAKR
KRPT
PAMU
BURD
ALSE
ODME
MELE
MERS
SEKI
NICO
KDRL
DORT
SAKZ
MARS
GAZI
SENK
KATZ
TELA
BARG
MATR
HELW
MEST
20 mm/yr
km
0
200
400
600
(b)
Greece
Black Sea
Caucasus
North Anatolian fault
Turkey
Aegean Sea
Hellenic trench
Mediterranean Sea
Dead Sea transform
Arabia
km
0
200
400
600
46°
44°
42°
40°
38°
36°
34°
32°
30°
20°
25°
30°
35°
40°
45°

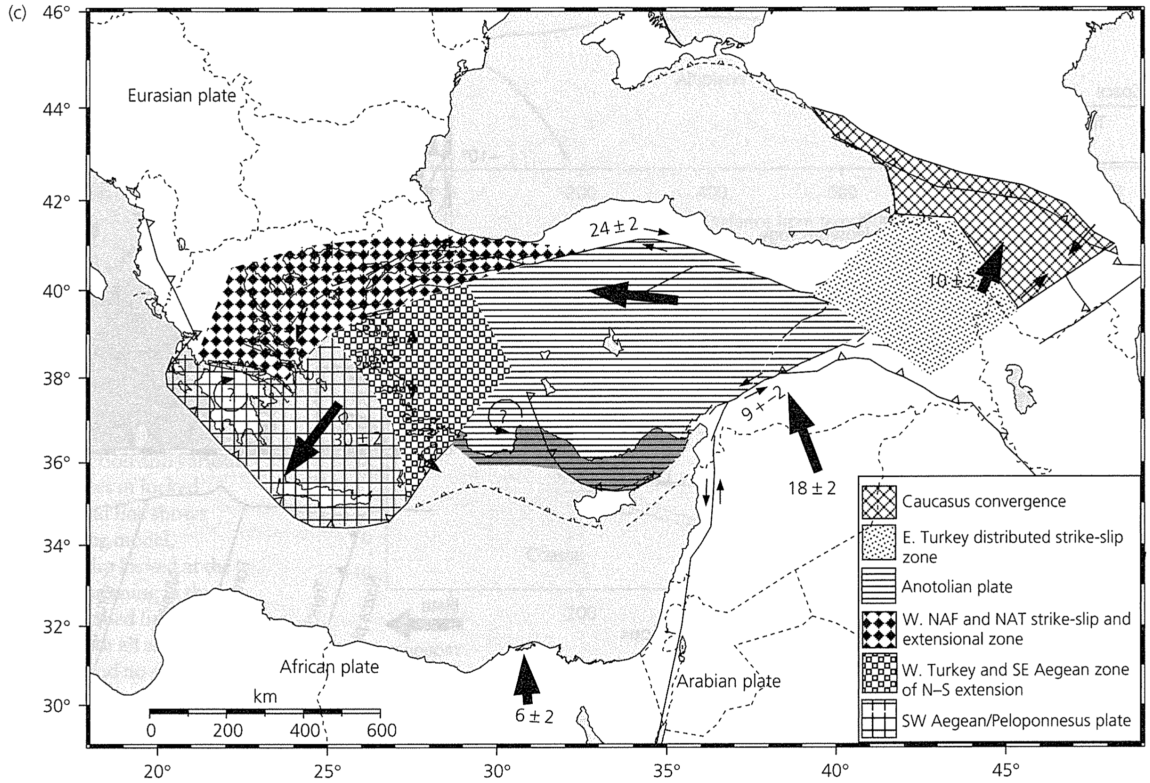


Fig. 5.6-8 GPS observations of motions relative to Eurasia (a), focal mechanisms (b), and tectonic interpretation (c) for a portion of the Africa–Arabia–Eurasia plate collision zone. Note strike-slip along the North Anatolian fault, extension in western Anatolia and the Aegean region, and compression in the Caucasus mountains. Rates are in mm/yr. (McClusky *et al.*, 2000. *J. Geophys. Res.*, *105*, 5695–5719, copyright by the American Geophysical Union.)

anics. The collision process is thought to involve a complex interplay between forces due directly to the collision, gravitational forces due to the resulting uplift and crustal thickening, and forces from the resulting mantle flow. Earthquake depths and studies of seismic velocity, attenuation, and anisotropy are providing data on crustal thicknesses, thermal and mechanical structures, and mantle flow. For example, *P*-wave travel time tomography shows high velocity under the presumably cold Himalayas, which contrasts with low velocity under Tibet. These and other seismological data are consistent with the idea that Tibet deforms easily during the collision.

An equally complicated situation occurs in the eastern Mediterranean collision zone involving the African, Arabian, and Eurasian plates. Combining GPS and focal mechanism data shows the complex motions. Figure 5.6-8 (a) shows the motions of sites in the western Mediterranean relative to Eurasia. Northern portions of Arabia move approximately N40°W, consistent with global plate motion models. Western Turkey rotates as the Anatolian plate about a pole near the Sinai peninsula. Anatolia is thus "squeezed" westward between Eurasia and northward-moving Arabia (Fig. 5.6-8, c).[2] The motion across the North Anatolian fault, about 25 mm/yr, gives rise to large right-lateral strike-slip earthquakes (Fig. 5.6-8, b) such as the 1999 M_s 7.4 Izmit earthquake, which occurred about 100 km east of Istanbul and caused more than 30,000 deaths. To the west, the data show interesting deviations from a rigid Anatolian plate. The increasing velocities toward the Hellenic trench, where the Africa plate subducts below Crete and Greece, show that western Anatolia and the Aegean region are under extension, consistent with the normal fault mechanisms. This region may be being "pulled" toward the arc, perhaps by an extensional process similar to oceanic back-arc spreading, as the trench "rolls back" (Section 5.2.4). By contrast, eastern Turkey is being driven north-ward into Eurasia, causing compression that appears as the thrust fault earthquakes in the Caucasus mountains. The Dead Sea transform separates Arabia from the region to the west, sometimes viewed as the Sinai microplate. Strike-slip motion along this fault gives rise to the earthquakes mentioned in the Bible that repeatedly destroyed famous cities like Jericho.

5.6.2 *Seismic, aseismic, transient, and permanent deformation*

The examples in the previous section illustrate that earthquakes give powerful insights into the crustal deformation shaping the

[2] Consider a melon seed squeezed between a thumb and a forefinger.

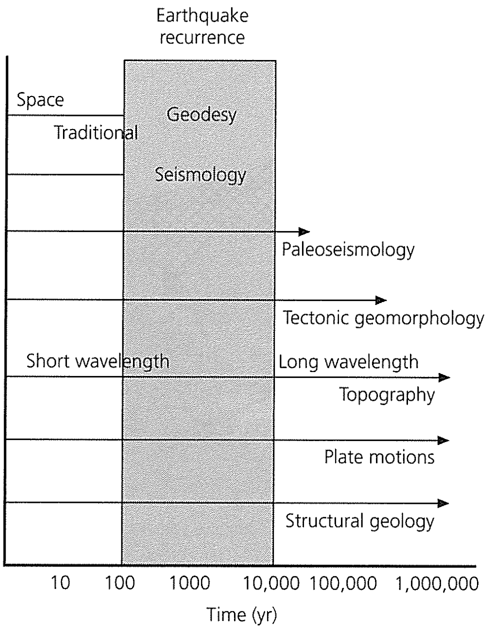


Fig. 5.6-9 Schematic illustration of how crustal deformation on various time scales is observed by different techniques.

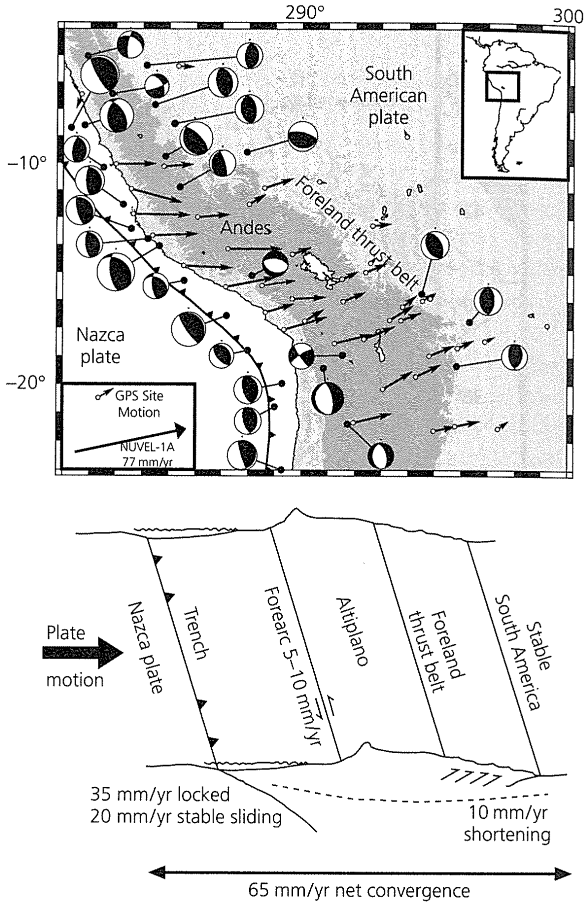


Fig. 5.6-10 *Top*: GPS site velocities relative to stable South America (Norabuena *et al.*, 1998. *Science, 279*, 358–62, copyright 1998 American Association for the Advancement of Science), and selected earthquake mechanisms in the boundary zone. Rate scale is given by the NUVEL-1A vector. *Bottom*: Cross-section showing approximate velocity distribution inferred from GPS data. (Stein and Klosko, 2002. From *The Encyclopedia of Physical Science and Technology*, ed. R. A. Meyers, copyright 2002 by Academic Press, reproduced by permission of the publisher.)

continents. Other approaches to studying this deformation, including various geodetic and geological means, sample the deformation in different ways on various time scales (Fig. 5.6-9). Hence, considerable attention goes into understanding how what we see with these different techniques are related. For example, as discussed earlier (Sections 4.5.4, 5.4.3), in many places only part of the plate motion seems to occur as earthquakes, and the rest takes place as aseismic slip. A related question is how the deformation shown by earthquakes, which has a time scale of a few years, is related to the longer-term deformation that is recorded by topography and the geologic record.

To explore these ideas, consider the distribution of motion within the boundary zone extending from the stable interior of the oceanic Nazca plate, across the Peru–Chile trench to the coastal forearc, across the high Altiplano and foreland thrust belt, and into the stable interior of the South American continent. Figure 5.6-10 shows GPS site velocities relative to stable South America, which would be zero if the South American plate were rigid and all motion occurred at the trench plate boundary. However, the site velocities are highest near the coast and decrease relatively smoothly from the interior of the Nazca plate to the interior of South America.

Figure 5.6-10 (*bottom*) shows an interpretation of these data. In this model, about half of the plate convergence (approximately 35 mm/yr) is locked at the subduction interface, causing elastic strain of the overriding plate that will be released in large interplate thrust earthquakes (Section 4.5.4) like those whose focal mechanisms are shown. Thus the locked fraction of the plate motion corresponds to the seismic slip rate, perhaps via a process in which only a fraction of the interface is locked at any time. Approximately 20 mm/yr of the plate motion occurs by stable sliding at the trench, which does not deform the overriding plate. This portion of the plate motion corresponds to aseismic slip. The rest occurs across the sub-Andean foreland fold-and-thrust belt, causing permanent shortening and mountain building, as shown by the inland thrust fault mechanisms. This portion of the plate motion would be considered aseismic slip if we considered only the fraction of the plate motion that appears in the trench seismic moment release, whereas in reality it occurs as inland deformation. These interpretations come from analyzing the GPS data in the convergence direction relative to the stable interior of South America (Fig. 5.6-11). If all the convergence were locked on the interplate thrust fault, the predicted rates would exceed those observed within about 200 km of the trench. However, if only about half of the predicted convergence goes into locking the fault, the predicted rates near the trench are less, because only the portion of the slip locked at the interface deforms the overriding plate. Similarly, the data farther than about 300 km

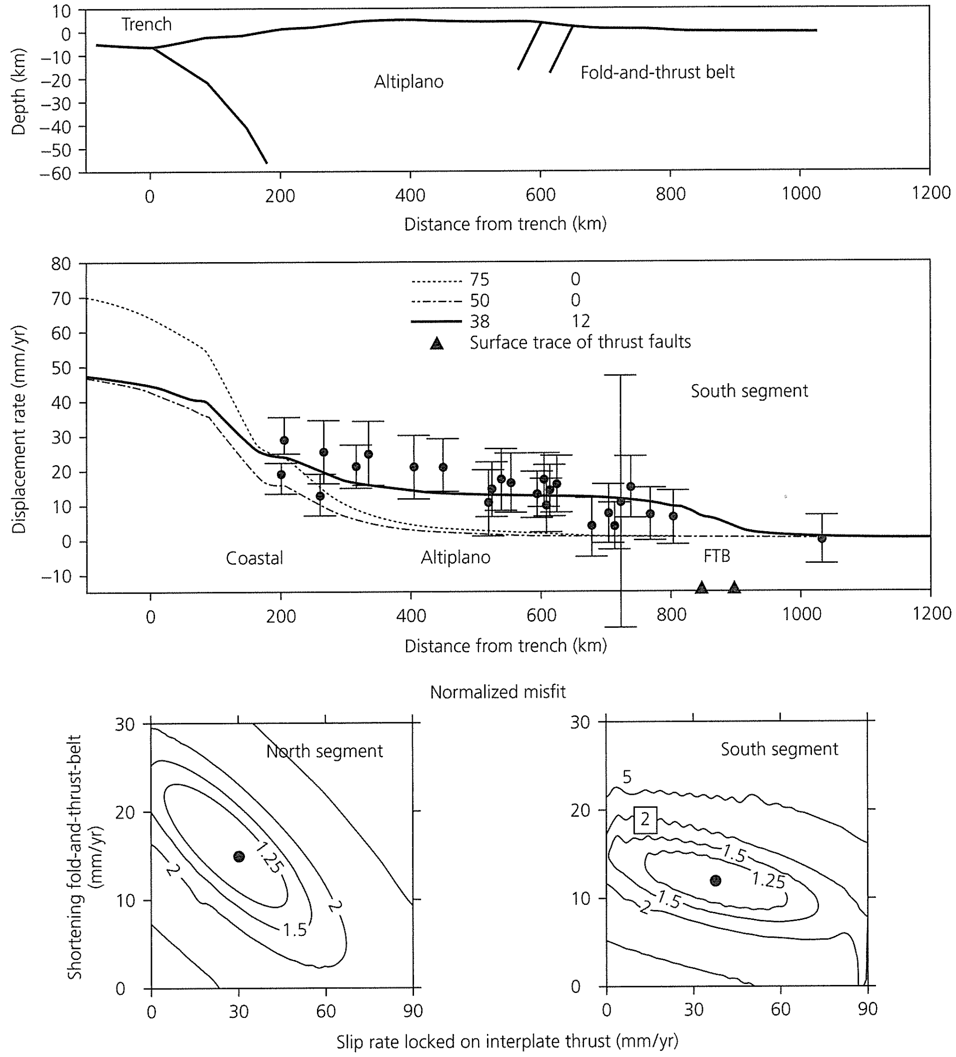


Fig. 5.6-11 Derivation of the model in Fig. 5.6-10 (*bottom*). *Top*: Model geometry, assuming partial slip locked at the plate boundary and shortening in the eastern Andes. *Center*: GPS site velocities in the convergence direction and various models, given by the rates of locked slip and shortening. Solid line shows predictions of best-fitting model, including both partial slip locked at the plate boundary and shortening in the eastern Andes. Short dashed line shows predictions of model with all slip locked on the plate boundary and no shortening. Long-short dashed line shows predictions of model with no shortening and partial slip locked on the plate boundary equal to the sum of best-fitting slip and shortening. *Bottom*: Contour plot showing misfit to the data as a function of the slip rate locked on the plate boundary and shortening rate in the eastern Andes. The best fits (dots) occur for about 30–40 mm/yr of locking and about 10–20 mm/yr shortening. (Norabuena *et al.*, 1998. *Science, 279*, 358–62, copyright 1998 American Association for the Advancement of Science.)

from the trench are better fit by assuming that about 10 mm/yr motion is locked on thrust faults in the eastern Andes. The locking and shortening rates are the best-fit parameters for this simple model, which does not include other possible complexities such as deformation in the Altiplano.

The idea that about 40% of the plate motion at the trench occurs by aseismic slip seems plausible, because studies using the history of large earthquakes at trenches often estimate that only about half the slip occurs seismically (Fig. 5.4-30). Given the problems of estimating source parameters of earthquakes from historical data, it is encouraging that the geodetic answer seems similar.

The relation between the shortening rate in the thrust belt inferred from GPS data and that implied by the earthquakes can also be studied. Assessing the seismic slip rate is a little more complicated than for transform faults (Section 5.3.3) or subduction zone thrust faulting (Section 5.4.3), because in continental deformation zones earthquakes occur over a distributed volume, rather than on a single fault, and have diverse focal mechanisms. Thus we sum the earthquakes' moment tensors (Section 4.4) to estimate a seismic strain rate tensor[3] using

$$\dot{e}_{ij} = \sum M_{ij}/(2\mu V t), \qquad (1)$$

where t is the time interval, and μ is the rigidity. V, the assumed seismic source volume, the product of the length and width of the zone of seismicity and the depth to which seismicity extends. For example, the thrust belt can be assumed to be approximately 2000 km long, 250 km wide, and faulting extends to about 40 km depth. We can then diagonalize the result and consider the eigenvalue associated with the P axis. Scaling this value by the assumed zone width gives an estimate of the shortening rate. The resulting value, less than 2 mm/yr, is significantly less than the approximately 10 mm/yr indicated by the

[3] Strain rates are often written using a dot to indicate the time derivative.

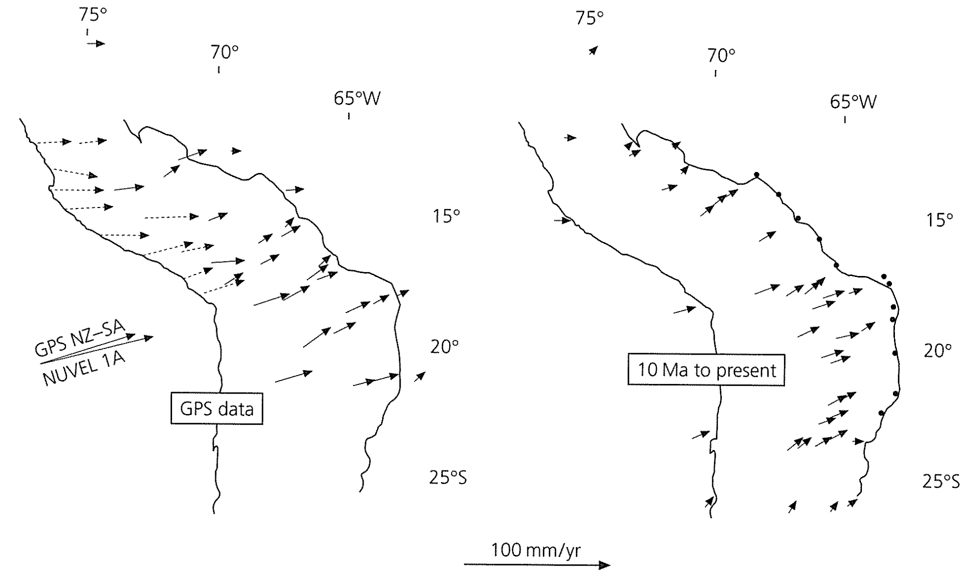


Fig. 5.6-12 Comparison of shortening across the Andes with respect to stable South America from GPS data (*left*) and geological studies (*right*). The dashed GPS vectors reflect elastic strain due to the earthquake cycle at the trench, and are not directly comparable to the permanent shortening in the geological data. Motion decreases toward the eastern extent of the mountain range, shown by the solid line. The geological vectors are largest at about 18°S and decrease to the north and south, showing how the variation in shortening that built the Andes bent them and made them widest about this point. (Hindle *et al.*, 2002.)

GPS data. Thus, even given the usual problem that the seismic history is short and may have missed the largest earthquakes, an effect one can attempt to correct for using earthquake frequency–magnitude data (Section 4.7.1), it looks like much of the shortening occurs aseismically.

An interesting question is how what we see today with earthquakes and GPS data relates to what occurs over geologic time. Figure 5.6-12 shows the results of geological studies, in which the arrows indicate the deformation that occurred over the past 10 Myr as the Andes formed. The directions and rates are similar to what are seen today, suggesting that the mountain building process has occurred relatively uniformly, although there have been some rate changes.[4]

Putting all this together gives some ideas about how the different measures of crustal deformation are related in this area. The first issue involves the relative amounts of seismic and aseismic deformation. It appears that about half of the plate motion at the trench occurs seismically. Similar fractions are also seen in other subduction zones (Fig. 5.4-30), implying that stable sliding at trenches is relatively common. Moreover, only about 10–20% of the shortening in the foreland thrust belt appears to occur seismically. Thus aseismic, and presumably permanent, deformation of rocks in the thrust belt seems like a major phenomenon. Similar results have also been observed for other continental deformation zones (Fig. 5.6-13). The next issue is that of permanent versus transient deformation. In the model of Fig. 5.6-11, the deformation of the South American plate due to the locked slip at the trench is transient, and will be released in the upcoming large trench earthquake. However, it seems likely that the deformation of the foreland thrust belt is permanent, and goes into faulting and folding rocks. Over time, this permanent displacement adds up (Fig. 5.6-12) to build the mountains.

Similar studies are going on around the world, and should lead to an improved understanding of the partitioning between seismic, aseismic, transient, and permanent deformation. Models are being developed to explore these issues (Section 5.7), which are important both for understanding continental evolution and for earthquake hazard assessment, because an apparent seismic moment deficit could indicate either overdue earthquakes or aseismic deformation.

5.6.3 Continental intraplate earthquakes

Another important application of earthquake studies deals with the internal deformation of the continental portions of the major plates. Although idealized plates would be purely rigid, intraplate earthquakes reflect the important and poorly understood tectonic processes of intraplate deformation. As in the oceans (Section 5.5.1), there appears to be a hierarchy of places that have such earthquakes. There are areas like the East African rift that can be thought of as either slow-moving plate boundaries or intraplate deformation, less active zones associated with either fossil structures or other processes like hot spots, and then intraplate earthquakes that are not easily correlated with any particular structure or cause.

One example is the New Madrid area in the central USA, which had large earthquakes in 1811–12 and has small earthquakes today. Other continental interiors, including Australia, western Europe, and India, have also had significant intraplate earthquakes. Because motion in these zones is at most a few mm/yr, compared to the generally much more rapid plate boundary motions, seismicity is much lower (Fig. 5.6-14) and thus harder to study. This difficulty is compounded by the fact that, unlike at plate boundaries, where plate motions give insight into why and how often earthquakes occur, we have little

[4] The similarity of the focal mechanism, GPS, and geological data illustrates the principle of *uniformitarianism*, that studying present processes gives insight into the past, a tenet of geology since Lyell and Hutton's seminal work almost two centuries ago.

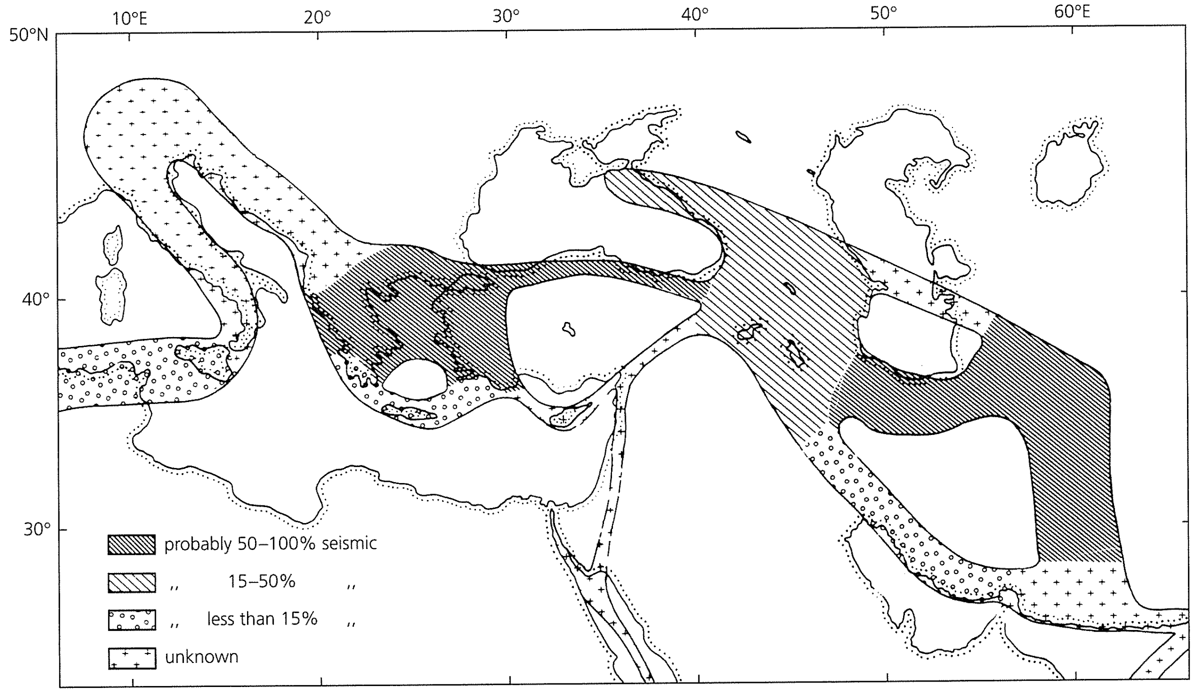


Fig. 5.6-13 Estimates of seismic deformation fractions for areas in the Mediterranean and Middle East. Seismicity appears to account for most or all of the deformation in western Turkey, Iran, and the Aegean, much of the deformation in the Caucasus and eastern Turkey, and little of the deformation in the Zagros and the Hellenic trench. (Jackson and McKenzie, 1988.)

idea of what causes intraplate earthquakes, and no direct way to estimate how often they should occur. As a result, progress in understanding these earthquakes is much slower than for earthquakes on plate boundaries, and key issues may not be resolved for a very long time.

Geodetic data illustrate the challenge. For example, comparison of the absolute velocities of GPS sites in North America east of the Rocky Mountains to velocities predicted by modeling these sites as being on a single rigid plate shows that the interior of the North American plate is rigid at least to the level of the average velocity residual, less than 1 mm/yr (Fig. 5.6-15). Similar results emerge from studies across the New Madrid zone itself and for the interiors of other major plates, showing that plates thought to have been rigid on geological time scales are quite rigid on decadal scales. For example, 1 mm/yr motion spread over 100 or 1000 km distance corresponds to strain rates of 10^{-8} and 10^{-9} yr^{-1} (3×10^{-16} and 3×10^{-17} s^{-1}), respectively. Because the geodetic data include measurement errors due to effects including instabilities of the geodetic markers, it seems likely that the tectonic strains are even smaller. However, over long enough time, even such small motions can accumulate enough slip for large earthquakes to occur.

This idea is consistent with what is known about large intraplate earthquakes. Although there is little seismological data for such events because they are rare, insight can be obtained from combining the seismological data with geodetic, paleoseismological, and other geological and geophysical data. For example, intensities estimated from historical accounts of the 1811–12 New Madrid earthquakes (Fig. 1.2-4) suggest magnitudes in the low 7 range. Paleoseismic studies (Section 1.2) indicate that several previous large earthquakes, presumably comparable to those of 1811–12, occurred 500–800 years apart. Thus, in 500–1000 years (Fig. 5.6-16, *top*) steady strain accumulation less than 2 mm/yr could provide up to 1–2 m of motion available for future earthquakes, suggesting that they would be about magnitude 7. A similar view comes from considering the earthquake history for the area. As discussed in Section 4.7.1, earthquakes of a given magnitude are approximately ten times less frequent than those one magnitude unit smaller. Thus, although the instrumental data contain no earthquakes with magnitude greater than 5, both these and a historical catalog in which magnitudes were estimated from intensity data can be extrapolated to imply that a magnitude 7 earthquake would occur about once every 1400 ± 600 years (Fig. 5.6-16, *bottom*). Hence, as expected, major intracontinental earthquakes occur substantially less frequently than comparable plate boundary events (Fig. 5.6-17). However, because of the lower attenuation in continental interiors (Section 3.7.10), such earthquakes can cause greater shaking than ones of the same magnitude on a plate boundary (Fig. 1.2-5).

Such earthquakes are generally thought to be due to the reactivation of preexisting faults or weak zones in response to either local or intraplate stresses. The New Madrid earthquakes, for example, are thought to occur on faults associated

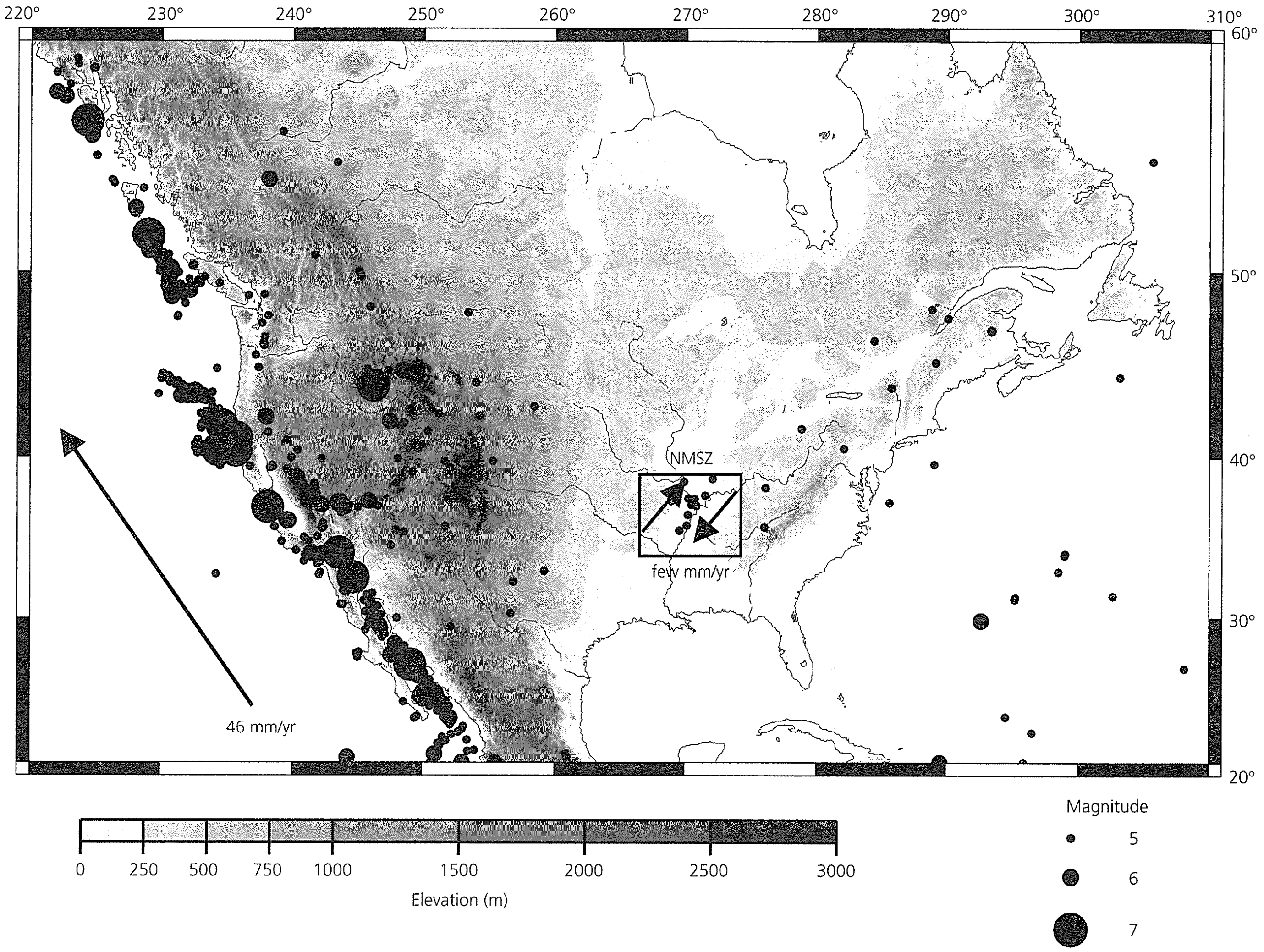


Fig. 5.6-14 Seismicity (magnitude 5 or greater since 1965) of the continental portion of the North American plate and adjacent area. Seismicity and deformation are concentrated along the Pacific–North America plate boundary zone, reflecting the relative plate motion. The remaining eastern portion of the continent, approximately that east of 260°, is much less seismically active. Within this relatively stable portion of the continent, seismicity, and thus presumably deformation, are concentrated in several zones, most notably the New Madrid seismic zone. (Weber *et al.*, 1998. *Tectonics, 17*, 250–66, copyright by the American Geophysical Union.)

with a Paleozoic failed continental rift, now buried beneath thick sediments deposited by the Mississippi river and its ancestors (Fig. 5.6-18). As a result, the faults are not exposed at the surface, so most ideas about them are based on inferences from seismology and other data. The intraplate stress field has been studied by combining focal mechanism and fault orientations with data from drill holes and *in situ* stress measurements (Fig. 5.6-19). In general, the eastern USA shows a maximum horizontal stress oriented NE–SW, consistent with the predictions of the stresses due to plate driving forces. Similar stress maps are being developed for other areas and are being used to investigate both intraplate deformation and plate driving forces. As noted in Section 3.6.5, it appears that seismic anisotropy in the lower continental crust may reflect the stress field that acted during a major tectonic event such as mountain building.

An intriguing question is why intraplate stresses cause earthquakes on particular faults, given that many weak zones could serve this purpose. Geological and paleoseismic data, together with the absence of significant fault-related topography, suggest that individual intraplate seismic zones may be active for only a few thousands of years, so intraplate seismicity migrates. This possibility is akin to that suggested for intermittent oceanic intraplate earthquake swarms. If so, there is nothing special about New Madrid or the other concentrations of intraplate seismicity we observe now — these zones will die off and be replaced by others. Moreover, there are enough tectonic structures available that (typically small) earthquakes will occur almost randomly throughout continental interiors.

A special case of this phenomenon occurs at passive continental margins, where continental and oceanic lithospheres join. Although these areas are in general tectonically inactive,

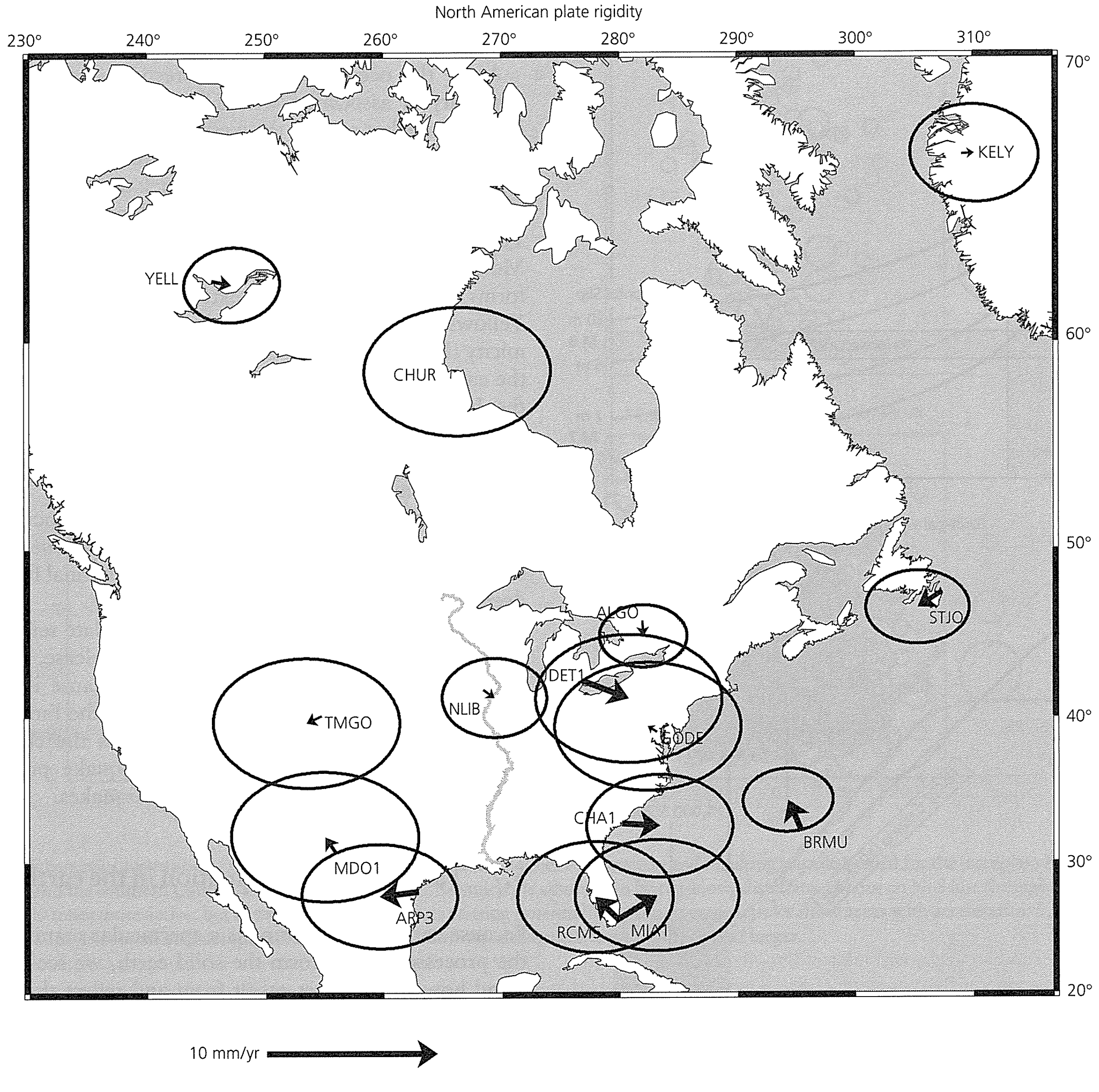


Fig. 5.6-15 Locations of continuously recording GPS sites used to estimate a Euler vector for the presumably stable portion of North America. For each, the misfit between the observed velocity and that predicted for a single plate is shown. The average misfit is less than 1 mm/yr, showing that eastern North America is quite rigid. (Newman *et al.*, 1999. *Science, 284*, 619–21, copyright 1999 American Association for the Advancement of Science.)

magnitude 7 earthquakes can occur, as on the eastern coast of North America (Fig. 5.6-20). Such earthquakes may be associated with stresses, including those due to the removal of glacial loads, which reactivate faults remaining from the original continental rifting (Fig. 5.6-1). Although such earthquakes are observed primarily on previously glaciated margins, they also occur on nonglaciated passive margins, perhaps due to sediment loading. In some cases large sediment slides occur, as was noted for the 1929 M_s 7.2 earthquake on the Grand Banks of Newfoundland, because the slides broke trans-Atlantic telephone cables and generated a tsunami that caused 27 fatalities.[5] An interesting unresolved question is whether tectonic faulting is required for such earthquakes, or whether the slump itself can account for what is seen on seismograms. Some studies find that the seismograms are best fit by a double-couple fault source, whereas others favor a single force consistent with the slump (Fig. 4.4-3). The issue is important because slumps occur in the sedimentary record along many passive

[5] These deaths account for all but one of Canada's known earthquake fatalities to date, although this situation could change after a large Cascadia subduction zone earthquake.

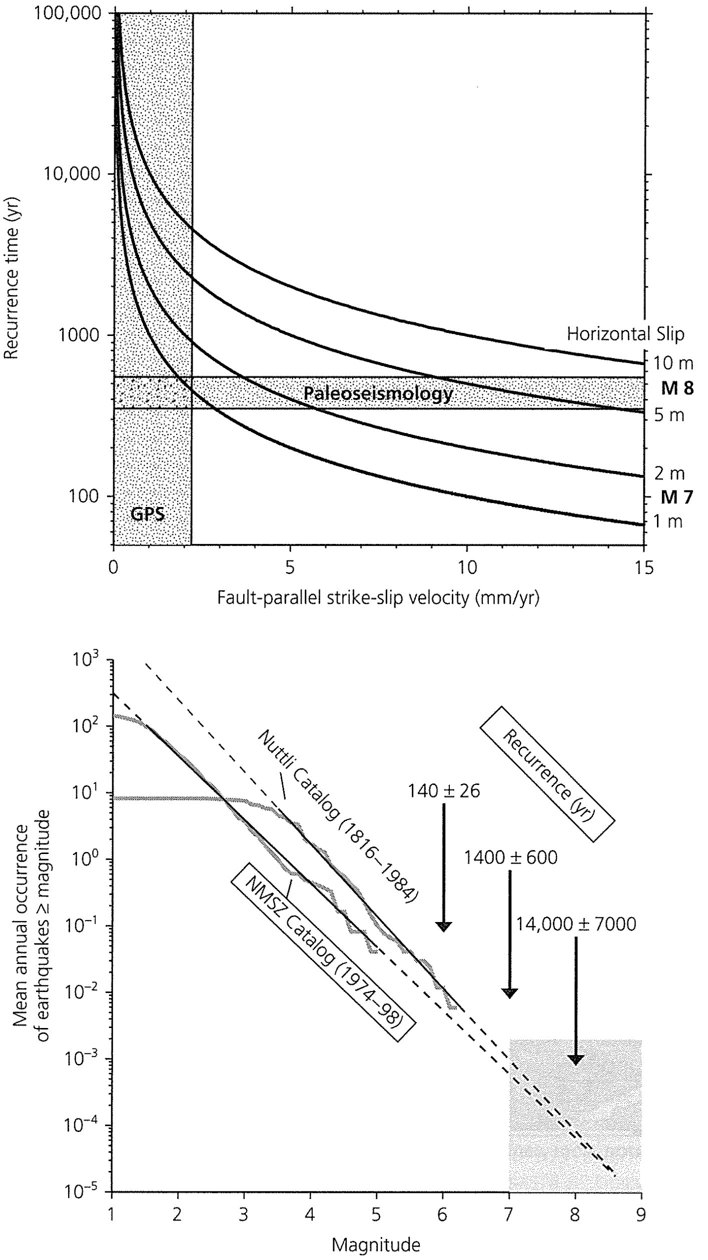


Fig. 5.6-16 *Top*: Relation between interseismic motion and the expected recurrence of large New Madrid earthquakes. The recurrence estimates from paleoseismic studies and geodetic data are jointly consistent with slip in the 1811–12 earthquakes of about 1 m, corresponding to a low magnitude 7 earthquake. *Bottom*: Earthquake frequency–magnitude data for the New Madrid zone. Both the instrumental and historic (1816–1984) data predict a recurrence interval of about 1000 yr for magnitude 7 earthquakes. (Newman *et al.*, 1999. *Science, 284*, 619–21, copyright 1999 American Association for the Advancement of Science.)

margins, even those that have not been recently deglaciated. Stresses associated with the removal of glacial loads may also play a role in causing earthquakes within continental interiors such as the northeastern USA and eastern Canada. It has also been suggested that the huge 1998 Balleny Island intraplate earthquake (Section 5.5.1) may have been triggered by stresses due to the shrinking Antarctic ice cap.

As in the oceans, another interesting class of intraplate seismicity is associated with hot spots. The area near the Yellowstone hot spot in the western USA shows an intriguing pattern of seismicity along the margins of the Snake River plain (Fig. 5.6-21), which is the volcanic track the hot spot produced as the North American plate moved over it (Fig. 5.2-8). This seismicity, which includes the 1959 M_s 7.5 Hebgen Lake, Montana,[6] and 1983 M_s 7.3 Borah Peak, Idaho, earthquakes, forms a parabolic pattern extending southwestward from Yellowstone itself. It thus stands out from the regional seismicity (Fig. 5.2-3) associated with the extensional tectonics of the eastern portion of the Basin and Range province, termed the Intermountain Seismic Belt. The absence of seismicity along the track itself seems likely to be a consequence of the thermal and magmatic perturbations produced by the hot spot, although the specific mechanism is still under discussion. Seismic tomography (Fig. 5.6-21) shows a low-velocity anomaly in the crust and upper mantle under Yellowstone itself, presumably due to partial melting and hydrothermal fluids, and a deeper anomaly that persists along the track.

In summary, although continental intraplate seismicity is a minor fraction of global seismic moment release, it has both scientific and societal interest precisely because it is rare. It provides one of our few ways of studying the limits of plate rigidity and intraplate stresses, and poses the challenge of deciding the appropriate level of earthquake preparedness for rare, but potentially destructive, earthquakes.

5.7 Faulting and deformation in the earth

Because earthquake faulting is a spectacular manifestation of the processes that deform the solid earth, we seek to understand how earthquakes result from and reflect this deformation. Valuable insight comes from laboratory experiments and theoretical models for the behavior of solid materials. Although the experiments and models are much simpler than the complexities of the real earth, they allow us to think about key features. Seismology and geophysics thus exploit research devoted to material behavior by a range of disciplines, including engineering, materials science, and solid state physics. We touch only briefly on some basic ideas, and more information can be found in the references at the end of the chapter.

5.7.1 Rheology

Materials can be characterized by their *rheology*, the way they deform. In seismology we typically take a continuum

[6] This earthquake triggered an enormous landslide that buried a campground, causing 28 deaths and dammed the Madison River, forming Quake Lake. These dramatic effects are still visible today and make the site well worth visiting. A visitor center and parking lot are built on the slide.

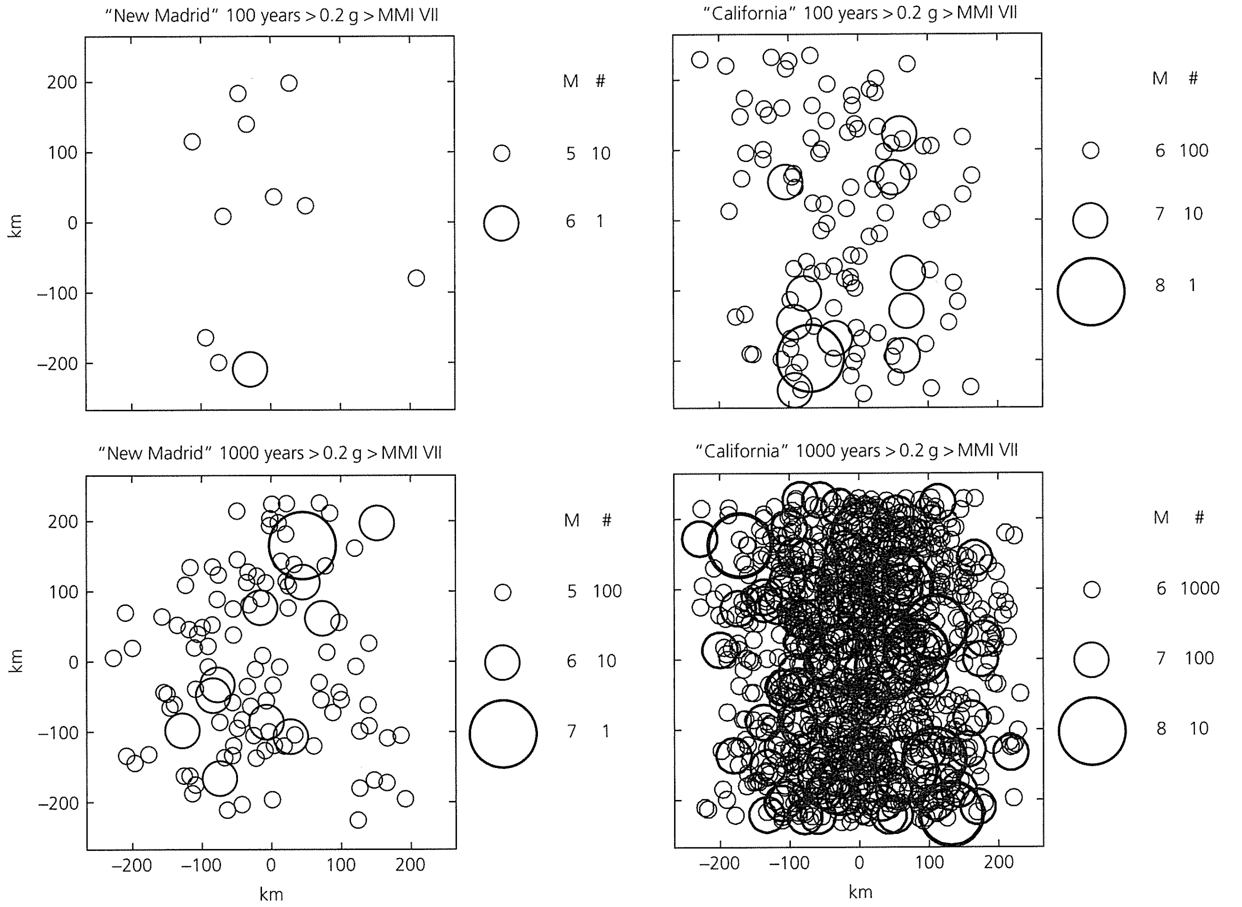


Fig. 5.6-17 Schematic illustration of the relation between the recurrence times of seismicity and resulting seismic hazard for the intraplate New Madrid seismic zone and the southern California plate boundary zone. Seismicity is assumed to be randomly distributed about an N–S line through 0, with California 100 times more active, but New Madrid earthquakes causing potentially serious damage (circles show areas with acceleration 0.2 g or greater, Table 1.2-4) over an area comparable to that for a California earthquake one magnitude unit larger.

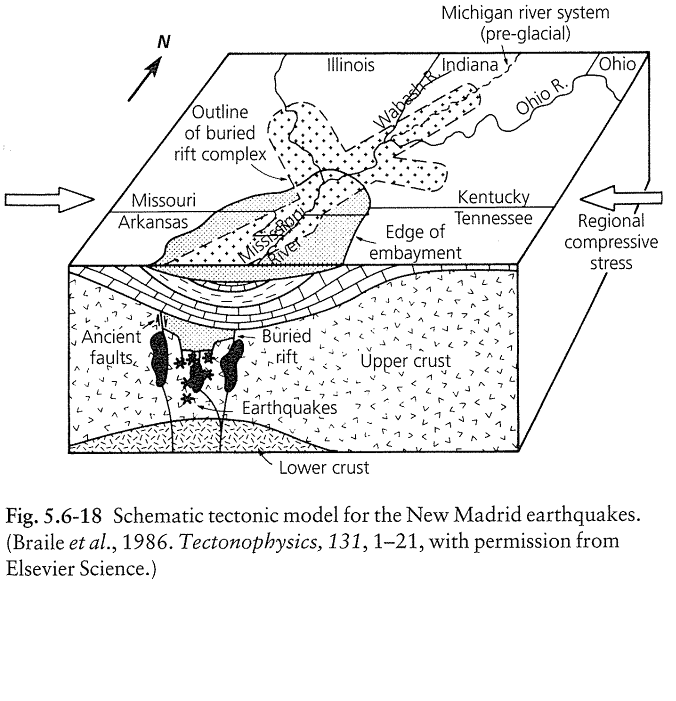


Fig. 5.6-18 Schematic tectonic model for the New Madrid earthquakes. (Braile *et al.*, 1986. *Tectonophysics, 131*, 1–21, with permission from Elsevier Science.)

approach, considering the earth to be a continuous deformable material. This means that we focus on its aggregate behavior (Section 2.3) rather than on how its behavior is determined by what happens at a microscopic scale.

To do this, consider the strain that results from compressing a rock specimen. The simplest case is shown in Fig. 5.7-1a. For small stresses, the resulting strain is proportional to the applied stress, so the material is purely *elastic*. Elastic behavior happens when seismic waves pass through rock, because the strains are small (Section 2.3.8). However, once the applied stress reaches a value σ_f, known as the rock's *fracture strength*, the rock suddenly breaks. Such *brittle* fracture is the simplest model for what happens when an earthquake occurs on a fault. Thus brittle fracture — a deviation from elasticity — generates elastic seismic waves.

Other materials show a change in the stress–strain curve for increasing stresses (Fig. 5.7-1b). For stresses less than the *yield stress*, σ_o, the material acts elastically. Thus, if the stress is released, the strain returns to zero. However, for stresses greater

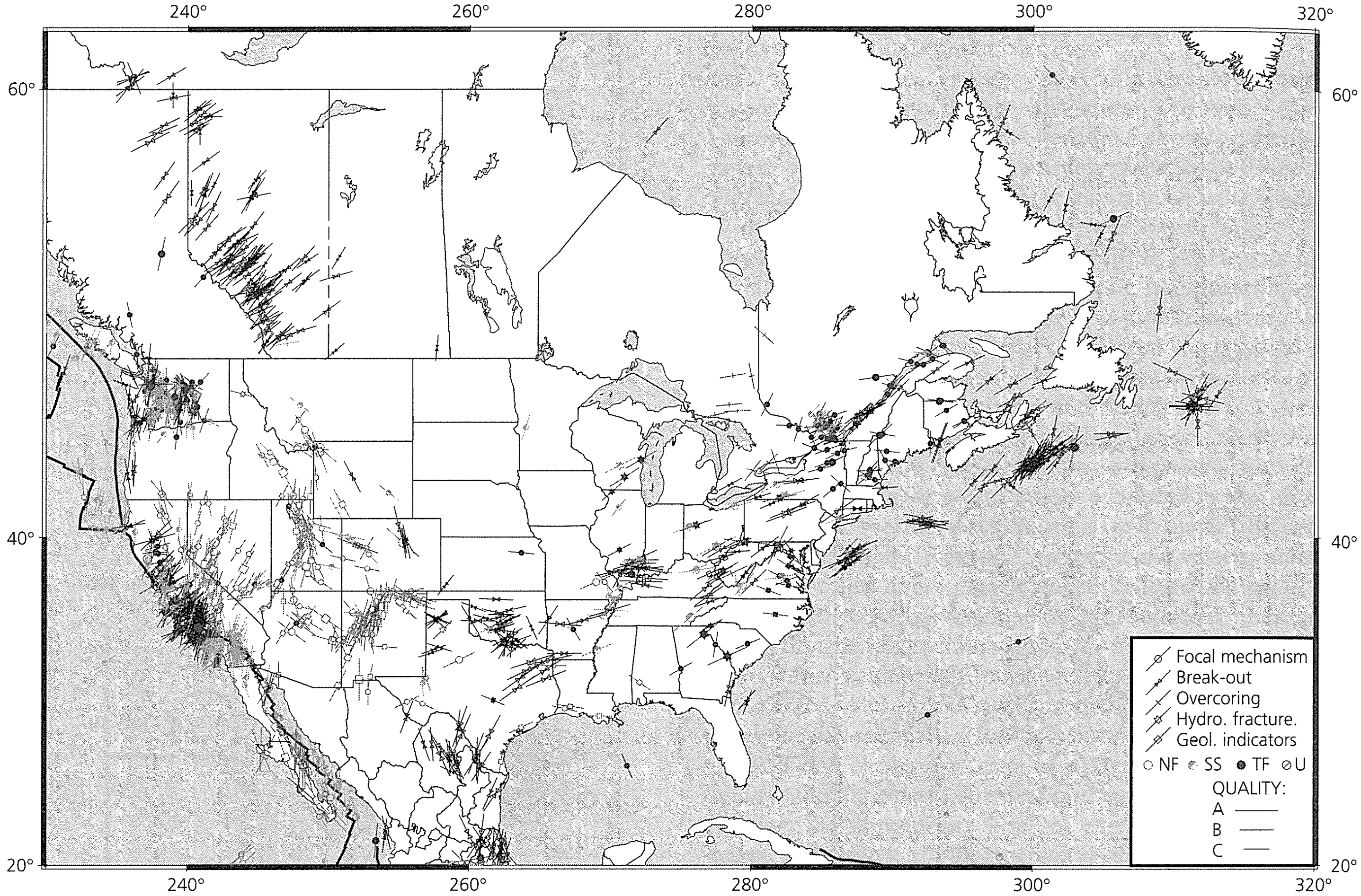


Fig. 5.6-19 Stress map for North America. (World Stress Map project, 2000. Courtesy of the US Geological Survey.)

than the yield stress, releasing the stress relieves the elastic portion of the strain, but leaves a permanent deformation (Fig. 5.7-1c). If the material is restressed, the stress–strain curve now includes the point of the permanent strain. The material behaves as though its elastic properties were unchanged, but the yield strength has increased from σ_0 to σ_0'. The portion of the stress–strain 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*. A common approximation is to treat ductile materials as *elastic-perfectly plastic*: stress is proportional to strain below the yield stress and constant for all strains when stress exceeds the yield stress (Fig. 5.7-2).

An important result of laboratory experiments is that at low pressures rocks are brittle, but at high pressures they behave ductilely, or flow. Figure 5.7-3 shows experiments where a rock is subjected to a compressive stress σ_1 that exceeds a confining pressure σ_3. For confining pressures less than about 400 MPa the material behaves brittlely — it reaches the yield strength, then fails. For higher confining pressures the material flows ductilely. These pressures occur not far below the earth's surface — as discussed earlier, 3 km depth corresponds to 100 MPa pressure — so 800 MPa is reached at about 24 km. This experimental result is consistent with the idea that the strong lithosphere is underlain by the weaker asthenosphere.

A related phenomenon is that materials behave differently at different time scales. A familiar example is that although an asphalt driveway is solid if one falls on it, a car parked on it during a hot day can sink a little ways into it. On short time scales the driveway acts rigidly, but on longer time scales it starts to flow as a *viscous fluid*. This effect is crucial in the earth, because the mantle is solid on the time scale needed for seismic waves to pass through it, but flows on geological time scales.

5.7.2 *Rock fracture and friction*

The first question we address is how and when rocks break. In the brittle regime of behavior, the development of faults and the initiation of sliding on preexisting faults depend on the applied stresses.

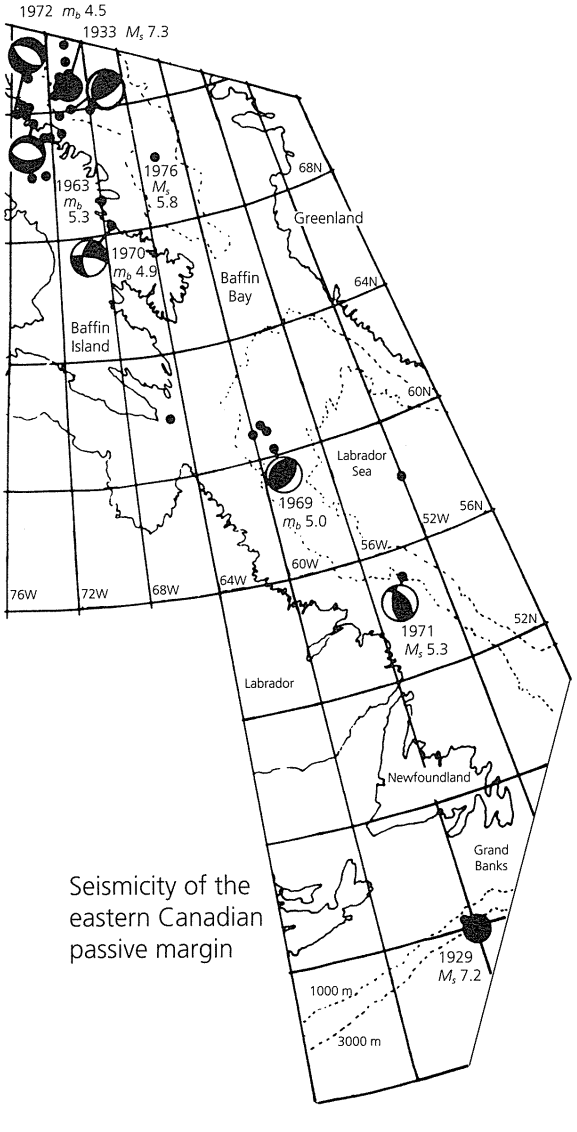


Fig. 5.6-20 Earthquakes along the passive continental margin of eastern Canada. These earthquakes may have occurred on faults remaining from continental rifting. (Stein *et al.*, 1979. *Geophys. Res. Lett.*, *6*, 537–40, copyright by the American Geophysical Union.)

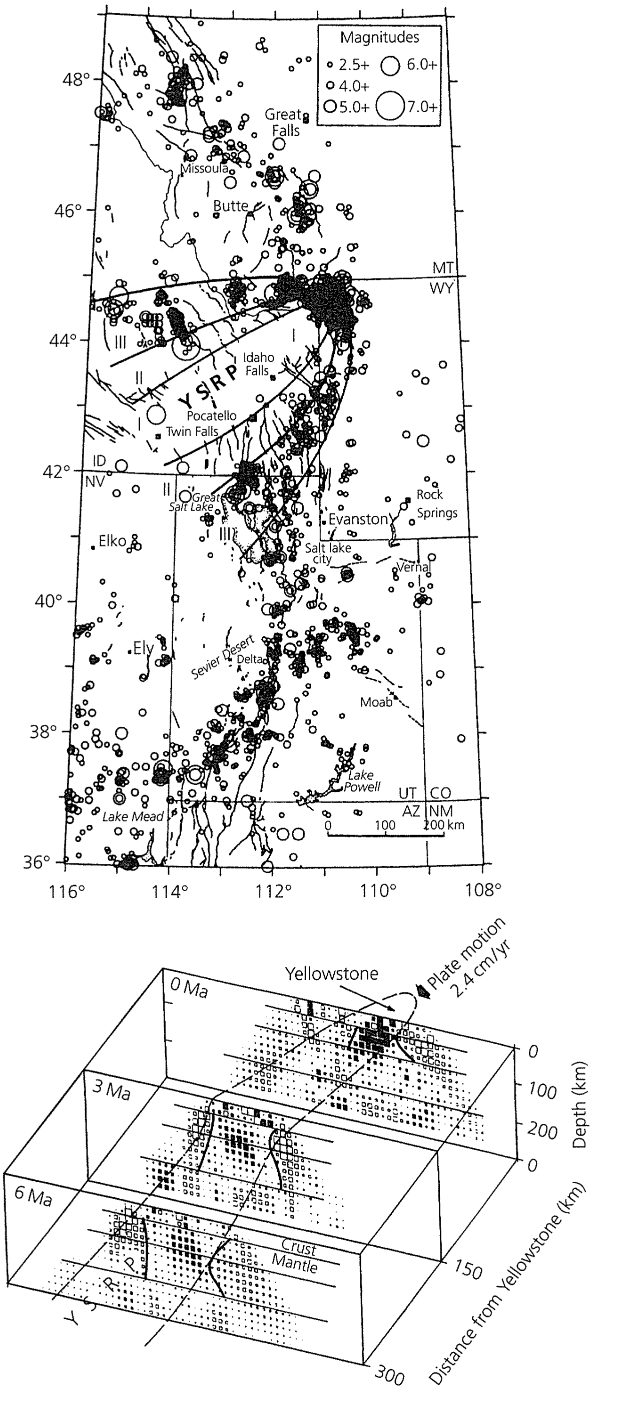


Fig. 5.6-21 *Top*: Seismicity (1900–85) of the Intermountain area of the western USA. Superimposed on the regional seismicity are earthquakes forming a parabola along the margins of the Yellowstone–Snake River plain (YRSP), the volcanic track of the Yellowstone hot spot. *Bottom*: *P*-wave velocities across the hot spot track, shown by squares scaled in size to the differences from a uniform-velocity model. The largest symbols are ±3%, with dark and open symbols showing low and high velocities. (Smith and Braile, 1994. *J. Volcan. Geotherm. Res.*, *61*, 121–87, with permission from Elsevier Science.)

Given a stress field specified by a stress tensor, we use the approach of Section 2.3.3 to find the variation in normal and shear stress on faults of various orientations. For simplicity, we consider the stress in two dimensions. If the coordinate axes ($\hat{e}_1$, $\hat{e}_2$) are oriented in the principal stress directions, the stress tensor is diagonal,

$$\sigma_{ij} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}. \qquad (1)$$

To find the stress on a plane whose normal $\hat{e}'_1$ is at an angle of θ from $\hat{e}_1$, the direction of σ_1 (Fig. 5.7-4), we transform the stress

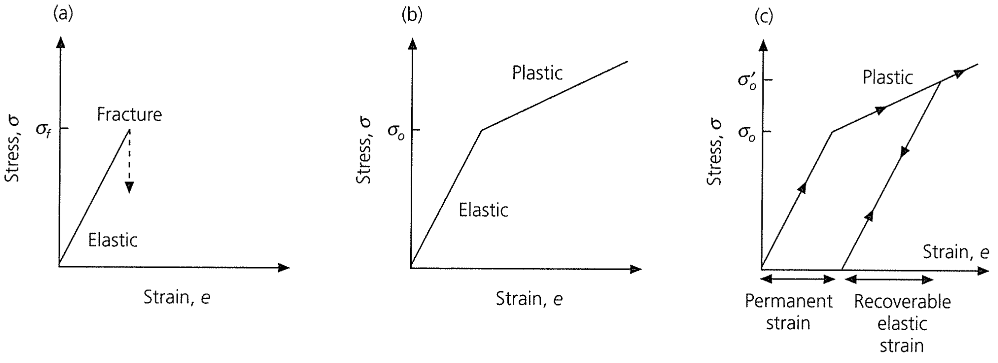


Fig. 5.7-1 (a): A material is perfectly elastic until it fractures when the applied stress reaches σ_f. (b): A material undergoes plastic deformation when the stress exceeds a yield stress σ_o. (c): A permanent strain results from plastic deformation when the stress is raised to σ'_o and released.

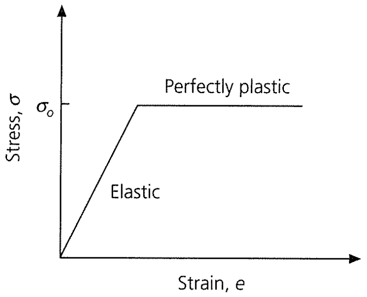


Fig. 5.7-2 An elastic–perfectly plastic rheology, which is a commonly used approximation for the behavior of ductile materials.

tensor from the principal axis coordinate system to a new coordinate system using the transformation matrix (Section 2.3.3)

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{2}$$

so that the stress in the new (primed) system is

$$\sigma'_{ij} = A\sigma A^T = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$
$$= \begin{pmatrix} \sigma_1 \cos^2\theta + \sigma_2 \sin^2\theta & (\sigma_2 - \sigma_1)\sin\theta\cos\theta \\ (\sigma_2 - \sigma_1)\sin\theta\cos\theta & \sigma_1 \sin^2\theta + \sigma_2\cos^2\theta \end{pmatrix}. \tag{3}$$

The normal and shear stresses on the plane vary, depending on the plane's orientation. The normal stress component, denoted by σ, is

$$\sigma = \sigma'_{11} = \sigma_1\cos^2\theta + \sigma_2\sin^2\theta = \frac{(\sigma_1+\sigma_2)}{2} + \frac{(\sigma_1-\sigma_2)}{2}\cos 2\theta, \tag{4a}$$

and the shear component, denoted by τ, is

$$\tau = \sigma'_{12} = (\sigma_2 - \sigma_1)\sin\theta\cos\theta = \frac{(\sigma_2-\sigma_1)}{2}\sin 2\theta. \tag{4b}$$

Figure 5.7-4 shows σ and τ as functions of θ for the case of σ_1 and σ_2 negative ($|\sigma_1| > |\sigma_2|$), which corresponds to compression at depth in the earth. A graphic way to show these is with *Mohr's circle*, a plot of σ versus τ (Fig. 5.7-5). Values for all different planes lie on a circle centered at $\sigma = (\sigma_1 + \sigma_2)/2$, $\tau = 0$, with radius $(\sigma_2 - \sigma_1)/2$. The point on the circle with angle 2θ, measured clockwise from the $-\sigma$ axis, gives the σ, τ values on the plane whose normal is at angle θ to σ_1.[1]

Laboratory experiments on rocks under compression show that fracture occurs when a critical combination of the absolute value of shear stress and the normal stress is exceeded. This relation, known as the *Coulomb–Mohr failure criterion*, can be stated as

$$|\tau| = \tau_o - n\sigma, \tag{5}$$

where τ_o and n are properties of the material known as the *cohesive strength* and *coefficient of internal friction*. (The minus sign reflects the convention that compressional stresses are negative.) The failure criterion plots as two lines in the τ–σ plane, with τ axis intercepts $\pm\tau_o$ and slope $\pm n$ (Fig. 5.7-6). If the principal stresses are σ_1, σ_2, such that Mohr's circle does not intersect the failure lines, the material does not fracture. However, given the same σ_2 but a higher σ'_1, Mohr's circle intersects the line, and the material breaks.

The failure lines show how much shear stress, τ, can be applied to a surface subject to a normal stress σ before failure occurs. The cohesive strength is the minimum (absolute value) shear stress for failure. The coefficient of internal friction indicates the additional shear stress sustainable as the normal stress increases. Thus, deeper in the crust, where the pressure and hence normal stress are higher, rocks are stronger, and higher shear stress is required to break them.

The failure lines and Mohr's circle show on which plane failure occurs for a given stress state. To find θ, the angle between the plane's normal and the maximum compressive stress (σ_1) direction, we write the failure lines as

$$|\tau| = \tau_o - \sigma\tan\phi, \tag{6}$$

[1] Following the seismological convention of compressive stresses being negative, Mohr's circle is shown for $\sigma < 0$. The opposite convention is often used in rock mechanics, e.g. Figs. 5.7-3 and 5.7-10.

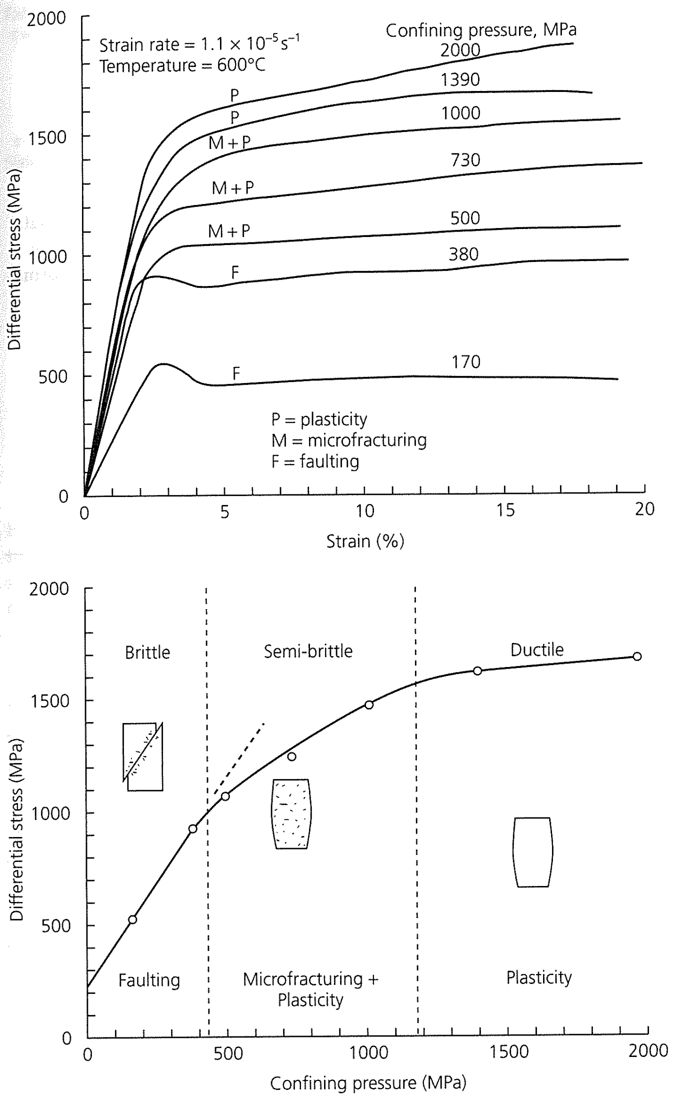


Fig. 5.7-3 Results of an experiment in which rocks are subjected to a compressive stress σ_1 greater than the confining pressure σ_3. *Top*: Differential stress ($\sigma_1 - \sigma_3$) versus strain (compare to Figs 5.7-1 and 2) curves for various confining pressures. *Bottom*: Ultimate strength ($\sigma_1 - \sigma_3$ at 10% strain rate, from top) for various confining pressures. For low (< 400 MPa) confining pressures, the material fractures, and its strength increases with pressure. For higher pressures, the material is ductile, and its strength increases only slowly with pressure. A semi-brittle transition regime, in which both microfractures and crystal plasticity occur, separates the brittle and ductile regimes. (Kirby, 1980. *J. Geophys. Res., 85*, 6353–63, copyright by the American Geophysical Union.)

where $n = \tan\phi$, and ϕ, the *angle of internal friction*, is formed by extending the failure line to the σ axis (Fig. 5.7-7). Fracture occurs at point F, where the failure line is tangent to Mohr's circle. Considering the right triangle AFB, we see that

$$\phi = 2\theta - 90°, \quad \text{so} \quad \theta = \phi/2 + 45°. \tag{7}$$

For example, in introducing the relation between fault plane solutions and crustal stresses in Section 2.3.5, we made the simplest assumption that fracture occurs at 45° to the principal stress axes, corresponding to the case $\phi = 0°$, $n = 0$, $\theta = 45°$. Physically, this means that the normal stress has no effect on the strength of the rock. However, rocks typically have n about 1, so $\phi = 45°$, $\theta = 67.5°$, and the fault plane is closer (22.5°) to the maximum compression (σ_1) direction (Fig. 5.7-8). This idea is important when using P and T axes of focal mechanisms to characterize stress directions.

Figure 5.7-7 also shows how to find the stresses when fracture occurs. Consider the point T on the failure line such that $\overline{T\sigma_2}$ is perpendicular to the σ axis. Because the angle $AT\sigma_2$ is θ (triangles AFT and Aσ_2T are congruent),

$$\overline{T\sigma_2} = A\sigma_2 \cot\theta, \tag{8}$$

or, since $\overline{A\sigma_2} = (\sigma_2 - \sigma_1)/2$,

$$\overline{T\sigma_2} = \frac{(\sigma_2 - \sigma_1)}{2}\cot\theta. \tag{9}$$

Similarly,

$$\overline{T\sigma_2} = \tau_o - \sigma_2 \tan\phi \tag{10}$$

(the minus sign is because σ_2 is negative), so

$$\frac{(\sigma_2 - \sigma_1)}{2}\cot\theta = \tau_o - \sigma_2 \tan\phi. \tag{11}$$

This relation can be written in terms of the angle of the fracture plane, using Eqn 7 and trigonometric identities,

$$\tan\phi = -\cot 2\theta = \frac{-1}{\tan 2\theta} = \frac{\tan^2\theta - 1}{2\tan\theta}, \tag{12}$$

yielding

$$\sigma_1 = -2\tau_o \tan\theta + \sigma_2 \tan^2\theta. \tag{13}$$

We will use this relation between the stresses when fracture occurs to estimate the maximum stresses in the crust.

Similar analyses show when the shear stress is high enough to overcome friction and cause sliding on a previously existing fault. The results are similar to those for a new fracture in unbroken rock, except that at low stress levels the preexisting fault has no cohesive strength. Thus slip on the fault occurs when $|\tau| = -\mu\sigma$, where μ is the *coefficient of sliding friction*, which can be expressed by an *angle of sliding friction*

$$\tan\alpha = \mu. \tag{14}$$

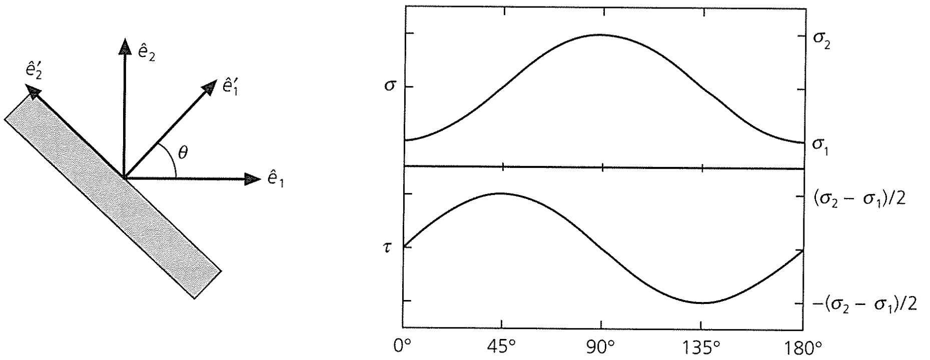


Fig. 5.7-4 *Left*: Geometry of a plane with normal $\hat{e}'_1$, oriented at an angle θ from $\hat{e}_1$, the direction of the maximum compressive stress σ_1. *Right*: Normal stress, σ, and shear stress, τ, as functions of the angle θ.

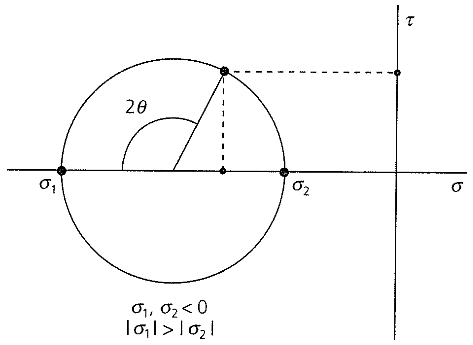


Fig. 5.7-5 Mohr's circle: Given a state of stress described by principal stresses σ_1 and σ_2, the normal stress, σ, and the shear stress, τ, for planes of all orientations lie on a circle with radius $(\sigma_2 - \sigma_1)/2$. The point on the circle with angle 2θ, measured clockwise from the $-\sigma$ axis, gives σ and τ on a plane whose normal is at an angle θ from the direction of σ_1.

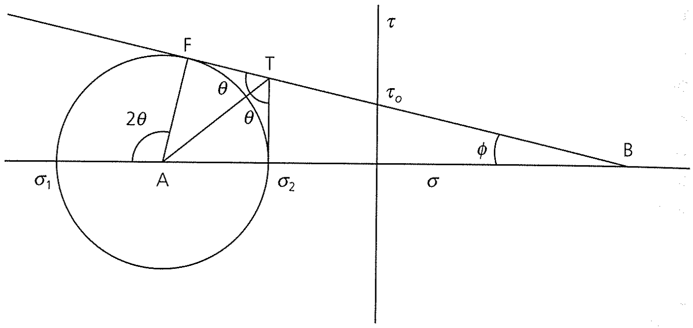


Fig. 5.7-7 Fracture occurs at point F, where a material's failure line, characterized by its cohesive strength, τ_o, and angle of internal friction, ϕ, is tangent to Mohr's circle. Hence θ is the angle of the plane on which fracture occurs, and F gives the stresses at fracture. Point A is the center of Mohr's circle, B is where the failure line intersects the σ axis, and $\overline{T\sigma_2}$ is perpendicular to the σ axis. For simplicity, only the upper failure line for $\tau > 0$ is shown in this and subsequent figures.

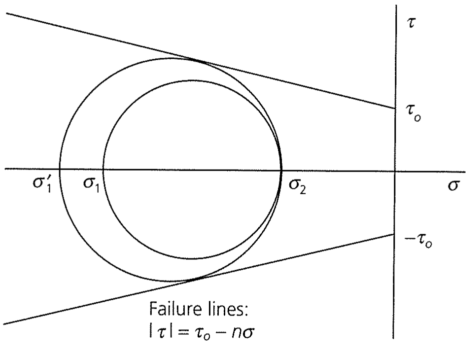


Fig. 5.7-6 The Coulomb–Mohr failure criterion assumes that a material fractures when Mohr's circle intersects the failure line.

Figure 5.7-9 shows the Mohr's circle representation of a rock with preexisting faults. In addition to the failure line, there is a frictional sliding line corresponding to

$$\tau = -\mu\sigma = -\sigma \tan \alpha. \tag{15}$$

Because the sliding line starts at the origin, it is initially below the failure line. Assume that the stresses are large enough that Mohr's circle touches the failure line at the point yielding frac-

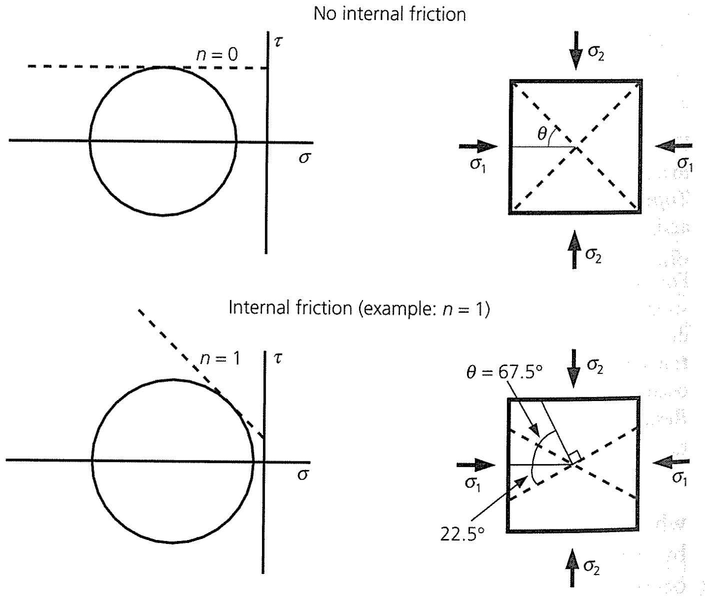


Fig. 5.7-8 With no internal friction, fracture occurs at an angle of 45°. For $n = 1$, the fracture angle is 67.5°, and the fault plane is closer (22.5°) to the maximum compression (σ_1) direction.

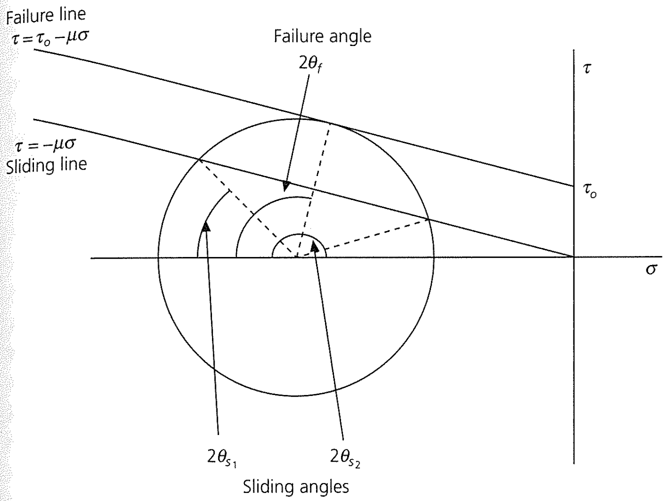


Fig. 5.7-9 Mohr's circle for sliding on a rock's preexisting faults. A new fracture would form at an angle θ_f, given by the failure line. However, slip will occur on a preexisting fault if there are any with angles between θ_{s1} and θ_{s2}, given by the intersection of the circle with the frictional sliding line.

ture on a plane corresponding to an angle θ_f. Similarly, the frictional sliding line intercepts the circle at two points, corresponding to angles θ_{s1} and θ_{s2}. Thus the rock can fail in several ways. If there are preexisting faults with angles θ_{s1} or θ_{s2}, slip on these faults may occur. Alternatively, a new fracture may form on the plane given by θ_f. However, because this fracture occurs at higher shear stress than is needed for frictional sliding on the preexisting faults, sliding is favored over the formation of a new fracture. Thus, if the stress has gradually risen to this level, sliding on preexisting faults would probably have prevented a new fracture from forming.

This effect can have seismological consequences. The simplest way to use focal mechanisms to infer stress orientations is to assume that the earthquakes occurred on newly formed faults. However, if the rock had been initially faulted, the earthquakes may have occurred on preexisting faults. In the representation of Fig. 5.7-9, if faults exist with normals oriented between θ_{s1} and θ_{s2} to the maximum compressive stress, slip on these faults will occur rather than the formation of a new fracture. Thus the inferred stress direction will be somewhat inaccurate. For example, the thrust focal mechanisms along the Himalayan front (Fig. 5.6-6) or eastern Andean foreland thrust belt (Fig. 5.6-10) have fault planes that rotate as the trend of the mountains changes, suggesting that the fault planes are controlled by the existing structures, so the P axes only partially reflect the stress field. A similar pattern appears for T axes along the East African rift (Fig. 5.6-2). In general, stress axes inferred from many fault plane solutions in an area seem relatively coherent (Fig. 5.6-19). Thus we assume that the crust contains preexisting faults of all orientations, so the average stress orientation inferred from the focal mechanisms is not seriously biased.

At this point, it is worth noting other complexities. Both the failure and sliding curves may be more complicated than straight lines. These curves, known as Mohr envelopes, can be derived from experiments at various values of stress. Additional complexity comes from the fact that water and other fluids are often present in rocks, especially in the upper crust. The fluid pressure, known as the *pore pressure*, reduces the effect of the normal stress and allows sliding to take place at lower shear stresses. This effect is modeled by replacing the normal stress σ with $\bar{\sigma} = \sigma - P_f$, known as the *effective normal stress*, where P_f is the pore fluid pressure.[2] Because the pore pressure is defined as negative, the effective normal stress is reduced (less compressive). Similarly, effective principal stresses taking into account pore pressure,

$$\bar{\sigma}_1 = \sigma_1 - P_f \quad \text{and} \quad \bar{\sigma}_2 = \sigma_2 - P_f, \tag{16}$$

are used in the fracture theory.

The relations we have discussed can be used to estimate the maximum stresses that the crust can support. Laboratory experiments (Fig. 5.7-10) for sliding on existing faults in a variety of rock types find relations sometimes called *Byerlee's law*:

$$\begin{aligned} &\tau \approx -0.85\bar{\sigma}, \quad |\bar{\sigma}| < 200 \text{ MPa} \\ &\tau \approx 50 - 0.6\bar{\sigma}, \quad |\bar{\sigma}| > 200 \text{ MPa}. \end{aligned} \tag{17}$$

These relations, written in terms of the normal and shear stresses on a fault, can be used to infer the principal stress as a function of depth. To do so, we write the minimum compressive stress as σ_3, because we are in three dimensions. We assume that the crust contains faults of all orientations, and that the stresses cannot exceed the point where Mohr's circle is tangent to the frictional sliding line, or else sliding will occur (Fig. 5.7-11). At shallow depths where $|\bar{\sigma}| < 200$ MPa, Eqn 17 shows that $\tau_0 = 0$. Thus Eqn 13, the relation between the stresses when fracture occurs, yields

$$\bar{\sigma}_1 = \bar{\sigma}_3 \tan^2 \theta_s. \tag{18}$$

Using Eqn 7 for the case of frictional sliding,

$$\theta_s = \alpha/2 + 45°, \tag{19}$$

and the values in Eqn 17 give

$$\mu = \tan \alpha = 0.85, \quad \alpha \approx 41°, \quad \theta_s \approx 66°, \quad \tan^2 66° \approx 5, \tag{20}$$

so the stresses are related by

$$\bar{\sigma}_1 \approx 5\bar{\sigma}_3. \tag{21}$$

At greater depths, where $|\bar{\sigma}| > 200$ MPa, $\alpha \approx 31°$ and $\theta_s = 60.5°$, so the stresses are related by

$$\bar{\sigma}_1 \approx -175 + 3.1\bar{\sigma}_3. \tag{22}$$

[2] The role of pore pressure in making sliding easier can be seen by trying to slide an object across a dry table and then wetting the table.

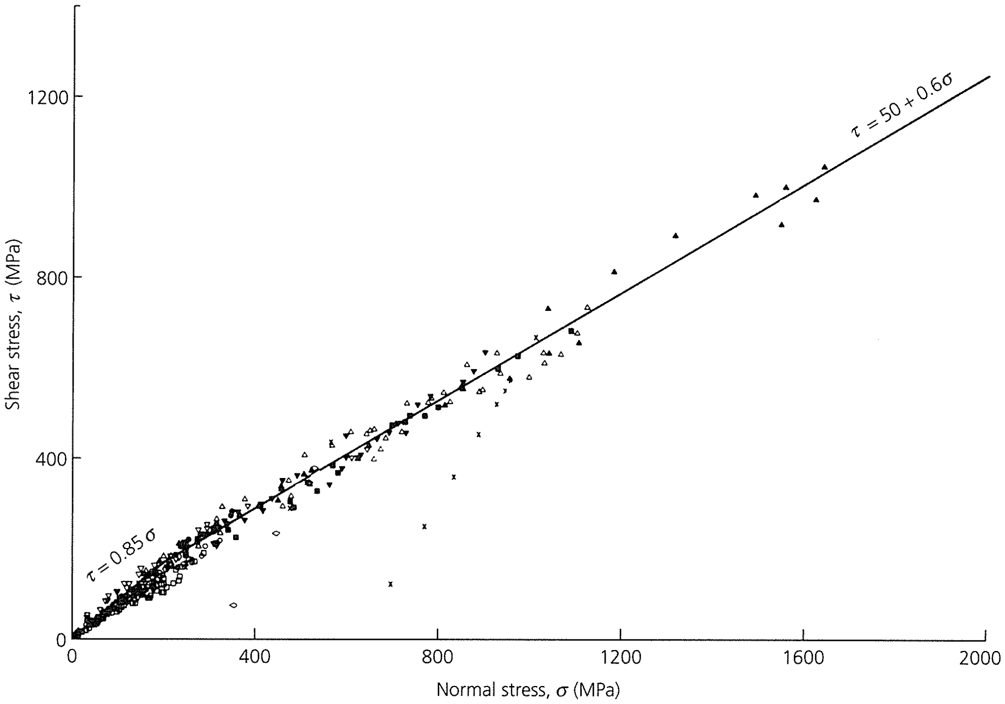


Fig. 5.7-10 Shear stress versus normal stress for frictional sliding, compiled for various rock types. Compressive stress is positive. (Byerlee, 1978. *Pure Appl. Geophys.*, *116*, 615–26, reproduced with the permission of Birkhauser.)

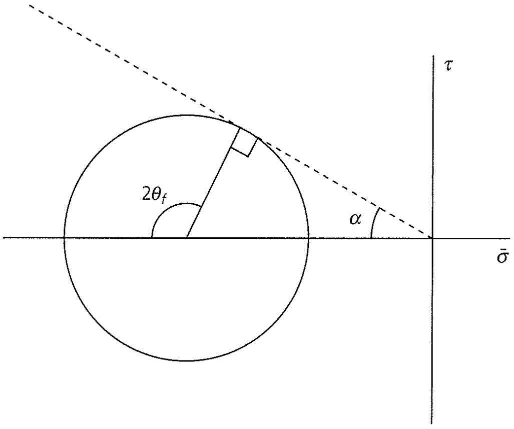


Fig. 5.7-11 Mohr's circle and sliding line for $|\bar{\sigma}| < 200$ MPa. If the lithosphere contains fractures in all directions, the stresses cannot exceed those at the point where Mohr's circle is tangent to the sliding line, because sliding would occur.

We assume that one principal stress, σ_1 or σ_3, is the vertical stress due to the lithostatic pressure as a function of depth (z),

$$\sigma_V = -\rho g z. \tag{23}$$

The other principal stress, which must be horizontal, is denoted σ_H. The pore pressure $P_f(z)$ is unknown. One common assumption is that the rock is dry, so $P_f(z) = 0$. Another is that the pore pressure is *hydrostatic*, which is equivalent to assuming that pores are connected up to the surface, so

$$P_f(z) = -\rho_f g z, \tag{24}$$

where ρ_f is the density of the fluid, which is usually water, with $\rho_f = 1$ g/cm^3. Alternatively, the pore pressure can be assumed to be a fixed fraction of the lithostatic pressure (Section 2.3.6).

We now can find the *strength* of the crust, defined by the maximum difference between the horizontal and vertical stresses that the rock can support. At shallow depths where $|\bar{\sigma}| < 200$ MPa, Eqn 21 shows that $\bar{\sigma}_1 = 5\bar{\sigma}_3$. There are two possibilities, depending on whether the vertical stress is the most ($\bar{\sigma}_1$) or least ($\bar{\sigma}_3$) compressive. If the vertical stress is the most compressive,

$$\sigma_V = \sigma_1, \quad \bar{\sigma}_1 = \sigma_V - P_f = -\rho g z - P_f(z)$$
$$\sigma_H = \sigma_3, \quad \bar{\sigma}_3 = \bar{\sigma}_1/5 = -(\rho g z + P_f(z))/5. \tag{25}$$

Alternatively, if the vertical stress is the least compressive,

$$\sigma_V = \sigma_3, \quad \bar{\sigma}_3 = \sigma_V - P_f = -\rho g z - P_f(z)$$
$$\sigma_H = \sigma_1, \quad \bar{\sigma}_1 = 5\bar{\sigma}_3 = -5(\rho g z + P_f(z)). \tag{26}$$

In the first case,

$$\sigma_H - \sigma_V = \sigma_3 - \sigma_1 = 0.8(\rho g z + P_f(z)), \tag{27}$$

corresponds to an extensional (positive) stress. In the second,

$$\sigma_H - \sigma_V = \sigma_1 - \sigma_3 = -4(\rho g z + P_f(z)) \tag{28}$$

corresponds to a compressive (negative) stress that is much greater in absolute value. Thus, at any depth, the crust can

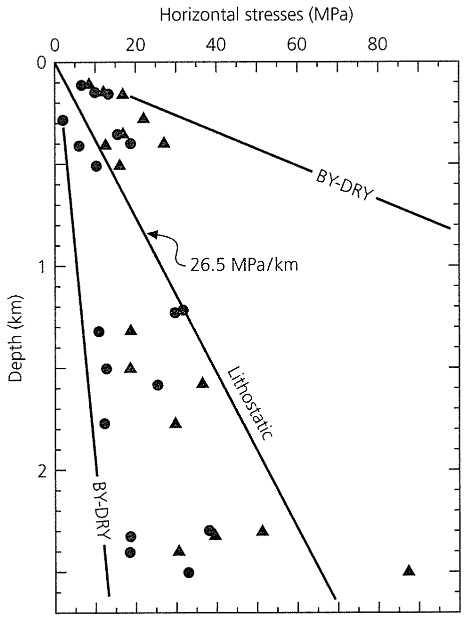


Fig. 5.7-12 Horizontal stresses measured in southern Africa. Dots are for horizontal stresses being the least compressive (σ_3), and triangles are for horizontal stresses being the most compressive (σ_1). The lithostatic stress gradient (26.5 MPa/km) is shown, along with Byerlee's law (BY) for zero pore pressure (DRY). The stronger line is for compression, and the weaker one is for extension. The observed stresses are within the maximum and minimum BY-DRY lines. (Brace and Kohlstedt, 1980. *J. Geophys. Res., 85*, 6248–52, copyright by the American Geophysical Union.)

support greater compressive deviatoric stress than extensional deviatoric stress (Fig. 5.7-12).

5.7.3 *Ductile flow*

When rocks behave brittlely, their behavior is not time-dependent; they either strain elastically or fail. By contrast, the deformation of ductile rock depends on time. A common model for the time-dependent behavior is a *Maxwell viscoelastic material*, which behaves like an elastic solid on short time scales and like a viscous fluid on long time scales. This model can describe the mantle because seismic waves propagate as though the mantle were solid, whereas postglacial rebound and mantle convection occur as though the mantle were fluid.

To see this difference, consider two types of deformation in one dimension. For an elastic solid subjected to elastic strain $e_E = e_{11}$,

$$\sigma = Ee_E, \tag{29}$$

where E is Young's modulus, and σ is σ_{11}. The simplest viscous fluid obeys

$$\sigma = 2\eta \frac{de_F}{dt}, \tag{30}$$

where η is the viscosity, and e_F is the fluid portion of the strain. This equation defines the viscosity, the property that measures a fluid's resistance to shear.[3]

We often think of an elastic material as a spring, which exerts a force proportional to distance. Thus stress and strain are proportional at any instant, and there are no time-dependent effects. By contrast, the viscous material is though of as a dashpot, a fluid damper that exerts a force proportional to velocity. Hence the stress and strain rate are proportional, and the material's response varies with time. These effects are combined in a viscoelastic material, which can be thought of as a spring and dashpot in series (Fig. 5.7-13). The combined elastic and viscous response comes from the combined strain rate

$$\frac{de}{dt} = \frac{de_E}{dt} + \frac{de_F}{dt} = \frac{1}{E}\frac{d\sigma}{dt} + \frac{\sigma}{2\eta}. \tag{31}$$

This differential equation, the rheological law for a Maxwell substance, shows how the stress in the material evolves after a strain e_o is applied at time $t = 0$ and then remains constant. At $t = 0$ the derivative terms dominate, so the material behaves elastically, and has an initial stress

$$\sigma_o = Ee_o. \tag{32}$$

For $t > 0$, $de/dt = 0$, so

$$\frac{d\sigma}{dt} = -\frac{E}{2\eta}\sigma, \tag{33}$$

whose integral is

$$\sigma(t) = \sigma_o \exp\left[-(Et/2\eta)\right]. \tag{34}$$

Thus stress relaxes from its initial value as a function of time (Fig. 5.7-13). A useful parameter is the *Maxwell relaxation time*,[4]

$$\tau_M = \frac{2\eta}{E} \approx \frac{\eta}{\mu}, \tag{35}$$

required for the stress to decay to e^{-1} of its initial value. For times less than τ_M the material can be considered an elastic solid, whereas for longer times it can be considered a viscous fluid.

For example, if the mantle is approximately a Poisson solid with $\mu \approx 10^{12}$ dyn/cm^2 and $\eta \approx 10^{22}$ poise, its Maxwell time is about 10^{10} s or 300 years. Although the viscosity is not that well known, so estimates of the Maxwell time vary, it is clear

[3] In familiar terms, viscosity measures how "gooey" a fluid is. Maple syrup is somewhat more viscous than water, and the earth's mantle is about 10^{24} times more viscous.

[4] Definitions of the Maxwell time vary, but always involve the ratio of the viscosity to an elastic constant.

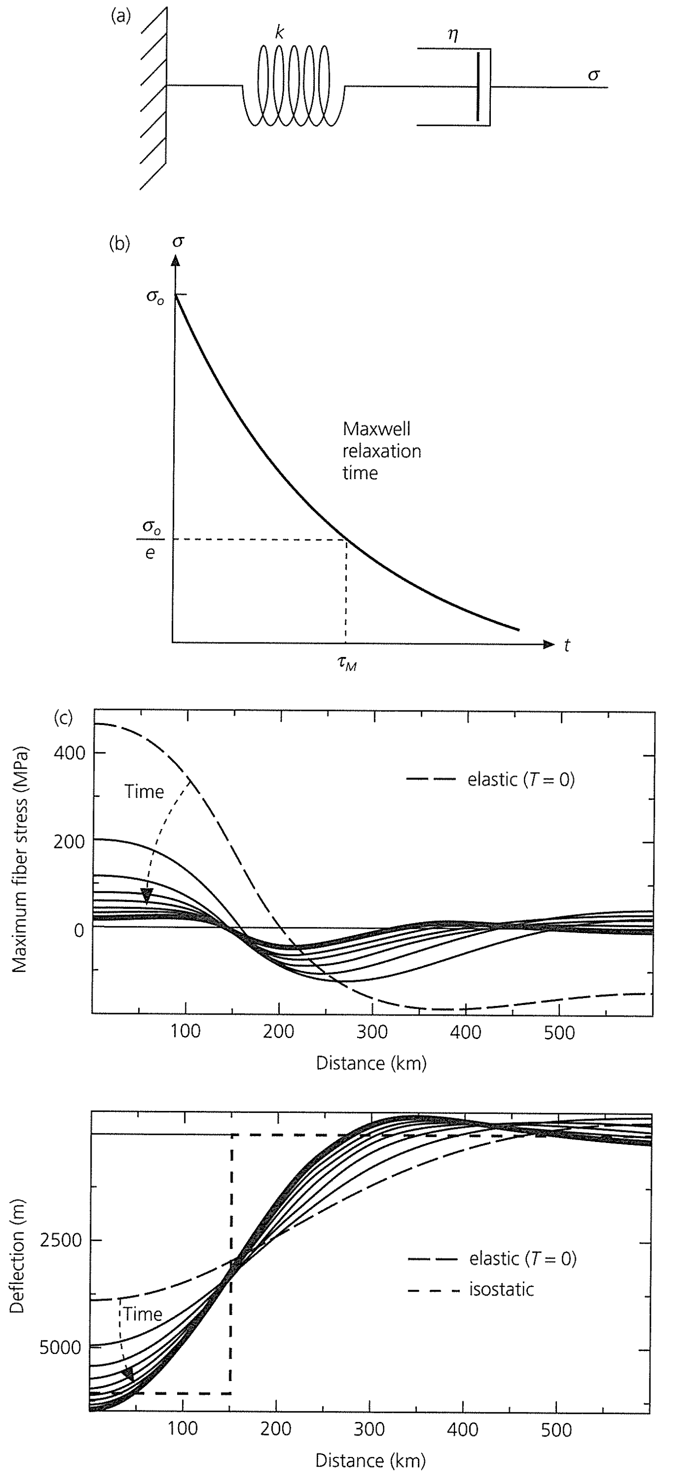


Fig. 5.7-13 (a) Model of a viscoelastic material as an elastic spring and viscous dashpot in series. (b) Stress response of a viscoelastic material to an applied strain. The Maxwell relaxation time, τ_M, is the time the stress takes to decay to e^{-1} of its initial value. (c) Evolution of the deflection and bending stress produced by a sediment load on a viscoelastic earth. At first the earth responds elastically, as shown by the long-dashed line, but with time it flows, so the deflection beneath the load deepens and the stresses relax. (Stein *et al.*, 1989, with kind permission from Kluwer Academic Publishers.)

that we can treat the mantle as a solid for seismological purposes and as a fluid in tectonic modeling. If we model the mantle as viscoelastic, then a load applied on the surface has an effect that varies with time. Figure 5.7-13c shows the effect of a 150 km-wide sediment load, as might be expected on a passive continental margin. Initially, the earth responds elastically, causing large flexural bending stresses. With time, the mantle flows, so the deflection beneath the load deepens and the stresses relax. In the time limit, the stress goes to zero, and the deflection approaches the isostatic solution, because isostasy amounts to assuming that the lithosphere has no strength. Stress relaxation may explain why large earthqukes are rare at continental margins, except where glacial loads have been recently removed (Fig. 5.6-20). Although the large sediment loads should produce stresses much greater than other sources of intraplate stress, including the smaller and less dense ice loads, the stresses produced by sediment loading early in the margin's history may have relaxed.

Laboratory experiments indicate that the rheology of minerals in ductile flow can be described by

$$\frac{de}{dt} = \dot{e} = f(\sigma)\, A \exp\,[-(E^* + PV^*)/RT], \tag{36}$$

where T is temperature, R is the gas constant, and P is pressure. $f(\sigma)$ is a function of the stress difference $|\sigma_1 - \sigma_3|$, and A is a constant. The effects of pressure and temperature are described by the *activation energy* E^* and the *activation volume* V^*. Observed values of $f(\sigma)$ are often fit well by assuming

$$f(\sigma) = |\sigma_1 - \sigma_3|^n$$

$$\dot{e} = |\sigma_1 - \sigma_3|^n A \exp\,[-(E^* + PV^*)/RT]. \tag{37}$$

The rheology of such fluids is characterized by a power law. If $n = 1$, the material is called *Newtonian*, whereas a non-Newtonian fluid with $n = 3$ is often used to represent the mantle. From Eqn 30 we see that for a Newtonian fluid the viscosity depends on both temperature and pressure:

$$\eta = (1/2A) \exp\,[(E^* + PV^*)/RT]. \tag{38}$$

Thus the viscosity decreases exponentially with temperature. This decrease is assumed to give rise to a strong lithosphere overlying a weaker asthenosphere, and the restriction of earthquakes to shallow depths.[5] For a non-Newtonian fluid, Eqn 30 gives the *effective viscosity*, the equivalent viscosity if the fluid were Newtonian.

We think of equations like Eqn 37 as showing the strength, or maximum stress difference $|\sigma_1 - \sigma_3|$, that the viscous material can support. This stress difference depends on temperature, pressure, strain rate, and rock type. The material

[5] Temperature-dependent viscosity is an effect familiar to automobile drivers in cold temperatures, when the engine and the transmission became noticeably sluggish.

is stronger at higher strain rates, and weakens exponentially with high temperatures. At shallow depths, the small pressure effect is often neglected, so the activation volume V^* is treated as zero. For example, a commonly used flow law for dry olivine is[6]

$$\dot{e} = 7\times 10^{4}\,|\sigma_1-\sigma_3|^{3}\,\exp\left(\frac{-0.52\ \text{MJ/mol}}{RT}\right)$$

$$\text{for}\quad |\sigma_1-\sigma_3| \le 200\ \text{MPa}$$

$$= 5.7\times 10^{11}\,\exp\left[\frac{-0.54\ \text{MJ/mol}}{RT}\left(1-\frac{[\sigma_1-\sigma_2]}{8500}\right)^{2}\right]$$

$$\text{for}\quad |\sigma_1-\sigma_3| > 200\ \text{MPa}, \qquad (39)$$

where $\dot{e}$ is in s^{-1}. Similarly, for quartz,

$$\dot{e} = 5\times 10^{6}\,|\sigma_1-\sigma_3|^{3}\,\exp\left(\frac{-0.19\ \text{MJ/mol}}{RT}\right)$$

$$\text{for}\quad |\sigma_1-\sigma_3| < 1000\ \text{MPa}. \qquad (40)$$

At a given strain rate, quartz is much weaker (can sustain a smaller stress difference) than olivine. Thus the quartz-rich continental crust should be weaker that the olivine-rich oceanic crust, an effect whose tectonic consequences are discussed next.

5.7.4 *Strength of the lithosphere*

The strength of the lithosphere as a function of depth depends upon the deformation mechanism. At shallow depths, rocks fail by either brittle fracture or frictional sliding on preexisting faults. Both processes depend in a similar way on the normal stress, with rock strength increasing with depth. However, at greater depths the ductile flow strength of rocks is less than the brittle or frictional strength, so the strength is given by the flow laws and decreases with depth as the temperatures increase. This temperature-dependent strength is the reason why the cold lithosphere forms the planet's strong outer layer.

To calculate the strength, a strain rate and a geotherm giving temperature as a function of depth are assumed. At shallow depths the strength, the maximum stress difference before frictional sliding occurs, is computed using Eqns 27 and 28. At some depth, the frictional strength exceeds the ductile strength allowed by the flow law, so for deeper depths the maximum strength is given by the flow law. Figure 5.7-14 shows a strength plot, known as a *strength envelope*, for a strain rate of $10^{-15}\ s^{-1}$ and a temperature gradient appropriate for old oceanic lithosphere or stable continental interior. In the frictional region, curves are shown for various values of λ, the ratio of pore pressure to lithostatic pressure. The higher pore pressures result in lower strengths. Ductile flow laws are shown for quartz and olivine, minerals often used as models for continental and oceanic rheologies. Strength increases with depth in the brittle region, due to the increasing normal stress, and then decreases with depth in the ductile region, due 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. The strength envelopes show that the lithosphere is stronger for compression than for tension in the brittle regime, but the two are symmetric in the ductile regime. Strength envelopes are often plotted using the rock mechanics convention of compression positive.

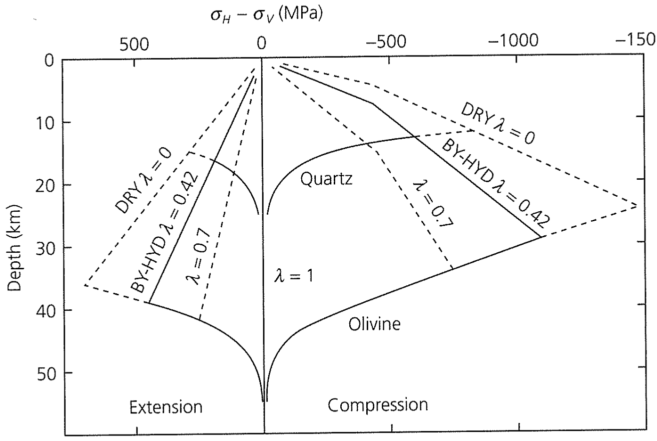


Fig. 5.7-14 Strength envelopes as a function of depth for various values of λ, the ratio of pore pressure to lithostatic pressure. BY-HYD lines are for Byerlee's law with hydrostatic pore pressure. At shallow depths, strength is controlled by brittle fracture; at greater depths ductile flow laws predict 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*). (Brace and Kohlstedt, 1980. *J. Geophys. Res., 85*, 6248–52, copyright by the American Geophysical Union.)

The actual distribution of strength with depth is probably more complicated, because the brittle–ductile transition occurs over a region of semi-brittle behavior that includes both brittle and plastic processes (Fig. 5.7-3). However, this simple model gives insight into various observations. In particular, we have seen that the depths of earthquakes in several tectonic environments seem to be limited by temperature. This makes sense, because for a given strain rate and rheology the exponential dependence on temperature would make a limiting strength for seismicity approximate a limiting temperature.

To see this, consider Fig. 5.7-15, which shows that as oceanic lithosphere ages and cools, the predicted strong region deepens. This result seems plausible because earthquake depths, seismic velocities, and effective elastic thicknesses imply that the strong upper part of the lithosphere thickens with age (Fig. 5.3-9). The strength envelopes are thus consistent with the observation that the maximum depth of earthquakes within

[6] Brace and Kohlstedt (1980).

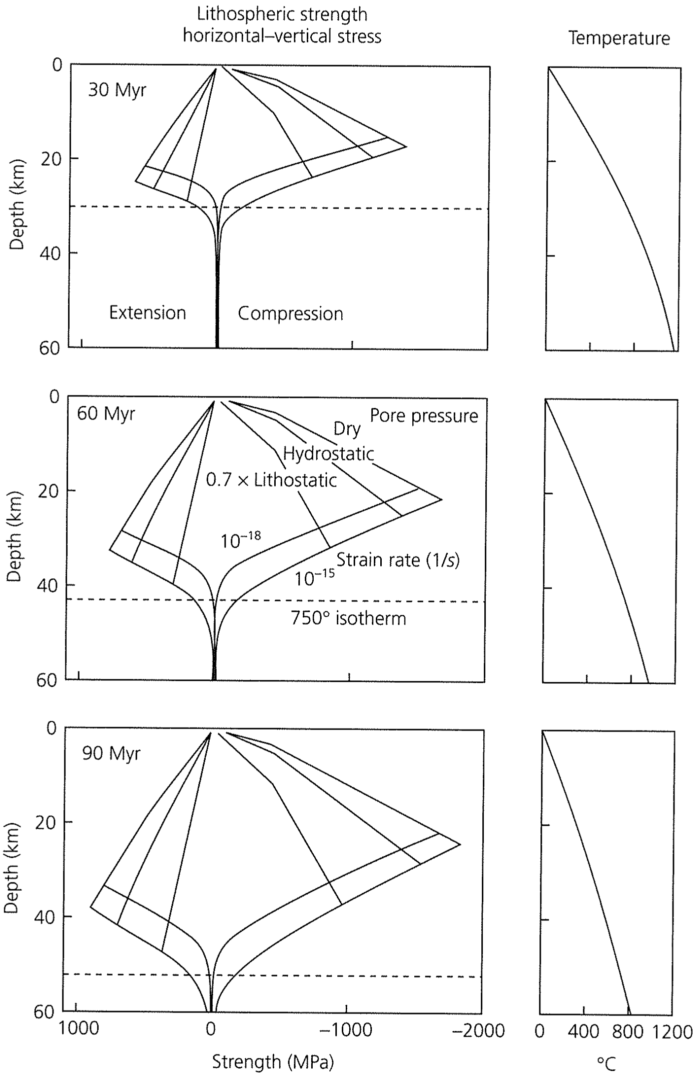


Fig. 5.7-15 Strength envelopes showing maximum stress difference (strength) as a function of depth for an olivine rheology, for geotherms (*right*) corresponding to cooling oceanic lithosphere of different ages. Strength in the brittle regime is reduced by higher pore pressure; strength in the ductile regime is reduced by lower strain rate. The depth range in which the material is strong enough for faulting increases with age. (Wiens and Stein, 1983. *J. Geophys. Res.*, *88*, 6455–68, copyright by the American Geophysical Union.)

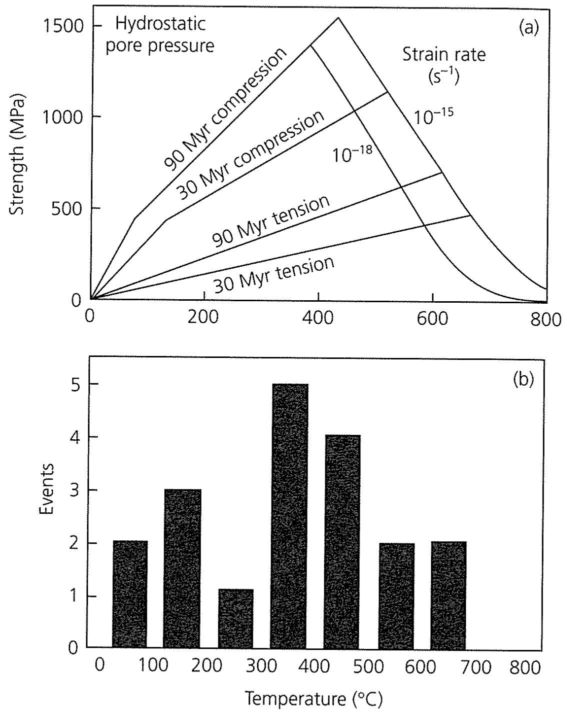


Fig. 5.7-16 Plots of strength and seismicity versus temperature. The strength envelopes explain the observation that intraplate oceanic seismicity occurs only above the 750°C isotherm. (Wiens and Stein, 1985. *Tectonophysics, 116*, 143–62, with permission from Elsevier Science.)

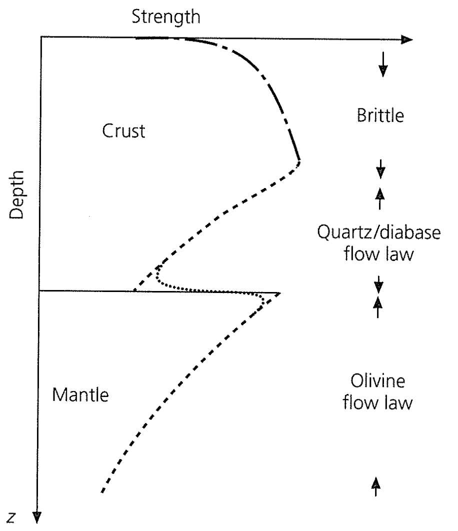


Fig. 5.7-17 Schematic strength envelope for continents. Below the ductile lower crust may be a stronger zone in the olivine-rich mantle. (Chen and Molnar, 1983. *J. Geophys. Res.*, *88*, 4183–4214, copyright by the American Geophysical Union.)

the oceanic lithosphere is approximately bounded by the 750°C isotherm (Fig. 5.7-16). These envelopes are drawn for strain rates of 10^{-15} and 10^{-18} s^{-1}, appropriate for slow deformation within plates. By contrast, a seismic wave with a period of 1 s, a wavelength of 10 km, and a displacement of 10^{-6} m corresponds to a strain rate of 10^{-10} s^{-1}. The successively greater effective elastic thicknesses, depth of the deepest earthquakes, and depth of the low-velocity zone are thus consistent with strength increasing with strain rate.

The strength envelopes give insight into differences between continental and oceanic lithospheres (Fig. 5.7-17). First, quartz is weaker than olivine at a given temperature (Fig. 5.7-14), consistent with the fact that the limiting temperature for continental seismicity is lower than for oceanic earthquakes (Fig. 5.7-18). Second, the strength profiles differ. The strength of oceanic lithosphere increases with depth and then decreases. However, in continental lithosphere we expect such a profile in

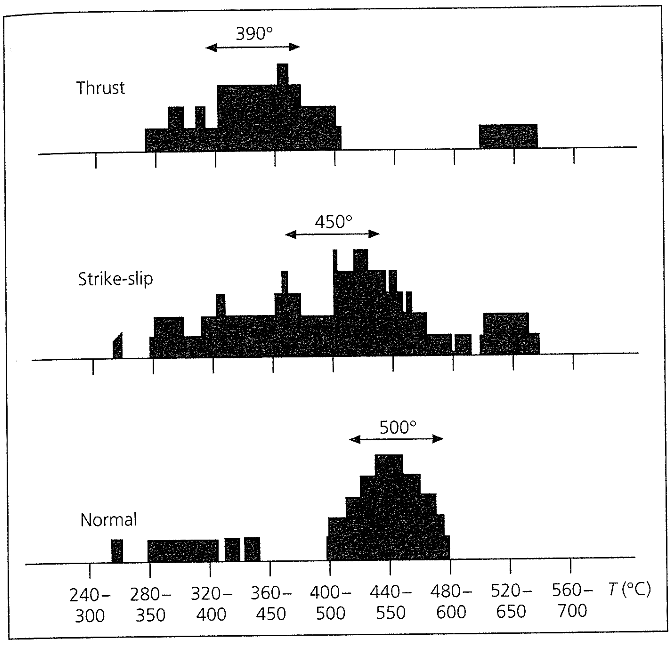


Fig. 5.7-18 Limiting temperatures for continental seismicity. These temperatures are much lower than those for oceanic lithosphere, since the quartz rheology in continents is much weaker than olivine. (Courtesy of J. Strehlau and R. Meissner.)

the quartz-rich crust, but also a second, deeper zone of strength below the Moho, due to the olivine rheology. This "jelly sandwich" profile including a weak zone may be part of the reason why continents deform differently than oceanic lithosphere. For example, some continental mountain building (Fig. 5.6-6) may involve crustal thickening in which slices of upper crust, which are too buoyant to subduct, are instead thrust atop one another. The weaker lower crust may also contribute in other ways to the general phenomenon that continental plate boundaries are broader and more complex than their oceanic counterparts (Fig. 5.2-4).

5.7.5 *Earthquakes and rock friction*

It is natural to assume that earthquakes occur when tectonic stress exceeds the rock strength, so a new fault forms or an existing one slips. Thus steady motion across a plate boundary seems likely to give rise to a cycle of successive earthquakes at regular intervals, with the same slip and stress drop (Fig. 5.7-19). However, we have seen that the earthquake process is more complicated. The time between earthquakes on plate boundaries varies (Fig. 1.2-15), although the plate motion causing the earthquakes is steady. Earthquakes sometimes rupture along the same segments of a boundary as in earlier earthquakes, and other times along a different set (Fig. 5.4-27). Moreover, many large earthquakes show a complicated rupture pattern, with some parts of the fault releasing more seismic energy than others (Fig. 4.5-10). Attempts to understand these complexities often combine two basic themes. Some of the complexity may be due to intrinsic randomness of the failure process, such that some small ruptures cascade into large earthquakes, whereas others do not (Section 1.2.6). Other aspects of the complexity may be due to features of rock friction.

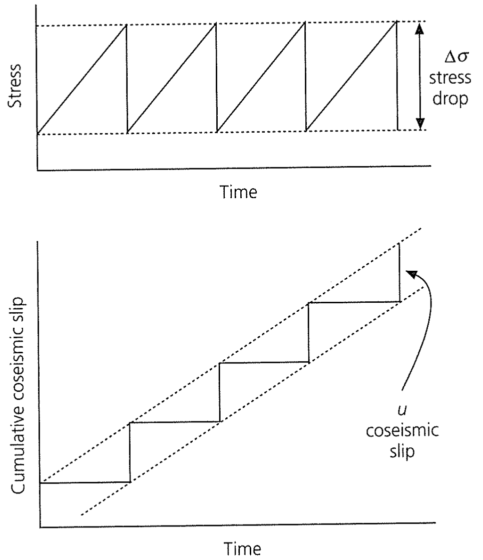


Fig. 5.7-19 Stress and slip history for an idealized earthquake cycle on a plate boundary, in which all earthquakes have the same stress drop and coseismic slip. (Shimazaki and Nakata, 1980. *Geophys. Res. Lett.*, *7*, 279–82, copyright by the American Geophysical Union.)

Interesting insight emerges from considering an experiment in which stress is applied until a rock breaks. When the fault forms, some of the stress is released, and then motion stops. If stress is reapplied, another stress drop and motion occur once the stress reaches a certain level. So long as stress is reapplied, this pattern of jerky sliding and stress release continues (Fig. 5.7-20).

This pattern, called *stick-slip*, looks like a laboratory version of what happens in a sequence of earthquakes on a fault. By this analogy, the stress drop in an earthquake relieves only part of the total tectonic stress, and as the fault continues to be loaded by tectonic stress, occasional earthquakes occur. The analogy is strengthened by the fact that at higher temperatures (about 300° for granite), stick-slip does not occur (Fig. 5.7-20). Instead, *stable sliding* occurs on the fault, much as earthquakes do not occur at depths where the temperature exceeds a certain value. Thus, understanding stick-slip in the laboratory seems likely to give insight into the earthquake process.

Stick-slip results from a familiar phenomenon: it is harder to start an object sliding against friction than to keep it going once it is sliding. This is because the *static friction* stopping the object from sliding exceeds the *dynamic friction* that opposes motion once sliding starts.[7] To understand how this difference

[7] This effect is the basis of cross-country skiing, where loading one ski makes it grip the snow, while unloading the other lets it glide.

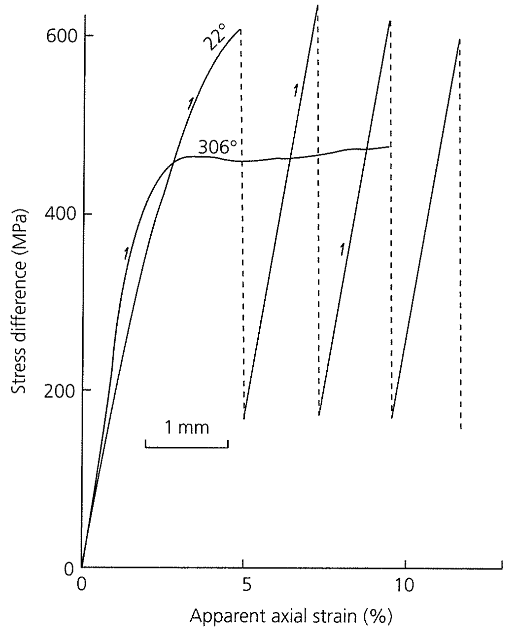


Fig. 5.7-20 Force versus slip history for a rock sample. At low temperature, so long as stress is reapplied, a stick-slip pattern of jerky sliding and stress release continues. By contrast, stable sliding occurs at high temperature. (Brace and Byerlee, 1970. *Science, 168*, 1573–5, copyright 1970 American Association for the Advancement of Science.)

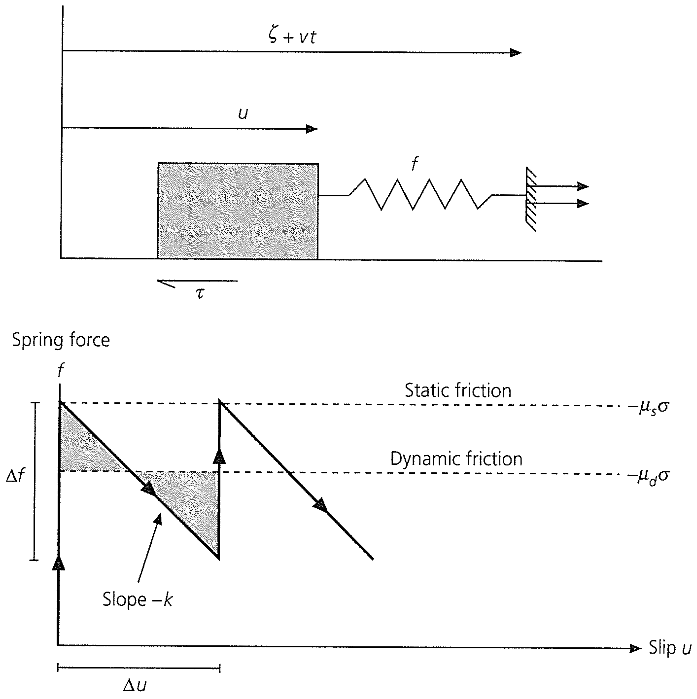


Fig. 5.7-21 A simple spring and slider block analog for stick-slip as a model for earthquakes. The slider is loaded by force f due to the spring end moving at velocity v. Before sliding, the block is retarded by a static friction force $\tau = -\mu_s\sigma$, but once sliding starts, the friction force decreases to $-\mu_d\sigma$. A series of slip events occur, each with slip Δu and force change (stress drop) Δf.

causes stick-slip, and get insight into stick-slip as a model for earthquakes, consider the experiment in Fig. 5.7-21. It turns out that if an object is pulled across a table with a rubber band, jerky stick-slip motion occurs.[8] Thus a steady load, combined with the difference in static and dynamic friction, causes an instability and a sequence of discrete slip events.

We analyze this situation assuming that a block (sometimes called a slider) is loaded by a spring that applies a force f proportional to the spring constant (stiffness) k and the spring extension. If the loading results from the spring's far end moving at a velocity v, the spring force is

$$f = k(\zeta + vt - u), \tag{41}$$

where u is the distance the block slipped, and ζ is the spring extension when sliding starts at $t = 0$. This motion is opposed by a frictional force $|\tau| = -\mu\sigma$ equal to the product of σ, the compressive (negative) normal stress due to the block's weight, and the friction coefficient, μ. By Newton's second law that force equals mass times acceleration,

$$m\frac{d^2u}{dt^2} = f - \tau = k(\zeta + vt - u) + \mu\sigma. \tag{42}$$

However, the block starts sliding only once the spring force exceeds the frictional force, so just before sliding starts at $t = 0$,

$$0 = k\zeta + \mu_s\sigma, \tag{43}$$

where μ_s is the static friction coefficient. For simplicity, assume that at the instant sliding starts, the friction drops to its dynamic value μ_d, and

$$m\frac{d^2u}{dt^2} = k(\zeta - u) + \mu_d\sigma. \tag{44}$$

Subtracting Eqn 43 from Eqn 44 gives

$$m\frac{d^2u}{dt^2} = -ku + (\mu_d - \mu_s)\sigma = -ku + \Delta\mu\sigma, \tag{45}$$

which we can use as the equation of motion for the block's slip history $u(t)$ if the loading rate v is slow enough to ignore during the slip event.

A solution to Eqn 45, with initial conditions $u(0) = 0$ and $\frac{du(0)}{dt} = 0$, is

[8] We suggest trying this experiment.

$$u(t) = \frac{\Delta\mu\sigma}{k}(1 - \cos\omega t) \quad \text{(slip)},$$

$$\frac{du(t)}{dt} = \frac{\Delta\mu\sigma}{\sqrt{km}}\sin\omega t \quad \text{(velocity)},$$

$$\frac{du^2(t)}{dt^2} = \frac{\Delta\mu\sigma}{m}\cos\omega t \quad \text{(acceleration)}, \tag{46}$$

where $\omega = \sqrt{k/m}$. As shown, the block starts slipping because the spring force exceeds the friction force. During the slip event, the spring force decreases as the spring shortens, until it becomes less than the friction force and the block slows and eventually stops. The block 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.

It is interesting to think of analogies between this model of slip events and earthquakes. The slip event's duration t_D, analogous to an earthquake rise time (Section 4.3.2), satisfies

$$\frac{du(t_D)}{dt} = 0, \quad t_D = \frac{\pi}{\omega} = \pi\sqrt{m/k}. \tag{47}$$

The total slip during the event is

$$\Delta u = u(t_D) = 2\Delta\mu\sigma/k, \tag{48}$$

and the drop in the spring force, which is analogous to an earthquake stress drop (Section 4.6.3), is

$$\Delta f = 2\Delta\mu\sigma. \tag{49}$$

Thus the rise time depends on the spring constant, but not on the difference between static and dynamic friction. However, the total slip and stress drop depend upon the friction difference. None of these depend upon the loading rate, which is analogous to the rate of plate motion causing earthquakes on a plate boundary. But the loading rate determines the time between successive slip events. Thus, in the plate boundary analogy, the time between large earthquakes depends on the plate motion rate, but their slip and stress drop depend on the frictional properties of the fault and the normal stress. Hence faster-slipping boundaries would have more frequent large earthquakes, but the slip and stress drop in them would not be greater than on a slower boundary with similar frictional properties and normal stress.

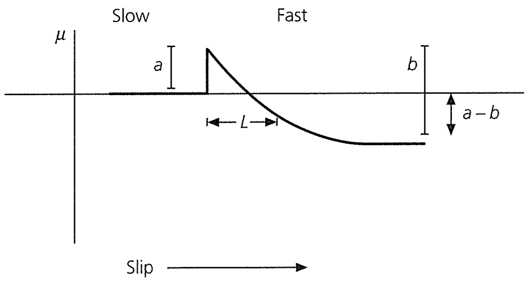


Fig. 5.7-22 Evolution of friction in a simple rate- and state-dependent model. If the slip rate increases by a factor of e, friction increases by a, and then decreases as slip progresses to a steady-state value $a - b$. (After Scholz, 1990. Reprinted with the permission of Cambridge University Press.)

Laboratory experiments show that the difference between static and dynamic friction is more complicated than the constant values 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- and state-dependent friction* with a variable coefficient of sliding friction, μ, are used to describe these effects. In a simple model of this sort,

$$\mu = [\mu_0 + b\psi + a\ln(v/v^*)], \tag{50}$$

where μ_0 is the coefficient of static friction. The friction depends on the slip rate v, normalized by a rate v^*, and a state variable ψ that represents the slip history

$$\frac{d\psi}{dt} = -(v/L)[\psi + \ln(v/v^*)], \tag{51}$$

where L is an experimentally determined characteristic distance. The friction also depends on a and b, which characterize the material.

Figure 5.7-22 illustrates how friction evolves. If the slip rate increases by a factor of e, the friction increases by a, and then decreases as slip progresses, reaching a new steady-state value. With time, ψ reaches a steady-state value given by Eqn 51,

$$0 = -(v/L)[\psi_{ss} + \ln(v/v^*)], \quad \psi_{ss} = -\ln(v/v^*). \tag{52}$$

The steady state friction (Eqn 50) is

$$\mu_{ss} = [\mu_0 + b\psi + a\ln(v/v^*)] = [\mu_0 + (a - b)\ln(v/v^*)], \tag{53}$$

and varies with slip rate as

$$\frac{d\mu_{ss}}{d\ln v} = (a - b), \tag{54}$$

so after the slip velocity change, the net friction change is $(a - b)$. If $(a - b)$ is negative, the material shows velocity weakening, which permits earthquakes to occur by stick-slip. However, for $(a - b)$ positive, the material shows velocity strengthening, and stable sliding is expected. Laboratory results (Fig. 5.7-20) show that $a - b$ for granite changes sign at about 300°, which should be the limiting temperature for earthquakes. Thus the frictional

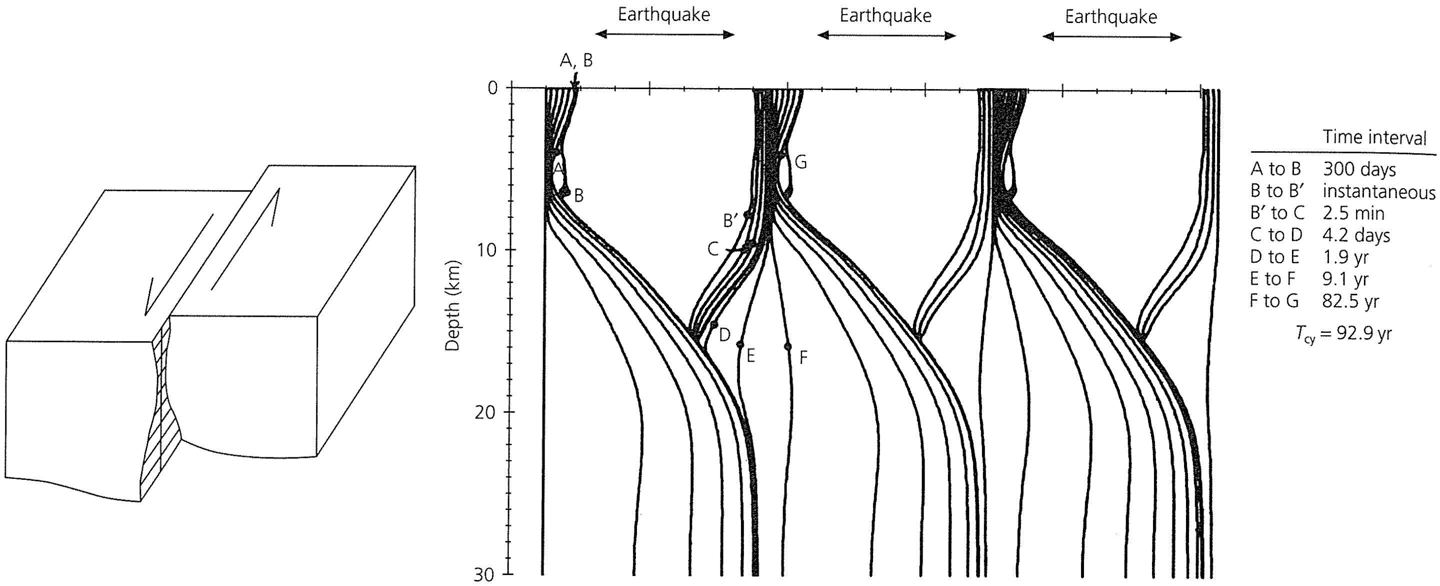


Fig. 5.7-23 Earthquake cycle for a model in which a strike-slip fault with rate- and state-dependent frictional properties is loaded by plate motion. The slip history for three cycles as a function of depth and time is shown by the lines, each of which represents a specific time. Steady motion occurs at depth, and stick-slip occurs above 11 km. (After Tse and Rice, 1986. *J. Geophys. Res., 91*, 9452–72, copyright by the American Geophysical Union.)

model predicts a maximum depth for continental earthquakes similar to that predicted by the rock strength arguments.

These results can be used to simulate the earthquake cycle, using fault models analogous to the simple slider model (Fig. 5.7-21). Figure 5.7-23 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 is described by rate- and state-dependent frictional properties as a function of depth, such that stick-slip occurs above 11 km. Initially from time A to 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 times B′ to 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 5.7-24 shows this idea schematically for the slider model in Fig. 5.7-21. Assume that after an earthquake cycle, the compressive normal stress σ on the slider is reduced. This "unclamping" reduces the frictional force resisting sliding, so it takes less time for the spring force to rise again to the level needed for the next slip event. Conversely, increased compression "clamps" the slider more, and so increases the time until the next slip event. In addition, by Eqn 49, the stress drop in the slip event changes when σ changes.

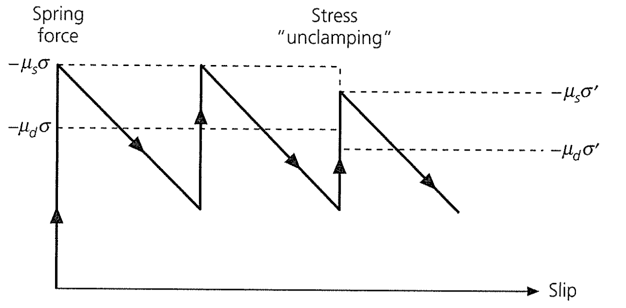


Fig. 5.7-24 Modification of a slider block model (Fig. 5.7-21) to include the effects of changes in normal stress. Reduced normal stress ($|\sigma| < |\sigma'|$) reduces the frictional force, and so "unclamps" the fault and decreases the time until the next slip event.

For earthquakes, the analogy implies that earthquake occurrence on a segment of a fault may reflect changes in the stress on the fault resulting from earthquakes elsewhere. This concept is quantified using the Coulomb–Mohr criterion (Eqn 5) that sliding can occur when the shear stress exceeds that on the sliding line (Fig. 5.7-9), or $\tau > \mu\sigma$. We can thus define the Coulomb failure stress

$$\sigma_f = \tau + \mu\sigma \qquad (55)$$

such that failure occurs when σ_f is greater than zero. Whether a nearby earthquake brings a fault closer to or further from failure is shown by the change in Coulomb failure stress due to the earthquake,

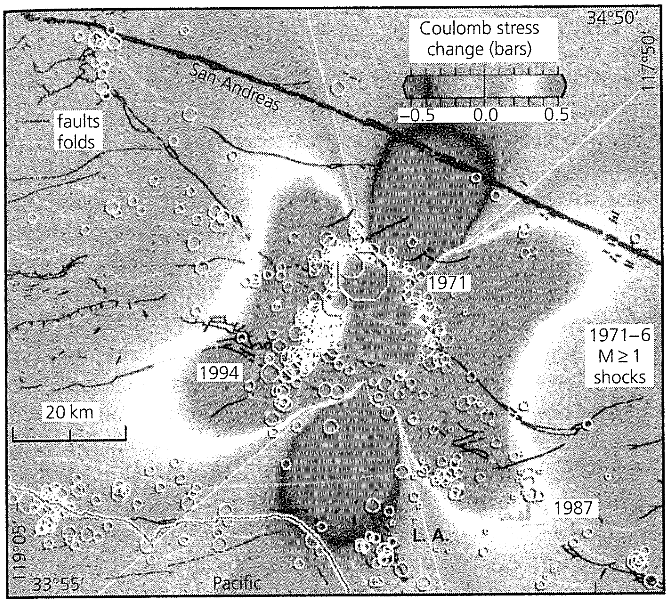


Fig. 5.7-25 Predicted changes in Coulomb failure stress due to the 1971 San Fernando earthquake. The Whittier Narrows and Northridge earthquakes subsequently occurred in regions where the 1971 earthquake increased the failure stress. (Stein *et al.*, 1994. *Science, 265*, 1432–5, copyright 1994 American Association for the Advancement of Science.)

$$\Delta\sigma_f = \Delta\tau + \mu\Delta\sigma. \qquad (56)$$

Failure is favored by positive $\Delta\sigma_f$, which can occur either from increased shear stress τ or a reduced normal stress (compression is negative, so $\Delta\sigma > 0$ favors sliding).

Some earthquake observations provide support for this idea. Figure 5.7-25 shows the predicted Coulomb failure stress changes in the Los Angeles region due to the 1971 (M_s 6.6) San Fernando earthquake. The stress change pattern reflects the earthquake's focal mechanism, thrust faulting on a NW–SE-striking fault (Fig. 5.2-3). Two moderate earthquakes, the 1987 Whittier Narrows (M_L 5.9) and 1994 Northridge (M_w 6.7) earthquakes subsequently occurred in regions where the 1971 earthquake increased the failure stress, suggesting that the stress change may have had a role in triggering the earthquakes. A similar pattern has been found after other earthquakes, and some studies have found that aftershocks are concentrated in regions where the mainshock increased the failure stress. Stress triggering may explain why successive earthquakes on a fault sometimes seem to have a coherent pattern. For example, the 1999 M_s 7.4 Izmit earthquake on the North Anatolian fault (Fig. 5.6-8) appears to be part of a sequence of major (M_s 7) earthquakes over the past 60 years, which occurred successively further to the west, and hence closer to the metropolis of Istanbul.

An intriguing feature of such models is that the predicted stress changes are of the order of 1 bar, or only 1–10% of the typical stress drops in earthquakes (Section 4.6.3). Such small stress changes should only trigger an earthquake if the tectonic stress is already close to failure. However, as in the slider model (Fig. 5.7-24), stress changes can affect the time until the tectonic stress is large enough to produce earthquakes. It has been argued that the 1906 San Francisco earthquake reduced the failure stress on other faults in the area, causing a "stress shadow" and increasing the expected time until the next earthquake on these faults. This is consistent with the observation that during the 75 years before the 1906 earthquake, the area had 14 earthquakes with M_w above 6, whereas only one occurred in the subsequent 75 years. Such analyses may help improve estimates of the probability that an earthquake of a certain size will occur on a given fault during some time period. To date, such estimates have large uncertainties (Section 4.7.3), in part because of the large variation in the time intervals between earthquakes. Stress loading models, some of which incorporate rate- and state-dependent friction because simple Coulomb friction does not predict large enough changes in recurrence time, may explain some of the variations and thus reduce these uncertainties.

This discussion brings out the importance of understanding the state of stress on faults. On this issue, the friction models give some insight, but major questions remain. Earthquake stress drops estimated from seismological observations are typically less than a few hundred bars (tens of MPa). Yet, the expected strength of the lithosphere (e.g., Fig. 5.7-14–16) is much higher, in the kilobar (hundreds of MPa) range. The laboratory results (Fig. 5.7-20) and frictional models (Fig. 5.7-21) suggest an explanation for this difference, because in both the stress drop during a slip event is only a fraction of the total stress.

However, the frictional models do not explain an intriguing problem called the "San Andreas" or "fault strength" paradox. As noted in Section 5.4.1, a fault under shear stress τ slipping at rate v should generate fractional heat at a rate equal to τv. Thus, if the shear stresses on faults are as high (kbar or hundreds of MPa) as expected from the strength envelopes, significant heat should be produced. But little if any heat flow anomaly is found across the San Andreas fault (Fig. 5.7-26), suggesting that the fault is much weaker than expected. A similar conclusion emerges from consideration of stress orientation data. Although the Coulomb–Mohr model predicts that the maximum principal stress directions inferred from focal mechanisms, geological data, and boreholes should be about 23° from the San Andreas fault (Fig. 5.7-8), the observed directions are essentially perpendicular to the fault (Fig. 5.6-19), implying that the fault acts almost like a free surface. To date, there is no generally accepted explanation for these observations. The most obvious one is that the effective stress on the fault is reduced by high pore pressure, but there is discussion about whether pressures much higher than hydrostatic pressure could be maintained in the fault zone. An alternative explanation, that the fault zone is filled by low-strength clay-rich fault gouge, faces the difficulty that experiments on such material find that it has normal strength unless pore pressures are high.

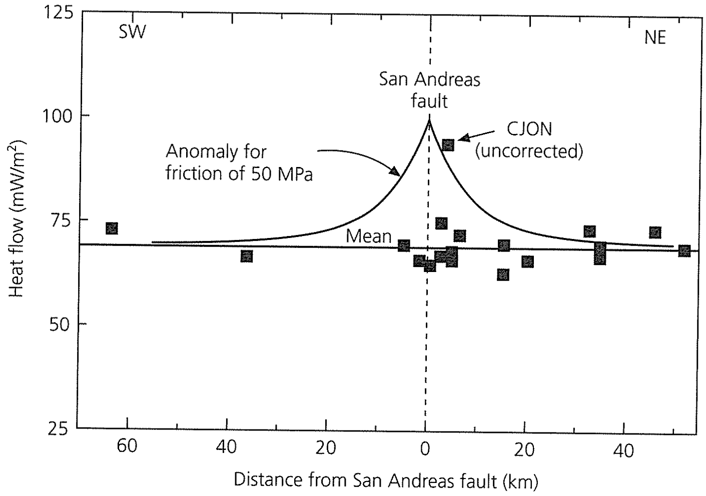


Fig. 5.7-26 Observed (squares) heat flow across the San Andreas fault. The elevated heat flow predicted by shear heating (solid line) is not observed, except for one point (CJON, Cajon pass), where alternative interpretations are possible, implying that the fault is weak. (Lachenbruch and Sass, 1988. *Geophys. Res. Lett.*, *15*, 981–4, copyright by the American Geophysical Union.)

In summary, ideas based on rock friction are providing important insights into earthquake mechanics. Although many issues remain unresolved, and some attractive notions remain to be fully demonstrated, rock friction seems likely to play a growing role in addressing earthquake issues.

5.7.6 *Earthquakes and regional deformation*

The large, rapid deformation in earthquakes is often part of a slow deformation process occurring over a broader region. As discussed in Section 5.6.2, there often appear to be differences between the seismic, aseismic, transient, and permanent deformations sampled by different techniques on different time scales. Experimental and theoretical ideas about rheology and lithospheric dynamics are being used to investigate the relation between earthquakes and the regional deformations that produce them.

We have seen that earthquakes often reflect deformation distributed over a broad plate boundary zone. In this case, we can think of the lithosphere as a viscous fluid and use earthquakes as indicators of its deformation. This idea is like the physical model (Fig. 5.6-7) that used deformable plasticine as an analogy for the deformation of Asia resulting from the Himalayan collision. Figure 5.7-27 shows such an analysis for part of the Pacific–North America plate boundary zone in the western United States. The deformation is assumed to result from a combination of forces due to the transform plate boundary and forces due to the potential energy of elevated topography, which tends to spread under its own weight. To test this idea, a continuous velocity field has been interpolated from space-geodetic, fault slip, and plate motion data (Figs 5.2-3 and 5.6-3). The velocity field is treated as being due to the motion of a viscous fluid, and is converted to a strain rate tensor field. This is then compared to the magnitude of the stress tensor inferred

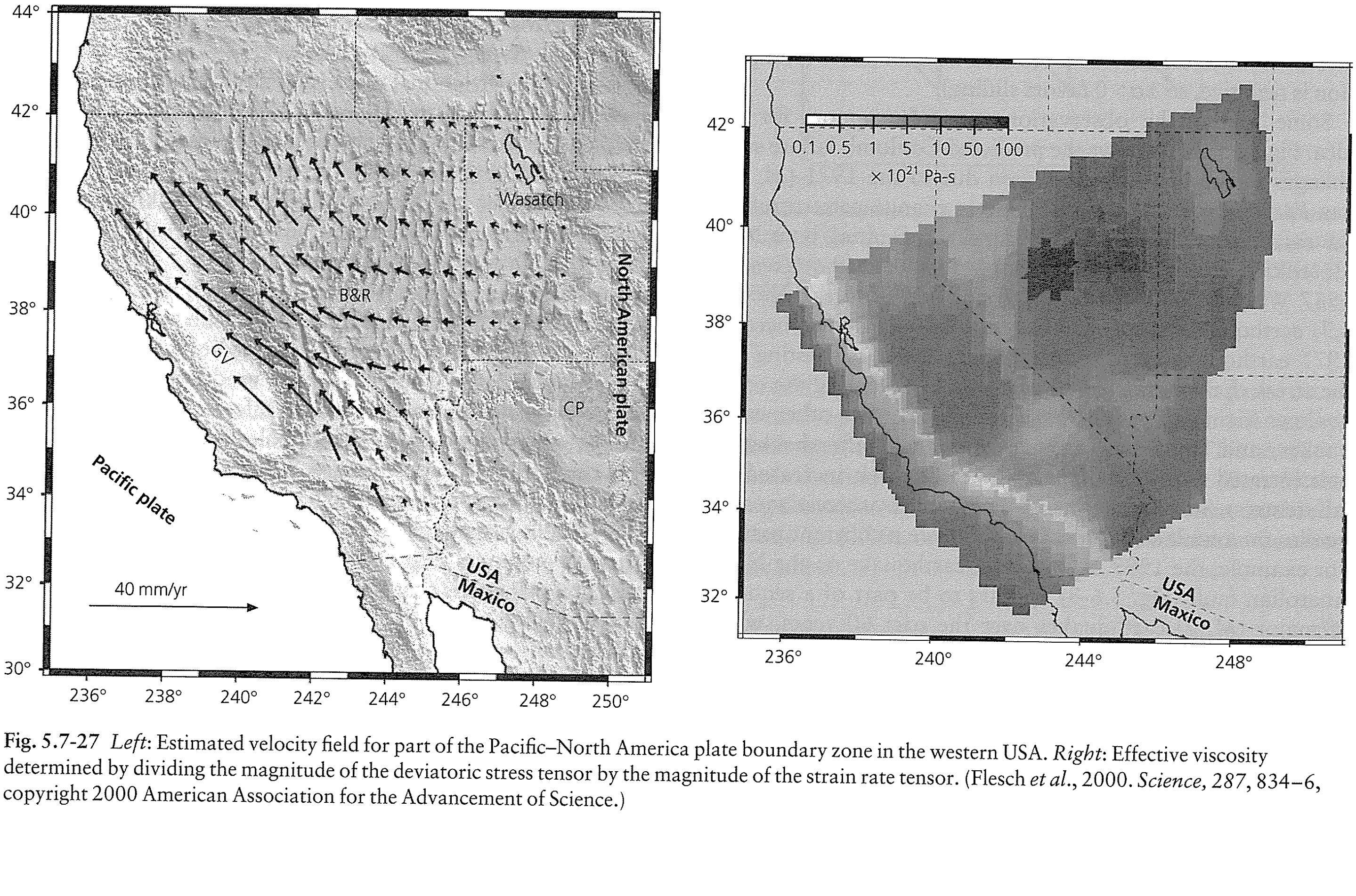


Fig. 5.7-27 *Left*: Estimated velocity field for part of the Pacific–North America plate boundary zone in the western USA. *Right*: Effective viscosity determined by dividing the magnitude of the deviatoric stress tensor by the magnitude of the strain rate tensor. (Flesch *et al.*, 2000. *Science*, *287*, 834–6, copyright 2000 American Association for the Advancement of Science.)

from topography and plate boundary forces. The ratio of stress to strain rate at any point, which is the vertically averaged effective viscosity, varies significantly. Low values along the San Andreas fault and western Great Basin show that the strain rates are relatively high for the predicted stress, consistent with a weak lower crust. The Great Valley–Sierra Nevada block has little internal deformation, and thus acts relatively rigidly and appears as a high-viscosity region. Summing seismic moment tensors (Section 5.6.2) yields a seismic strain rate averaging about 60% of the inferred total strain. As discussed earlier, this discrepancy may indicate some aseismic deformation or that the 150 years of historical seismicity is too short for a reliable estimate.

Viscous fluid models can be used to study how the lithosphere deforms on different time scales. For example, as noted in Section 5.6.2, GPS data across the entire Nazca–South America plate boundary zone show faster motion than is inferred from structural geology or topographic modeling. The difference probably occurs because the GPS data record instantaneous velocities that include both permanent deformation and elastic deformation that will be recovered during future earthquakes, whereas the lower geological rates reflect only the permanent deformation. This can be modeled by representing the overriding South American plate using a simple one-dimensional system of a spring, a dashpot, and a pair of frictional plates (Fig. 5.7-28). This system approximates the behavior of the crust: the spring gives the elastic response over short periods, the dashpot gives the viscous response over geological time scales, and the frictional plates simulate the thrust faulting earthquake cycle at the trench. As plate convergence compresses the system, the stress $\sigma(t)$ increases with time until it reaches a yield strength σ_y, when an earthquake occurs, stress drops to σ_b, and the process repeats. Displacement accumulates at a rate v_0 except during earthquakes, when the displacement drops by an amount Δu. The topography and geologic data record the averaged long-term shortening rate v_c shown by the envelope of the sawtooth curve, whereas GPS data record the higher instantaneous velocity v_0. The instantaneous velocity thus results from the portion of the plate motion locked at the trench that deforms the overriding plate elastically (Fig. 4.5-14) and is released as seismic slip in interplate earthquakes. By contrast, the aseismic slip component at the trench has no effect because it does not contribute to locking on the interface and deformation of the overriding plate. Similar models are being explored for other regions where deformation appears to vary on different time scales.

Viscous fluid models are also used to analyze other aspects of the earthquake cycle. For example, Fig. 5.7-29 shows the strain rate near portions of the San Andreas fault compared to the time since the last great earthquake on that portion of the fault. Postseismic motion seems to continue for a period of years after an earthquake and then slowly decays, presumably due to the steady interseismic motion. A similar picture emerges from GPS and other geodetic results following large trench thrust faulting earthquakes. For a number of years, sites near the trench on the overriding plate move seaward, showing postseismic motion consistent with the earthquake focal mechanism (Fig. 4.5-15). Eventually, however, the sites resume the landward interseismic motion usually seen at trenches (Fig. 5.6-10). Such observations are challenging to interpret, because postseismic afterslip on or near a fault can have effects at the surface similar to viscoelastic flow of the asthenosphere (Fig. 5.7-29), but offer the prospect of improving our understanding of both earthquake processes and the rheology of the lithosphere and the asthenosphere. A tantalizing possibility is that the viscous asthenosphere permits stress waves generated by large earthquakes to travel slowly for large distances and contribute to earthquake triggering.

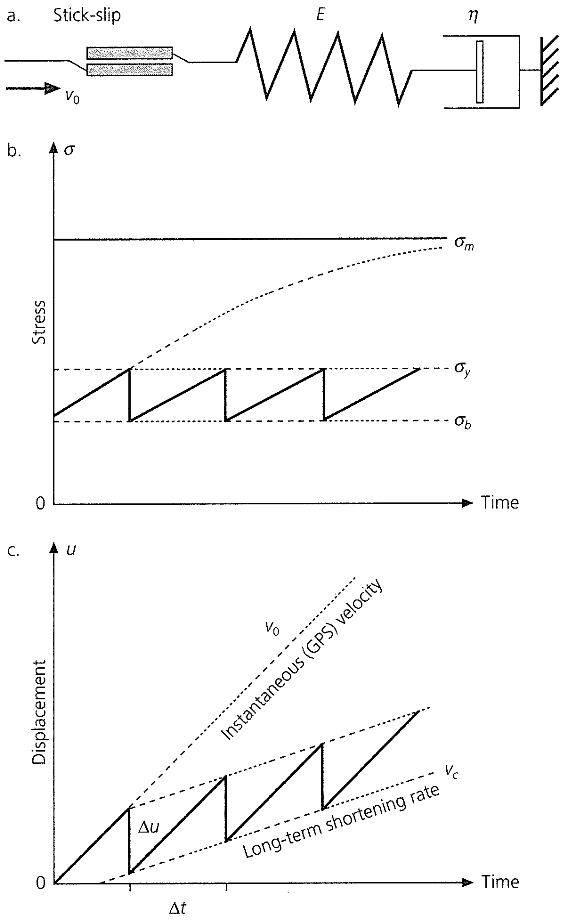


Fig. 5.7-28 a: Model for a viscoelastic–plastic crust to describe the response of the overriding South American plate to the subduction of the Nazca plate. The dashpot represents a viscous body modeling the permanent deformation, the spring represents an elastic body modeling the transient deformation, and the frictional plates represent the earthquake cycle at the trench. b: Stress evolution for the model. c: Displacement history for the model. Displacement accumulates at the instantaneous rate v_0 except during earthquakes, when slip Δu occurs. GPS data record a gradient starting at v_0 from the trench, whereas the envelope of the displacement curve v_c is the long-term shortening rate reflected in geological records and topography. (Liu *et al.*, 2000. *Geophys. Res. Lett., 18*, 3005–8, copyright by the American Geophysical Union.)

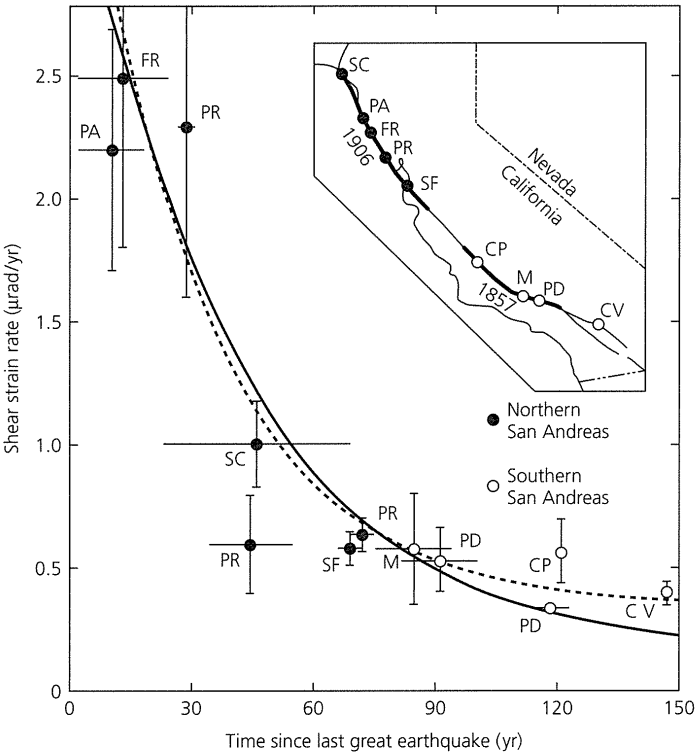


Fig. 5.7-29 Shear strain rate near portions of the San Andreas fault compared to the time since the last great earthquake. The data are similar to the predictions of two alternative models: viscoelastic stress relaxation (solid curve) and aseismic postseismic slip beneath the earthquake fault plane (dashed line). (Thatcher, 1983. *J. Geophys. Res., 88*, 5893–902, copyright by the American Geophysical Union.)

Further reading

Given the comparatively recent discovery of plate tectonics, its importance for most aspects of geology, and its crucial role in the earthquake process, many excellent sources, a few of which are listed here, offer more information about this chapter's topics.

The dramatic development of plate tectonics is discussed from the view of participants by Menard (1986) and in Cox's (1973) collection of classic papers. Basic ideas in plate tectonics are treated in most introductory and structural geology texts. More detailed treatments include Uyeda (1978), Fowler (1990), Kearey and Vine (1990), and Moores and Twiss (1995). Cox and Hart (1986) present the basic kinematic concepts, and global plate motion models are discussed by Chase (1978), Minster and Jordan (1978), and DeMets *et al.* (1990).

Thermal and mechanical aspects of plate tectonics are discussed by Turcotte and Schubert (1982) and Sleep and Fujita (1997). Mid-ocean ridge tectonics and structure are discussed by Solomon and Toomey (1992) and Nicolas (1995). The thermal evolution of oceanic lithosphere is discussed by Parsons and Sclater (1977) and Stein and Stein (1992); McKenzie (1969) presents the subduction zone thermal model we follow. Papers in Bebout *et al.* (1996) cover many aspects of subduction, and Kanamori (1986) reviews subduction zone thrust earthquakes. Lay (1994) treats the nature and fate of subducting slabs, and deep earthquakes are reviewed by Frohlich (1989), Green and Houston (1995), and Kirby *et al.* (1996b). For a derivation of the ridge push force see Parsons and Richter (1980); Wiens and Stein (1985) discuss its application to oceanic intraplate stresses. Yeats *et al.* (1997) cover a wide variety of topics about the relation of earthquakes to regional geology. Rosendahl (1987) reviews continental rifting. Papers in Gregersen and Basham (1989) treat aspects of passive margin and continental interior earthquakes with emphasis on postglacial effects.

Concepts in continental deformation are treated by Molnar (1988) and England and Jackson (1989); Gordon (1998) gives an overview of plate rigidity and diffuse plate boundaries. Applications of space geodesy to tectonics are reviewed by papers in Smith and Turcotte (1993) and by Dixon (1991), Gordon and Stein (1992), and Segall and Davis (1997). Many GPS data and results, including an overview brochure, can be found on the University NAVSTAR Consortium WWW site http://www.unavco.org. Stress maps and their interpretations are discussed by Zoback (1992) and other papers in the same journal issue; stress maps are available at the World Stress Map project WWW site http://www-wsm.physik.uni-karlsruhe.de.

Mantle plumes in general are reviewed by Sleep (1992); Nataf (2000) and Foulger *et al.* (2001) discuss seismic imaging of plumes; Smith and Braile (1994) discuss the Yellowstone hot spot; and Stein and Stein (1993) discuss oceanic hot spot swells. Papers in Peltier (1989) treat many aspects of mantle convection; Silver *et al.* (1988) explore the relationship between subduction, convection, and mantle structure; and Christensen (1995) reviews the effects of phase transitions on mantle convection. The heat engine perspective on global tectonics is discussed by Stacey (1992), and Ward and Brownlee (2000) summarize the arguments advocating a crucial role for plate tectonics in the origin and survival of life on Earth.

Topics involving rock mechanics, flow, and their tectonic applications are discussed by Jaeger (1970), Weertman and Weertman (1975), Jaeger and Cook (1976), Turcotte and Schubert (1982), Kirby (1983), Kirby and Kronenberg (1987), and Ranalli (1987). Scholz (1990) and Marone (1998) cover topics dealing with the relation of rock mechanics to earthquakes, with special emphasis on rock friction. Our treatment of the slider model for faulting follows Scholz (1990). Related topics, including issues of continental deformation and fault strength, are also treated by papers in Evans and Wong (1992). Stein (1999) summarizes the concept of stress triggering of earthquakes.

Problems

1. Assume that Pacific–North America plate motion along the San Andreas fault occurs at 35 mm/yr.
   - (a) If all this motion occurs seismically in earthquakes about 22 years apart, which is a typical recurrence interval for the Parkfield fault segment, how much slip would you expect in the earthquakes? From Fig. 4.6-7, estimate likely fault lengths and magnitudes for such earthquakes.
   - (b) Give similar estimates if the earthquakes occur about 132 years apart, as at Pallett Creek.
2. Assume that all the earthquakes in the Pallett Creek sequence (Fig. 1.2-15) involved 4 m of seismic slip. Using the time interval from the present to the 1857 earthquake, calculate the seismic slip rate on this portion of the San Andreas fault. Next, do so by averaging the recurrence intervals for the past two earthquakes (1857 and 1812), the past three, and so on for the entire earthquake history. What are the implications of this simple experiment for seismic slip estimates? What other sources of uncertainty should also be considered, and how might they affect this estimate?
3. (a) Use Table 5.2-1 to find the rate that the Juan de Fuca plate subducts beneath North America at 46°N, 125°W.
   - (b) If all this motion occurs in large earthquakes, how often would you expect an earthquake if the slip in each were 5 m? How would this estimate change if the slip were 10 or 20 m?
   - (c) How would the answers to (b) change if only 25% or 50% of the plate motion occurred by seismic slip?
   - (d) Paleoseismic observations and historic records of a tsunami imply that this subduction zone has had very large earthquakes approximately 500 years apart. Suggest some possibilities in view of parts (a)–(c). How might you attempt to distinguish between them?
   - (e) The crust subducting at this trench is about 10 million years old. Given the convergence rate and the observations from other trenches in Fig. 5.4-30, what might you infer about the moment magnitude of the largest earthquake expected here? Find the corresponding seismic moment and suggest a plausible fault geometry and amount of slip that would also be consistent with the paleoseismic and plate motion observations.
4. For rigid plates, Eqn 5.2.10 shows that we can find the *angular* velocity vector of one plate from the sum of two others. Show that at a point we can also do this for the *linear* velocity vectors.
5. The news media sometimes ask "How large would the largest possible earthquake be?" Estimate the seismic moment and moment magnitude by assuming that all the trenches in the world (48,000 km) slip at the same time, that 10 m of slip occurs, and the fault width is 250 km.
6. Estimate the thermal Reynolds number R defined in Eqns 5.3.19 and 5.4.3, assuming that $\kappa = 10^{-6}$ m^{2}s^{-1}. What does this estimate imply about the processes of plate cooling and subduction?
7. Assume that oceanic lithosphere has a thermal conductivity of 3.1 Wm^{-1}°C^{-1}.
   - (a) Find the heat flow for old oceanic lithosphere, assuming a linear temperature gradient (Fig. 5.3-8), a basal temperature of 1450°C, and a plate thickness of 95 km.
   - (b) How would this value change for a basal temperature of 1350°C and plate thickness 125 km?
   - (c) If the lithosphere under a midplate region were thinned to 50 km while the basal temperature remained 1350°, what would the heat flow be, assuming a linear temperature gradient?
8. A way to get insight into the physics of subduction is to use a classic result from fluid mechanics, called Stokes' problem, which describes the terminal velocity v at which a sphere of radius a and density ρ sinks due to gravity in a fluid with viscosity η and lower density ρ'. The result is $v = 2ga^2(\rho - \rho')/9\eta$. Estimate the subduction velocity of a slab assuming the slab is a sphere with radius equal to half its thickness. To do this, estimate the density contrast from the thermal model (Eqn 5.4.14) and a coefficient of thermal expansion $\alpha = 3 \times 10^{-5}$ °C^{-1}. Use a mantle viscosity from Section 5.5.3. Because this is a back-of-the-envelope calculation, there is no correct answer, but you should be able to come up with something reasonable (within an order of magnitude or two of reality).
9. The result that a subducting slab that reaches the core should still be thermally distinct (Fig. 5.4-5) may seem surprising. For another estimate, use the one-dimensional cooling equation in Section 5.3.2 to estimate how long a slab should need to warm up to 90% of the ambient lowermost mantle temperatures, assuming that it were immediately transported to the base of the mantle and that $\kappa = 10^{-6}$ m^{2}s^{-1}.
10. Using the definition of the slab pull force (Eqn 5.4.15):
    - (a) Write the force in terms of the age of the subducting plate.
    - (b) Explain whether this force would be greater or smaller, and why, for increased values of subducting plate age, coefficient of thermal expansion, and thermal diffusivity.
11. Assume that in a subducted slab the depth of the spinel–perovskite phase transition deepens from its usual 660 km outside the slab to 700 km, and that the core of the slab is 800° colder than the surrounding mantle. What is the Clapeyron slope of the phase change?
12. The surface of Venus is much hotter (450°C) than that of Earth. If Venus had plate tectonics and the rocks were similar, so that the temperature gradient in old lithosphere there were the same as on Earth, how would the thickness of the "oceanic" lithosphere differ? How would the slab pull and ridge push forces differ? What other differences might you expect?
13. Express the ratio of the slab pull (Eqn 5.4.15) and ridge push (Eqn 5.5.6) forces. Explain why this ratio depends on thermal diffusivity. Estimate this ratio near a trench where old oceanic lithosphere is subducting, assuming that $\kappa = 10^{-6}$ m^{2}s^{-1}.
14. To see if momentum can be responsible for the Indian plate's northward motion long after its collision with Asia began, estimate the momentum of the Indian plate and that of an ocean liner, and compare the two.
15. Use Mohr's circle to show why
    - (a) Rocks at depth do not fracture under lithostatic pressure alone.
    - (b) The deviatoric stress needed for fracture increases at greater depth.
16. Suppose that a rock is stressed close to its brittle limit. Show graphically which will make the rock fracture sooner: (a) increasing σ_1 or (b) decreasing σ_2 by the same amount (assume a two-dimensional case where σ_1 and σ_2 are both negative, and internal friction exists).
17. Suppose that the fracture line for a particular rock is $\tau = 80 - 0.5\sigma$, where stresses are in MPa. What angle would the normal to a fracture plane make with σ_1? If σ_1 is 400 MPa at failure, what is σ_2?
18. For the slider block earthquake model in Section 5.7.5:
    - (a) Derive an expression for the time between successive slip events.

(b) Sketch the force–slip diagram for two different spring constants, and use the sketch to explain how the slip and force drop in a slip event change and why.

(c) For the slider block model, formulate a quantity analogous to an earthquake's seismic moment, and explain why it depends on each term. What is the major difference between this quantity and the seismic moment?

(d) Recall the observation (Fig. 4.6-11) that earthquake stress drops are similar for a wide range of earthquakes. If the slider block model is relevant, what does this imply?

(e) What conditions might correspond to aseismic slip, which could be viewed as the limit of a continuous series of very small slip events?

Computer problems

C-1. (a) Write a subroutine to compute the rate and azimuth of plate motion at a point, given the location and an Euler vector in the form (pole latitude, longitude, magnitude).

(b) Use the Euler vector in Table 5.2-1 to test your program on the San Andreas and Aleutian site examples in Section 5.2.1.

C-2. (a) Find the rate and azimuth of Cocos–North America plate motion at 18.3°N, 102.5°W.

(b) This location is the epicenter of a large 1985 Mexican earthquake, whose mechanism had nodal planes whose strike and dip are (127°, 81°) and (288°, 9°). Infer from the tectonics of the Middle American trench which plane was the fault plane. Using the methods of Section 4.2, determine the azimuth of slip during the earthquake. How does this compare to your predicted azimuth?

C-3. (a) Write a subroutine to add and subtract two Euler vectors given in the form (pole latitude, longitude, magnitude). The output should be a Euler vector in the same form.

(b) Use your program to determine the absolute Euler vector for the Pacific plate using Table 5.2-1.

(c) Determine the rate and azimuth of absolute plate motion at Hawaii (Fig. 5.2-7). Compare the direction to the Hawaiian–Emperor seamount chain.

C-4. Write a program to plot the temperature distribution in the oceanic lithosphere as a function of age using the cooling half-space thermal model (Eqn 5.3.4). Compute erf(s) (Eqn 5.3.3) using either available software or numerical integration as discussed in problem 4C-6.

C-5. (a) Write a program to plot the temperature distribution in a subducting slab using the analytic thermal model (Eqn 5.4.3). Compute it for a plate subducting at 80 mm/yr at an angle of 45°. Make assumptions that seem reasonable and justify them.

(b) Change the program to make the age of the subducting plate a parameter and generate temperature fields for different slabs, as in Fig. 5.4-6.

(c) Using the results of (b) and Fig. 5.4-4, estimate a temperature above which deep earthquakes are not observed.

6 Seismograms as Signals

We shall introduce the concepts of signal and noise. We define the signal as the desired part of the data and the noise as the unwanted part. Our definition of signal and noise is subjective in the sense that a given part of the data is "signal" for those who know how to analyze and interpret the data, but it is "noise" for those who do not. For example, for many years the times of the first arrivals of P- and S-waves were the only signals conveyed by an earthquake, and the rest of the seismogram, such as surface waves and coda waves, had to be considered as useless until appropriate methods of interpretations were found.

Thus, through the application of a new technique to old data, an analyst can experience a moment of discovery as joyful as a data gatherer does using a new observational device.

Aki and Richards, *Quantitative Seismology*, 1980

6.1 Introduction

Seismology uses various techniques to study the displacement field as a function of position and time associated with elastic waves in the earth, and to draw inferences from it about the nature of seismic sources and the earth. Although some techniques depend on specific aspects of seismic waves in the earth, others rely on general properties of functions of space and time.

We thus often use a class of techniques known as *signal processing* or *time series analysis*. Signal processing considers functions of time or space, also called series or signals, in general terms without regard to the specific physics involved. As a result, many wave propagation subjects, including seismology, radar, sonar, and optics, can be treated in similar ways. The signals can have different forms. For example, in seismology, we can treat either a continuous (*analog*) record of ground motion or the *digital* data that result from representing the ground motion as being *sampled* at discrete intervals, providing numbers that can be manipulated using a computer.

In general terms, we can think of *filtering* a signal, or applying some operation that modifies the signal. We have already discussed several examples. A seismometer is a filter, in that it yields a record of ground motion that differs from the actual ground motion. Similarly, processes in the earth such as dispersion or attenuation have effects that can be described as a filter acting on the wave field. We can also consciously apply filters to enhance parts of a seismogram or seismic wave field and suppress others. In this chapter we extend these ideas by considering mathematical approaches that are common to such applications and then seeing how these approaches give additional insight into the physical processes. We discuss some basic concepts and provide references at the end of the chapter for more extensive treatments.

6.2 Fourier analysis

6.2.1 Fourier series

In many applications, we use an approach based on the idea that any time series can be decomposed into the sum or integral of harmonic waves of different frequencies, using methods known as *Fourier analysis*. We derived the properties of seismic waves using a harmonic wave, a sinusoid of a single frequency, and noted that any wave could be treated as the sum of harmonic waves. Thus we showed that waves on a string could be viewed as the sum of the string's normal modes, or standing waves (Section 2.2.5), and that waves in a spherical earth can be written as the sum of the earth's normal modes (Section 2.9). This concept is especially useful when the components with various frequencies behave differently. For example, surface waves of different frequencies have different apparent velocities (Section 2.8) and seismic wave attenuation varies with frequency (Section 3.7). Similarly, we will see shortly that seismometers respond differently to ground motion of different frequencies. Fourier analysis lets us decompose the signal into harmonic waves, consider each harmonic wave separately, and then recombine the harmonic waves. Thus we use this approach to analyze situations where the effect of the earth or a seismometer can be described by a filter. We also use Fourier

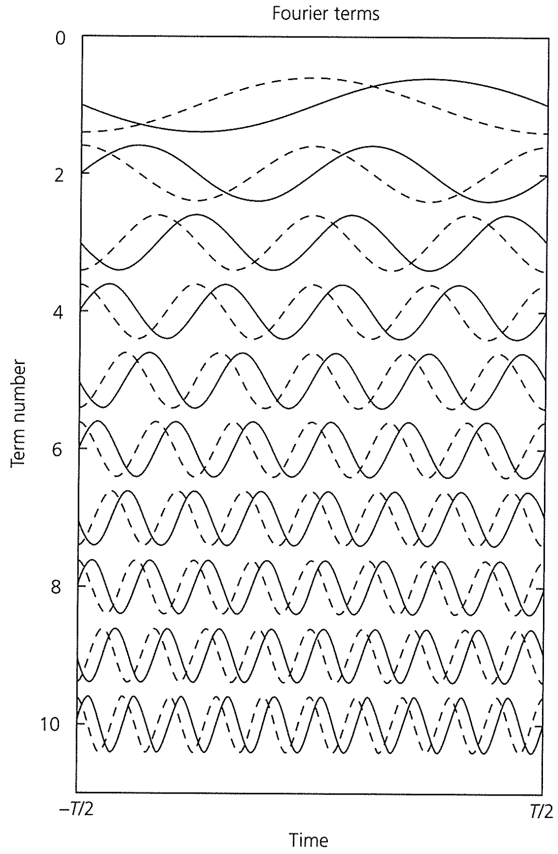


Fig. 6.2-1 Successive terms of a Fourier series. Solid lines are $\sin(2n\pi t/T)$; dashed lines are $\cos(2n\pi t/T)$.

analysis to filter a signal when the part that interests us overlaps with a part that does not in the time or space domains, but the two can be separated in the frequency or wavenumber domains.

We first consider the decomposition of a signal with a finite duration into a *Fourier series*, or sum of harmonic components with different frequencies. We will see later that as the duration of the signal becomes infinite, the Fourier series becomes the *Fourier transform* integral.

The Fourier series for an arbitrary function of time $f(t)$ defined over the interval $-T/2 < t < T/2$ is

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi t}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi t}{T}\right). \tag{1}$$

This series decomposes $f(t)$ into a sum of Fourier terms that are sine and cosine functions with different periods, because $\sin(2n\pi t/T)$ and $\cos(2n\pi t/T)$ are periodic with period T/n, or frequency n/T (Fig. 6.2-1). Larger values of n correspond to shorter periods, or higher frequencies. For $n = 0$, the cosine term equals 1 for all values of t, and there is no sine term, because it would be zero.

The sine and cosine Fourier terms are a set of *orthogonal functions*, which means that the integral of the product of two different ones over the interval from $-T/2$ to $T/2$ is always zero:

$$\int_{-T/2}^{T/2} \sin\left(\frac{2m\pi t}{T}\right) \sin\left(\frac{2n\pi t}{T}\right) dt = \frac{T}{2}\,\delta_{mn}(1 - \delta_{m0}), \tag{2}$$

$$\int_{-T/2}^{T/2} \cos\left(\frac{2m\pi t}{T}\right) \cos\left(\frac{2n\pi t}{T}\right) dt = \frac{T}{2}\,\delta_{mn}(1 + \delta_{m0}), \tag{3}$$

$$\int_{-T/2}^{T/2} \cos\left(\frac{2n\pi t}{T}\right) \sin\left(\frac{2m\pi t}{T}\right) dt = 0 \quad \text{for all } m, n, \tag{4}$$

where the Kronecker delta, δ_{mn}, equals 1 for $m = n$ and 0 otherwise (Eqn A.3.37). For the special case $m = n = 0$, the integral in Eqn 2 is zero, and the integral in Eqn 3 is twice the value for any other $m = n$.[1]

To express the Fourier series for a given function, we solve for the coefficients a_n and b_n by multiplying both sides of Eqn 1 by the appropriate sine or cosine term and integrating from $-T/2$ to $T/2$. For example, to find the coefficient a_k, where k is some particular integer, we multiply by $\cos(2k\pi t/T)$ and integrate to get

$$\int_{-T/2}^{T/2} \cos\left(\frac{2k\pi t}{T}\right) f(t)\, dt =$$

$$\int_{-T/2}^{T/2} \cos\left(\frac{2k\pi t}{T}\right) \left[a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi t}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi t}{T}\right)\right] dt. \tag{5}$$

By the orthogonality relations (Eqns 2–4), the only term in the sums on the right-hand side whose contribution to the integral is nonzero is $\cos(2\pi kt/T)$, so the equation simplifies to

$$\int_{-T/2}^{T/2} \cos\left(\frac{2k\pi t}{T}\right) f(t)\, dt = a_k \int_{-T/2}^{T/2} \cos^2\left(\frac{2k\pi t}{T}\right) dt = \frac{T}{2} a_k (1 + \delta_{k0}), \tag{6}$$

which shows that the coefficient a_k is

$$a_k = \frac{2 - \delta_{k0}}{T} \int_{-T/2}^{T/2} \cos\left(\frac{2k\pi t}{T}\right) f(t)\, dt. \tag{7}$$

[1] The proofs of Eqns 2–4 are left for the problems.

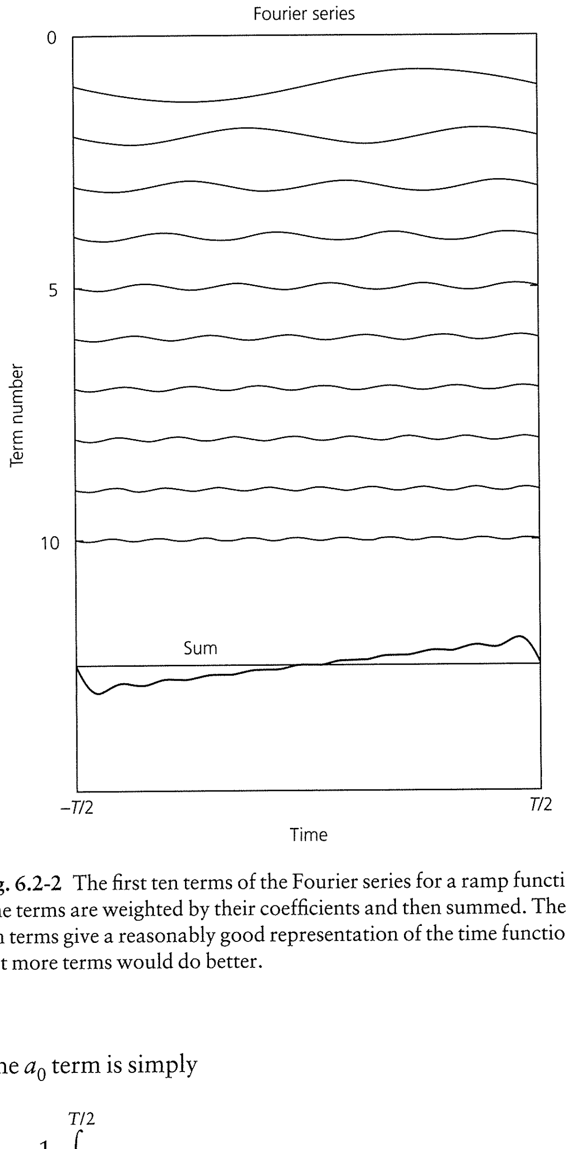


Fig. 6.2-2 The first ten terms of the Fourier series for a ramp function. The terms are weighted by their coefficients and then summed. The first ten terms give a reasonably good representation of the time function, but more terms would do better.

The a_0 term is simply

$$a_0 = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,dt, \tag{8}$$

which corresponds to the average value of the function. The coefficients of the sine terms are found similarly by

$$b_k = \frac{2}{T}\int_{-T/2}^{T/2} \sin\left(\frac{2k\pi t}{T}\right) f(t)\,dt. \tag{9}$$

Mathematically, what we have done is to consider the function $f(t)$ as being in a vector space whose basis vectors (Section A.3.6) are the sine and cosine Fourier terms. The coefficients a_k and b_k are the components that describe the particular vector $f(t)$. Thus, multiplying each basis function by the appropriate coefficient and then summing yields the function. Similarly, the operation of finding the coefficients using the integrals in Eqns 7–9 corresponds to finding each component of a vector by taking the scalar product with the appropriate unit basis vector (Eqn A.3.27).

Figure 6.2-2 illustrates this idea for a ramp function $f(t) = t/T$. Performing the integrations in Eqns 7–9 gives $a_k = 0$ and $b_k = (-1)^{k+1}/k\pi$. The cosine terms are zero, because the function is odd ($f(t) = -f(-t)$), whereas cosine is an even function ($f(t) = f(-t)$). Conversely, if the function were even, the Fourier series would include only cosine terms. Adding up the first ten sine terms reproduces the ramp reasonably well. If more terms were used, the ramp would be reproduced even better. The terms with small k are longer-period functions, and so describe the long-period features of the time series, whereas those with larger k reproduce the shorter-period features.

We used the Fourier series to express waves on a string as the sum of the string's normal modes (Section 2.2.5). Each normal mode has a spatial eigenfunction, which is a Fourier term, and an eigenfrequency. The amplitude of each Fourier term depends on the source that generated the waves, so different waves are represented by differently weighted sums of the Fourier terms. For the string the Fourier series described the variation of a function in space along a finite string, whereas here we use it to describe the variation of a function of time over a finite period. Because waves are functions of both time and space, Fourier analysis can be used for either variable or both. Fourier series are also used in other geophysical applications to represent functions that vary in space or time over finite domains. For example, we used Fourier series to describe the temperature fields in cooling oceanic lithosphere (Eqn 5.3.19) and in subducting plates (Eqn 5.4.3).

6.2.2 *Complex Fourier series*

The Fourier series (Eqn 1) can be written in a simpler form. First, we use the angular frequencies $\omega_n = 2n\pi/T$, expand the sine and cosine functions into complex exponentials, and regroup terms as

$$f(t) = a_0 + \frac{1}{2}\sum_{n=1}^{\infty}[(a_n - ib_n)e^{i\omega_n t} + (a_n + ib_n)e^{-i\omega_n t}]. \tag{10}$$

Then we use the definitions of the coefficients in Eqns 7–9, again expanding the sine and cosine functions into complex exponentials:

$$(a_n - ib_n)/2 = \frac{1}{T}\int_{-T/2}^{T/2} [\cos\omega_n t - i\sin\omega_n t] f(t)\,dt$$

$$= \frac{1}{T}\int_{-T/2}^{T/2} e^{-i\omega_n t} f(t)\,dt$$

$$(a_n + ib_n)/2 = \frac{1}{T}\int_{-T/2}^{T/2} [\cos \omega_n t + i \sin \omega_n t] f(t)\, dt$$

$$= \frac{1}{T}\int_{-T/2}^{T/2} e^{i\omega_n t} f(t)\, dt. \quad (11)$$

Next, we define

$$F_n = (a_n - ib_n)/2, \quad F_0 = a_0, \quad \text{and} \quad F_{-n} = (a_n + ib_n)/2, \quad (12)$$

so that the Fourier series becomes

$$f(t) = F_0 + \sum_{n=1}^{\infty} F_n e^{i\omega_n t} + \sum_{n=1}^{\infty} F_{-n} e^{-i\omega_n t}. \quad (13)$$

Because $-\omega_n = -2n\pi/T = \omega_{-n}$ and F_{-n} is the complex conjugate of F_n, $(F_{-n} = F_n^*)$, the negative exponentials can be written

$$\sum_{n=1}^{\infty} F_{-n} e^{-i\omega_n t} = \sum_{n=-1}^{-\infty} F_n e^{i\omega_n t}. \quad (14)$$

Making these substitutions in Eqn 10 yields the Fourier series in complex number form:

$$f(t) = \sum_{n=-\infty}^{\infty} F_n e^{i\omega_n t}, \quad (15)$$

with components

$$F_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t) e^{-i\omega_n t} dt. \quad (16)$$

6.2.3 *Fourier transforms*

The complex Fourier series, which represents a function of time in terms of a sum over discrete angular frequencies ω_n, can be extended into the *Fourier transform* that represents the function as an integral over a continuous range of angular frequencies. Thus, although we used the Fourier series to describe the discrete normal modes of a finite string and the earth, we use the Fourier transform in most seismological applications, because we regard the waves as continuous functions of angular frequency.

To do this, we write Eqn 15 as

$$f(t) = \sum_{n=-\infty}^{\infty} F_n e^{i\omega_n t} \Delta n \quad (17)$$

(because $\Delta n = 1$), and define the difference between the successive angular frequencies

$$\Delta\omega = (2\pi/T)\Delta n \quad (18)$$

so that

$$\Delta n = (T\Delta\omega)/(2\pi) \quad (19)$$

and

$$f(t) = \sum_{n=-\infty}^{\infty} F_n (T/2\pi) e^{i\omega_n t} \Delta\omega. \quad (20)$$

Next, we let the period T over which $f(t)$ is defined go to infinity, so that the angular frequencies ω_n become close enough that the discrete ω_n can be replaced by the continuous variable ω. As a result, $\Delta\omega$ becomes $d\omega$, and the sum becomes an integral. We assert (note the difference between seismology and mathematics texts) that this can be done such that the product TF_n remains finite and can be replaced by the continuous function of angular frequency $F(\omega)$. The Fourier series (Eqn 20) becomes the integral

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega, \quad (21)$$

and the expression for the coefficients (Eqn 16) becomes

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt. \quad (22)$$

Equation 22 is called the *Fourier transform*, and Eqn 21 is the *inverse Fourier transform*. These can be defined in alternate ways by interchanging the signs on the exponentials and placing the $1/2\pi$ before either integral.

It may seem strange that by starting with a real function of time $f(t)$ we obtain the transform $F(\omega)$, which is a complex function of angular frequency. The idea of negative angular frequencies may also seem disturbing. In a sense the two offset each other – we obtain a real time function by integrating a complex transform over both positive and negative angular frequencies.

An important feature of the transform and inverse transform is that their dimensions are different. For example, if $f(t)$ is a seismogram that has the dimensions of displacement, its transform $F(\omega)$ has the dimensions of displacement multiplied by time (from the dt term). Thus, if $f(t)$ gives ground motion in centimeters, $F(\omega)$ gives the transform of ground motion in centimeter-seconds.

The Fourier transform, a complex-valued function of angular frequency, can be written in terms of two real-valued functions of angular frequency:

$$F(\omega) = |F(\omega)|\, e^{i\phi(\omega)}, \quad (23)$$

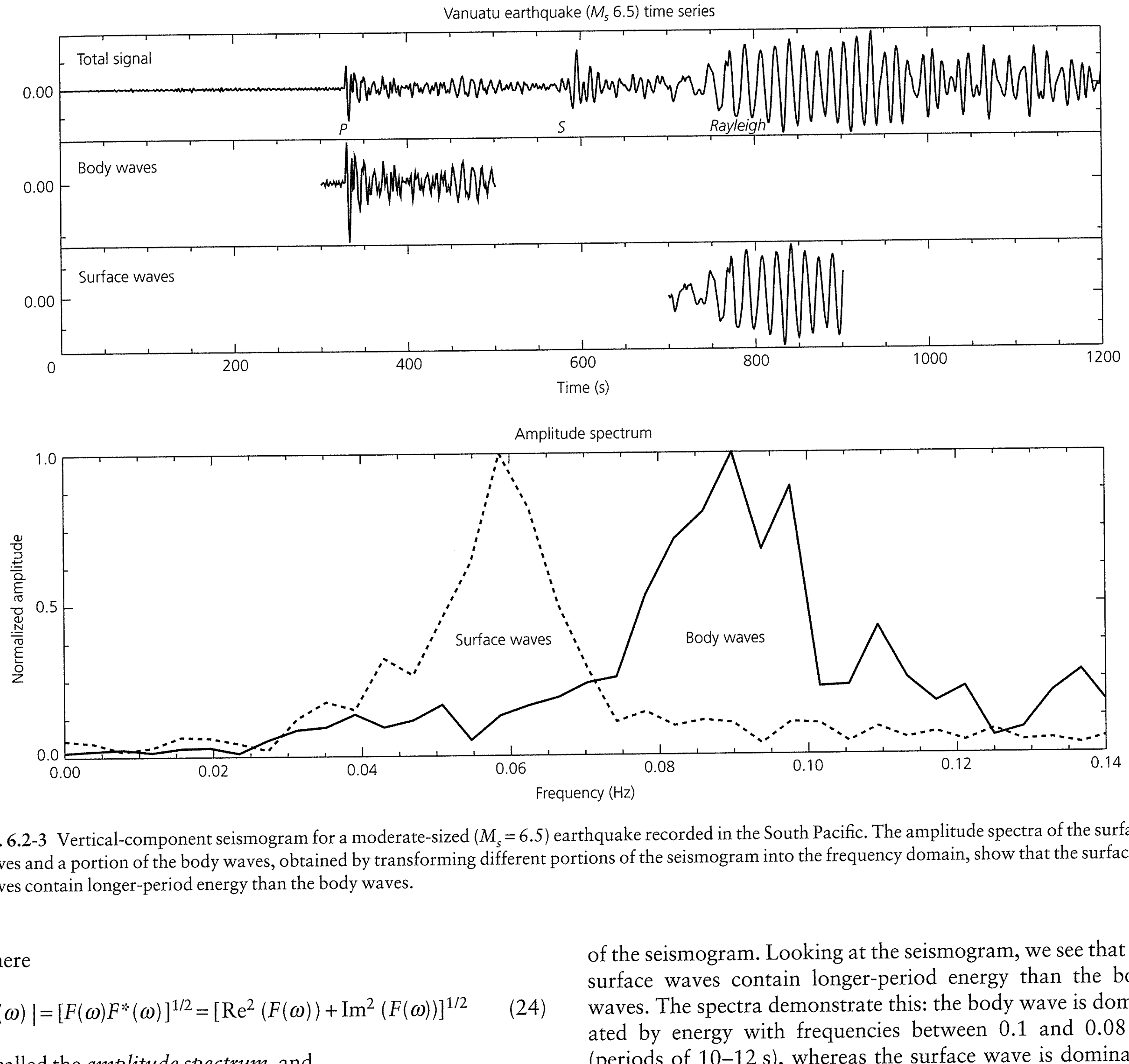


Fig. 6.2-3 Vertical-component seismogram for a moderate-sized ($M_s = 6.5$) earthquake recorded in the South Pacific. The amplitude spectra of the surface waves and a portion of the body waves, obtained by transforming different portions of the seismogram into the frequency domain, show that the surface waves contain longer-period energy than the body waves.

where

$$|F(\omega)| = [F(\omega)F^*(\omega)]^{1/2} = [\mathrm{Re}^2\,(F(\omega)) + \mathrm{Im}^2\,(F(\omega))]^{1/2} \qquad (24)$$

is called the *amplitude spectrum*, and

$$\phi(\omega) = \tan^{-1}(\mathrm{Im}\,(F(\omega))/\mathrm{Re}\,(F(\omega))) \qquad (25)$$

is the *phase spectrum*.[2]

Both the amplitude and the phase spectra are needed to fully represent the transform, which is also called the complex spectrum. In many applications only the amplitude spectrum is shown, because it indicates how the energy (the square of the amplitude) in the time series depends on frequency. Figure 6.2-3 shows a seismogram for a moderate-size earthquake, together with amplitude spectra for the body and surface wave portions of the seismogram. Looking at the seismogram, we see that the surface waves contain longer-period energy than the body waves. The spectra demonstrate this: the body wave is dominated by energy with frequencies between 0.1 and 0.08 Hz (periods of 10–12 s), whereas the surface wave is dominated by energy with frequencies between 0.07 and 0.05 Hz (periods of 14–20 s). For comparison, Fig. 6.2-4 shows data for a much larger earthquake. The seismogram, from an instrument designed to record at long periods, covers seven days after the earthquake. The large oscillations with periods of about 90,000 s are tides within the solid earth. Superimposed on these is the signal due to the earthquake. The portion of the amplitude spectrum shown indicates the presence of energy at long periods (0.002 Hz corresponds to 500 s period). The energy is concentrated at discrete peaks, corresponding to the earth's normal modes.

The Fourier transform $F(\omega)$ is another way of representing the time series $f(t)$. We speak of $f(t)$ as being in the "time domain," and $F(\omega)$ as being in the "frequency domain." The

[2] The notations Re and Im indicate the real and imaginary portions of a complex number (Section A.2).

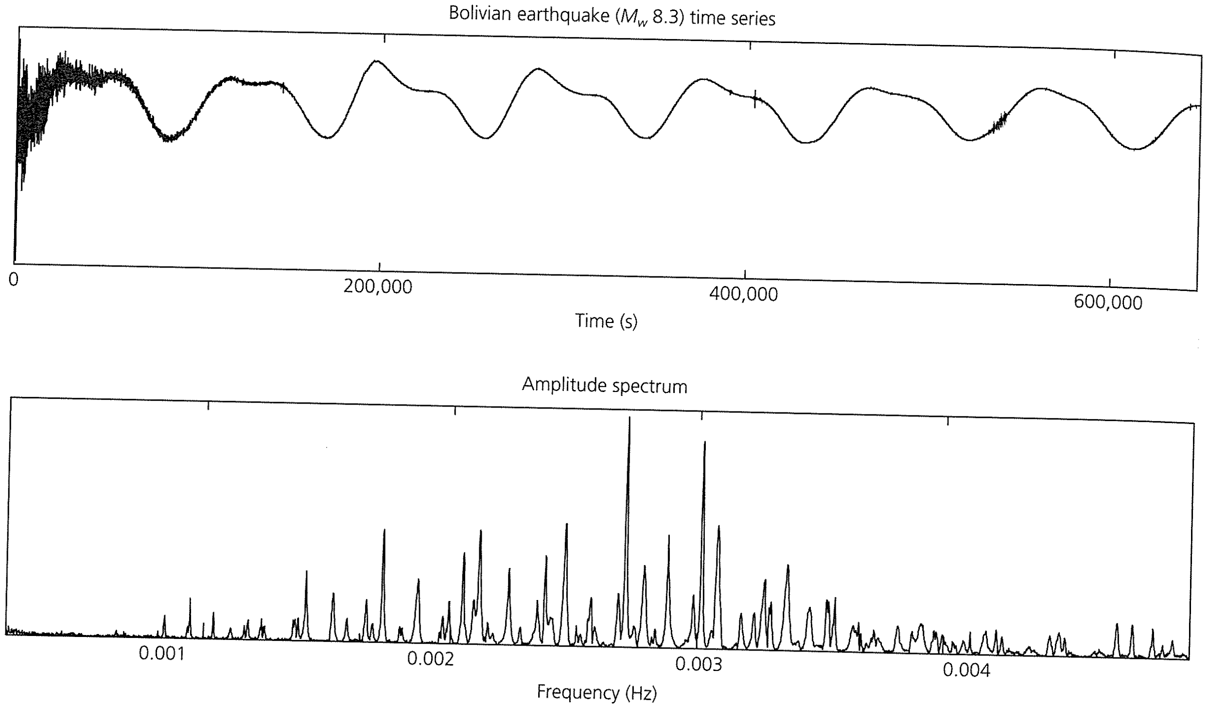


Fig. 6.2-4 Vertical-component seismogram and amplitude spectrum for the great ($M_w = 8.3$) 1994 Bolivian deep earthquake recorded in Arizona. The time series extends for days after the earthquake, showing the solid earth tide and the signal due to the earthquake. The earth's normal modes appear as peaks in the amplitude spectrum.

two representations are equivalent, because we can easily convert data from one domain to the other without losing any information. We will see that some methods of analyzing seismograms are more easily conducted in the frequency domain, and that there is a relation between time and frequency domain operations.

The Fourier transform and inverse transform relate a function of time $f(t)$ and its transform $F(\omega)$, a function of angular frequency. Similar relations apply between other pairs of variables. In seismology, the other commonly used pair is distance and wavenumber. Because the wavenumber is the spatial frequency (Section 2.2.2), it is related to distance in the same way that angular frequency is related to time. Hence, there are applications in which a double Fourier transform is taken to convert a set of seismograms, which describe displacement as a function of space and time, into a function of wavenumber and frequency (Section 3.3.5). A triple Fourier transform can similarly be taken for data in two space dimensions and time.

6.2.4 *Properties of Fourier transforms*

The Fourier transform has a number of interesting properties that we often use, whose proofs are left for the problems.

(1) The Fourier transform is linear: if $F(\omega)$ and $G(\omega)$ are the transforms of $f(t)$ and $g(t)$, then $(aF(\omega) + bG(\omega))$ is the transform of $(af(t) + bg(t))$. This property makes the Fourier transform useful in filtering, because it permits us to treat a signal as the sum of several signals, knowing that the transform will be the sum of their transforms.

(2) The Fourier transform of a purely real time function has the symmetry

$$F(-\omega) = F^*(\omega). \tag{26}$$

Thus for seismograms (which are real because the motion of the ground is purely real), the values of the transform for the negative frequencies can be found from those for positive frequencies. Hence, in filtering seismograms, we can operate on only the positive frequencies and compute the value of the transform at the negative frequencies by taking the conjugate, thus saving computer time and storage space.

(3) The Fourier transform of a time series shifted in time is found by changing the phase of the transform: if the transform of $f(t)$ is $F(\omega)$, the transform of $f(t-a)$ is $e^{-i\omega a}F(\omega)$. In analyzing seismograms it is arbitrary what time we choose as the origin; the amplitude spectrum stays the same, and the phase changes in a simple way. This makes sense, because in the absence of attenuation a wave keeps its shape but changes in phase as it propagates. Similarly, shifting a Fourier transform in frequency causes a phase change in the corresponding time series: the inverse transform of $F(\omega - a)$ is $e^{iat}f(t)$. These relations are sometimes called *shift theorems*.

(4) The Fourier transform of the derivative of a time function is found by multiplication: $(i\omega)F(\omega)$ is the transform of

$df(t)/dt$. Similarly, $(i\omega)^n F(\omega)$ is the transform of $d^n f(t)/dt^n$. This makes differentiation easy on a computer, and is an easy way to change a displacement record into velocity, or velocity into acceleration. This property also makes it easy to solve differential equations (e.g., Eqn 3.7.8) using the Fourier transform, an approach that is often posed as using a sinusoidal trial solution. Hence we sometimes write and operate on the wave equation using the Fourier transform of the wave field (Eqns 2.2.34, 3.3.74).

(5) The total energy in a Fourier transform is the same as that in the time series:

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega, \tag{27}$$

a relation known as *Parseval's theorem*. This relation arises because the time series and its Fourier transform are equivalent representations.

6.2.5 *Delta functions*

In using Fourier transforms, we often need to describe a signal that is concentrated at a single time or frequency. This is done using the Dirac delta function, an entity that is not truly a function, but rather a generalized function that is the limit of a sequence of continuous functions. The delta function can be defined in several ways, each of which offers a different insight into its nature.

A delta function at $t = t_0$, written $\delta(t - t_0)$, is defined as the limit of a Gaussian function that keeps the area constant (= 1) as the width (σ) narrows and the height, $1/\sigma\sqrt{2\pi}$, increases (Fig. 6.2-5):

$$\delta(t - t_0) = \lim_{\sigma \to 0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{t - t_0}{\sigma}\right)^2\right]. \tag{28}$$

Thus the Dirac delta function is a continuous function analogous to the Kronecker delta symbol, δ_{ij} (Eqn A.3.37) which is a function of two discrete variables, i and j. An alternative definition comes from defining the delta function by how it behaves when integrated, a property called "sifting." This is defined as

$$f(t_0) = \int_{-\infty}^{\infty} f(t)\delta(t - t_0)\, dt. \tag{29}$$

Thus the delta function at $t = t_0$ "sifts out" the value of a function at time t_0 if it is multiplied by the function and integrated over all time.

A third definition comes from considering a step, or Heaviside, function $H(t - t_0)$ that is 0 for time before $t = t_0$ and equal to 1 afterwards (Fig. 6.2-5). The delta function $\delta(t - t_0)$ is the derivative of the step, because it is zero except at t_0, when it goes to infinity. Because the delta function is located where its argument is zero, $\delta(t_0 - t)$ is at time t_0, whereas $\delta(t + t_0)$ is at time $-t_0$.

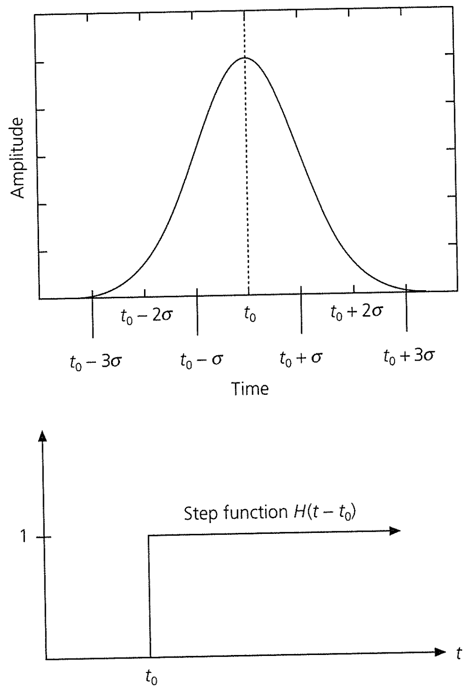


Fig. 6.2-5 Definitions of a delta function at $t = t_0$. *Top*: $\delta(t - t_0)$ is the limit of a Gaussian function with width σ. The area stays equal to 1 as the width narrows and the height increases. *Bottom*: $\delta(t - t_0)$ is the derivative of a step function $H(t - t_0)$ at time $t = t_0$, which is zero at all times except near t_0, when it goes to infinity.

To find the Fourier transform of the delta function, we use the definition of the transform (Eqn 22) with $f(t) = \delta(t - t_0)$,

$$F(\omega) = \int_{-\infty}^{\infty} \delta(t - t_0) e^{-i\omega t} dt = e^{-i\omega t_0}, \tag{30}$$

and evaluate the integral by the sifting property (Eqn 29). If the delta function is at time zero,

$$F(\omega) = \int_{-\infty}^{\infty} \delta(t) e^{-i\omega t} dt = 1. \tag{31}$$

Similarly, for a delta function at $t = t_0$, the amplitude spectrum (Eqn 24) is also

$$|F(\omega)| = (e^{-i\omega t_0} e^{i\omega t_0})^{1/2} = 1, \tag{32}$$

but the phase spectrum (Eqn 25) is

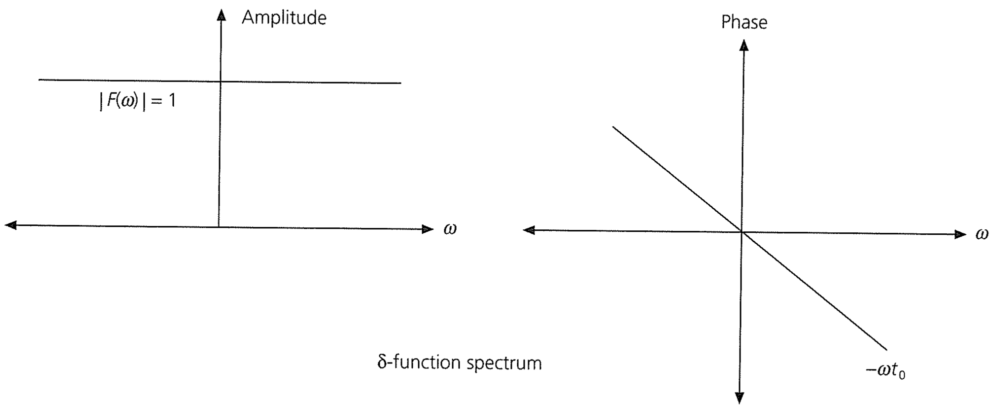


Fig. 6.2-6 The Fourier transform of a delta function, $\delta(t-t_0)$, is $e^{-i\omega t_0}$. Its amplitude spectrum has unit amplitude at all frequencies, and its phase spectrum has a slope of $-t_0$.

$$\phi(\omega) = -\omega t_0, \tag{33}$$

as shown in Fig. 6.2-6. This example illustrates one of the Fourier transform properties noted in Section 6.2.4, that shifting a function by a time t_0 changes its transform by $e^{-i\omega t_0}$.

The delta function's amplitude spectrum has unit amplitude at all frequencies. Another way to see this is to write the inverse transform, using Eqn 21,

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega t_0} e^{i\omega t} d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega(t-t_0)} d\omega = \delta(t-t_0), \tag{34}$$

which shows that the delta function is an integral or sum of sinusoids of all frequencies. These are in phase only at time t_0, giving a large amplitude, and are out of phase at all other times, giving a zero amplitude (Fig. 6.2-7).

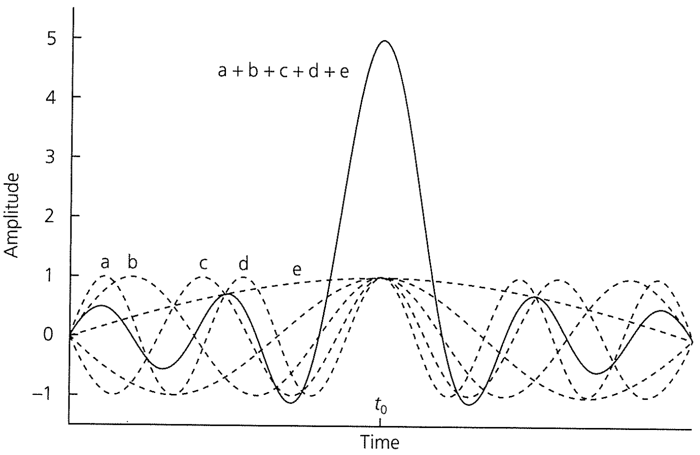


Fig. 6.2-7 Because the Fourier transform of a delta function has unit amplitude at all frequencies, it corresponds to the sum of sinusoids of all frequencies. These are in phase only at time t_0, giving a large amplitude, and are out of phase at all other times, giving zero amplitude. In this example, five sinusoids (dashed lines *a–e*) with unit amplitude (cos $[(2n+1)(t-t_0)]$) are summed (solid line), giving a peak of amplitude 5 at t_0.

Although so far we have discussed delta functions only in the time domain, they are also useful in the frequency domain. The properties of the frequency domain delta functions are analogous to those in the time domain. A delta function at angular frequency ω_0, $\delta(\omega-\omega_0)$, has an inverse transform of

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta(\omega-\omega_0) e^{i\omega t} d\omega = \frac{1}{2\pi} e^{i\omega_0 t}. \tag{35}$$

Thus we can express the delta function in terms of its Fourier transform,

$$\delta(\omega-\omega_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega_0 t} e^{-i\omega t} dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(\omega_0-\omega)t} dt, \tag{36}$$

showing that it is the integral, or sum, of sinusoids that are in phase only at frequency ω_0.

Delta functions in angular frequency give the spectra of sinusoids with a single frequency. For example, a cosine with frequency ω_0, given by

$$f(t) = \cos \omega_0 t = (e^{i\omega_0 t} + e^{-i\omega_0 t})/2, \tag{37}$$

has a Fourier transform of

$$F(\omega) = \frac{1}{2}\int_{-\infty}^{\infty} [e^{i\omega_0 t} + e^{-i\omega_0 t}] e^{-i\omega t} dt = \frac{1}{2}\int_{-\infty}^{\infty} [e^{i(\omega_0-\omega)t} + e^{-i(\omega_0+\omega)t}] dt. \tag{38}$$

By Eqn 36, this is the sum of two delta functions in the frequency domain,

$$F(\omega) = \pi[\delta(\omega-\omega_0) + \delta(\omega+\omega_0)]. \tag{39}$$

Thus the amplitude spectrum of the cosine time function in Eqn 37 consists of two delta functions, one at ω_0 and one at

$-\omega_0$. If the time function were a sine rather than a cosine, the amplitude spectrum would be the same, but the phase spectrum would be different. Given the relation between the transforms of functions shifted in time discussed in the previous section, this makes sense, because a sine function is a time-shifted cosine, and vice versa.

This example illustrates one of the reasons for using Fourier transforms. The frequency domain description of the function is simpler, because a large number of points are needed to accurately describe the cosine as a function of time, but only two complex numbers, the values of the transforms at $\pm\omega_0$, are needed to describe it as a function of frequency. Time series more complicated than a pure cosine are often more easily described in the frequency domain, and processes that act on the time series are also often more easily represented in the frequency domain. In such cases, it is common to work in the frequency domain and then use the inverse transform to generate the final time series.

6.3 Linear systems

Among the uses of Fourier analysis in seismology is modeling different factors affecting a seismogram. First, a seismogram is a record of ground motion that includes the effect of the seismometer. Furthermore, the ground motion combines the effects of the seismic source and the elastic and anelastic earth structure along the propagation path (Section 4.3). To characterize the combined effects of these different factors, we use the idea of a *linear system*, a general representation of any device or process that takes an input signal and modifies it. This representation treats these processes as mathematical operators transforming an input signal into an output signal.

6.3.1 *Basic model*

A linear system is one in which if input signals $x_1(t)$ and $x_2(t)$ produce output signals $y_1(t)$ and $y_2(t)$, the combined input $(Ax_1(t) + Bx_2(t))$ yields $(Ay_1(t) + By_2(t))$ (Fig. 6.3-1). We have previously referred to this feature as the principle of *superposition*. Fortunately, the earth generally behaves this way in transmitting seismic waves. As a result, linear system models are used in a wide variety of seismological applications. Fourier analysis is a natural tool for studying linear systems because the Fourier transform has these same linear properties (Section 6.2.4).

We characterize a linear system by its response to an impulsive delta function in time (Fig. 6.3-2). This *impulse response* $f(t)$ can be used to find the response of the system to an arbitrary input signal. Viewed in the frequency domain, the *impulse*, whose spectral amplitude is equal to 1 at all frequencies, gives rise to an output $F(\omega)$, which is the transform of the impulse response, sometimes called the *transfer function*. Thus, if the input signal is an arbitrary signal $x(t)$, with transform $X(\omega)$, the resulting output spectrum is just the input spectrum times the spectrum of the impulse response,

$$Y(\omega) = X(\omega)F(\omega). \tag{1}$$

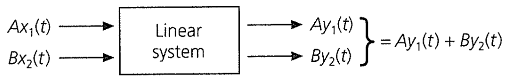


Fig. 6.3-1 Definition of a linear system.

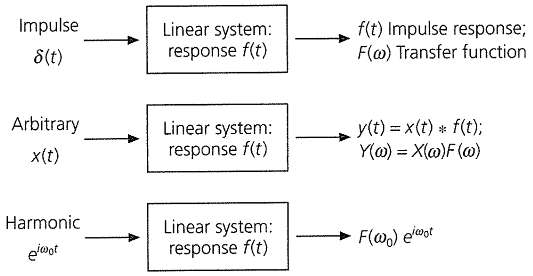

Impulse $\delta(t)$
Linear system: response $f(t)$
$f(t)$ Impulse response; $F(\omega)$ Transfer function
Arbitrary $x(t)$
Linear system: response $f(t)$
$y(t) = x(t) * f(t)$; $Y(\omega) = X(\omega)F(\omega)$
Harmonic $e^{i\omega_0 t}$
Linear system: response $f(t)$
$F(\omega_0)\, e^{i\omega_0 t}$

Fig. 6.3-2 Characterization of a linear system by its impulse response $f(t)$ and transfer function $F(\omega)$.

Because the transforms are generally complex numbers, the phase as well as the amplitude of the input signal is usually modified.

The output in the time domain $y(t)$ can be found by inverting the transform,

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)F(\omega)e^{i\omega t}d\omega. \tag{2}$$

To see that this works, note that for the impulse $x(t) = \delta(t)$, $X(\omega) = 1$, and $y(t) = f(t)$. This equation gives another way to think of the impulse response. For a harmonic input signal of unit amplitude $e^{i\omega_0 t}$, whose transform is the delta function in frequency

$$X(\omega) = 2\pi\delta(\omega - \omega_0), \tag{3}$$

the output is

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} 2\pi\delta(\omega - \omega_0)F(\omega)e^{i\omega t}d\omega = F(\omega_0)e^{i\omega_0 t}, \tag{4}$$

a harmonic signal of the same frequency with the amplitude of the transfer function at that frequency.

It is interesting to consider the relation between the input time function, the impulse response, and the output time function. To do this, we expand Eqn 2 by writing out the transforms of $X(\omega)$ and $F(\omega)$,

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x(\tau)e^{-i\omega\tau}d\tau\right]\left[\int_{-\infty}^{\infty} f(\tau')e^{-i\omega\tau'}d\tau'\right]e^{i\omega t}d\omega, \tag{5}$$

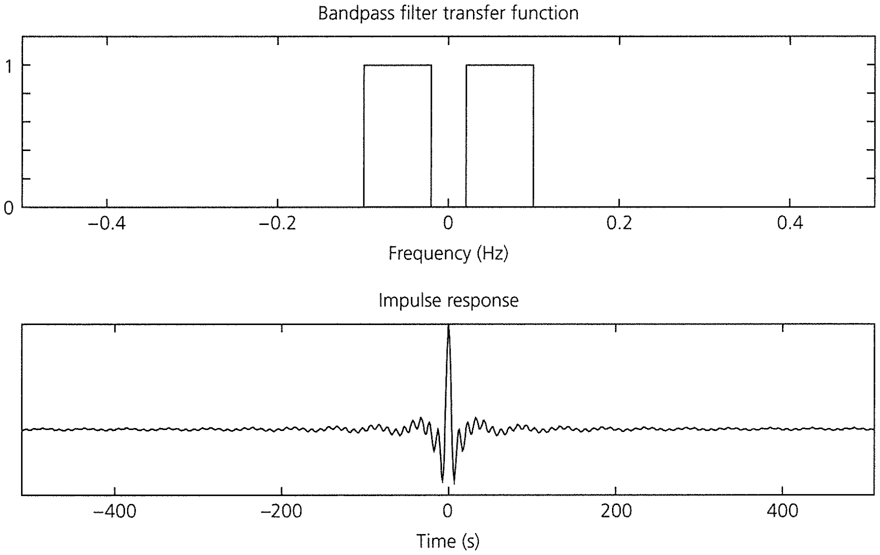


Fig. 6.3-3 A simple bandpass filter specified in the frequency (*top*) and time (*bottom*) domains.

and regrouping terms,

$$y(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(\tau) f(\tau') \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t-\tau'-\tau)} d\omega \right] d\tau d\tau'. \tag{6}$$

Using the inverse transform of the delta function (Eqn 6.2.34),

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega(t-\tau'-\tau)}\, d\omega = \delta(t - \tau' - \tau), \tag{7}$$

we eliminate the frequency integral and obtain

$$y(t) = \int_{-\infty}^{\infty} x(\tau) \left[\int_{-\infty}^{\infty} f(\tau')\delta(t - \tau' - \tau) d\tau' \right] d\tau. \tag{8}$$

Finally, carrying out the inner integration using the sifting property of the delta function (Eqn 6.2.29) yields

$$y(t) = \int_{-\infty}^{\infty} x(\tau) f(t - \tau) d\tau. \tag{9}$$

This integral operation, known as the *convolution* of the functions $x(t)$ and $f(t)$, is often written as

$$y(t) = x(t) * f(t). \tag{10}$$

The output of a linear system is thus the convolution of the input signal and the impulse response. Comparison of Eqns 10 and 1 shows the relation between operations in the two domains: convolution in the time domain corresponds to multiplication in the frequency domain. The reverse is also true: frequency domain convolution corresponds to time domain multiplication.

We thus have two different ways of implementing any operation that can be characterized by a linear system. The effect that the system has on an input signal is specified either by the impulse response in the time domain or by its transform, the transfer function in the frequency domain. For example, to filter a seismogram so that only a certain range of frequencies remains, we can filter in either the frequency or time domains. To do this in the frequency domain, we can define a simple *bandpass filter*, a function which is 1 in the frequency range of interest and 0 for all other frequencies. Figure 6.3-3 (*top*) shows the amplitude spectrum of the filter, whose phase spectrum is defined as zero for all frequencies. To perform the filtering, we multiply this function by the Fourier transform of the seismogram, point by point for all frequencies, and take the inverse transform of the result. The resulting filtered seismogram has only the desired frequencies. Alternatively, however, we could find the impulse response of the bandpass filter by taking the inverse Fourier transform of the amplitude spectrum in the top of Fig. 6.3-3, and filter the data by convolving this impulse response (Fig. 6.3-3, *bottom*) with the seismogram in the time domain.

A few points about this simple filter are worth noting. First, although it is typical to plot the transfer function only for the positive frequencies, the filter is also defined for negative frequencies, to ensure that the resulting signal is real (Section 6.2.4). Second, the peculiar appearance of the impulse response makes sense when we recall that the impulse response describes what comes out of the filter when a delta function comes in (Fig. 6.3-2). The delta function's amplitude spectrum is constant for all frequencies, but only some of these frequencies are transmitted through the filter. The lack of high frequencies is particularly noticeable, and results in the noncausal impulse response beginning before time zero. We noted a similar phenomenon in Section 3.7.8, where anelasticity acted as a

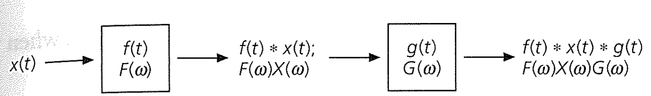


Fig. 6.3-4 When a signal goes through two linear systems in succession, the net output is the convolution of the impulse responses in the time domain, or the product of the transfer functions in the frequency domain.

filter, removing high frequencies and thus making the waveforms noncausal unless the effects of physical dispersion were included. Third, this filter has sharp "corners" at the edges of the passband, although in real applications the corners are smoothed for reasons we discuss shortly.

Because the same effect can be achieved by either time domain or frequency domain filtering, the choice of domain can be made for convenience. Surprisingly, the operations of taking transforms and inverse transforms are sufficiently fast in computation that it generally makes sense to filter in the frequency domain. An attraction of this method is that filters are usually easier to specify in the frequency domain, because it is clear which are the desired and undesired parts of the signal. For example, in Fig. 6.3-3 (*bottom*), the corresponding time domain filter is difficult to visualize intuitively. Similarly, the transfer function, or instrument response, of a seismometer is more easily specified in the frequency domain, as we will discuss in Section 6.6.

6.3.2 *Convolution and deconvolution modeling*

Linear system ideas are so pervasive in seismology that we discussed them in applications such as reflection seismology (Section 3.3.6) and earthquake source studies (Section 4.3) before we justified them mathematically. One reason why these models are so useful is that they are easily generalized to multiple linear systems, so quite complicated physical effects can be described. Specifically, if a signal $x(t)$ goes through two linear systems in succession (Fig. 6.3-4), with impulse responses $f(t)$ and $g(t)$, the net output is either a convolution in the time domain,

$$y(t) = x(t) * f(t) * g(t), \tag{11}$$

or the product of the transfer functions in the frequency domain

$$Y(\omega) = X(\omega)F(\omega)G(\omega). \tag{12}$$

We can extend this to an arbitrary number of linear systems.

A common application is to think of a seismogram as the output resulting from sending a source signal through a set of linear systems. In the simplest case, the seismogram $u(t)$ can be written in terms of three basic effects,

$$u(t) = x(t) * g(t) * i(t), \tag{13}$$

where $x(t)$ is the source signal, $g(t)$ is the response of an operator representing the effects of earth structure along the path of the seismic waves, and $i(t)$ is the impulse response of the seismometer.

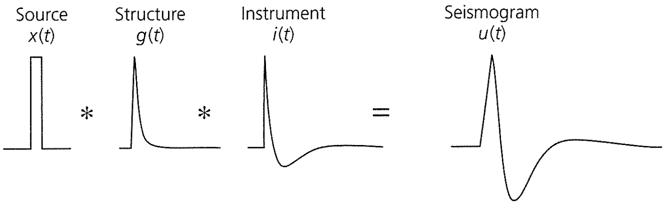


Fig. 6.3-5 A seismogram can be modeled as the convolution of the source signal with operators representing the effects of earth structure and the seismometer. This can be done in the time domain as a set of convolutions, $u(t) = x(t) * g(t) * i(t)$, or in the frequency domain as a set of multiplications, $U(\omega) = X(\omega)G(\omega)I(\omega)$. (After Chung and Kanamori, 1980. *Phys. Earth Planet. Inter.*, *23*, 134–59, with permission from Elsevier Science.)

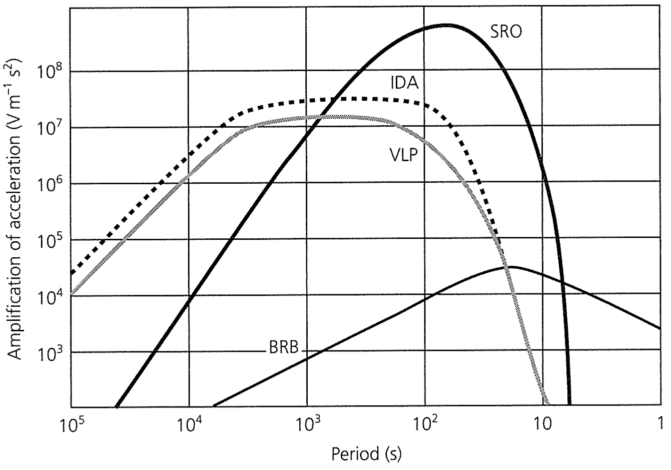


Fig. 6.3-6 Transfer functions for various seismometers, some of which are discussed in Section 6.6. SRO is the Seismic Research Observatory, IDA is International Deployment of Accelerometers, VLP is Very Long Period, and BRB is Broadband. Transfer functions are the frequency domain equivalents of the time domain instrument response shown in Fig. 6.3-5 as $i(t)$.

Figure 6.3-5 shows a simple example: a seismogram resulting from the convolution of a trapezoidal source function representing the signal emitted by an earthquake with operators giving the effects of earth structure and the seismometer. Each operator can be specified in either domain. For example, the time domain impulse response of a seismometer reflects the fact that its transfer function depends on frequency (Fig. 6.3-6). Once the different effects are characterized by their response in the time or frequency domain, the seismogram due to their combined effects can be obtained.

Convolution can be used to describe the response of a system in space as well as time. For example, probabilistic earthquake hazard maps like Fig. 1.2-3 can be viewed as two-dimensional convolutions in space of an assumed distribution of earthquake sources with an impulse response like Fig. 1.2-5 giving the expected ground motion as a function of earthquake magnitude and distance.

Often the impulse response is defined in both space and time. This is the basic approach used to find the response of the earth to a seismic source (Chapter 4). The displacement at a point $\mathbf{x}$ and time t is

$$u(\mathbf{x}, t) = \iint G(\mathbf{x}-\mathbf{x}'; t-t') f(\mathbf{x}', t') dt'dV', \tag{14}$$

where $G(\mathbf{x} - \mathbf{x}'; t - t')$ is the *Green's function*,[1] the impulse response to a source at position $\mathbf{x}'$ and time t', and $f(\mathbf{x}', t')$ is the distribution of seismic sources. Thus the integral gives the total response due to the distribution of sources. In most cases the source is limited in space and time, so the integral is done over the source region. Often the source is at a point in space or time, so $f(\mathbf{x}', t')$ contains delta functions and is easily integrated using the sifting property. A nice feature of this formulation is that the principle of reciprocity, which says that the source and the receiver can be interchanged, emerges directly. The Green's function in Eqn 14 is for a laterally homogeneous medium, so the response depends only on the distance between the source and the receiver. In a general medium Eqn 14 becomes

$$u(\mathbf{x}, t) = \iint G(\mathbf{x}, t; \mathbf{x}', t') f(\mathbf{x}', t')\, dt'dV'. \tag{15}$$

When a system is described by a convolution, we can examine the effects of the different contributing factors using *deconvolution*. We start with the output and one of the time series that were convolved to form it, and then find the other. For example, in Section 3.3.6 we discussed using seismic reflection data to obtain the sharpest resolution of reflectors in the earth. We assumed that a seismogram $s(t)$ results from convolution of a source pulse, or wavelet, $w(t)$, and an earth structure operator, $r(t)$. $r(t)$, known as a reflector series, is presumed to be a set of delta functions with positions corresponding to the travel time for a reflection from an interface and amplitudes corresponding to the amplitude of the reflected arrival. Thus

$$s(t) = w(t) * r(t) \quad \text{and} \quad S(\omega) = W(\omega)R(\omega). \tag{16}$$

If the travel time differences between the arrivals corresponding to individual reflectors are shorter than the duration of the wavelet, interference can occur, giving a complicated signal. Hence it would be desirable to have a delta function source wavelet whose Fourier transform is simply 1, so that the seismogram would equal the reflector series. Although a physical source wavelet is not a delta function, we simulate such a wavelet by creating an *inverse filter*[2] $w^{-1}(t)$, which, when convolved with the wavelet, yields a delta function:

$$w^{-1}(t) * w(t) = \delta(t). \tag{17}$$

As we saw in Section 3.3.6, the Fourier transform of the inverse filter is just $1/W(\omega)$, so deconvolution can be done by dividing the Fourier transforms

$$S(\omega)/W(\omega) = R(\omega). \tag{18}$$

This sometimes works well, but can be problematic at frequencies where the source wavelet spectrum $W(\omega)$ is small (causing $R(\omega)$ to go to infinity), so a minimum amplitude threshold can be set.

As an alternative, inverse filters can be designed in the time domain to compress the source wavelet into a function as close to a delta function as possible. This approach is a special case of the general problem of finding a shaping filter that converts a given input into a given output. We will shortly discuss another approach, which relies not on the convolution, but on the related cross-correlation operator.

Deconvolution is also used in other applications. A conceptually similar one is modeling seismograms from a distant earthquake as a sum of secondary arrivals generated when the upcoming wave encounters interfaces below the receiver (Fig. 6.3-7). The vertical component is assumed to represent the direct arrival, and is used as a Green's function that is deconvolved from a horizontal component to find a *receiver function* characterizing the structure. The receiver function corresponds to the reflector series in this geometry. Another application of deconvolution is to take seismograms and deconvolve the effects of the seismometer to find the true ground motion, or deconvolve a seismogram to try to find the source pulse due to an earthquake (Section 4.3.3).

6.3.3 *Finite length signals*

We have seen that the Fourier transform describes a signal as the sum of harmonic signals with different frequencies. One important limitation is that the Fourier transform requires integration over all time. In reality, we only have data over a finite interval of time.

To see how this affects our results, consider a *window* function $b(t)$ which selects part of the data. Its effect on the data $f(t)$ is represented by multiplying $f(t)$ by $b(t)$. We then ask how the Fourier transform of the function, including the effect of the window

$$G(\omega) = \int_{-\infty}^{\infty} b(t) f(t) e^{-i\omega t} dt, \tag{19}$$

is related to the transform of the original function, $F(\omega)$.

[1] The same entity is commonly termed a Green's function in physical problems and an impulse response in time series analysis. In seismology the terms are used essentially interchangeably.

[2] The notation $w^{-1}(t)$ does not mean $1/w(t)$.

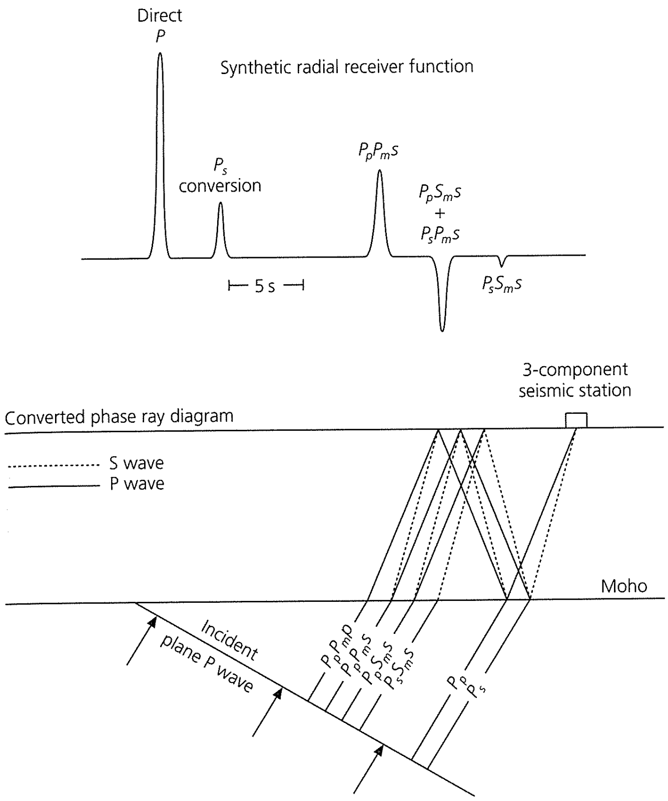


Fig. 6.3-7 Schematic diagram of the receiver function approach. The receiver function, derived by deconvolving the vertical component from a horizontal component, should have arrivals corresponding to the times of seismic wave phases generated when the upcoming wave encounters interfaces below the receiver and amplitudes reflecting the amplitudes of these waves. The receiver function can be used to study the depths of the interfaces and the velocity contrast there. Because a horizontal component is used, the phases predicted involve *P*-to-*S* conversions and their reverberations, as described by the nomenclature used to identify phases (e.g., *PpPms*). Owens *et al.*, 1987. © Seismological Society of America. All rights reserved.)

This question can be answered by writing $b(t)$ and $f(t)$ using their inverse transforms,

$$G(\omega)=\int_{-\infty}^{\infty}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}B(\omega')e^{i\omega' t}d\omega'\right]\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega'')e^{i\omega'' t}d\omega''\right]e^{-i\omega t}dt$$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}B(\omega')\left[\int_{-\infty}^{\infty}F(\omega'')\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-i\omega t+i\omega' t+i\omega'' t}dt\right)d\omega''\right]d\omega', \tag{20}$$

recognizing that the inner integral is the Fourier transform of a delta function in frequency (Eqn 6.2.36),

$$G(\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}B(\omega')\left[\int_{-\infty}^{\infty}F(\omega'')\delta(\omega-\omega'-\omega'')d\omega''\right]d\omega', \tag{21}$$

and using the sifting property (Eqn 6.2.29) to obtain

$$G(\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}B(\omega')F(\omega-\omega')d\omega'=\frac{1}{2\pi}B(\omega)*F(\omega). \tag{22}$$

Thus the effect of multiplying a time series by a window function is that the spectrum of the time series is convolved with the spectrum of the window function. This is an example of the fact that just as convolution in the time domain corresponds to multiplication in the frequency domain, so multiplication in the time domain corresponds to convolution in the frequency domain.

To see the effect of windowing on the spectrum, consider the simplest window function, a "boxcar" which describes taking only the data in a certain time interval (Fig. 6.3-8),

$$\begin{aligned} b(t)&=1 \quad \text{for } -T<t<T,\\ &=0 \quad \text{otherwise.}\end{aligned} \tag{23}$$

Its Fourier transform is

$$B(\omega)=\int_{-T}^{T}e^{-i\omega t}dt=-\frac{e^{-i\omega t}}{i\omega}\bigg|_{-T}^{T}=\frac{2\sin\omega T}{\omega}=\frac{2T\sin\omega T}{\omega T}, \tag{24}$$

whose amplitude spectrum $|B(\omega)|$ has a characteristic shape with a central lobe and smaller side lobes, and equals zero where $x=\omega T=2n\pi$. The width of the central lobe is $2\pi/T$. This $|(\sin x)/x|$ curve, sometimes called a *sinc* function, is convolved with, and thus modifies, the spectrum $|F(\omega)|$.

For example, if $f(t)$ is a sine wave (Fig. 6.3-9a) whose amplitude spectrum is described by two delta functions, convolution with $B(\omega)$ yields the spectrum of a finite length sine wave, two sinc functions. Thus, taking a finite length of record "smears" the delta functions of the infinite length record's spectrum into broader peaks with side lobes (Figs 6.3-9b). Taking longer records (increasing T) yields sharper spectra (more like the delta function), because the width of the central lobe of the sinc function is proportional to $1/T$.

This effect has an important consequence for analyzing signals containing different frequencies, as shown in Fig. 6.3-9c for a time series with two frequencies. For shorter record lengths (Figs 6.3-9d and e), the spectral peaks broaden until they start to overlap and cannot be resolved separately. Once the width in frequency of the central lobe of the sinc function exceeds the separation between the two spectral peaks (Figs 6.3-9e), they cannot be resolved. Thus the frequency resolution, the minimum separation in frequency for which

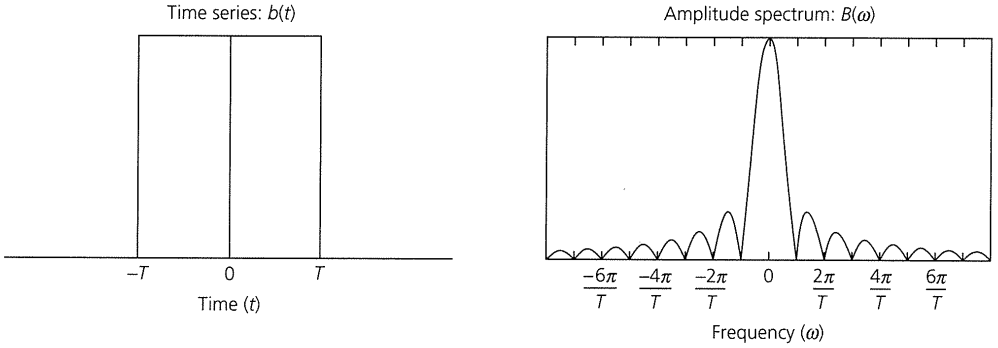


Fig. 6.3-8 Time and frequency domain representations of the simplest window function, a "boxcar" that selects only the data in a certain time interval (*left*). The amplitude spectrum (*right*) has a central peak and smaller side lobes.

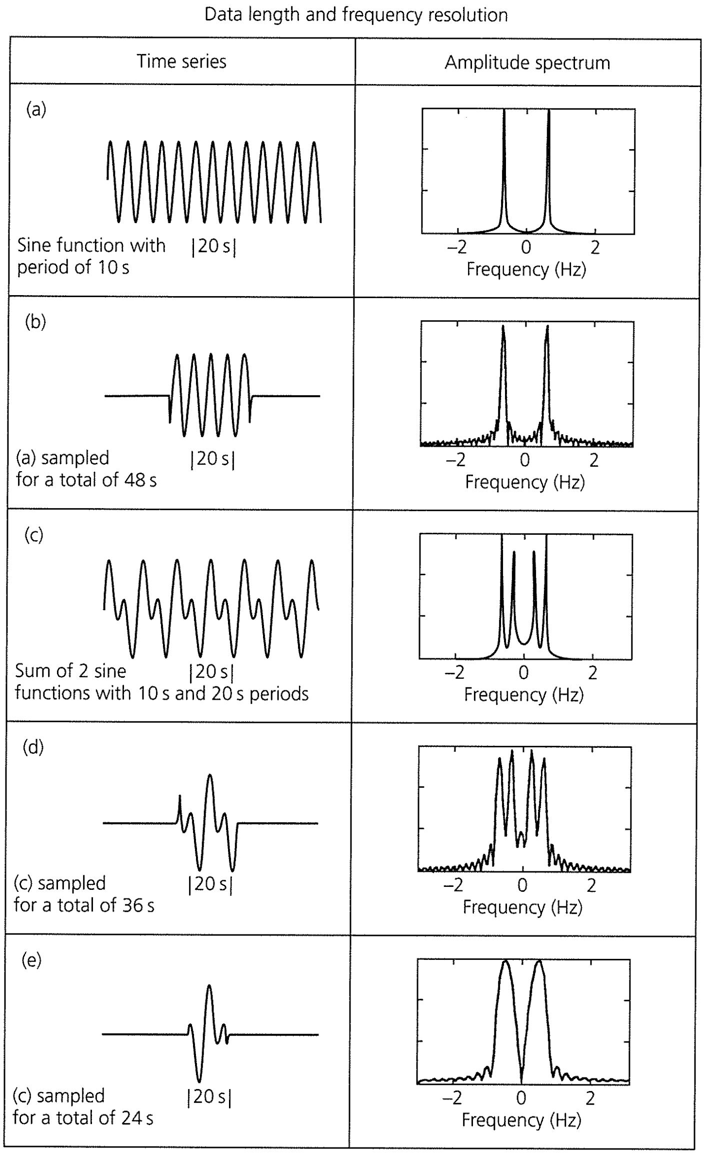


Fig. 6.3-9 Effects of finite data length on the spectrum. The spectrum of the sine wave in (a) is "smeared" by taking a short data window (b). For a time series with two frequencies (c), shorter record lengths cause the spectral peaks to broaden (d) until they start to overlap and cannot be resolved separately (e).

two peaks can be resolved, is proportional to the reciprocal of the record length.

This relation between signals in the time and frequency domains demonstrates a fundamental principle. By taking a finite length portion of a time function, we broaden and distort its spectrum in a predictable way. The reverse occurs in the frequency domain; taking a finite portion of the spectrum distorts the time function, as we discussed in considering Fig. 6.3-3. For example, because a seismometer only responds to ground motion in a certain frequency range, the resulting seismogram is a somewhat distorted record of the ground motion. Similarly, physical processes like anelasticity (Section 3.7.8) and diffraction (Section 2.5.10) that remove high frequencies distort the resulting waveforms.

Thus we have an "uncertainty principle" that the product of the "widths" in the two domains is constant; for a time domain record with duration T, the resolution in the frequency domain is proportional to $1/T$. Perfect resolution in frequency requires infinite record length in time, and infinite bandwidth in frequency is needed to represent a time function exactly. These properties are general features of Fourier transform pairs, so also apply to distance and wavenumber.[3]

The sinc or $|\sin x/x|$ function, which we used to represent taking a finite portion of a time series, appears in other similar applications. We saw that diffraction through a slit, in which only part of a wave front is transmitted, is described by a sinc function (Fig. 2.5-18). The sinc function also describes the spectrum of waves radiated from a finite fault (Section 4.6.2).

In real cases, we do not have infinite lengths of data. Moreover, it is not always desirable to take more data. For example, the signal of interest on a seismogram eventually decays into the noise due to attenuation, or is interferred with by a different signal. We seek the best resolution of the spectrum of the signal of interest, but as the record length increases, the noise has a greater effect and increasingly contaminates the spectrum. We thus select a compromise record length and try to obtain the best spectrum. This issue arises in estimating seismic attenuation, which broadens spectral peaks (Section 3.7.7) in a way

[3] The uncertainty principle also appears in quantum physics, where the position and momentum of a particle form a Fourier transform pair. Thus, the better we know a particle's position, the less we know about its momentum, and vice versa.

similar to that of finite record length. Longer records broaden the peaks less, and so give better estimates of attenuation up to the point where the effects of noise degrade the estimates.

Though we can never get around the problem of finite record length, it can be ameliorated by using a different window function than a boxcar. A window function whose "corners" are less "sharp," known as a *taper*, reduces the size of the side lobes and thus the distortion. One simple such function, a cosine taper, is a boxcar with smoother ends:

$$
\begin{aligned}
W(t) &= \frac{1}{2}\left[1 + \cos\frac{\pi(t + T - T_1)}{T_1}\right] && \text{for } -T < t < -T + T_1 \\
&= 1 && \text{for } -T + T_1 \le t \le T - T_1 \\
&= \frac{1}{2}\left[1 + \cos\frac{\pi(t - T + T_1)}{T_1}\right] && \text{for } T - T_1 < t < T \\
&= 0 && \text{for other times.} \qquad (25)
\end{aligned}
$$

The parameter T_1 is the tapered fraction of the half-length T. Figure 6.3-10 illustrates the effect of tapering data, by comparing the spectra of two windows of the same length. The side lobes for the tapered window are reduced.

Such a taper is often applied in the time domain to data, with $T_1/T \approx 0.1$, before taking spectra. Similarly, bandpass filters are often tapered in the frequency domain. In the frequency domain, a pure bandpass filter is two boxcar functions for the positive and negative frequencies in the passband (Fig. 6.3-3). The corresponding inverse transform thus looks like a sinc function, and causes "ringing," analogous to the side lobes, in the time domain. The ringing can be reduced by tapering the response at the edges of the passbands. For the same reason, the spectrum of a theoretical (synthetic) seismogram computed in the frequency domain is tapered before the inverse Fourier transform is used to produce a synthetic seismogram in the time domain.

This example brings out the general point that, in filtering data, we make certain choices depending on our goals and accept the consequences. There are no absolute criteria for what is best. For example, tapering a filter in the frequency domain reduces the ringing that can produce spurious noncausal arrivals, at the price of distorting the spectrum and waveform. We will see in Section 6.6.5 that this issue appears in designing digital seismometers.

6.3.4 Correlation

Often we want to measure how similar two signals are. A common application is identifying a reflected arrival by finding the portion of a seismogram that most resembles a direct arrival or a function that we believe represents the source. To do this, we define the part of the signal we seek to identify as $f(t)$, the remaining portion of the seismogram as $x(t)$, and form the integral

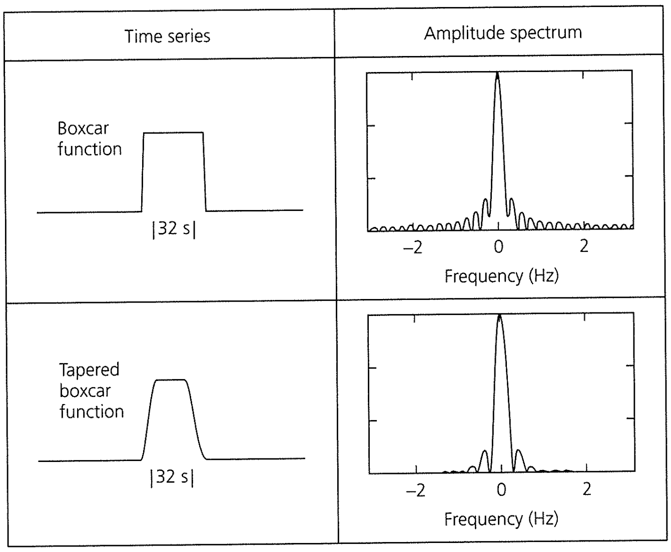


Fig. 6.3-10 Comparison of the spectra of two windows of the same length. The side lobes for the tapered window are reduced, but the central peak is less sharp.

$$
C(L) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) f(t + L)\,dt. \qquad (26)
$$

$C(L)$, the *cross-correlation* of $x(t)$ and $f(t)$, measures the similarity between $f(t)$ and later portions of $x(t)$ by shifting $f(t)$ by different *lag times*, L, and evaluating the integral of the product as a function of L. The lag for which $C(L)$ is maximum is the time shift that makes the two functions most similar. Although T formally goes to infinity, we set T to an appropriate value, because the data exist only in a finite time range. Thus the $1/T$ factor is a normalization, which is often neglected. Cross-correlation and convolution are similar operations, the major difference being the sign of the time shift.

Figure 6.3-11 shows an example of applying cross-correlation to determine the travel time difference between direct S and SS phases. The SS phase should be similar to S, once S is corrected to include the effects of the additional attenuation on the longer ray path and the $\pi/2$ phase shift due to the surface reflection (Section 3.5.1). Direct S is selected on the seismogram, corrected, and then cross-correlated with the rest of the seismogram. The peak in the cross-correlation gives the lag that measures the arrival time difference between the two phases. Another application of cross-correlation is in exploration seismology, where an assumed Vibroseis source signal is cross-correlated with seismograms, giving peaks at times when reflections occur (Section 3.3.6). In these applications, the cross-correlation is being used to identify reflections, much as could be done by deconvolution, because the cross-correlation is similar to the convolution.

A special case of the cross-correlation is the *auto-correlation*, the cross-correlation of a time series with itself

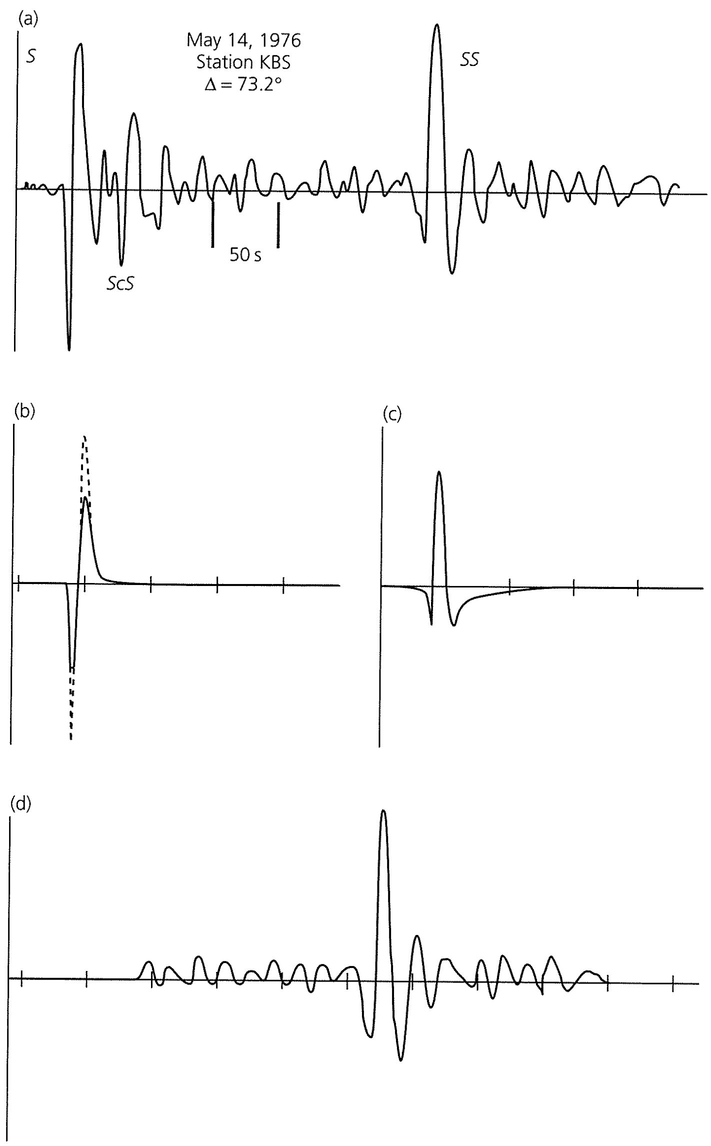


Fig. 6.3-11 Application of the cross-correlation to determine the travel time difference between direct *S* and reflected *SS* phases on a seismogram (a). The direct *S* phase (dashed line in (b)) is corrected for attenuation (solid line in (b)), phase-shifted (c), and then cross-correlated with the rest of the seismogram (d). The peak in the cross-correlation gives the lag that measures the arrival time difference between the two phases. (Kuo *et al.*, 1987. *J. Geophys. Res.*, 92, 6421–36, copyright by the American Geophysical Union.)

$$R(L) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t)f(t+L)dt. \tag{27}$$

Not surprisingly, the auto-correlation is maximum at zero lag and is an even function of the lag (Figs 6.3-12 and 3.3-30). When the cross-correlation is used to identify reflections (Figs 6.3-11 and 3.3-31), it makes the seismogram look like the auto-correlation of the signal near the reflection.

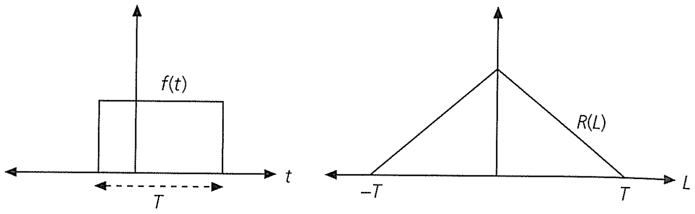


Fig. 6.3-12 Illustration, for a boxcar function, that the auto-correlation is maximum at zero lag and is an even function of the lag.

The auto-correlation is significant in the theory of filtering because it is related to the amplitude spectrum. To see this, consider a function $f(t)$ that is zero except between $-T/2$ and $T/2$. The auto-correlation

$$R(L) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t)f(t+L)dt \tag{28}$$

can be expanded using the inverse Fourier transform and using the time shift theorem (Section 6.2.4),

$$R(L) = \lim_{T\to\infty} \frac{1}{2\pi T} \int_{-T/2}^{T/2} f(t) \left[\int_{-\infty}^{\infty} F(\omega)e^{i\omega(t+L)}d\omega \right] dt$$

$$= \lim_{T\to\infty} \frac{1}{2\pi T} \int_{-\infty}^{\infty} F(\omega)e^{i\omega L} \left[\int_{-T/2}^{T/2} f(t)e^{i\omega t}dt \right] d\omega$$

$$= \lim_{T\to\infty} \frac{1}{2\pi T} \int_{-\infty}^{\infty} F(\omega)F(-\omega)e^{i\omega L}d\omega$$

$$= \lim_{T\to\infty} \frac{1}{2\pi T} \int_{-\infty}^{\infty} |F(\omega)|^2 e^{i\omega L}d\omega, \tag{29}$$

where the last step uses the fact that $F(-\omega) = F^*(\omega)$. Thus, if we define the *power spectrum*, a normalized version of the amplitude spectrum,

$$P(\omega) = \lim_{T\to\infty} \frac{1}{T} |F(\omega)|^2, \tag{30}$$

we see that the auto-correlation is the inverse Fourier transform of the power spectrum:

$$R(L) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |P(\omega)| e^{i\omega L}d\omega. \tag{31}$$

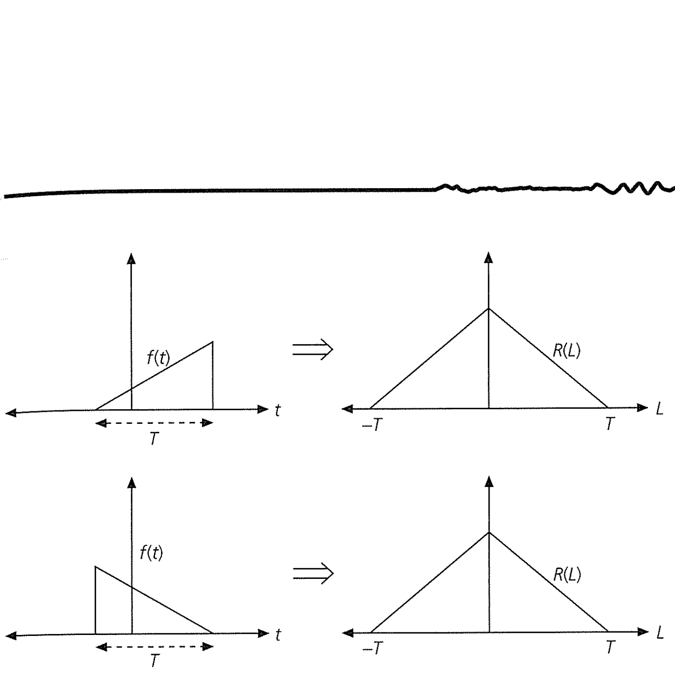


Fig. 6.3-13 Illustration showing that a function has the same auto-correlation if it is reversed in time.

As a result, the auto-correlation of a function contains information only about its amplitude spectrum, but not about its phase. Functions with the same amplitude spectrum but different phase spectra have the same auto-correlation. For example, a function has the same auto-correlation if it is reversed in time (Fig. 6.3-13).

6.4 Discrete time series and transforms

The analysis of seismic data using Fourier transforms requires computers. Thus the ground motion, a continuous function of time, is represented by a signal consisting of the ground motion measured, or *sampled*, at discrete points in time. Early seismometers, which recorded on paper wrapped around a rotating drum, yielded continuous analog seismograms which were digitized to create a discretized seismogram. Modern seismometers typically record the ground motion as a set of amplitude values measured repeatedly over a constant interval, such as 40 times per second (40 sps, "samples per second"). To work with digitized seismograms, the transforms and other mathematical operations that we formulated in Section 6.3 as continuous functions of time are replaced by discretized versions. Working with the discretized data is the subject of *digital* signal processing, whose basic ideas we discuss next.

6.4.1 *Sampling of continuous data*

The operation of sampling a signal at intervals Δt can be represented by multiplying the signal by a series of delta functions (Section 6.2.5) in time spaced Δt apart, called a *Dirac comb* or *Shah* function (Fig. 6.4-1):

$$\nabla(t;\Delta t) \equiv \sum_{n=-\infty}^{\infty} \delta(t-n\Delta t). \qquad (1)$$

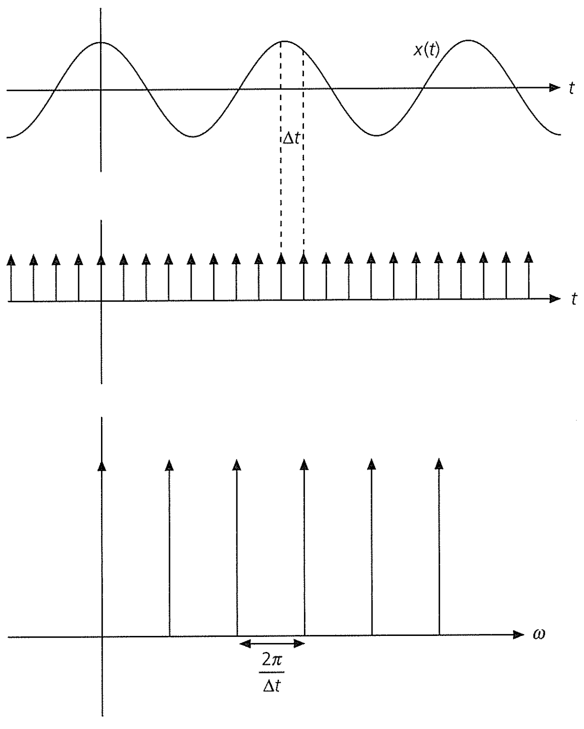


Fig. 6.4-1 Sampling a signal at intervals Δt (*top*) is described by multiplying the signal by a series of delta functions that are spaced Δt apart in time (*center*), called a Dirac comb. The transform of a Dirac comb spaced at Δt in time is a comb spaced $2\pi/\Delta t$ in angular frequency (*bottom*).

To see what this does to the spectrum of the signal being sampled, consider the Fourier transform of the Dirac comb,

$$\int_{-\infty}^{\infty} \nabla(t;\Delta t)e^{-i\omega t}dt = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(t-n\Delta t)e^{-i\omega t}dt = \sum_{n=-\infty}^{\infty} e^{-i\omega n\Delta t}, \qquad (2)$$

which was evaluated using the sifting property of the delta function (Eqn 6.2.29). It turns out that although the Fourier transform of a single delta function is a complex exponential, the transform of a Dirac comb is another Dirac comb. To see this, note that because $\nabla(t;\Delta t)$ is periodic with period Δt, it can be expanded in a complex Fourier series (Section 6.2.2),

$$\nabla(t;\Delta t) = \sum_{m=-\infty}^{\infty} F_m e^{i\omega_m t} \quad \text{for } \omega_m = 2m\pi/\Delta t, \qquad (3)$$

whose coefficients are given by

$$F_m = \frac{1}{\Delta t}\int_{-\Delta t/2}^{\Delta t/2} \nabla(t;\Delta t)e^{-i\omega_m t}dt = \frac{1}{\Delta t}\int_{-\Delta t/2}^{\Delta t/2} \sum_{n=-\infty}^{\infty} \delta(t-n\Delta t)e^{-i\omega_m t}dt. \qquad (4)$$

Because in the interval $(-\Delta t/2, \Delta t/2)$ only one delta function, $\delta(t-0)$, occurs, the Fourier coefficients are

$$F_m = \frac{1}{\Delta t}\int_{-\Delta t/2}^{\Delta t/2} \delta(t) e^{-i\omega_m t} dt = \frac{1}{\Delta t} e^{i\omega_m 0} = \frac{1}{\Delta t}; \tag{5}$$

so the Fourier series for the Dirac comb is

$$\nabla(t; \Delta t) = \frac{1}{\Delta t}\sum_{m=-\infty}^{\infty} e^{i2m\pi t/\Delta t}. \tag{6}$$

Now, consider a Dirac comb in the *frequency* domain, $\nabla(\omega; 2\pi/\Delta t)$, which consists of delta functions spaced $2\pi/\Delta t$ apart in angular frequency,

$$\nabla(\omega; 2\pi/\Delta t) \equiv \sum_{n=-\infty}^{\infty} \delta(\omega - n2\pi/\Delta t). \tag{7}$$

Its inverse transform can be evaluated using the sifting property to yield

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \nabla(\omega; 2\pi/\Delta t) e^{i\omega t} d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(\omega - n2\pi/\Delta t) e^{i\omega t} d\omega$$
$$= \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} e^{i2n\pi t/\Delta t}, \tag{8}$$

which is just $\Delta t/2\pi$ times the Fourier series for $\nabla(t; \Delta t)$ (Eqn 6). Thus the transform of a Dirac comb spaced at Δt in time is $(2\pi/\Delta t)\nabla(\omega; 2\pi/\Delta t)$, a comb spaced $2\pi/\Delta t$ in angular frequency with an amplitude of $2\pi/\Delta t$ (Fig. 6.4-1).

The effects of sampling the signal $x(t)$ at times Δt can be found by writing the sampled signal $\underline{x}(t)$ as the product of the signal and the Dirac comb in time,

$$\underline{x}(t) = x(t)\nabla(t; \Delta t). \tag{9}$$

Because multiplication in the time domain corresponds to convolution in the frequency domain, the transform of the sampled signal, $\underline{X}(\omega)$, can be written as

$$\underline{X}(\omega) = X(\omega) * (2\pi/\Delta t)\, \nabla(\omega; 2\pi/\Delta t). \tag{10}$$

Hence $X(\omega)$ is convolved with the Dirac comb, causing the spectrum of the sampled signal $\underline{X}(\omega)$ to be *periodic* in angular frequency with period $(2\pi/\Delta t)$.

To see what this does, suppose that the signal $x(t)$ is *band limited* such that its spectrum $X(\omega)$ is zero outside the principal angular frequency band $-\pi/\Delta t < \omega < \pi/\Delta t$, the range between the first delta functions on either side of the origin (Fig. 6.4-2a). Thus, after sampling, the adjacent $X(\omega)$ do not overlap (Fig. 6.4-2b), and the spectrum of the sampled time series is the same as that of the original time series in the principal frequency range.

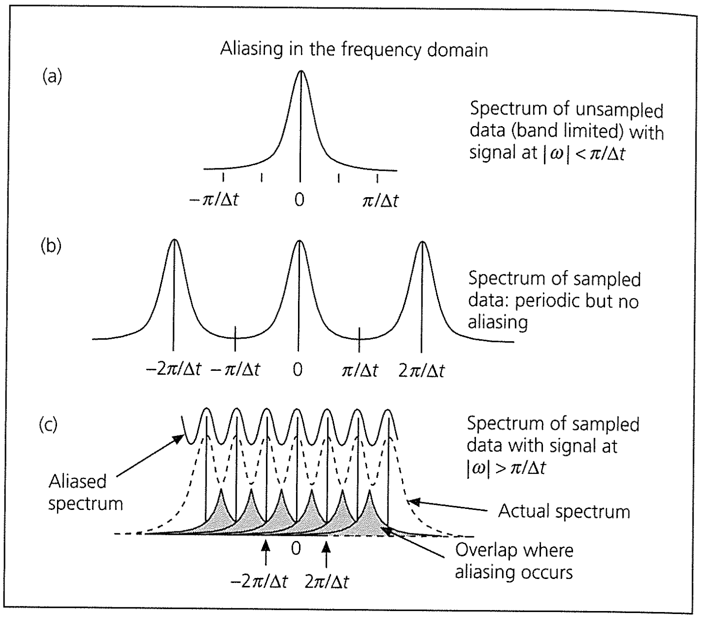

Aliasing in the frequency domain
(a)
Spectrum of unsampled data (band limited) with signal at $|\omega| < \pi/\Delta t$
$-\pi/\Delta t$
0
$\pi/\Delta t$
(b)
Spectrum of sampled data: periodic but no aliasing
$-2\pi/\Delta t$
$-\pi/\Delta t$
0
$\pi/\Delta t$
$2\pi/\Delta t$
(c)
Spectrum of sampled data with signal at $|\omega| > \pi/\Delta t$
Aliased spectrum
Actual spectrum
0
Overlap where aliasing occurs
$-2\pi/\Delta t$
$2\pi/\Delta t$

Fig. 6.4-2 Effect of sampling on the frequency amplitude spectrum. The spectrum of the unsampled signal (a) is convolved with a Dirac comb, making the spectrum of the sampled signal periodic in angular frequency with period $(2\pi/\Delta t)$. If the spectrum of the unsampled signal is zero outside the principal angular frequency band $-\pi/\Delta t < \omega < \pi/\Delta t$, the range between the first delta functions on either side of the origin, the spectrum of the sampled signal is the same as that of the original signal in this frequency range (b). Otherwise the spectra overlap after convolution (c), a phenomenon called aliasing that makes the sampled spectrum inaccurate.

On the other hand, if $X(\omega)$ is not limited to this range, the spectra overlap after sampling, so that two adjacent spectra both contribute at these frequencies (Fig. 6.4-2c). The effect of the periodicity is that for angular frequencies $|\omega| > \pi/\Delta t$, or frequencies $|f| > 1/(2\Delta t)$, the spectrum is inaccurate, because the overlap area is *folded* into the principal frequency range. This phenomenon, called *aliasing*, can be avoided by sampling the signal sufficiently densely that the spectra do not overlap. This requires that the sampling interval Δt be such that the corresponding frequency, known as the *Nyquist frequency*,

$$f_N = 1/(2\Delta t) \quad \text{or} \quad \omega_N = \pi/\Delta t, \tag{11}$$

is higher than the highest-frequency component of the signal, so that the spectrum is correctly resolved. The shorter the sampling interval, the higher the Nyquist frequency, the larger the interval over which the spectrum is periodic, and thus the higher the frequency below which the spectrum is correctly resolved. In practice, it is desirable to sample even more densely, perhaps four or more times, than the Nyquist criterion. As we sample more densely, the sampled signal becomes a better representation of the signal, and its spectrum becomes a better representation of the true spectrum.

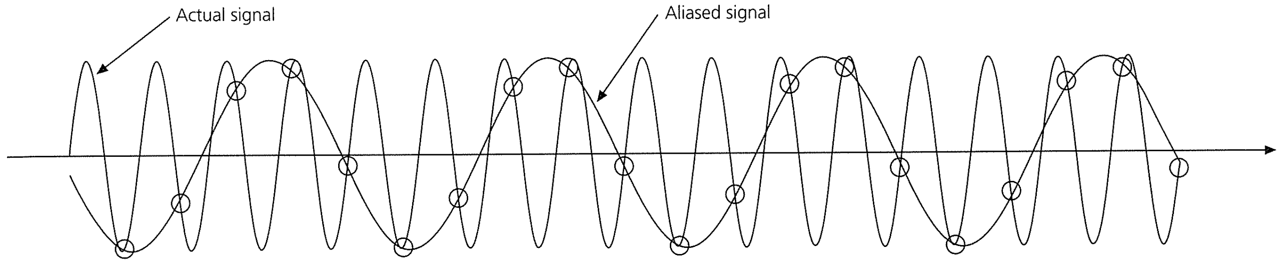


Fig. 6.4-3 In the time domain, aliasing can be viewed by noting that at least two samples per wavelength are needed to reconstruct a sinusoid accurately. Any higher frequencies are aliased into lower ones. In this case, sampling a sine wave at a sampling interval of four-fifths of the period of the wave results in an aliased signal with a period that is four times greater.

Another way to see these ideas is to note that at least two samples per wavelength are needed to reconstruct a sinusoid accurately. Any higher frequencies are aliased into lower ones (Fig. 6.4-3).[1] Aliasing occurs when the data are sampled, and once this occurs, the data cannot be "unaliased." As a result, seismic data are filtered with an analog *anti-aliasing* filter to remove frequencies above the Nyquist frequency before sampling to produce a digital seismogram.

6.4.2 The discrete Fourier transform

We now consider the Fourier transform of a sampled time series. If the function $f(t)$ is sampled at N time points that are Δt apart, the function can be represented as

$$f(t) = f(n\Delta t) \quad \text{for } n = 0, 1, \dots, N-1. \tag{12}$$

To make subsequent derivations easier, we require N to be an even number. The Fourier transform integral,

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, \tag{13}$$

can be written as a summation:

$$F(\omega) = \Delta t \sum_{n=0}^{N-1} f(n\Delta t)e^{-i\omega n\Delta t}. \tag{14}$$

This transform is a continuous function of ω that we approximate using its values at discrete frequency points. Because sampling produces a spectrum that is periodic in angular frequency with period $2\pi/\Delta t$, or twice the Nyquist angular frequency ω_N, we divide this interval into N points as

$$F(\omega) = F(k\Delta\omega) \quad \text{for } k = 0, 1, \dots, N-1, \tag{15}$$

with

$$\Delta\omega = 2\omega_N/N = 2\pi/N\Delta t = 2\pi/T, \tag{16}$$

where $T = N\Delta t$ is the total length of the data in time, sometimes called the record length. This sampled Fourier transform of a sampled time series is called the *Discrete Fourier Transform* (DFT):

$$F(k\Delta\omega) = \Delta t \sum_{n=0}^{N-1} f(n\Delta t)e^{-ik\Delta\omega n\Delta t} = \Delta t \sum_{n=0}^{N-1} f(n\Delta t)e^{-ikn2\pi/N}. \tag{17}$$

The DFT gives values at angular frequencies

$$0, \Delta\omega, 2\Delta\omega, \dots (N/2)\Delta\omega, \dots (N-1)\Delta\omega. \tag{18}$$

The second half of the values represent angular frequencies greater than $(N/2)\Delta\omega$, which equals the Nyquist angular frequency. These points correspond to the negative angular frequencies, wrapped around to follow the positive angular frequencies. For example, the first point after the Nyquist angular frequency occurs for angular frequency

$$\begin{aligned}(N/2+1)\Delta\omega &= (N/2)\Delta\omega + \Delta\omega = \omega_N + \Delta\omega \\ &= -\omega_N + \Delta\omega = -\left(\frac{N}{2} - 1\right)\Delta\omega,\end{aligned} \tag{19}$$

where we use the fact that the spectrum is periodic with period $2\omega_N$. Each successive point corresponds to an increment of $-\Delta\omega$. Thus, we can consider the DFT to give values at angular frequencies

$$0, \Delta\omega, 2\Delta\omega, \dots \left(\frac{N}{2} - 1\right)\Delta\omega, \omega_N, -\left(\frac{N}{2} - 1\right)\Delta\omega, \dots, \\ -2\Delta\omega, -\Delta\omega. \tag{20}$$

Graphically, we can think of folding the second half of the DFT about zero frequency to give the values of the spectrum at the negative frequencies (Fig. 6.4-4).

[1] An illustration of sampling issues is that in Western films, wagon wheels sometimes appear to rotate backwards, stop, or rotate only slowly forward. These effects result from differences between the wheels' rotation rate and the movie cameras' sampling rate, typically 24 frames per second.

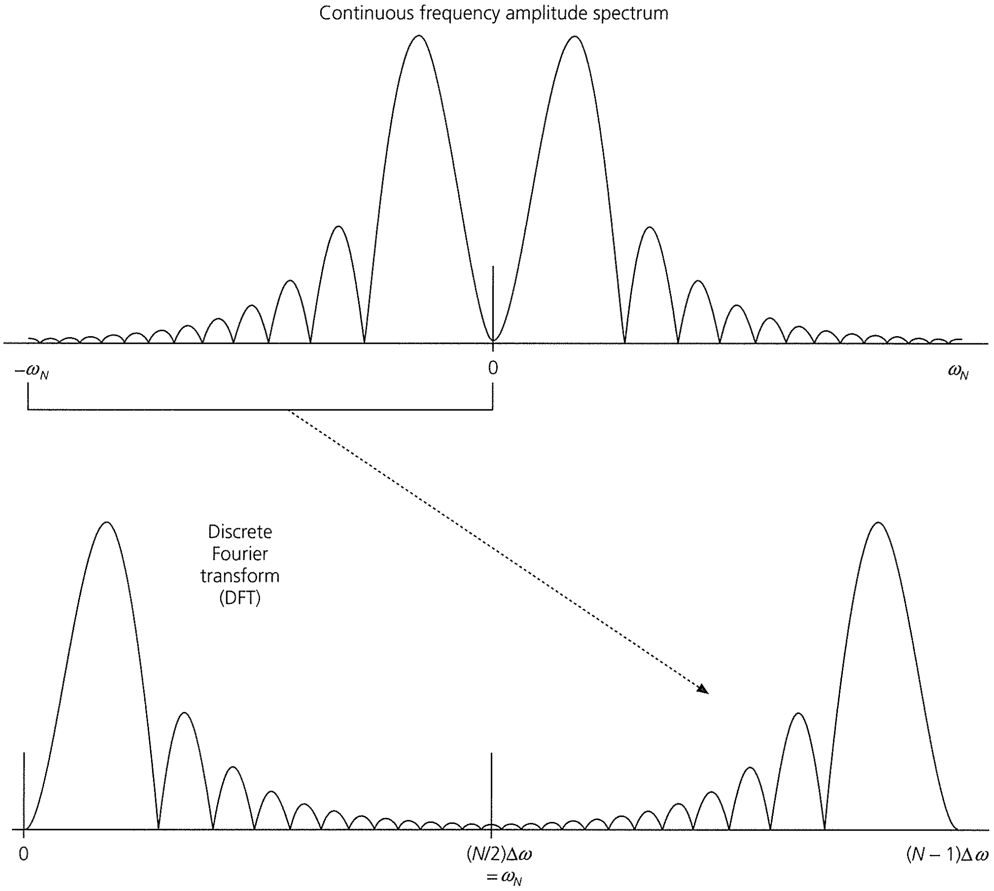


Fig. 6.4-4 Due to the periodicity of the discrete Fourier transform, the second half of the values of the frequency amplitude spectrum, at angular frequencies greater than the Nyquist angular frequency $(N/2)\Delta\omega$, represents the negative angular frequencies.

The fact that the DFT is the sampled spectrum of a sampled time series has two interesting consequences. The *highest* angular frequency that can be resolved is the Nyquist, which depends inversely on the sampling rate in time, because $\omega_N = \pi/(\Delta t)$. On the other hand, the *resolution* in frequency, given by the spacing between successive angular frequency points, $\Delta\omega = 2\pi/(N\Delta t)$, depends inversely on $T = N\Delta t$, the total record length.

For example, to resolve the singlets making up the normal mode multiplet ${}_0S_2$ (Fig. 2.9-16), we would like a frequency resolution of at least 0.0001 cycles/minute, or 1.7×10^{-6} s^{-1}. This requires data extending for $1/1.7 \times 10^{-6}$ s, or more than 160 hours, after the earthquake. However, because the mode's period is 54 minutes, a seismogram sampled every few minutes would be adequate and give a manageable number of data points. We need, however, to prevent aliasing due to surface and body waves that have periods of tens to hundreds of seconds. An easy way to do this would be to start the analysis a day or so after the earthquake, when the shorter-period waves have decayed due to attenuation. This approach uses the earth's anelasticity as a natural anti-aliasing filter. By contrast, reflection seismology requires high temporal resolution to resolve closely spaced interfaces, so reflection data are sampled at high rates such as 250 times per second after an anti-aliasing filter is applied.

By analogy to the DFT, we write the inverse DFT (IDFT) by approximating the inverse Fourier transform integral

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega \tag{21}$$

in the same way, which gives

$$\begin{aligned} f(n\Delta t) &= \frac{1}{2\pi}\sum_{k=0}^{N-1} F(k\Delta\omega)e^{i(k\Delta\omega)(n\Delta t)}\Delta\omega \\ &= \frac{\Delta\omega}{2\pi}\sum_{k=0}^{N-1} F(k\Delta\omega)e^{ikn2\pi/N} \\ &= \frac{1}{N\Delta t}\sum_{k=0}^{N-1} F(k\Delta\omega)e^{ikn2\pi/N}. \end{aligned} \tag{22}$$

An interesting feature of the IDFT comes from the fact that it samples the spectrum at discrete frequencies $\Delta\omega$. Sampling the time series at Δt causes the phenomenon of aliasing, because the spectrum is periodic in angular frequency with period $2\pi/(\Delta t)$. By analogy, sampling the frequency spectrum at $\Delta\omega$ makes the time series periodic with a period of

$$\frac{2\pi}{\Delta\omega} = \frac{2\pi}{2\pi/(N\Delta t)} = (N\Delta t) = T, \qquad (23)$$

which is equal to the original record length.[2] This *wraparound* phenomenon can be important, as we shall see when discussing the use of DFTs to carry out convolutions.

6.4.3 Properties of DFTs

For simplicity, we write the DFT and the inverse DFT implicitly assuming a unit sampling interval, $\Delta t = 1$, and define

$$F(k) \equiv F(k\Delta\omega) = \sum_{n=0}^{N-1} f(n)e^{-2\pi ikn/N} \quad \text{for } k \text{ and } n = 0, 1, \ldots, N-1 \qquad (24)$$

$$f(n) \equiv f(n\Delta t) = \frac{1}{N}\sum_{k=0}^{N-1} F(k)e^{2\pi ikn/N} \quad \text{for } k \text{ and } n = 0, 1, \ldots, N-1. \qquad (25)$$

The two equations are very similar in form and are easy to evaluate — the forward and inverse transforms differ only in the sign of the exponential and the $1/N$ normalization. This is especially clear if we define the complex exponential as $W = e^{-2\pi i/N}$, so the definitions of the DFT and IDFT become

$$F(k) = \sum_{n=0}^{N-1} f(n)W^{kn} \quad \text{and} \quad f(n) = \frac{1}{N}\sum_{k=0}^{N-1} F(k)W^{-kn}. \qquad (26)$$

The terms with the complex exponential are periodic in N,

$$W^{kn} = W^{(N+k)n} = W^{k(N+n)}, \qquad (27)$$

so the DFT and IDFT can be defined for all integers k, n, j as

$$f(n) = f(jN + n), \quad F(k) = F(jN + k). \qquad (28)$$

A formal statement of the relation between the negative and positive frequencies can also be given as

$$f(-n) = f(N - n), \quad F(-k) = F(N - k). \qquad (29)$$

We used this relation when we explained how the second half of the DFT corresponds to negative frequencies (Fig. 6.4-4).

Using these definitions, we can show that the discrete transforms have properties that we discussed for the continuous transforms in Section 6.2.4:[3]

(1) The DFT and IDFT are linear: if $A(k)$ and $B(k)$ are the transforms of time series $a(n)$ and $b(n)$, then $\alpha A(k) + \beta B(k)$ is the transform of $\alpha a(n) + \beta b(n)$. Thus we can use the discrete transforms to model linear systems.

(2) The DFT of a real time series (i.e., one for which $f(n) = f^*(n)$) has the symmetry

$$F(-k) = F(N - k) = F^*(k). \qquad (30)$$

Thus, as with the continuous transform, the values for the negative frequencies are the conjugates of those for the positive frequencies.

(3) Shifting a time series in time simply changes the phase of the DFT: if the transform of $f(n)$ is $F(k)$, the DFT of $f(n - j)$ is $W^{kj}F(k)$. Similarly, shifting a Fourier transform in frequency changes the phase of the IDFT: the inverse transform of $F(k - m)$ is $W^{-mn}f(n)$.

6.4.4 The fast Fourier transform (FFT)

For these concepts to be useful, the transforms and inverse transforms must be evaluated on a computer. Moreover, it only makes sense to carry out filtering using Fourier transforms if the transform and inverse transform operations are relatively quick. It turns out that an elegant algorithm known as the Fast Fourier Transform (FFT) provides a fast way of carrying out the DFT and IDFT.

The time a computer needs to carry out an algorithm depends on how many arithmetic operations are needed. We would expect that evaluating all N points in the DFT, each of which is the sum of the N terms in the series, would require approximately N^2 operations. The FFT algorithm, however, requires a much smaller number of operations, approximately $N \log_2 N$. The difference is substantial; for $N = 4096$, $N^2 = 16{,}777{,}216$, but $N \log_2 N = 49{,}152$ — about 340 times fewer! As a consequence, the introduction of the FFT made digital signal processing common in seismology and many other disciplines.

Entire books have been written about the FFT, so we only briefly sketch the approach here. The underlying idea is that a simple method can be used to compute the transform of a series of points by splitting it in half. We take a series with N points,

$$f(n) \quad \text{for } n = 0, 1, \ldots, N-1 \qquad (31)$$

and form two subseries, one with the odd-numbered points and one with the even-numbered points:

$$a(n) = (f(0), f(2), f(4), \ldots) = f(2n) \quad \text{for } n = 0, 1, \ldots, N/2 - 1,$$
$$b(n) = (f(1), f(3), f(5), \ldots) = f(2n+1). \qquad (32)$$

The DFTs of the two subseries are

$$A(k) = \sum_{n=0}^{N/2-1} a(n)e^{-4\pi ikn/N} \quad \text{and}$$

[2] Because of this periodicity, the record length is considered to be $N\Delta t$ rather than $(N-1)\,\Delta t$.

[3] As for the continuous transforms, the proofs are left for the problems.

$$B(k) = \sum_{n=0}^{N/2-1} b(n)e^{-4\pi ikn/N}, \tag{33}$$

where k goes from 0 to $N/2 - 1$, and the factor of 4 comes from the fact that the subseries lengths are $N/2$.

The DFT of the original series can be written in terms of the DFTs of the subseries,

$$\begin{aligned} F(k) &= \sum_{n=0}^{N-1} f(n)e^{-2\pi ikn/N} \\ &= \sum_{n=0}^{N/2-1} [a(n)e^{-2\pi ik(2n)/N} + b(n)e^{-2\pi ik(2n+1)/N}] \\ &= A(k) + e^{-2\pi ik/N}B(k) \quad \text{for } k = 0, 1, \ldots, N/2 - 1, \end{aligned} \tag{34}$$

giving the first $N/2$ points of $F(k)$. The second $N/2$ points come from replacing k by $k + N/2$,

$$F(k + N/2) = A(k + N/2) + e^{-2\pi i(k+N/2)/N} B(k + N/2), \tag{35}$$

and noting that, because the DFTs of the subseries are periodic with a period equal to their length, $N/2$,

$$A(k + N/2) = A(k) \quad \text{and} \quad B(k + N/2) = B(k). \tag{36}$$

Because the exponential can be written as

$$e^{-2\pi i(k+N/2)/N} = e^{-\pi i}\, e^{-2\pi ik/N} = -e^{-2\pi ik/N}, \tag{37}$$

the second half of the transform can be found from the first, using

$$F(k + N/2) = A(k) - e^{-2\pi ik/N}B(k). \tag{38}$$

In terms of $W = e^{-2\pi i/N}$, the expressions for the two parts of the transform (Eqns 34 and 38) have the simple form of

$$F(k) = A(k) + W^kB(k) \quad \text{and} \quad F(k + N/2) = A(k) - W^kB(k). \tag{39}$$

This method is called *doubling* — finding the transform of an N-point series from the transforms of its two $N/2$-point subseries. Doubling can be applied recursively, because we can find the transform of each $N/2$-point series from that of two $N/4$-point series, etc. Ultimately, a series of length $N = 2^n$ can be evaluated via $n = \log_2 N$ such stages. In the final stage, the transform of each 2-point series is found from two 1-point series, but the transform of a 1-point series is itself. Various methods can be used to further speed up operations.

Thus, to obtain the FFT of a time series, we treat the data points as N 1-point series, use doubling to form $(N/2)$ 2-point series, and so on until the final N-point transform. The same FFT algorithm can also be used to take the inverse transform. Commonly, the same computer program is used for both forward and inverse FFTs, except that the sign of the exponential must be changed and the $1/N$ normalization remembered (the last being a traditional bane of students).

In using the FFT to transform data as part of a filtering operation, the factor of $1/N$ may be included at any step in the process. Often, however, we use the FFT to obtain the Fourier transform of a time series, and compare this to a result derived in the frequency domain, such as an analytic expression for a synthetic seismogram as a function of ω. In this case, we have to consider the units of both the forward and the inverse DFT. The forward DFT is an approximate way of evaluating the Fourier transform integral (Eqn 13), in which the differential dt is replaced by the difference Δt. Thus, the FFT results are multiplied by Δt. Similarly, the IDFT approximates the inverse transform integral (Eqn 21), with the differential $d\omega$ replaced by the difference $\Delta\omega$. Hence the results from inverting the FFT are multiplied by $\Delta\omega/(2\pi)$. The product of these two factors is $\Delta\omega\Delta t/2\pi = 1/N$, as expected.

This discussion assumes that the series length N is a power of 2. If this is not the case, a number of zeroes necessary to obtain a power of 2 can be added to the end of the time series. Such *zero padding* has the effect of sampling the spectrum more densely, because the sample interval is unchanged, but the frequency interval $\Delta\omega = 2\pi/(N\Delta t)$ decreases. Despite the denser sampling, the real resolution in frequency is not increased beyond that resulting from the real (nonzero) data length. Instead, smooth interpolation is done within the range of actual resolution $\Delta\omega_{\text{real}} = 2\pi/T_{\text{nonzero}}$.

Finally, it is worth distinguishing between the DFT and the FFT. The DFT is the discrete approximation to the Fourier transform which has the periodic properties we have discussed. The FFT is a clever method for computing the DFT with many fewer operations.

6.4.5 Digital convolution

As discussed in Section 6.3.2, the convolution is used in many seismological applications. This operation has some special features when carried out with discretized time series and their transforms.

Given two discrete time series with unit sample period, $x(m)$ with M points $x(0), x(1), \ldots, x(M - 1)$ and $f(n)$ with N points $f(0), f(1), \ldots, f(N - 1)$, the convolution in the time domain is written, by analogy to the integral definition, as

$$y(t) = x(t) * f(t) = \sum_{m=0}^{M-1} x(m)f(t - m). \tag{40}$$

We evaluate the summation for each value of t that yields a nonzero value. Because $f(n)$ is zero for n outside the range $(0, N - 1)$ and $x(m)$ is zero for m outside the range $(0, M - 1)$, there are $N + M - 1$ terms in the convolution, and $y(t)$ is defined for $t = 0, 1, \ldots, N + M - 2$. For example, if $N = 3$ and $M = 4$, the $3 + 4 - 1 = 6$ terms are

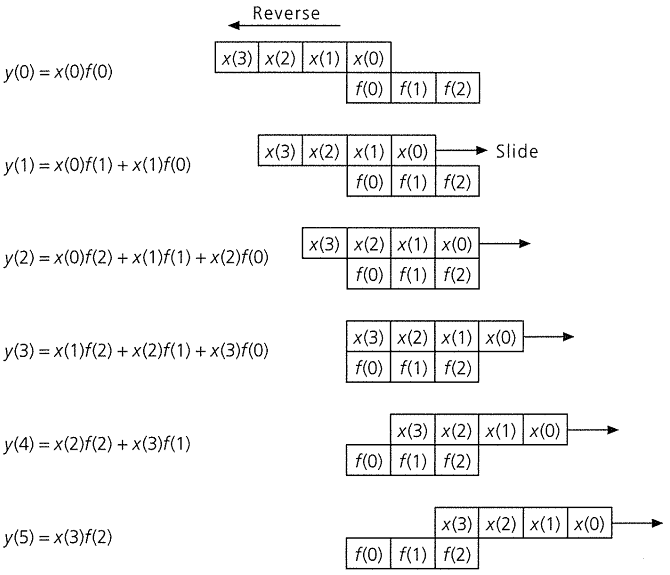

Reverse
y(0) = x(0)f(0)
x(3) x(2) x(1) x(0)
f(0) f(1) f(2)
y(1) = x(0)f(1) + x(1)f(0)
Slide
y(2) = x(0)f(2) + x(1)f(1) + x(2)f(0)
y(3) = x(1)f(2) + x(2)f(1) + x(3)f(0)
y(4) = x(2)f(2) + x(3)f(1)
y(5) = x(3)f(2)

Fig. 6.4-5 Schematic diagram of a time domain convolution of two sampled time series as a reverse, multiply, and slide operation.

$$
\begin{aligned}
&y(0)=x(0)f(0)\\
&y(1)=x(0)f(1)+x(1)f(0)\\
&y(2)=x(0)f(2)+x(1)f(1)+x(2)f(0)\\
&y(3)=x(1)f(2)+x(2)f(1)+x(3)f(0)\\
&y(4)=x(2)f(2)+x(3)f(1)\\
&y(5)=x(3)f(2). \qquad (41)
\end{aligned}
$$

We can think of this operation as reversing the order of $x(m)$ and sliding it past $f(n)$, while conducting all nonzero multiplications (Fig. 6.4-5).

These formulations show that the convolution has more terms than either of the time series being convolved. This has some interesting consequences if we do the convolution in the frequency domain. Because the data are sampled at discrete intervals, convolution in the frequency domain requires taking two discrete Fourier transforms, multiplying them, and then taking the inverse discrete Fourier transform. If $Y(k)$, $X(k)$, and $F(k)$ are the DFTs of $y(t)$, $x(m)$, and $f(k)$, then

$$Y(k)=X(k)F(k) \qquad (42)$$

gives the complex spectrum at each angular frequency. This brings out an important point; all the DFTs must be defined at the same frequencies. For a time series of length N with unit sample period ($\Delta t=1$), the angular frequencies in the DFT are

$$k\Delta\omega=k2\pi/N \quad \text{for } k=0,1,\ldots,N-1. \qquad (43)$$

Because $x(m)$ and $f(n)$ have different lengths, the points in the two transforms would correspond to different angular frequencies. To avoid this, the two time series are extended with zeroes at their ends, so that their lengths equal the same power of 2.

A further point to bear in mind is that the time series corresponding to the convolution is longer than either of the two series that are convolved. If the number of points in the DFT is less than this length, a wraparound phenomenon similar to aliasing occurs when we invert the transform, due to the periodicity resulting from the sampled transform. The two time series thus need to be extended to a length at least that of their convolution before their DFTs are taken.

6.5 Stacking

Seismology uses data to estimate quantities that describe the earth and seismic sources. Ideally these estimates are both *accurate* and *precise*. Accuracy measures the deviation of the estimate from its true value, whereas precision measures the repeatability of individual estimates. Hence the accuracy depends on systematic errors that bias groups of estimates, whereas the precision depends on random errors that affect individual estimates. Estimates can be precise but inaccurate, or accurate but imprecise. For example, an estimate of an earthquake's location depends on the quality of the travel time data used and the accuracy of the velocity model. High-quality travel time data, together with an incorrect velocity model, can yield a location that is precise in that the data are well fit and so imply small uncertainty, but inaccurate in that the resulting location is not where the earthquake occurred. In such a case the true uncertainty exceeds the formal uncertainty inferred from how well the model fits the data. Conversely, an accurate velocity model and poor travel time data can give a location that is accurate in that it is close to where the earthquake occurred, but imprecise in that the location has a large uncertainty and there are large misfits to the data.

Approaches to improving the accuracy and precision of estimates are often couched in terms of measuring a quantity like the length of a table. Accuracy is improved by using different measuring tools, ideally calibrated against each other. Precision is improved by making multiple measurements, ideally by different people. We follow such approaches for the earth when possible, but face additional complexities. For example, an earthquake is a nonrepeatable experiment, so we cannot make additional measurements. We can use different techniques, but still face difficulties. A case in point is that estimates of an earthquake's depth from travel times and waveform modeling are only partially independent. Both can be biased similarly by incorrect assumptions about the near-source velocity, but the travel times are independent of the assumed source mechanism, and the waveform modeling (which depends on relative arrival times) would not be biased by an error in the absolute timing of individual seismograms.

A further complexity is that different methods can measure related but not identical entities: the earthquake depth ranges inferred from travel times, waveform modeling, aftershock locations, and geodesy differ somewhat, because each measures related but not identical quantities.

Most discussions of these issues focus on random errors because they are easy to estimate from the scatter of measurements. However, it is worth bearing in mind that systematic errors not included in these error estimates can be more significant, as discussed in Section 1.1.2. Systematic errors can come about in surprising ways and have subtle and crucial effects. For example, we have noted that velocity heterogeneities can perturb ray paths and thus bias earthquake focal mechanisms (Section 3.7.3); attenuation variations can bias estimates of the yields of nuclear explosions (Section 1.2.8); errors in the paleomagnetic time scale can bias estimates of plate motions (Section 5.2.2); and effects including an undetected earthquake can change estimates of earthquake recurrence from paleoseismology (Section 1.2.5). Systematic biases are difficult to detect, but sometimes are identified from discrepancies between different approaches. For example, the discrepancy between earth models derived from body waves and those from normal modes suggests physical dispersion due to anelasticity (Section 3.7.8), and the discrepancy between oceanic Love and Rayleigh wave velocities points toward anisotropy (Section 3.6.5). Hence, when data are discordant, as in the differences in earthquake frequency–magnitude relations derived from seismological and paleoseismic data (Section 4.7.1), systematic bias is one possible cause.

In this section, we develop some general ideas about errors and consider some examples. Our focus is one of the most useful methods for improving estimates from seismological data: *stacking*, or taking multiple measurements and averaging them. We do this either by averaging measurements such as travel times from different seismograms, or by adding many seismograms and then estimating parameters. This process has two effects. First, it improves precision by reducing the effects of random noise in the data. Second, if the data are averaged in specific ways, the precision, and perhaps accuracy, can be improved by suppressing some features of the data and thus enhancing desired features.

6.5.1 *Random errors*

We seek to estimate a quantity x from multiple measurements, each of which gives a value x_i due to noise and the limitations of the measurements. With enough measurements, a pattern generally emerges in which the values x_i are distributed about a value x'. If we neglect systematic errors of measurement, we can estimate the value of x from the measured values x_i and say something about how this estimate is related to the unknown true value of x.

For this purpose we view the measured values x_i as random samples from a parent distribution described by the probability density function $p(x)$ that gives the probability of observing a certain value. For example, in Section 4.7.3 we treated the occurrence of earthquakes as samples from a parent distribution of recurrence times. That example illustrated that in most applications it is not clear what the most suitable parent distribution is. It is common to assume that the parent distribution is a Gaussian distribution, also called the "normal distribution," because it often describes the frequencies at which very different phenomena occur. A famous result called the *central limit theorem* shows that this is because a sum of random numbers approaches a Gaussian distribution even if the random numbers are derived from other probability distributions.

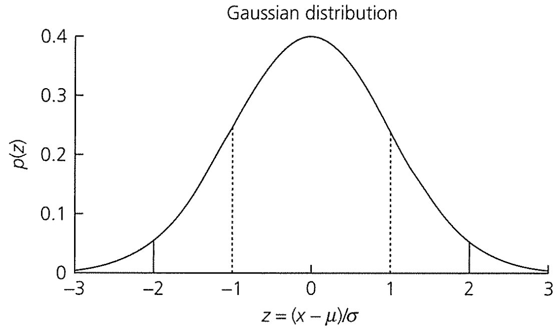


Fig. 6.5-1 Probability density function for a Gaussian distribution with mean μ and standard deviation σ. Ranges within one and two standard deviations of the mean are shown by vertical lines.

For a Gaussian distribution, the probability that the i^{th} measurement would yield a value in the interval $x_i \pm dx$, in the limit as $dx \to 0$, is

$$p(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]. \tag{1}$$

The distribution is thus characterized by two parameters: the mean, μ, and the standard deviation, σ. The most probable measurement is the mean value, and values on either side of it are less likely the further from the mean they are. The distribution is often written as a function of the normalized variable $z = (x - \mu)/\sigma$,

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\,[-z^2/2]. \tag{2}$$

Figure 6.5-1 shows the familiar "bell curve" that results.

A common application is to estimate how likely a measurement is to be within a range z from the mean. To do this, we integrate the probability density function to find the cumulative probability

$$A(z) = \int_{-z}^{z} p(y)dy = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} \exp\,[-y^2/2]dy. \tag{3}$$

For $z = 1$, we get $A(z) = 0.68$, indicating that there is a 68% probability that a measurement will be within one standard deviation of the mean. Similarly, $A(2) = 0.95$ and $A(3) = 0.997$, indicating a 95% probability that a measurement will be within two standard deviations of the mean, and a greater than 99% probability that it will be within three standard deviations. We used such ideas in estimating earthquake probabilities (Section 4.7.3).

We expect that if we made an infinite number of measurements (samples) without any systematic biases, a histogram of the measurements would look like the parent distribution. The mean of the observed values will be the mean of the distribution

$$\mu = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right], \tag{4}$$

and the *spread* of the measurements is the *variance* (standard deviation squared) of the distribution,

$$\sigma^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2\right]. \tag{5}$$

Thus, if the assumptions we have made are valid, the mean of a large number of measurements, μ, would be the value that we seek.

The difficulty in reality is that only a limited number of measurements are available to estimate μ. As a result, the actual mean μ' is not necessarily equal to μ. We thus ask what method of deriving μ' from the measurements gives the maximum likelihood that μ' is actually the mean of the parent distribution.

To find this, we assume that the parent distribution had mean μ' and standard deviation σ, so the probability that the i^{th} measurement would yield a value in the interval $x_i \pm dx$ in the limit as $dx \to 0$ is

$$p_i(\mu') = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i - \mu'}{\sigma}\right)^2\right]. \tag{6}$$

For N observations, the probability of observing a particular set of values x_i is the product of the probabilities that each individual measurement would have that particular value,

$$p(\mu') = \prod_{i=1}^{N} P_i(\mu') = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^N \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu'}{\sigma}\right)^2\right]. \tag{7}$$

The most probable value of μ' is the one that maximizes $p(\mu')$, the probability of obtaining the set of measurements actually found. To find this value, we set the derivative of the argument of the exponential equal to zero,

$$0 = \frac{d}{d\mu'}\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu'}{\sigma}\right)^2\right] = -\frac{1}{2}\sum_{i=1}^{N}\frac{d}{d\mu'}\left[\frac{x_i - \mu'}{\sigma}\right]^2, \tag{8}$$

which occurs for

$$\sum_{i=1}^{N}[x_i - \mu'] = 0, \tag{9}$$

or

$$\mu' = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{10}$$

This is not surprising — the average value of x_i is the best estimate of the mean. An interesting question is what is the standard deviation σ_N of this estimate of μ'? Specifically, how does the uncertainty associated with this estimate compare to the uncertainty of each individual measurement?

To answer this, we use the *propagation of errors*, a general method for finding the relation between the uncertainty in a function and the uncertainty in the variables that it depends on. If z is a function of multiple variables, then

$$z = f(u, v, \ldots), \tag{11}$$

and we have N measurements of $(u, v, \ldots)$. The mean value of the function is its value for the mean of the arguments,

$$\bar{z} = f(\bar{u}, \bar{v}, \ldots), \tag{12}$$

and its variance is

$$\sigma_z^2 = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2. \tag{13}$$

If we expand z in a Taylor series about its mean value,

$$z_i - \bar{z} = (u_i - \bar{u})\frac{\partial z}{\partial u} + (v_i - \bar{v})\frac{\partial z}{\partial v} + \ldots, \tag{14}$$

so

$$\begin{aligned}\sigma_z^2 &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left[(u_i - \bar{u})\frac{\partial z}{\partial u} + (v_i - \bar{v})\frac{\partial z}{\partial v} + \ldots\right]^2 \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left[(u_i - \bar{u})^2\left(\frac{\partial z}{\partial u}\right)^2 + (v_i - \bar{v})^2\left(\frac{\partial z}{\partial v}\right)^2 + \ldots\right. \\ &\quad \left. + 2(u_i - \bar{u})\frac{\partial z}{\partial u}(v_i - \bar{v})\frac{\partial z}{\partial v} + \ldots\right].\end{aligned} \tag{15}$$

To simplify this expression, we use the variances of each variable about its mean

$$\sigma_u^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} (u_i - \bar{u})^2 \quad \text{and} \quad \sigma_v^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} (v_i - \bar{v})^2 \qquad (16)$$

and the *covariances* that describe how fluctuations between variables are correlated:

$$\sigma_{uv}^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} (u_i - \bar{u})(v_i - \bar{v}). \qquad (17)$$

Substituting Eqns 16 and 17 into Eqn 15 gives

$$\sigma_z^2 = \sigma_u^2\left(\frac{\partial z}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial z}{\partial v}\right)^2 + \ldots + 2\sigma_{uv}^2\left(\frac{\partial z}{\partial u}\right)\left(\frac{\partial z}{\partial v}\right) + \ldots \qquad (18)$$

This relation, called the propagation of errors equation, illustrates that the extent to which the uncertainty in each variable contributes to the uncertainty in a function depends on the partial derivative of the function with respect to that variable. We often assume that the variations in the different variables are uncorrelated (which is not always the case), so we set the covariances equal to zero, and simplify the variance of z to

$$\sigma_z^2 = \sigma_u^2\left(\frac{\partial z}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial z}{\partial v}\right)^2 + \ldots \qquad (19)$$

This result is a general one that we have already mentioned in the context of estimating the uncertainty of geodetic rates (Eqn 4.5.8) and earthquake source parameters (Eqn 4.6.23).

In the specific application here, we consider the mean to be a function of the observations,

$$z = \mu' = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad (20)$$

so the error propagation equation can be used with $(u, v, \ldots) = x_i$. Assuming that the variables are independent, so their errors are uncorrelated, we get

$$\sigma_{\mu'}^2 = \sum_{i=1}^{N} \sigma_{x_i}^2\left(\frac{\partial \mu'}{\partial x_i}\right)^2 = \sum_{i=1}^{N} \sigma_{x_i}^2\left(\frac{1}{N}\frac{\partial}{\partial x_i}\sum_{i=1}^{N} x_i\right)^2 = \frac{1}{N^2}\sum_{i=1}^{N} \sigma_{x_i}^2. \qquad (21)$$

If all the observations have equal uncertainties ($\sigma_{x_i}^2 = \sigma^2$), then

$$\sigma_{\mu'}^2 = \sigma^2/N. \qquad (22)$$

Thus the variance of the mean is $1/N$ times the variance of the individual measurements. Hence making N measurements reduces the standard deviation of the mean by $1/\sqrt{N}$. This is the basic idea behind stacking; averaging multiple measurements of some quantity yields an estimate that has a smaller uncertainty than the individual measurements.

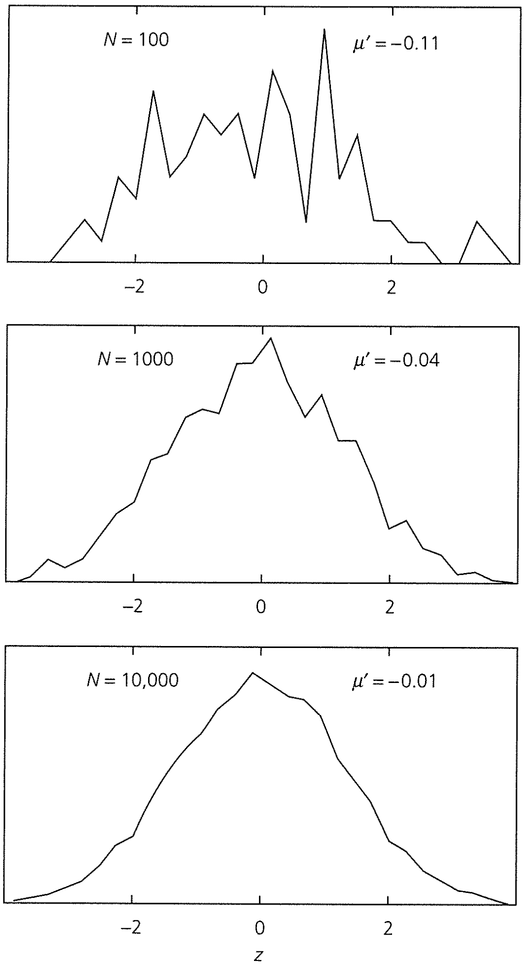


Fig. 6.5-2 Results of drawing N samples from a Gaussian parent distribution with mean zero and a unit standard deviation. For small numbers of samples, the observed distribution can look quite different from the parent distribution, and the sample mean μ' differs from that of the parent distribution. As the number of samples increases, the observed distribution looks increasingly like the parent distribution.

Figure 6.5-2 illustrates this idea. We assume that measurements of some quantity are described by a Gaussian parent distribution with a mean of zero, and we try to estimate this quantity with different numbers of samples. As the number of samples increases, the distribution of samples looks increasingly like the parent distribution, and the sample mean approaches the mean of the parent distribution. However, for a small number of samples, the observed distribution can look quite different from the parent distribution. This issue arises in studying earthquake recurrence, where the few samples available make it difficult to assess whether apparent differences in earthquake history (Section 4.7.1) are significant and what parent distributions and parameters should be used to estimate earthquake probabilities (Section 4.7.3).

This simple Gaussian model is widely used in analyzing data. We assume that each measurement includes the quantity of interest and some *noise*, defined as the portion of the signal that is not of interest. The noise thus reflects both true errors of measurement and processes not under consideration, all of which are assumed to be uncorrelated between measurements. To the extent that these assumptions are valid, stacking data will improve the signal. The random, uncorrelated noise idea often seems to be a good approximation. However, if noise is correlated between measurements, as can occur if the measurement equipment is biased or an "error" source is otherwise common to the measurements, the desired noise reduction will be less. For instance, the structure under a seismometer is studied by means of receiver functions that are derived using the radial and vertical components (Fig. 6.3-7), assuming that the noise on each is uncorrelated. However, noise due to microseismic activity (Section 6.6.3) will be correlated between components and hence can yield spurious layering.

6.5.2 *Stacking examples*

A simple stacking approach is to add seismograms at nearby stations, assuming that they contain a common signal of interest plus "noise" that differs between stations. The noise includes differences in the response of the seismometers and differences in the seismograms generated by the interaction between the upcoming waves and the crustal structure under each seismometer. If the seismometers and crustal structure are similar enough, stacking seismograms should reduce the noise and yield a better representation of the signal of interest than the individual seismograms.

An extension of this idea is used for seismograms at different places or times. If we know theoretically how the signal of interest varies as a function of position or time, we can correct the data to a common position or time and stack them. For example, in CMP stacking of reflection seismic data, traces with a common midpoint are shifted by a time corresponding to the travel time curve of a reflection and then stacked (Section 3.3.4). The reflected arrivals are in phase and thus enhanced, whereas other arrivals with different travel time curves are out of phase and thus suppressed. Although the undesired arrivals are not random noise, they are reduced relative to the reflected arrivals. Random noise in the data is also reduced.

This approach is also useful in observing deeper earth structures, such as mantle discontinuities (Section 3.5.3). Figure 6.5-3 shows an example of stacking large numbers of long-period transverse-component seismograms to enhance precursors to the *SS* arrivals. The precursors, $S_{410}S$, $S_{520}S$, and $S_{660}S$, are underside reflections from the discontinuities at 410, 520, and 660 km depths. However, these phases are weak and are not easily observed above the noise on individual seismograms. Stacking many records enhances these arrivals, allowing the depths of the discontinuities to be studied. Moreover, after removal of the theoretical signals of $S_{410}S$ and $S_{660}S$ (Fig. 6.5-3, *middle*), the stacked record shows the $S_{520}S$ arrival

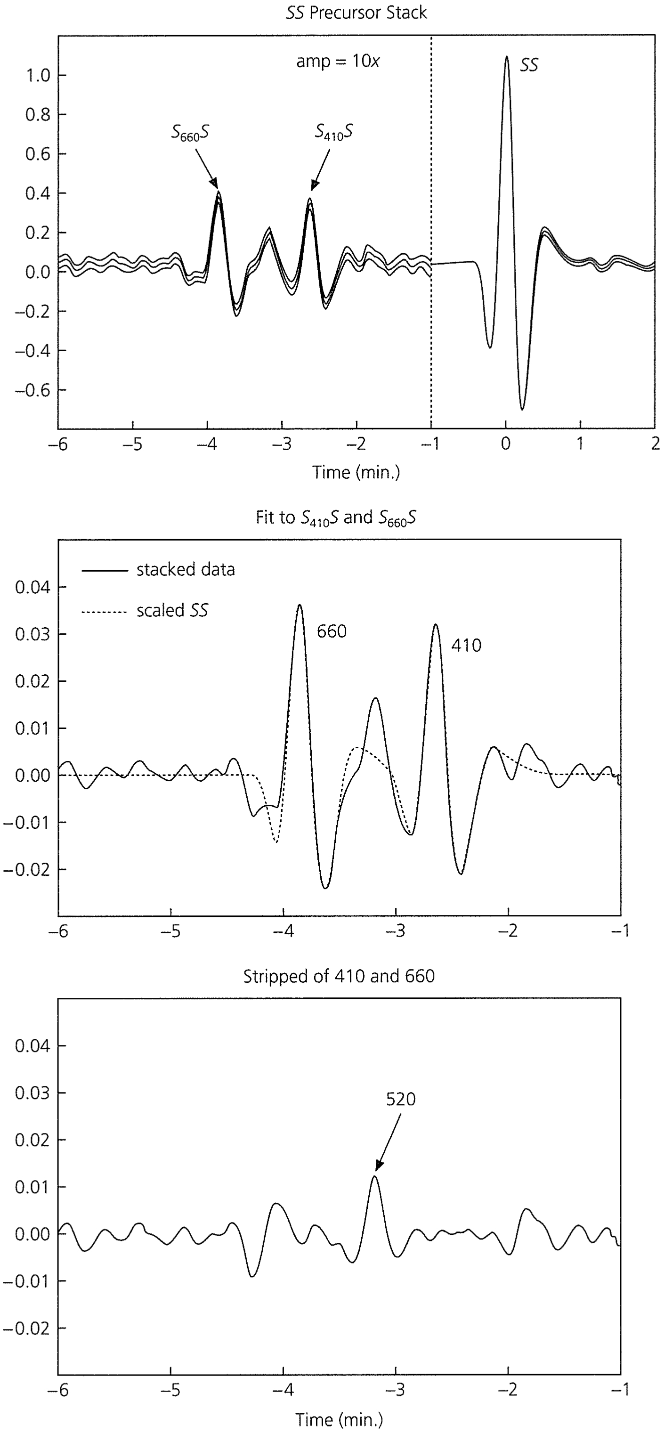


Fig. 6.5-3 Stacking long-period seismograms to identify the depth of mantle discontinuities by enhancing precursors to *SS*. The initial stack (*top*) shows the $S_{410}S$ and $S_{660}S$ underside reflections off the 410 km and 660 km discontinuities, magnified by a factor of 10. A theoretical signal generated from the *SS* wave (*center*) is subtracted from the observed stack to reveal the reflection from the 520 km discontinuity (*bottom*). (Shearer, 1996. *J. Geophys. Res.*, *101*, 3053–66, copyright by the American Geophysical Union.)

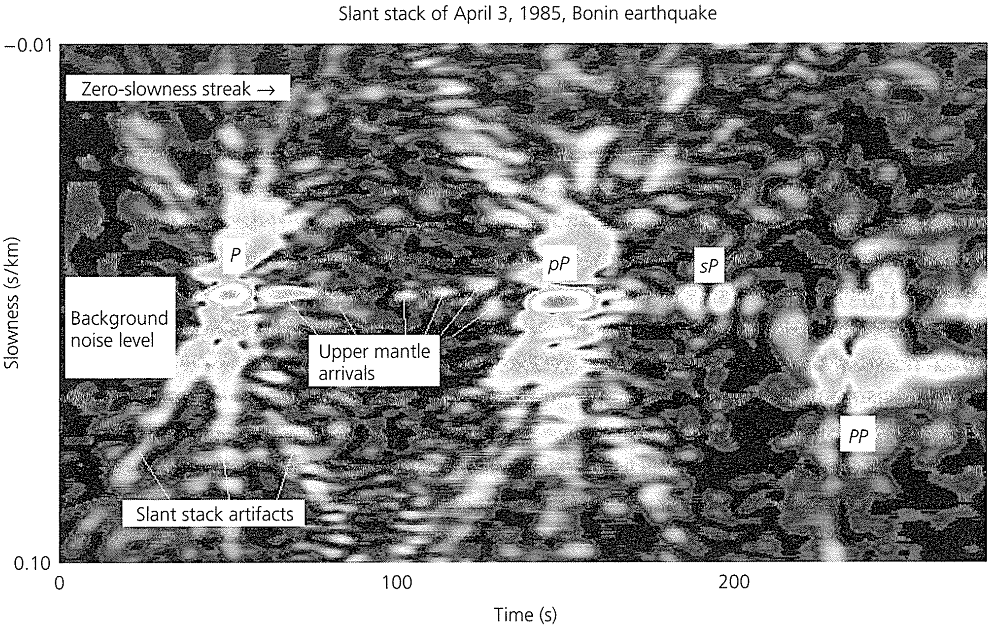


Fig. 6.5-4 Slant stack of seismograms at 279 stations for a deep (476 km) earthquake. The bull's-eyes are concentrations of seismic energy for particular arrivals. (Vidale and Benz, 1992. Reproduced with permission from *Nature*.)

(Fig. 6.5-3, *bottom*), which is weak due to the gradual velocity change at the 520 km discontinuity, and so rarely observed otherwise.

Mantle structures can also be observed with slant stacks (Section 3.3.5). The seismograms are stacked as functions of both time and slowness, so instead of getting a single seismogram, as in Fig. 6.5-3, we get a plot of seismic energy as a function of time and slowness. As shown in Fig. 6.5-4, arrivals occur as high-amplitude bull's-eyes. The *P* and *pP* arrivals have a slightly different slowness due to the small (about 1°) difference in incidence angles. The large arrivals create smeared features that are artifacts of the slant stacking.

Stacking is also used to enhance specific normal modes of the earth. The amplitudes of normal modes vary between stations, because they depend on spherical harmonics that are functions of latitude and longitude, which differ between individual modes (Section 2.9.3). Although simply stacking seismograms from different sites does not make spectral peaks stand out better, correcting for the theoretical variation in amplitude and phase for a given mode and then stacking enhances the mode of interest and suppresses others (Fig. 6.5-5).

Stacking can be applied to very large volumes of data. Figure 6.5-6 shows record sections generated with thousands of digitally recorded seismograms from different earthquakes and seismometers. The seismograms were rotated into vertical, radial, and transverse components, grouped by source–receiver distance, and then those within half-degree intervals were normalized to a common amplitude and stacked. The strong arrivals in the stacked record sections correspond to the major phases shown in the travel time curves. It is interesting to compare this analysis of global seismic data spanning large distance ranges with reflection seismic data analysis (Section 3.3.4). For reflection data, CMP stacking involves forming common

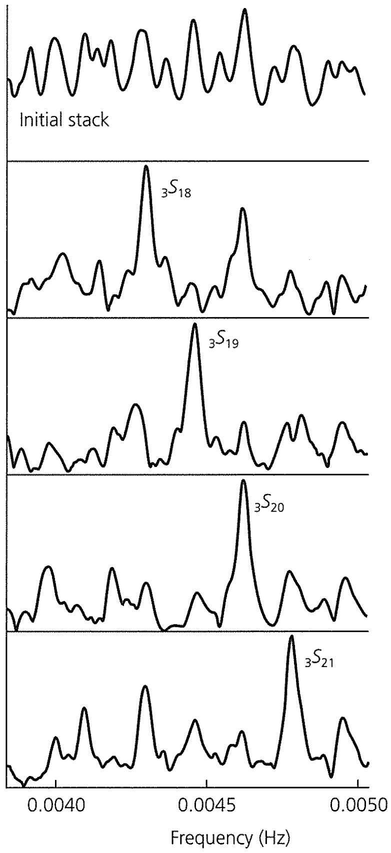


Fig. 6.5-5 Stacking long-period seismograms to enhance specific normal modes of the earth. Although a given mode multiplet is not enhanced by simply stacking seismograms from different sites (*top*), stacking using its predicted variation between sites enhances the multiplet and suppresses others (*lower panels*). (Mendiguren, 1973. *Science, 179*, 179–80, copyright 1973 American Association for the Advancement of Science.)

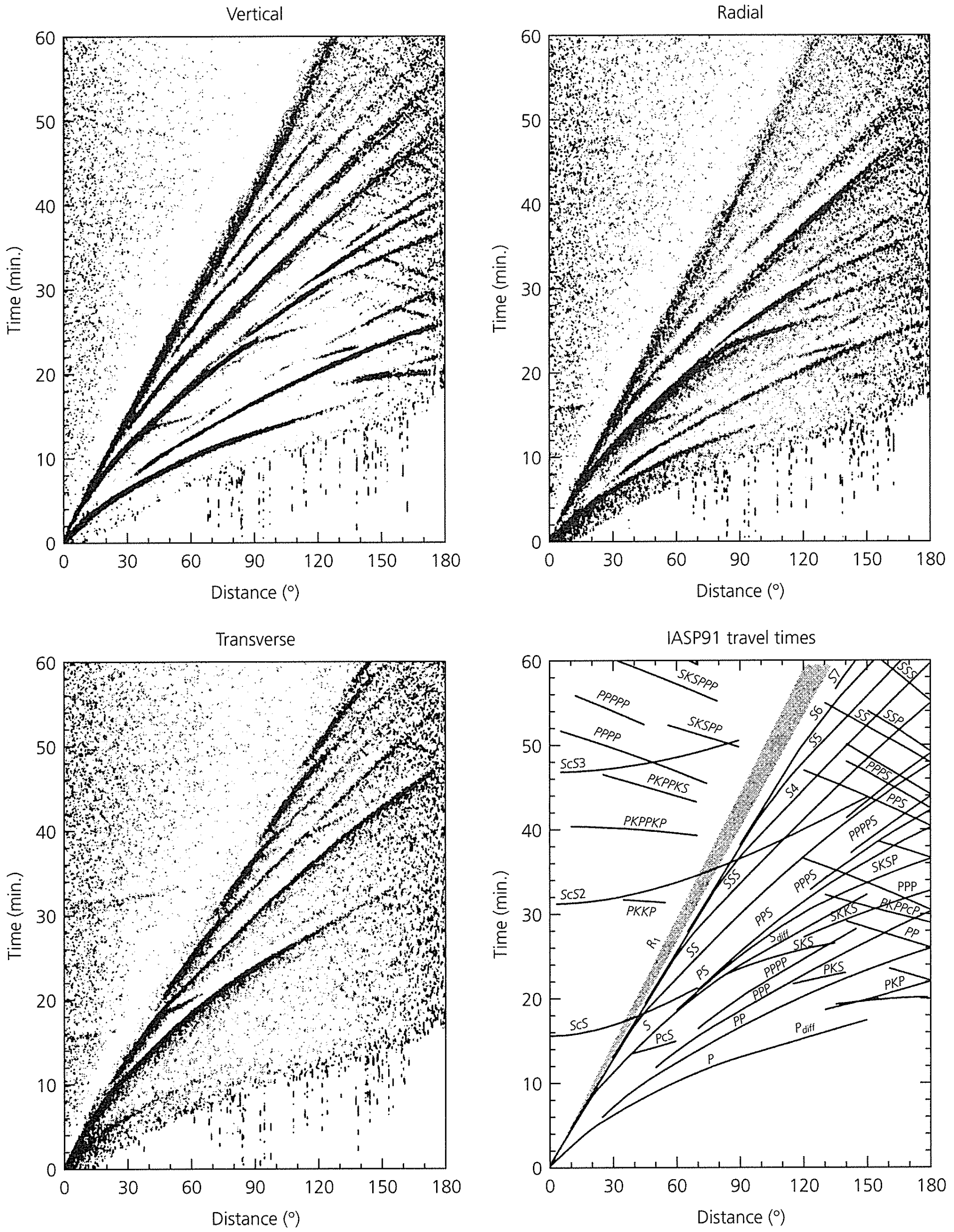


Fig. 6.5-6 Stacking of global seismograms to produce record sections. The three stacks, each for a different component, show distinct arrivals that can be compared to those predicted by the travel time curve for an earth model. (Astiz *et al.*, 1996. © Seismological Society of America. All rights reserved.)

midpoint gathers and stacking them over all source–receiver distances (offsets) (Fig. 3.3-18), to produce synthetic zero-offset traces on which reflected arrivals are enhanced. These traces are then shown together to produce a common midpoint section, a function of midpoint and time. By contrast, the global data are gathered by common offset, stacked for that offset, and then displayed as a function of offset and time. This operation only reduces noise, rather than enhancing specific arrivals, and so shows various arrivals (direct waves, reflections, surface waves, etc.). Another example was shown in Fig. 2.7.4, where many long-period seismograms were stacked to demonstrate the group and phase velocities of surface waves.

In these or other stacking operations, one possible source of systematic error is incorrect transformation of the data between different times or positions. Interestingly, in the very different cases just discussed, a common difficulty is lateral variation in structure. In the reflection example, structures may dip rather than be flat-lying, causing traces with common midpoints not to sample the same point on a reflector (Fig. 3.3-19). In the global travel time analysis, seismograms for the same source–receiver distance differ when the structure between the source and the receiver differs. An analogous effect occurs for normal modes due to deviations of the structure from spherical symmetry. Nevertheless, because in most cases structure varies primarily with depth, these stacking operations generally work well.

6.6 Seismometers and seismological networks

6.6.1 *Introduction*

Given what we have discussed about signal processing, we now introduce some ideas about *seismometry*, the design and development of seismic instrumentation. Although we informally call such systems seismometers, the *seismometer* is actually the sensor recording ground motion, and thus a key component of the entire *seismograph* system, which also contains amplifying, timing, and recording components. The product, a record of ground motion as a function of time, is a *seismogram*.

Following linear system theory, we note that a seismogram is not an exact representation of the ground motion. Seismograms depend upon the seismometer and the rest of the seismograph system, because the sensitivities of seismometers vary with the frequency of the motion recorded. Moreover, seismometers record ground motion as displacement, velocity, acceleration, or various combinations of these.[1]

Once recorded, distributing seismic data is crucial, because the data are of no use until they are available for study. Hence seismology has long been a leader among the sciences in developing public data distribution. This tradition began a century ago out of necessity. Unlike a geological field observation or a geochemical experiment, observations at many sites are needed to locate and study earthquakes, with the more data the better. Soon after seismometers became sensitive enough to teleseismically record earthquakes, arrival times were shared. The first major attempt to gather and publish seismically recorded arrival times was the bulletin of the Bureau Central International de Séismologie (BCIS), which began in 1904. The International Seismological Summary (ISS) began publication in 1913,[2] and eventually became the *Bulletin of the International Seismological Centre* (ISC), now an authoritative source of earthquake locations. Not only arrival times but also polarities and amplitudes were disseminated, enabling the study of magnitudes and focal mechanisms.

This sharing of data has been crucial to seismology's growth. In the modern era, the World Wide Standardized Seismograph Network (WWSSN), which started in 1962, was the first means of globally sharing full seismic waveform data. Today, high-quality digital global seismic data are available through the Federation of Digital Broad-Band Seismographic Networks (FDSN), of which the stations of the US-sponsored Incorporated Research Institutions for Seismology (IRIS) are a part. Data and results such as earthquake locations are also provided by national and regional data centers. Seismologists anywhere in the world need only a computer and access to the Internet to freely and conveniently obtain terabytes[3] of digital seismic data, software to look at it, and a great deal of other earthquake information. As much as any development in theory or seismometry, this free access to data and software is responsible for the remarkable growth of the field within the past century. Not only can scientists work more efficiently, but this openness has encouraged the sharing of data and models, and allowed comparison and testing of results.

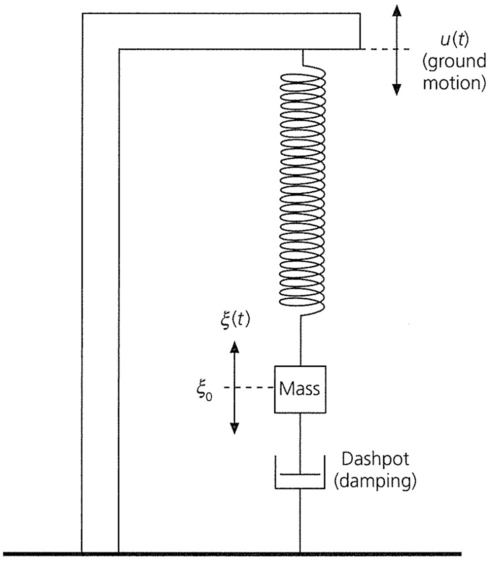


Fig. 6.6-1 Pendulum seismograph consisting of a mass, a spring, and a dashpot.

6.6.2 *The damped harmonic oscillator*

The basic problem of seismometry is how to measure the motion of the ground using an instrument that is also on the ground. The traditional solution is to use an inertial, known as a pendulum, system, so that the motion of the pendulum is out of phase with the ground motion. Three orthogonal seismometers (vertical, north–south, east–west) can give a three-dimensional record of ground motion. A schematic vertical seismometer is shown in Fig. 6.6-1. The key elements of the system are the mass, the spring, and a dashpot, or damping device. We consider such a system in general, without concern for the mechanics of how it is actually implemented.

This mechanical seismometer system is a damped simple harmonic oscillator. If the spring equilibrium length in the absence of ground motion is ξ_0, the spring exerts a force proportional to its extension from equilibrium as a function of time, $\xi(t) - \xi_0$, times a spring constant k. The dashpot, with damping constant d, exerts a force proportional to the velocity between the mass (m) and the earth. So, for a ground motion $u(t)$,

$$m\frac{d^2}{dt^2}[\xi(t) + u(t)] + d\frac{d\xi(t)}{dt} + k[\xi(t) - \xi_0] = 0. \tag{1}$$

[1] This is analogous to the way animals see differently; the electromagnetic radiation is the same, but human eyes respond slightly differently than those of bears (which are very nearsighted), and entirely differently from the hexagonally tiled eyes of flies.

[2] Its original name was the *Monthly Bulletin of the Seismological Committee of the British Association for the Advancement of Science.*

[3] One terabyte (Tbyte) equals 10^{12} bytes.

If we define $\xi(t) - \xi_0$ as $\xi(t)$, the displacement relative to the equilibrium position, Eqn 1 becomes

$$m\ddot{\xi} + d\dot{\xi} + k\xi = -m\ddot{u}, \tag{2}$$

or

$$\ddot{\xi} + 2\varepsilon\dot{\xi} + \omega_0^2\xi = -\ddot{u}, \tag{3}$$

where the single and double dots denote the first and second time derivatives, $\omega_0 = \sqrt{k/m}$ is the natural frequency of the undamped system, and the damping is described by $\varepsilon = d/(2m)$. This is a linear differential equation with constant coefficients that we encountered when we used a damped harmonic oscillator as a model for anelasticity (Section 3.7.5). Thus Eqn 3 is the inhomogeneous (forcing term) version of Eqn 3.7.8, where the damping term ε appeared as $\omega_0/2Q$. To solve it, we assume that

$$u(t) = e^{-i\omega t} \quad \text{and} \quad \xi(t) = X(\omega)e^{-i\omega t} \tag{4}$$

and substitute Eqn 4 into Eqn 3 to yield

$$X(\omega)(-\omega^2 - 2\varepsilon i\omega + \omega_0^2)e^{-i\omega t} = \omega^2 e^{-i\omega t}, \tag{5}$$

or

$$X(\omega) = -\omega^2/(\omega^2 - \omega_0^2 + 2\varepsilon i\omega), \tag{6}$$

which is the instrument response produced by a ground motion $e^{i\omega t}$.

$X(\omega)$ is complex and can be written in terms of the amplitude and phase responses

$$X(\omega) = |X(\omega)|e^{i\phi(\omega)}, \tag{7}$$

where

$$|X(\omega)| = \omega^2/[(\omega^2 - \omega_0^2)^2 + 4\varepsilon^2\omega^2]^{1/2}, \tag{8}$$

$$\phi(\omega) = -\tan^{-1}\frac{2\varepsilon\omega}{\omega^2 - \omega_0^2} + \pi. \tag{9}$$

As shown in Fig. 6.6-2, these functions have several interesting features. First, as the angular frequency of the ground motion, ω, approaches the natural frequency of the pendulum, ω_0, the amplitude response is large. This effect, called *resonance*, is like "pumping" a playground swing at its natural period. Thus the seismometer responds best to ground motion near its natural period.

For frequencies much greater than the natural frequency, $\omega \gg \omega_0$, $|X(\omega)| \to 1$, and $\phi(\omega) \to \pi$, so the seismometer records the ground motion, but with the sign reversed.[4] To see why this occurs, consider Eqn 3. For $\omega \gg \omega_0$, the $\ddot{\xi}$ term is the largest term on the left-hand side, so $\ddot{\xi}$ approximately equals $\ddot{u}$. Thus the seismometer responds to the ground *displacement*. On the other hand, for frequencies much less than the natural frequency, $\omega \ll \omega_0$, $|X(\omega)| \to \omega^2/\omega_0^2$, and $\phi(\omega) \to 0$. Hence, in this case the seismometer responds to *acceleration*, as can be seen from Eqn 3, because the $\omega_0^2\xi$ term is dominant, so ξ is proportional to $\ddot{u}$. The shape of the instrument response depends on the damping factor $h = \varepsilon/\omega_0$. For $h = 0$, the system is undamped, and the amplitude response is peaked around the resonant frequency, $\omega = \omega_0$. The seismometer amplifies ground motion with periods near its natural period. As damping is increased, the curve is smeared out. Thus the natural period and damping are used to design a seismometer to record ground motion in a particular period range.

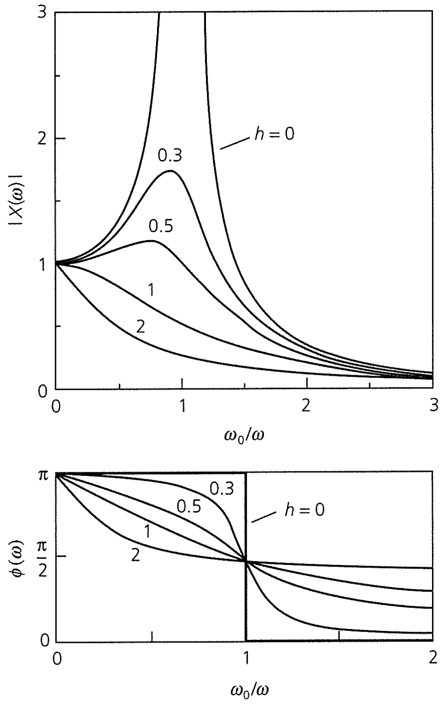


Fig. 6.6-2 Amplitude response $|X(\omega)|$ and phase delay $\phi(\omega)$ for a pendulum seismometer such as that shown in Fig. 6.6-1.

Figure 6.6-2 bears a strong resemblance to Fig. 3.7-13, which showed the frequency response for a damped harmonic oscillator as a function of Q. The plots are slightly different, in that Fig. 3.7-13 is plotted as a function of ω, and Fig. 6.6-2 is plotted as ω_0/ω. In addition, Fig. 6.6-2 is normalized to the value at $\omega_0/\omega = 0$. However, the curves convey the same information because h and Q are related as $h = 1/2Q$. The Q values in Fig. 3.7-13 of 5, 15, and 100 correspond to h values of 0.1, 0.03, and 0.005, all of which would plot close to the curve for $h = 0$ in Fig. 6.6-2.

[4] To see this, quickly jiggle an object hanging by a rubber band and note that its motion is out of phase with your hand.

6.6.3 Earth noise

An important consideration in designing seismometers is earth noise. A challenge of seismometry is to create sensors sensitive enough to record small teleseismic signals, given that noise sets a limit to the level of detection. Moreover, studies using seismic data in many applications must consider the signal-to-noise ratio.

Many factors contribute to seismic noise, including solar and lunar tides within the solid earth, fluctuations in temperature and atmospheric pressure, storms, human activities, and ocean waves. These factors are constantly at work, so the crust is continually reverberating. Most of the noise occurs at periods between 1 and 10 seconds. Such waves, called *microseisms*, are shown in Fig. 6.6-3 (*top*). Even before the first waves arrive from the earthquake shown, the seismogram shows a roughly constant level of seismic energy (*center*). The spectrum shows that most of this noise is in the frequency range of 0.1–0.2 Hz (periods of 5–10 s) (*bottom*). The primary source for these microseisms is thought to be ocean waves. Seismometers are noisier the closer they are to coastlines, so ocean island stations are among the noisiest.

How a seismometer is deployed has a great effect upon the noise that it records. Most sources of noise decrease away from the surface, so permanent seismometer installations are often in boreholes. For portable seismometers, burying them even half a meter beneath the surface greatly reduces noise from daily temperature fluctuations. Rain generates high frequency noise, and wind, coupled to the ground through the roots of swaying trees, can generate severe long-period noise. Human activity (trucks, trains, machinery, etc.) causes significant ground noise, so seismologists deploying temporary stations face a trade-off between the convenience (continuous power, security, constant temperature, no flooding) of building basements and the lower noise of remote sites.

6.6.4 Seismometers and seismographs

Seismometers record ground motions ranging from large high-frequency accelerations near an earthquake to small ultra-long-period normal mode signals. Because no single seismograph can do this, different instruments have evolved to handle the different *dynamic ranges* and *frequency ranges* of seismic waves.

Dynamic range is measured in decibels (dB), which increase by 20 for each order of magnitude increase in amplitude. Thus, if signal A_1 is five orders of magnitude larger than signal A_2, $A_1/A_2 = 10^5$, and the dynamic range is 100 dB. The displacements associated with a magnitude 2 earthquake may be as low as 10^{-10} m, whereas teleseismic displacements from a magnitude 8 earthquake may be on the order of 10^{-1} m, and displacements near a large earthquake can be much greater. Thus the dynamic range of seismometry is at least 180 dB. Similarly, the frequency range of seismometers spans seven orders of magnitude from Earth tides (0.000023 Hz) to ultra-high frequencies of greater than 200 Hz for very shallow structure investigations.

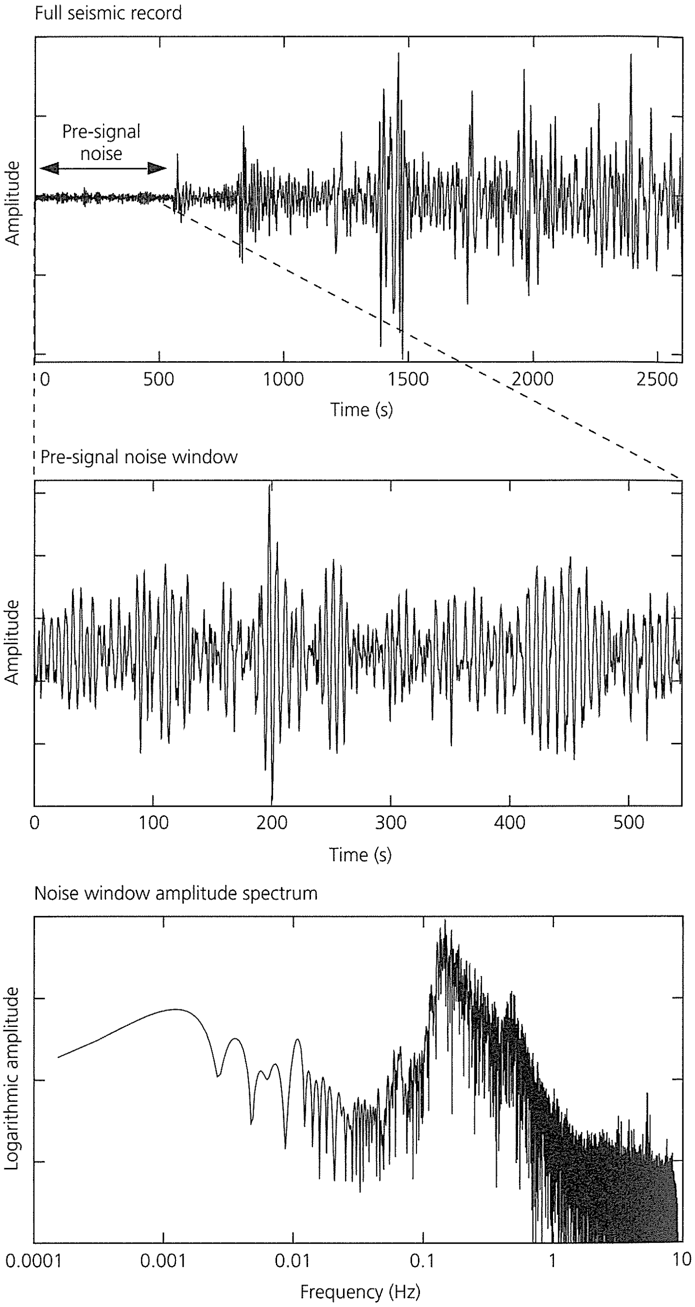
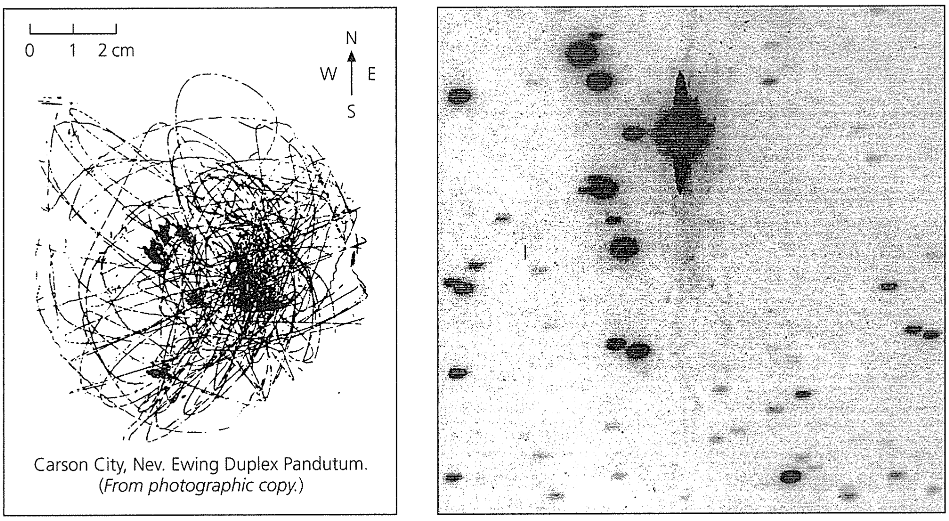


Fig. 6.6-3 Demonstration of seismic noise on a broadband seismogram in Hudson, New York, from an April 7, 1995, Tonga earthquake. *Top*: Seismic noise appears before the first arrival, which is P_{diff}. *Center:* Visual examination of the noise shows waves with a dominant period of about 5–6 s, called *microseisms*. *Bottom*: The spectrum of the noise has largest amplitude in the 5–10 s period range.

The earliest attempts to record the motions of earthquakes used *seismoscopes*, which differ from seismographs in that they record ground motion without time information. The first known seismoscope, built by the Chinese astronomer Chang



Carson City, Nev. Ewing Duplex Pandutum. (*From photographic copy.*)

Fig. 6.6-4 Two examples of seismoscope recordings, which show the amplitudes of motions without a record of time. *Left*: Seismogram of the great 1906 San Francisco earthquake, recorded by the Ewing duplex pendulum seismoscope in Carson City, Nevada. (Kanamori, 1988. Importance of historical seismograms for geophysical research, in *Historical Seismograms and Earthquakes of the world*, ed. W.H.K. Lee, H. Myers and K. Shimizaki, copyright 1988 by Academic Press, reproduced by permission of the publisher.) *Right*: Seismogram of a $m_b = 4.3$ earthquake in Hawaii, recorded as a telescope image at the Hawaii Telescope Observatory. The dark images are stars, and the lines emanating from the large star at the upper center of the image result from tilting of the telescope during the earthquake. (Courtesy of L. Meech.)

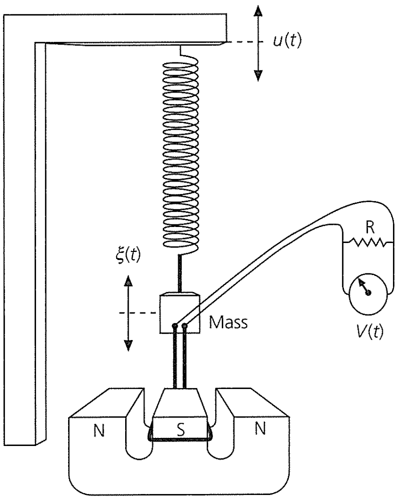


Fig. 6.6-5 Schematic illustration of an electromagnetic seismograph, in which the mass is coupled to an electromagnetic transducer. Motions of the mass move the coil through the magnetic field, generating an electric current. The voltage across the coil is proportional to the relative velocity between the mass and the magnet.

Heng in about AD 132, consisted of a pendulum inside a 6 ft-diameter jar. Eight dragons' heads with metal balls in their mouths were placed around the rim of the jar, so the balls would drop in the direction from which seismic waves arrived. Later seismoscopes included a pendulum etching a path on a bed of sand (A. Bina, 1751), a collection system for a bowl filled to the brim with mercury (A. Cavalli, 1784), and optical reflection off a basin of mercury (R. Mallet, 1851). Two very different seismoscope recordings are shown in Fig. 6.6-4.

Early seismometers, incorporating a record of the time-dependence of the ground motion, were purely mechanical instruments like that outlined in Section 6.6.2. Seismometry began with the designs of F. Cecchi around 1875, and developed rapidly through the work of seismologists like J. Milne, J. Ewing, and T. Gray. The first teleseismic recording was by a seismograph in Potsdam of a Japanese earthquake in 1889. By the start of the twentieth century a global network of more than 40 seismographs was in operation. Such instruments often produced excellent data but responded best to very large earthquakes because their magnifications were low, only about 100 times the actual ground motion.

Higher magnifications are achieved by using electromagnetic instruments, based on a design introduced by Galitzin in 1914 that is now common. The motion of the pendulum relative to the frame is measured by moving a coil attached to the mass through the magnetic field produced by a magnet fixed to the seismometer frame. The voltage produced in the coil is proportional to the time rate of change of the magnetic field, and thus to the velocity of the mass relative to the frame (Fig. 6.6-5). The sensitivity can be increased by feeding the output from this sensor into a galvanometer, a wire suspended by a thin fiber such that it is deflected by the current produced by the sensor (Fig. 6.6-6). A mirror is attached so that ground motion deflects the mirror and thus changes the position of a beam of light hitting a piece of photographic paper. The paper is mounted on a helical drum which turns once per hour.

Thus the response of an electromagnetic analog seismometer system is a combination of the pendulum, transducer (electromagnetic velocity sensor), and galvanometer responses. These are shown as log–log plots in Fig. 6.6-7. The pendulum response (Fig. 6.6-7a, b) is proportional to ω^2 for $\omega < \omega_s$, the pendulum frequency. The transducer response (Fig. 6.6-7c, d) is proportional to ω because it responds to the velocity, the derivative of displacement. The galvanometer response

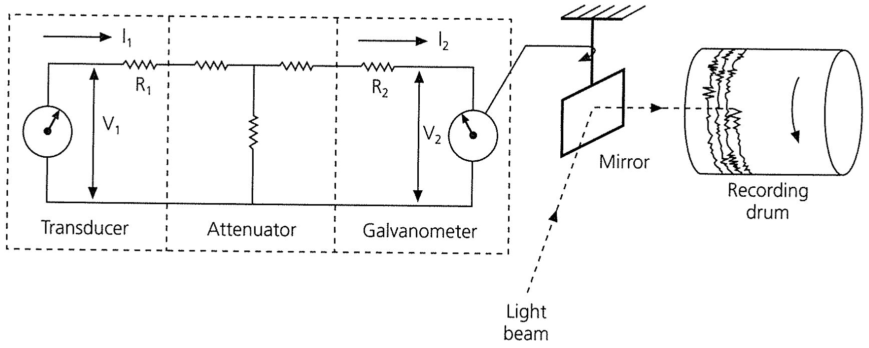


Fig. 6.6-6 Coupling of the transducer of an electromagnetic seismograph to a galvanometer, which deflects a mirror and thus a light beam, causing a time history of the voltage and thus the mass movements to be recorded on photographic paper. Timing pulses deflect the mirror to make minute and hour marks.

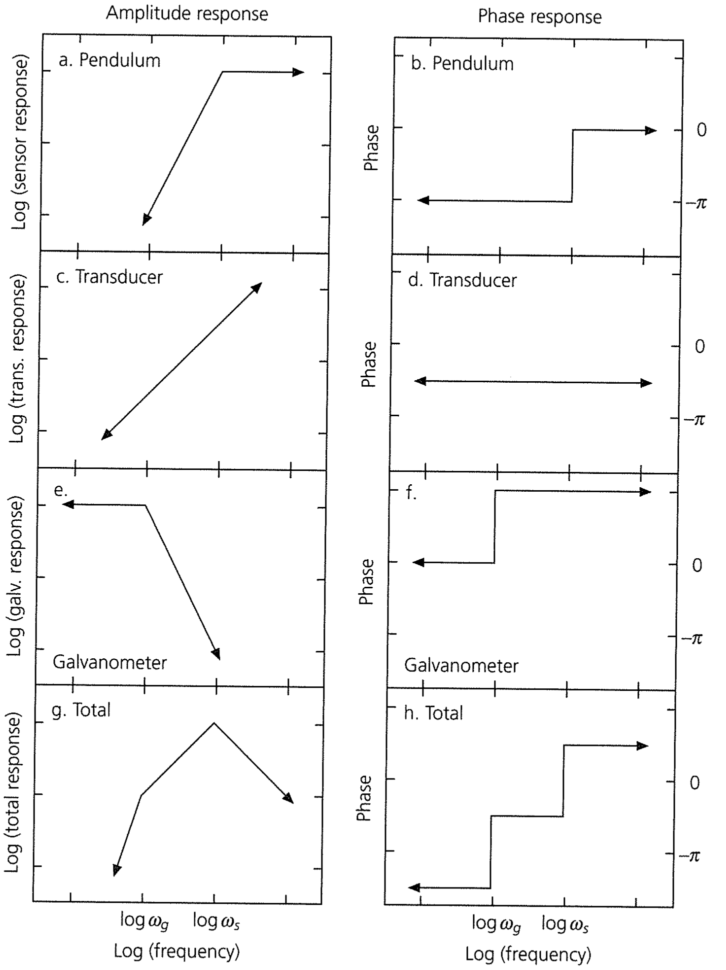


Fig. 6.6-7 Response of the components of an electromagnetic seismograph system. Left panels show the amplitude responses, and right panels show the phase responses. ω_s and ω_g are the pendulum and galvanometer frequencies.

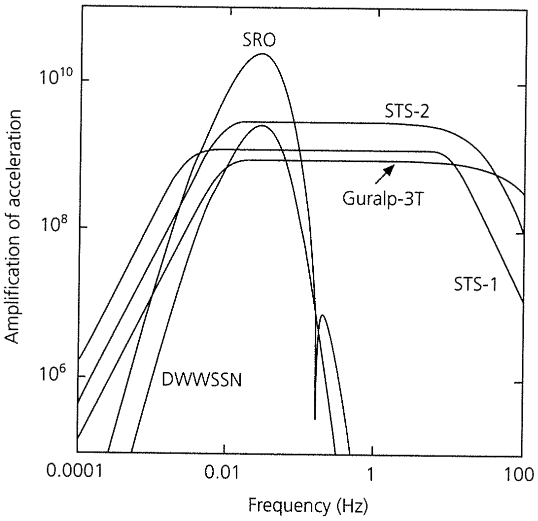


Fig. 6.6-8 Frequency domain instrument responses for several types of seismometers. The SRO and DWSSN sensors have responses peaked at long periods and so do not record high-frequency signals. The STS-1, STS-2, and Guralp-3T sensors are broadband seismometers with a flat response over a wide range of frequencies.

(Fig. 6.6-7e, f) falls off as ω^{-2} for $\omega > \omega_g$, the galvanometer frequency. The combined effect is shown in Fig. 6.6-7g, h. Thus, the response of an electromagnetic seismometer can be "shaped" by choosing the pendulum and galvanometer periods.

Two classic electromagnetic instruments used heavily for years were the World Wide Standardized Seismograph Network (WWSSN) long- and short-period instruments. The long-period (LP) instrument had a pendulum period of 15 s (30 s in some early versions) and a galvanometer period of 100 s. The short-period instrument had a 1 s pendulum and a 0.75 s galvanometer. Each WWSSN station had three LP and three SP instruments oriented to record ground motion in the vertical, east–west, and north-south directions. The resulting response curve of the LP instrument (labeled "DWWSSN" from when some of the WWSSN seismometers were converted to record digitally) is shown in Fig. 6.6-8. Instruments ran at several possible magnifications (gains). The two different instruments were designed to reduce the effects of seismic noise. The LP sensors had peak sensitivity in the 10–40 s range, making them ideal for long-period teleseismic studies. The SP sensors were peaked at around 1 s, a good period with which to pick the travel times of *P* waves.

A sample of the data is shown in Fig. 6.6-9. The record, covering 24 hours, has calibration pulses at the beginning, which can be used to check the amplitude and phase

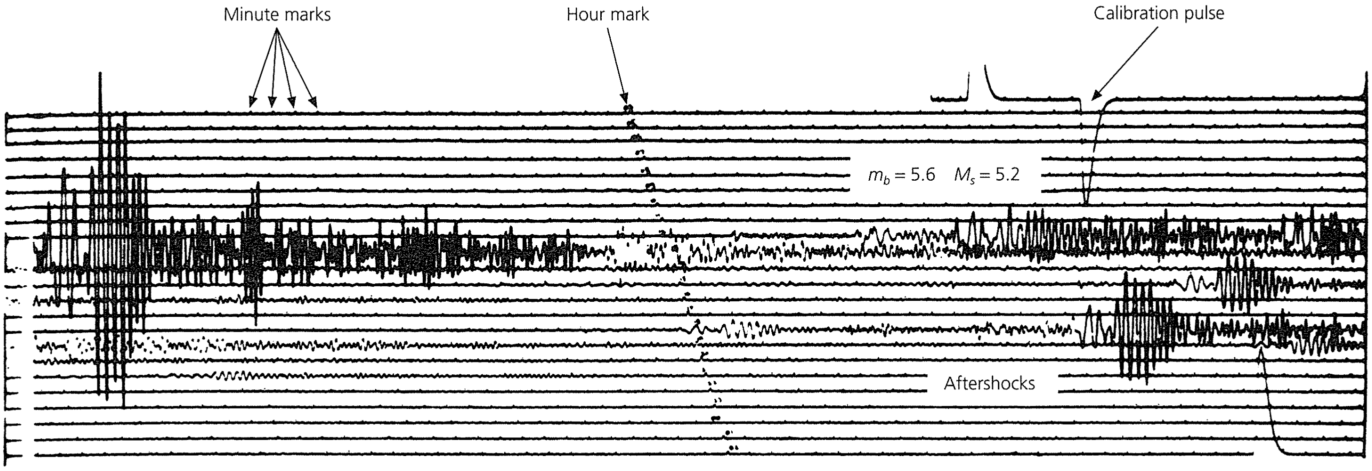


Fig. 6.6-9 Sample WWSSN seismogram, showing the long-period vertical component from an earthquake in the Indian Ocean, recorded 36° away in Pakistan.

calibration. Timing marks, generated by crystal clocks accurate to 1 part in 10^7 are placed at each minute (short mark) and each hour (longer mark). Every sixth hour has no hour mark. This timing allowed arrival times to be read accurately, and the calibration allowed studies using true amplitudes. The seismograms were microfilmed and made available to the seismological community.

Although many results discussed in this text were derived from such data, using WWSSN data was cumbersome. Microfiche records had be acquired, examined in a microfiche reader, copied, and refiled. The traces were then digitized by taping them to a special table that contained a grid of electromagnetic wires and then tracing the seismogram with a cursor. After digitization, the seismogram was interpolated to a desired sampling rate. The hand digitization added a source of error, as it was not always easy to follow the trace of interest, especially for large earthquakes where the surface waves could wrap around the seismic record for several hours. Because of the effort involved, entire Ph.D. dissertations might involve the analysis of only tens or hundreds of seismograms, a task that is now done in minutes to days.

The replacement of analog seismographs by digital broadband instruments has important advantages. The newer seismometers provide better data over a broader frequency band, and the digital data are available via magnetic tape, compact disk, or the Internet, making computer analysis much easier. Routine processing, such as rotating into radial and transverse components and making record sections, has become nearly trivial. Large volumes of data are available and can be processed easily. For example, as of 2000 the IRIS Data Management Center had over 7 Tbytes of digital data available over the Internet either immediately or with only the short delay needed for it to be read from mass storage systems.

Some of the technology involved in more recent seismograph systems is illustrated by one of the first digital seismological systems, the instrument used by the International Deployment of Accelerometers (IDA) shown schematically in Fig. 6.6-10. The sensor is a force-feedback gravimeter that detects vertical ground motion by the resulting change in gravity. The gravimeter mass is connected to the center plate of a capacitor whose outer two plates are fixed. As the mass moves, the voltage between the center plate and the outer plates is proportional to the displacement. A 5 kilohertz alternating voltage applied to the outer plates is amplitude-modulated (Section 2.8.1) by the lower-frequency seismic signal. The modulated signal is fed to an amplifier that generates a voltage proportional to the displacement of the mass. This signal then goes to an integrator circuit whose output is proportional to the acceleration of the mass. This is the seismic system's output, which is sampled

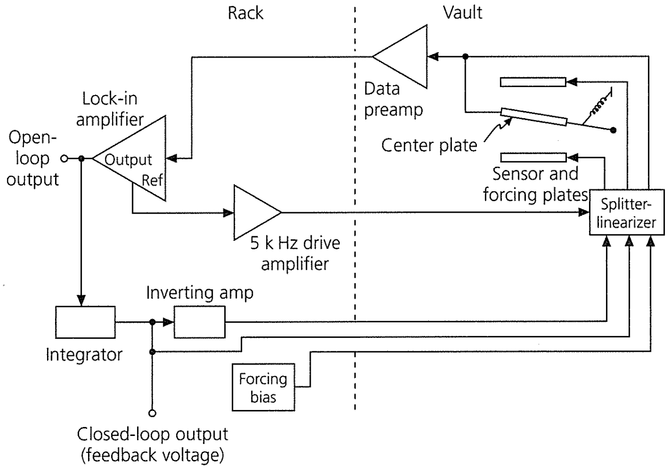


Fig. 6.6-10 Block diagram of the sensing and feedback electronics of an IDA gravimeter recording system. (Agnew *et al.*, 1976. *Eos Trans. Am. Geophys. Un.*, *57*, 180–8, copyright by the American Geophysical Union.)

once every ten seconds. The voltage is also fed back to the outer capacitor plates to stabilize the system and increase linearity. This force-feedback, an important feature of modern seismometers, provides a greater dynamic range because the mass does not move as far to record large amplitudes. Because this instrument can record a static displacement, it has a flat response out to frequencies approaching $\omega = 0$. Such long-period response is valuable for studying normal modes and large earthquakes.

The most versatile of the current digital seismometers are broadband systems that record over a very broad frequency range. At present, the primary broadband seismometers are the Streckheisen STS-1 and STS-2 and Guralp-3T, which use force-feedback technology to allow large dynamic and frequency ranges (Fig. 6.6-8). The advantages of such a broad frequency response are illustrated in Fig. 6.6-11. As shown, the seismogram can be filtered to isolate and give excellent records of two very different overlapping signals. These seismometers are very compact (the three-component STS-2 is the size of a bowling ball and weighs 20 lb)[5] but record with a flat response at over three orders of magnitude in frequency. The STS-1 is designed for permanent installation, whereas the STS-2 and Guralp-3T are robust enough to be used as portable instruments.

A variety of specialized seismic instruments are also used. *Strainmeters* are used to measure gradual displacements, especially near faults and volcanoes. Such instruments are technically challenging to build, and have taken unusual forms. For instance, an early strainmeter made by H. Benioff consisted of a quartz rod 24 m long, attached to the ground at one end, and extending through a capacitance transducer at the other. Strain rates as small as 10^{-15} s^{-1} could be recorded. A recent strainmeter with a hydraulic sensor achieves a strain sensitivity of 10^{-12} with a dynamic range of about 130 dB. Over longer distances, horizontal strains are observed using laser measurements between sites (often across faults) and space-geodetic techniques (Section 4.5), including the GPS satellite system and very long baseline radio interferometry.

At the other end of the spectrum of seismic instrumentation are strong-motion sensors that record strong shaking near an earthquake. Whereas strainmeters record minute displacements, strong-motion sensors, also called accelerometers, record accelerations up to 2 g without breaking or going off scale. For example, horizontal accelerations of 1.25 g were recorded 3 km from the 1971 San Fernando Valley earthquake, and vertical accelerations of 1.74 g were recorded 1 km from the 1979 Imperial Valley earthquake. Thus the seismometer pendulum frequency ω_0 is chosen to exceed the highest frequency of interest (about 20 Hz). These instruments are stable because the small pendulums make the accelerometers less susceptible to tilt and drift than longer-period instruments. A damping parameter (often 0.7 of the critical value) is chosen to

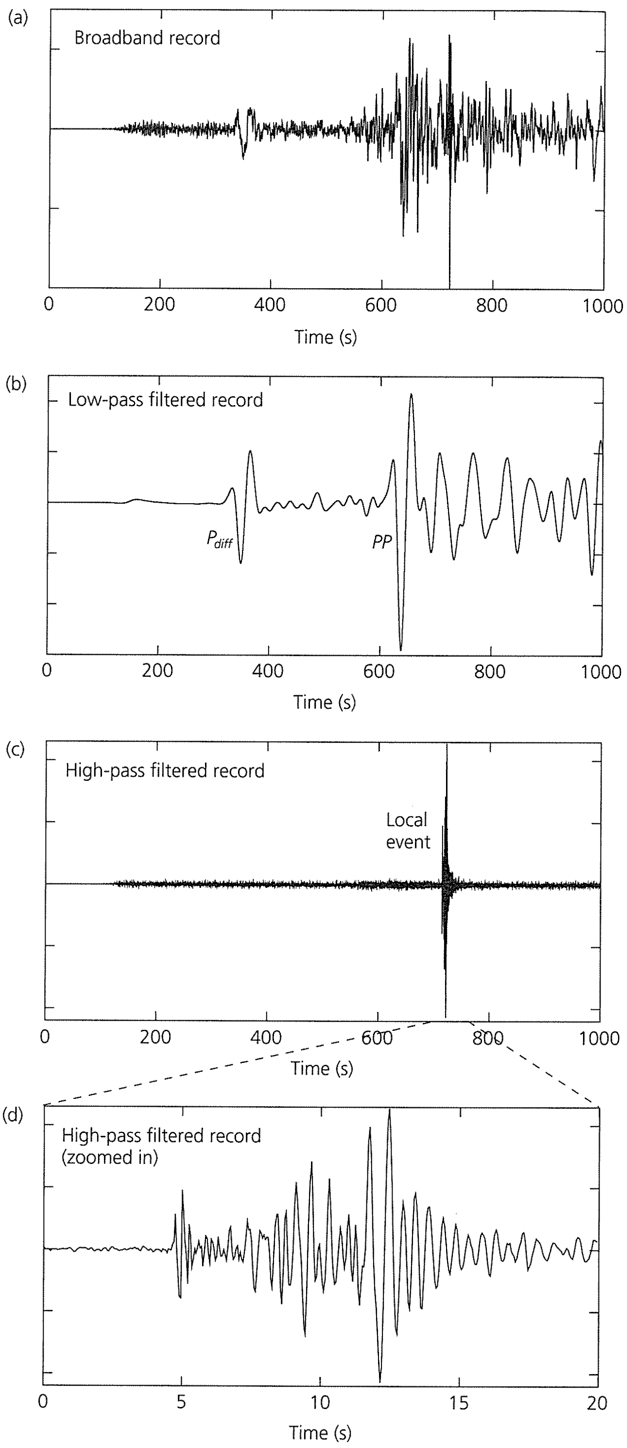


Fig. 6.6-11 STS-2 broadband seismogram recorded in Slippery Rock, PA, from a July 3, 1995, Tonga earthquake. Because the seismometer records a wide range of frequencies, the same seismogram can be used to study both local and teleseismic events. (a): The original broadband record. (b): The same record, low-pass filtered at a frequency of 0.03 Hz, showing the long-period teleseismic signals from the Tonga event. (c): The record high-pass filtered at 0.5 Hz, showing the high-frequency signals from a local event. (d): A zoom-in of the high-pass filtered record shows the full waveform of the local event. The $S - P$ time suggests that the event was 20 km away from the station, probably a local quarry blast.

[5] Before such technology, some mechanical seismometers built in the first half of the twentieth century weighed more than 20 tons because the large mass gave higher long-period magnification, as shown by Eqn 6.

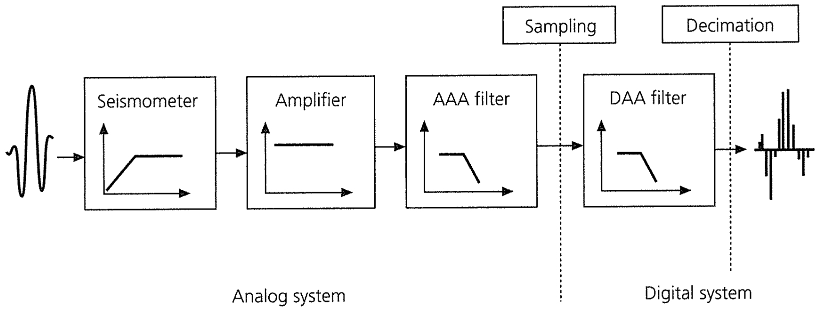


Fig. 6.6-12 Diagram showing the analog-to-digital (ADC) process. The analog part of the system consists of the generation of a seismic signal by the seismometer, its amplification, and analog anti-aliasing (AAA) filtering. The digital part of the system consists of sampling the AAA-filtered signal, filtering the signal further with a digital anti-aliasing (DAA) filter, and then decimating the signal to achieve the desired sampling rate. (Scherbaum, 1996, with kind permission from Kluwer Academic Publishers.)

give a response curve that is flat and directly proportional to ground acceleration from periods of zero to the natural period of the seismometer.

A major advance in seismometry has been in timing, which has long been a difficulty. In the early days of seismology, timing errors played a large part in the mislocation of earthquakes. However, seismometers now receive time signals from GPS satellites, whose atomic clocks are accurate to a billionth of a second. Similarly, although ocean bottom seismometers cannot receive GPS signals, accurate clocks for them are now available.

6.6.5 *Digital recording*

Although digital seismic data are easier to use than analog data, the conversion of continuous ground motion into a digital seismogram is not a trivial matter. Figure 6.6-12 shows how this is done. Ground motion, represented by the waveform at the left, is detected by the seismometer through the motion of the mass. This motion is converted into an analog electrical signal and then amplified. To avoid a spurious signal due to aliasing (Fig. 6.4-3), a combination of anti-aliasing filters is used. Many seismometers use an initial frequency domain low-pass filter as an analog anti-aliasing (AAA) filter. The filtered signal is then oversampled at a rate that is at least twice the frequency of the AAA filter in order to avoid aliasing. This signal is then convolved with a digital anti-aliasing (DAA) filter, often called a finite impulse response (FIR) filter, and finally resampled at twice the desired Nyquist frequency.

An example of a FIR filter is shown in Fig. 6.6-13a, with the resulting signal shown in Fig. 6.6-13c. The FIR filter maintains the shape of the pre-filtered signal, but introduces spurious noncausal arrivals that might be mistaken for early stages of earthquake rupture. These precursory signals result because the FIR filter's impulse response is an emergent signal. This effect can be removed by correcting the phase of the FIR filter to make it causal (Fig. 6.6-13b). This filter does not cause precursory signals (Fig. 6.6-13d), but the shapes of the waveforms are changed. We noted a similar phenomenon in Section 3.7.8, where anelasticity acted as a filter, removing high frequencies and making the waveforms noncausal unless the phase was changed. As discussed in Section 6.3.3, there is no perfect way

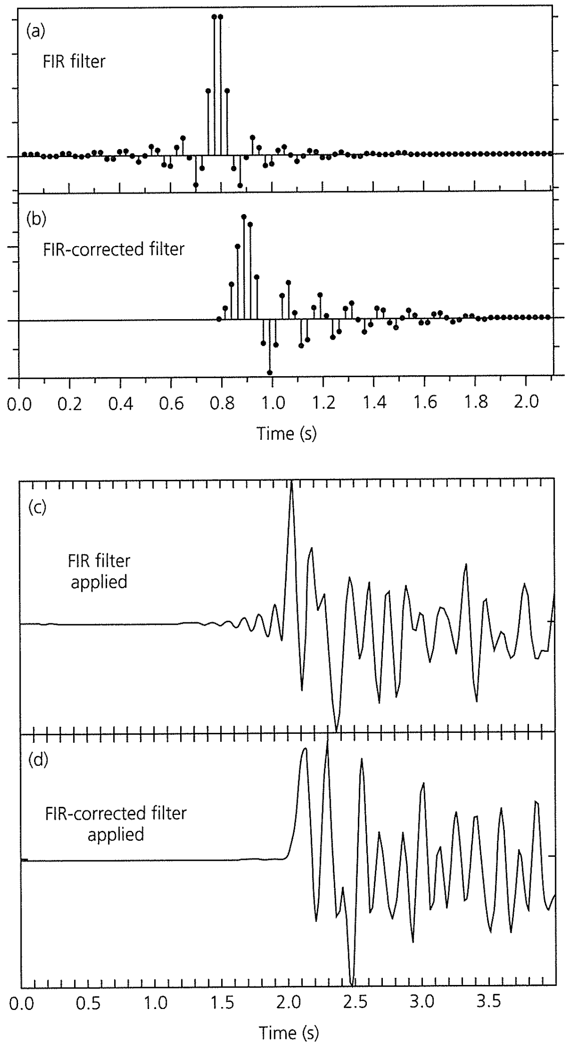


Fig. 6.6-13 Example of a FIR filter, a type of DAA filter, and its effects. When the FIR filter (a) is used for the digital anti-aliasing, the resulting signal (c) retains the wave shape of the original signal, but is preceded by high-frequency artifacts. When a phase-corrected FIR filter (b) is applied instead, the precursory effects vanish (d), but the seismic signal is phase-shifted from the original. (After Scherbaum, 1996, with kind permission from Kluwer Academic Publishers.)

to filter a seismic signal, so we decide what we seek and what we will accept as a consequence.

Because the seismogram depends on the instrument response that is convolved with the ground motion, obtaining the ground motion requires specifying the frequency response of the seismometer. This can be done by giving the amplitude and phase response as a list of the values at each frequency. A more compact representation gives the frequency response as a complex fraction like

$$T(i\omega) = \frac{\beta \prod_{j=1}^{L} (i\omega - z_j)}{\alpha \prod_{k=1}^{N} (i\omega - p_k)}. \qquad (10)$$

The fraction is described by a set of L complex *zeros* z_j at which the numerator is zero, N complex *poles* p_k at which the denominator is zero, and the constants β and α. Because the frequency terms $i\omega$ are always imaginary and the poles always contain a real part, the denominator never becomes zero, avoiding any singular values.

The instrument responses in Fig. 6.6-8 were calculated from the poles and zeroes of the seismometer responses. For example, the STS-1 response has three zeroes, all equal to (0, 0), and four poles, which come as complex conjugates: (−0.0123, 0.0123), (−0.0123, −0.0123), (−39.1800, 49.1200), (−39.1800, −49.1200). These poles provide the corner frequencies and determine the sharpness of the corners. Similarly, the DWWSSN response has five zeroes and 11 poles.

Seismometers record combinations of ground displacement, velocity, or acceleration, depending upon the application. In a strong-motion seismometer, the displacements may be greater than the size of the instrument itself, so accelerations are measured to keep signals on scale. This makes sense because accelerations are primarily responsible for damage to structures and so are considered in strong-motion studies. At the other end of the frequency spectrum, strainmeters are used to study slow tectonic displacements. In fact, if they measured accelerations, the signals would be so small as to be unusable. Most other branches of earthquake seismology fall in between, using the waves from distant earthquakes, and so use seismometers that record ground velocity.

Although different instruments record displacement, velocity, or acceleration, it is simple to convert between them. For instance, given a velocity record, the acceleration is found by taking the derivative of the seismogram, and the displacement record is found by integrating. This is easily done in the frequency domain, because if $F(\omega)$ is the Fourier transform of $f(t)$, then $i\omega F(\omega)$ is the transform of $df(t)/dt$, and $-\omega^2 F(\omega)$ is the transform of $d^2f(t)/dt^2$ (Section 6.2.4). Thus, a velocity seismogram can be converted to acceleration by multiplying the complex value of its transform at each frequency by $i\omega$, or to displacement by dividing by $i\omega$. Of the three, the displacement

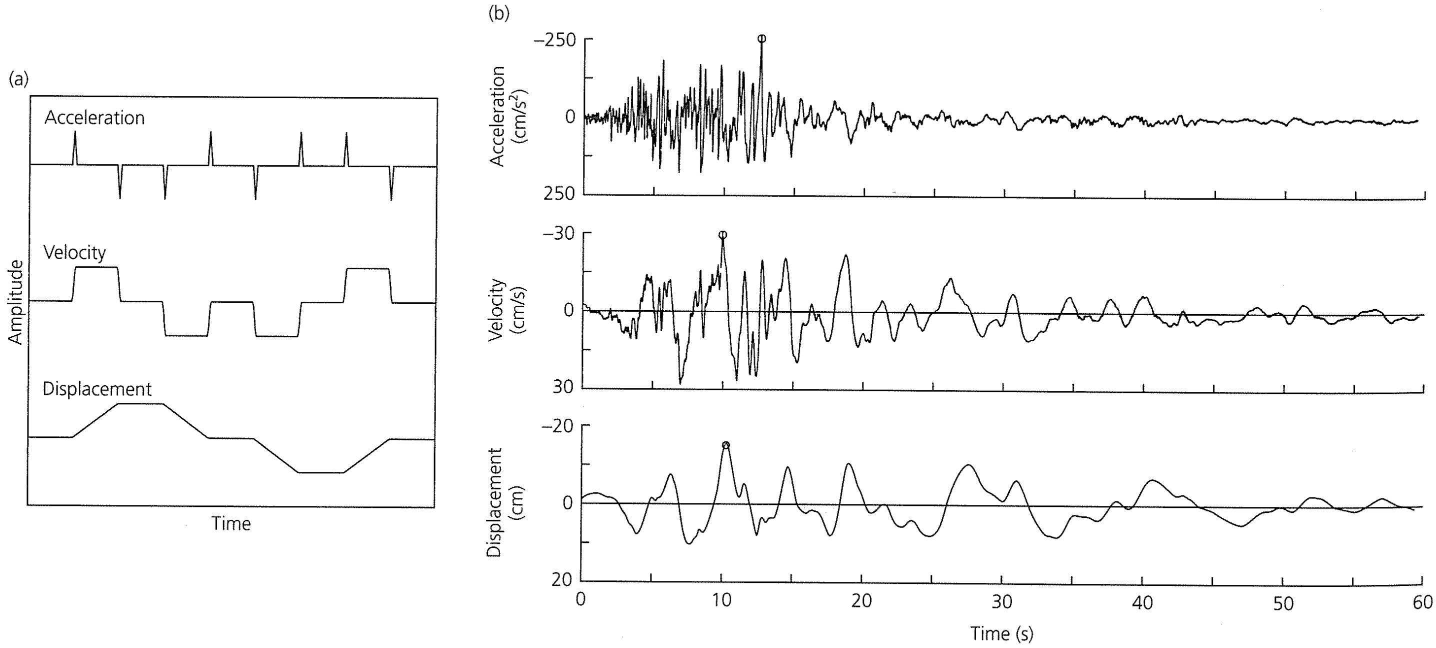


Fig. 6.6-14 Demonstration in the time domain of the relation between displacement, velocity, and acceleration. (a): A synthetic example, consisting of delta function-like acceleration pulses. The velocity and displacement signals are obtained through successive integrations of the accelerogram. (b): A real example, with an accelerogram recorded on the first floor of a building in Los Angeles during the 1971 San Fernando earthquake. The velocity and displacement records were obtained through successive integrations of the accelerogram. (Krinitzsky *et al.*, 1993. *Fundamentals of Earthquake Resistant Construction*. Copyright © 1993. Reprinted by permission of John Wiley & Sons, Inc.)

seismogram has the greatest power at low frequencies, and the acceleration seismogram has the greatest power at high frequencies. In general, displacements have lower frequencies than velocities, and velocities have lower frequencies than accelerations, because integration "smoothes" a signal, whereas differentiation makes it "rougher."[6]

Figure 6.6-14a illustrates this relation with three different versions of the same seismogram. If an accelerogram consists of high-frequency spikes (*top*), then smoother lower-frequency velocity (*center*) and displacement (*bottom*) traces result from integrating once and twice. Figure 6.6-14b shows this effect for a strong-motion seismogram of the 1971 San Fernando earthquake, where the velocity and acceleration records have higher frequencies than the displacement. It is common in earthquake engineering to show the response of a structure to ground motions using a plot that shows the displacement, velocity, and acceleration. Figure 6.6-15 shows this formulation for the data in Fig. 6.6-14b. This representation uses the relation between the Fourier transforms expressed above, so the velocity scale is vertical, whereas the acceleration and displacement scales have opposite slopes as a function of frequency.

6.6.6 *Types of networks*

Most seismic experiments require multiple seismometers that are deployed in networks or arrays. Different applications, such as studying regional and global earth structure, resource exploration, seismicity monitoring, or identifying nuclear tests, lead to different deployment geometries. In some cases a unique network of stations is used for a particular application, but often an existing network has a geometry that is a compromise for different objectives.

Although the division is somewhat artificial, deployments of seismometers are often divided into global networks, regional networks, and arrays. Global networks are used to study global patterns of seismicity, plate tectonics, mantle convection, and earth structure. For these purposes seismometers should ideally be spread evenly around the world. This means, however, that the station spacing is too sparse to resolve the entire wave field.[7] Instead, individual measurements at separate stations are combined for applications including locating earthquakes, 3-D tomography, and waveform analyses.

The antithesis of a global network is a local array, where a set of seismometers is deployed with a geometry chosen for a particular goal. Array data are often analyzed as a single entity, as in refraction and reflection studies (Sections 3.2 and 3.3). Other examples are arrays used to locate distant nuclear tests. Data from the array stations are stacked to track the propagation of the wave field across the array, so the wave vector shows the direction the waves came from and the distance they have traveled. One of several exceptions to this division between global networks and arrays is normal mode seismology, where all the stations of a global network are sometimes used as a single array.

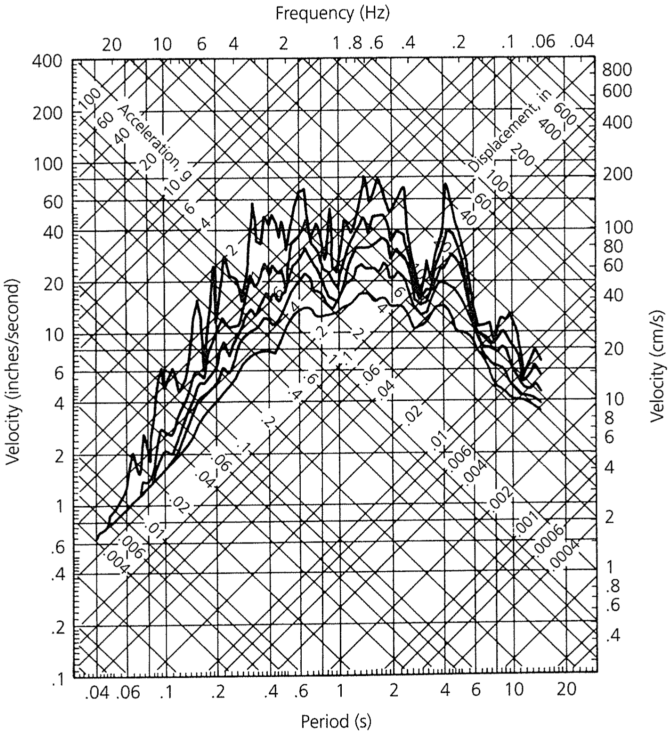


Fig. 6.6-15 Demonstration in the frequency domain of the relation between displacement, velocity, and acceleration. In this example, taken from the accelerogram in Fig. 6.6-14b, a site response spectrum of the building housing the strong-motion seismometer is given as displacement, velocity, and acceleration. The multiple curves show the amplitude of the building response at various levels of damping, with the undamped curve at the top, and successive levels of damping at 2%, 5%, 10%, and 20% of critical damping. (Krinitzsky *et al.*, 1993. *Fundamentals of Earthquake Resistant Construction*. Copyright © 1993. Reprinted by permission of John Wiley & Sons, Inc.)

Between global networks and arrays are regional networks, which usually focus on the seismicity or structure of a particular region. The data are sometimes analyzed with array techniques, but are more often combined as individual measurements (such as arrival times or amplitudes) in the same way as global network data.

6.6.7 *Global networks*

The global network of seismometers has a rich history. At the start of the twentieth century there were already seismometers

[6] An analogy might be to compare displacement and velocity to the topography and gradient of a mountain. A kilometer of topography over a horizontal wavelength of a meter would be very unusual, but a kilometer of topography over a longer wavelength of 5–10 km would be a normal mountain. Similarly, large vertical gradients are rare at the scale of mountains (El Capitan in Yosemite and the Jungfrau in Switzerland are exceptions), but common at the higher spatial frequency scale of meters, as where a path goes over a boulder.

[7] By analogy to time series, such undersampling is termed *spatial aliasing*.

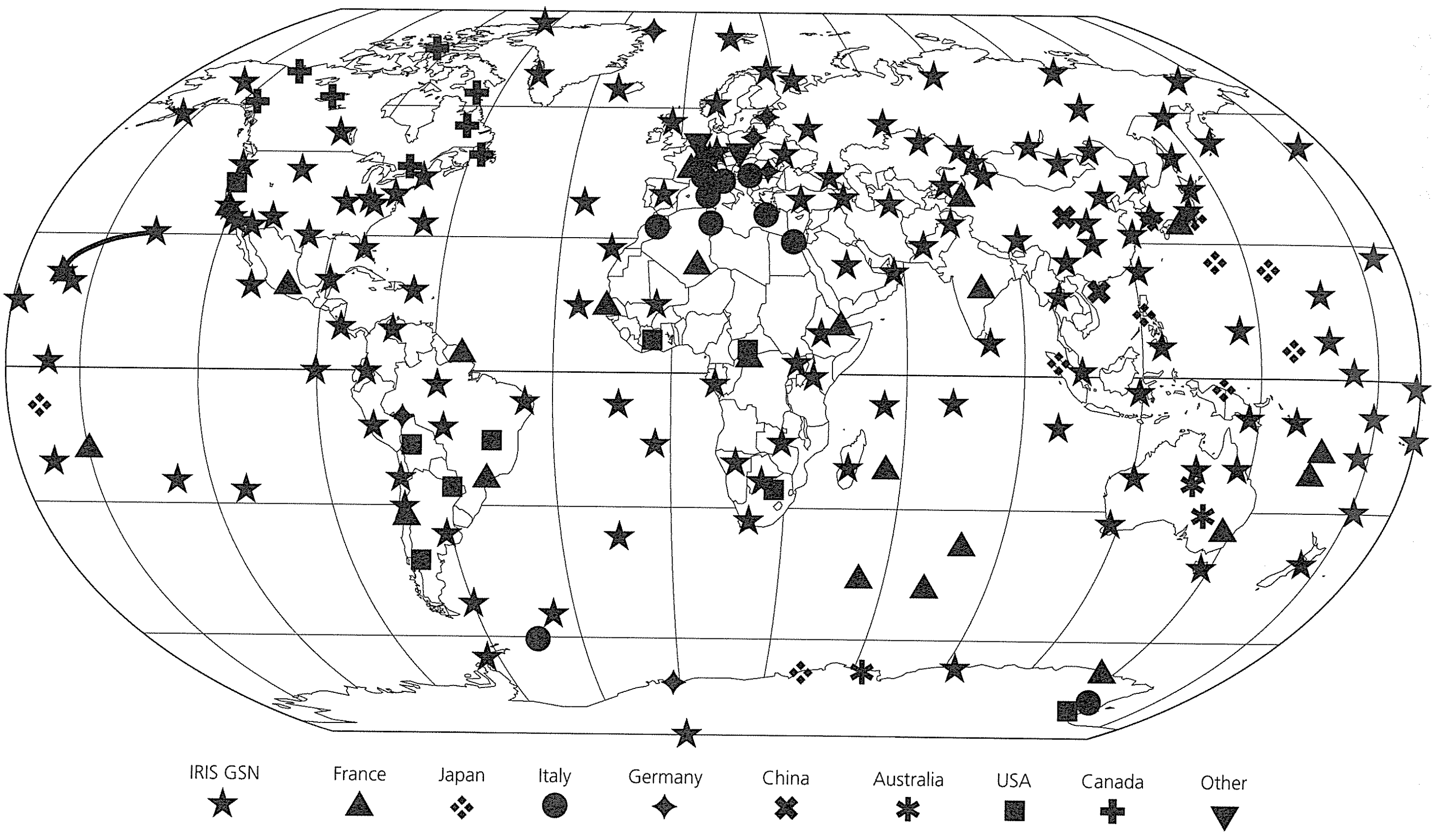


Fig. 6.6-16 Station map of the Federation of Digital Broad-Band Seismographic Networks (FDSN) as of 1999. (Courtesy of the Incorporated Research Institutions for Seismology.)

in locations around the world, operated by groups including many Jesuit institutions. Devastating earthquakes such as the 1906 San Francisco and 1923 Tokyo events spurred the installation of seismometers and the interchange of data. Bulletins of earthquake locations were published by several agencies, the most notable being the ISS/ISC bulletin (Section 6.6.1). By mid-century, the ISS received arrival times from several hundred stations for very large earthquakes. However, there were problems due to a lack of standardization. Different types of seismometers were used, with a wide range in the quality of the response, timing, and station operation practices. As a result, earthquake locations were often poor, and focal mechanisms, which require accurate information about polarities, were rarely derived.

These problems were largely solved with the creation of the World Wide Standardized Seismographic Network. WWSSN seismometers were standardized and had known responses. The network was installed, starting in 1961, to monitor nuclear testing within Eurasia, and had a high density of stations around the borders of the Soviet Union, China, and Eastern Europe. The WWSSN, which reached its peak of about 120 stations in the late 1960s, gave a great boost to geophysics. Several great earthquakes in the 1960s, such as the 1964 Alaska earthquake, provided excellent sources for seismic investigations. WWSSN data were crucial for advances in plate tectonics, earthquake source studies, and global velocity structure.

The first digital stations began to be deployed in the 1970s. Over the next two decades, the number of permanent digital seismometers increased gradually. Following the phase-out of the WWSSN, these became part of the Global Digital Seismic Network, the primary means of global broadband data collection between 1977 and 1986. The GDSN was enhanced by the network of IDA gravimeters, beginning in 1977, and by the French GEOSCOPE network, which has deployed broadband seismometers since 1982.

In 1986, the GDSN gave way to the IRIS Global Seismographic Network (GSN) program, which incorporates many borehole seismometers with an aim toward global coverage, with 128 stations spaced about 2000 km apart. These are extremely quiet, permanent broadband seismic stations of the highest quality. The GSN is part of a larger Federation of Digital Broad-Band Seismographic Networks (FDSN) that also includes the US National Seismographic Network (NSN) and networks from other countries including Canada (CNSN), China (CDSN), France (GEOSCOPE), Germany (GEOFON), Italy (MEDNET), Japan (Pacific 21), and Taiwan (BATS). FDSN station locations are shown in Fig. 6.6-16. Some FDSN stations are also part of the International Monitoring

System (IMS) network used to monitor nuclear testing (Section 1.2.8).

Although the present global network of broadband seismometers relies on land sites, it is hoped that the global network will soon include permanent ocean bottom seismometers (OBS), especially in the Southern Hemisphere, where there is much less land, and coverage is currently very uneven. Although OBS instruments are currently used mostly for temporary deployments, the technology is evolving to the point where permanent sites are practical.

An important aspect of the different networks of high-quality broadband seismometers is considerable standardization in data processing and formatting. All 7 terabytes of seismic data archived by the IRIS DMC[8] as of 2000 are available in a format called SEED (Standard for the Exchange of Earthquake Data), which is the standard for the FDSN. SEED data can be converted into whatever format an investigator requires.

It was not until the mid-1990s, more than 30 years after the start of the WWSSN, that the global number of permanent digital broadband seismometers surpassed the number of WWSSN stations at its heyday. However, digital data from all parts of the FDSN can be retrieved as if it were a single array, making it more powerful than the WWSSN for seismic analyses. Many stations now report in *real time* through satellite telemetry, so seismic signals arrive at data centers a fraction of a second after they occur, allowing better quality control. Efforts are being made to eventually have all GSN stations report in real time, which will be important for applications like tsunami warning. Software has been developed to take real-time data from different networks and display it on the Internet as if it were from a single array. Hence, anyone with a computer and access to the Internet will soon be able to examine global seismic data within seconds of them being recorded.

6.6.8 *Arrays*

For global networks, the precise configuration of individual stations is less important than the total coverage. However, the geometries of seismic arrays are optimized for certain investigations. Arrays can be linear, two-dimensional, and even three-dimensional, incorporating borehole seismometers (Fig. 7.3-8).

There is always a trade-off between the benefits of linear versus two-dimensional arrays. The same number of stations, and therefore cost and time for installation, provides greater resolution if deployed in a linear manner, but the resulting two-dimensional "slice" into the earth does not image the third dimension. Linear arrays have long been the mainstay of *active source* reflection and refraction experiments.[9] A marine linear array is easily deployed by towing hydrophones behind a ship, and similar linear deployments are used for land-based studies. These data are analyzed using techniques discussed in Sections 3.2 and 3.3.

Linear arrays are most useful if the structure being investigated varies most in one direction, as is often the case at plate boundaries. For instance, Fig. 5.3-10 (*bottom*) showed the seismic structure of the East Pacific rise obtained from an array of OBSs. Because the structure of the lithosphere changes much more significantly perpendicular to the ridge than parallel to it, most of the OBSs were deployed in a line crossing the ridge. Most of the remaining seismometers were placed in a second line, parallel to the first. Both lines were aligned along a great circle path to the seismogenic zones of Tonga and South America, so as to maximize the chance of obtaining good signals from distant earthquakes. Similarly, at subduction zones and transform faults structure varies more significantly across the plate boundary than along it, so refraction lines are often placed perpendicular to the boundary. For example, Fig. 3.2-17 showed a cross-section of the western US lithosphere perpendicular to the San Andreas fault that was derived from refraction surveys.

Two-dimensional arrays can create a three-dimensional image of a small region. As a result, two-dimensional arrays have been deployed around hot spots, rifts, plateaus, transform faults, and subduction zones to study their structure and tectonics. Reflection data are also now commonly gathered by two-dimensional surface deployments. An important contributor to this development has been advances in computers and graphics software that make it possible to analyze and model such data and display the resulting earth structure in a comprehensible fashion. Such three-dimensional images are of great importance in exploring for oil and gas and managing existing oil and gas fields.

Special two-dimensional arrays, often consisting of short-period vertical seismometers, have been used to monitor the locations and magnitudes of underground nuclear tests. The most ambitious such array was the circular Large Aperture Seismic Array (LASA), which operated in Montana from the mid-1960s until 1978. LASA was an array of arrays totaling 525 high-frequency vertical seismometers. Twenty-one clusters of 25 seismometers, each covering 7 km^2, were deployed with a total array diameter of 200 km (Fig. 6.6-17). A similar array is the Norwegian Seismic Array (NORSAR), built in 1971, with 22 sub-arrays spanning an area of 100 km^2. Part of NORSAR, the NORESS array, has 24 seismometers distributed within a 3 km-diameter circle. It has counterparts in northern Norway, Finland, and Germany. As with the WWSSN, arrays designed for nuclear monitoring have also been important for studies of earth structure. Array data can be stacked (Section 6.5), allowing small seismic signals to be extracted from noise. The characteristics of the inner core boundary were first quantified using stacked array data for *PKiKP* waves, which reflect at the boundary but are rarely identified on individual seismograms due to their small amplitudes.

[8] Because all data are duplicated in a sort order, and also stored off site, the computer storage needed is four times greater, or 28 Tbytes.

[9] *Active* experiments include their own seismic sources, as opposed to *passive* experiments using earthquake sources.

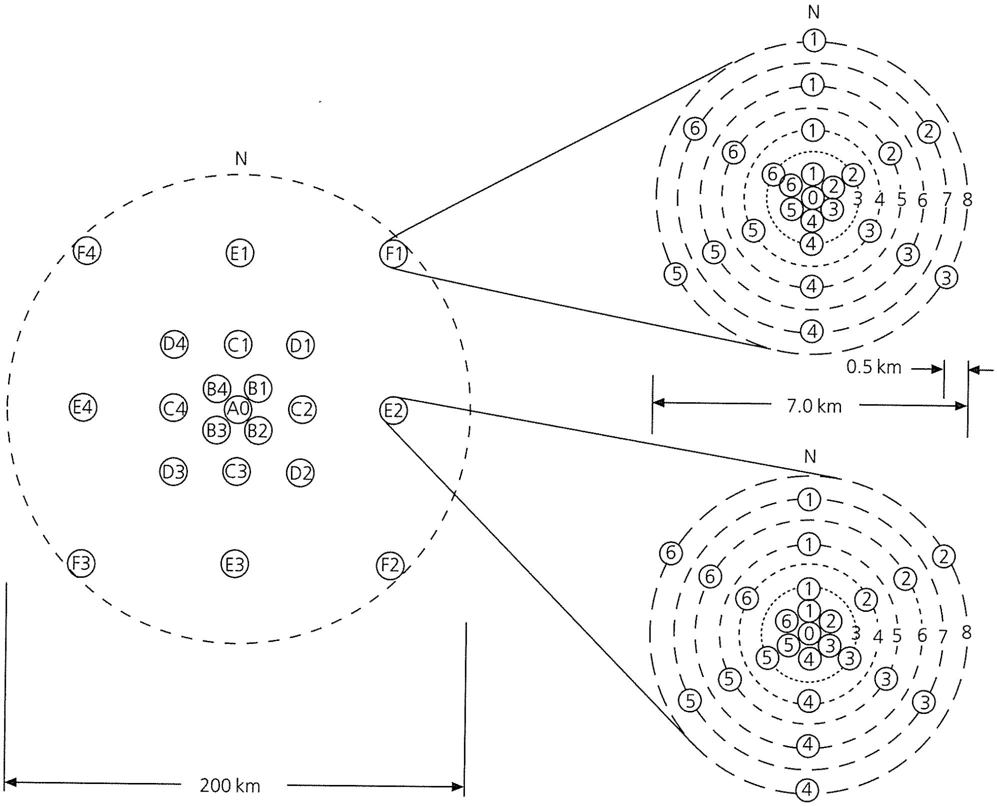


Fig. 6.6-17 Seismometer geometry of the Large Aperture Seismic Array (LASA). (Capon, 1969. *J. Geophys. Res.*, *74*, 3182–94, copyright by the American Geophysical Union.)

6.6.9 *Regional networks*

Regional networks, intermediate between global networks and arrays, are usually constructed to monitor local seismicity or volcanism. Including Alaska, Hawaii, and Puerto Rico, over 3200 seismic stations are part of more than 40 separate US networks (Fig. 6.6-18). Some have only a few stations, and some have hundreds. Many use short-period vertical sensors, but some use accelerometers. For example, the California Strong-Motion Instrumentation Program operates more than 400 accelerometers to provide data for earthquake engineers. Strong-motion data also provide excellent information on source properties because much of the seismic signal is severely attenuated at teleseismic distances. Some networks also incorporate broadband seismometers. For instance, as of 2000, the Southern California Seismographic Network operated 79 broadband stations in addition to its 163 short-period instruments. Regional network stations can also be valuable for earth structure studies, as shown in Fig. 6.6-19.

Many countries have regional networks. For instance, as of 1999, Japan had about 560 stations in operation. These stations have provided valuable data about the subduction process there, including the double seismic zones (Fig. 5.4-20) and *ScS*-to-*P* conversions at the slab top (Fig. 2.6-15).

Regional networks, like global networks, are continually being upgraded. In the USA there are efforts under way, as part of the Advanced National Seismic System (ANSS), to install more broadband and short-period seismometers, and to add about 6000 strong-motion sensors in urban areas at risk from damaging earthquakes. A very ambitious network planned is the USArray, which would have three different components operating simultaneously. First, the number of permanent broadband stations would be increased (Fig. 6.6-20, *left*). Second, 400 portable broadband seismometers would travel around the country. Over eight years, this "bigfoot" array would visit about 2000 sites in the continental USA, with an average station spacing of about 70 km, before going to Alaska and Hawaii (Fig. 6.6-20, *right*). Third, about 2400 seismometers (a mix of broadband, short-period, and high-frequency sensors) would be used as flexible arrays to accompany the moving array. As planned, USArray will be an array at the scale of a regional network. Data from the moving array will be available in near-real time, and can be processed using migration techniques to attain high-resolution imaging deep into the mantle.

Interestingly, because there is an increasing trend toward real-time telemetry for transmitting data from the sensors, seismology is moving toward a situation where data from global networks, regional networks, and many local arrays can be easily combined, largely eliminating the distinctions between networks. This development offers great scientific opportunities.

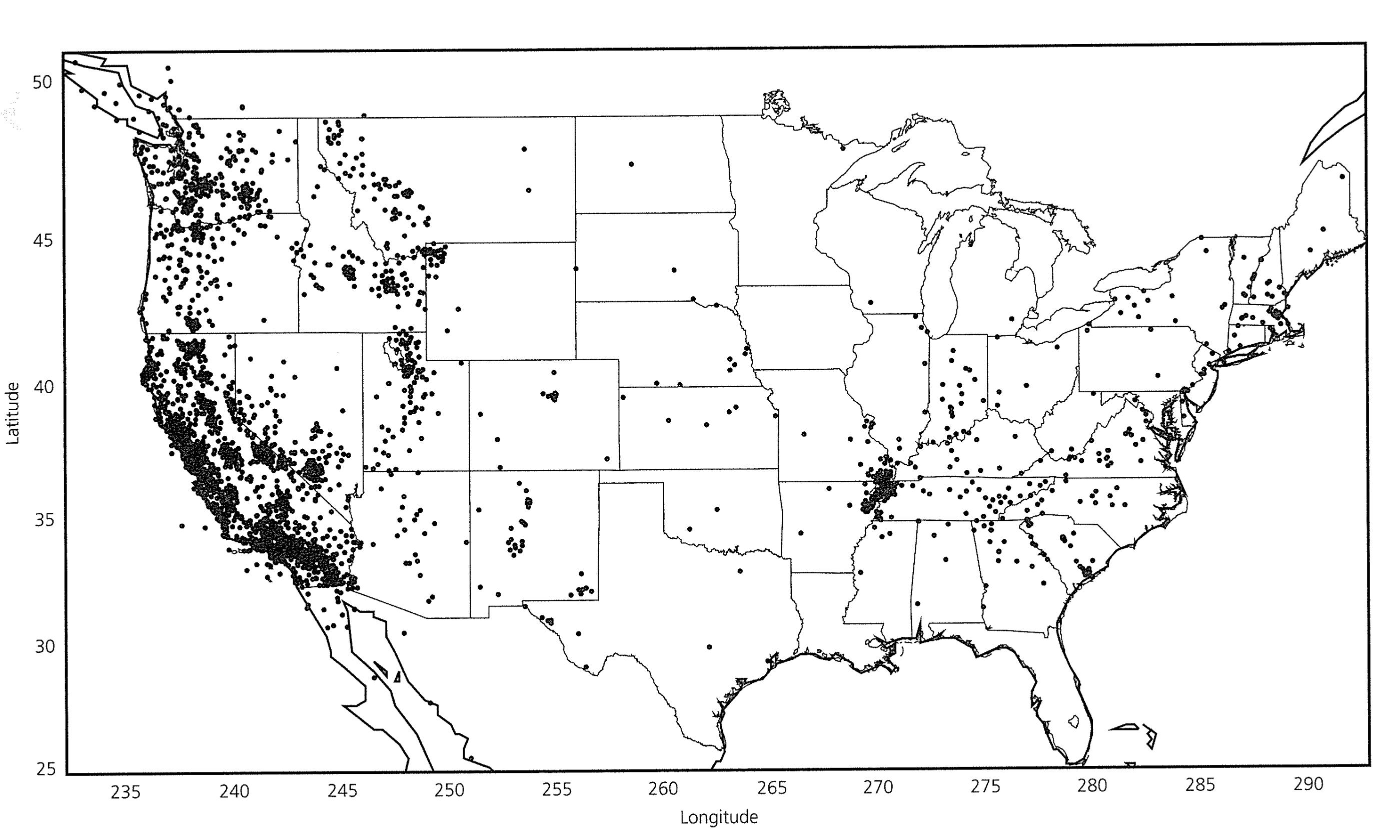


Fig. 6.6-18 Map of regional network seismometer stations in the continental USA as of 1999. Some networks are cooperatively operated with Canadian and Mexican institutions.

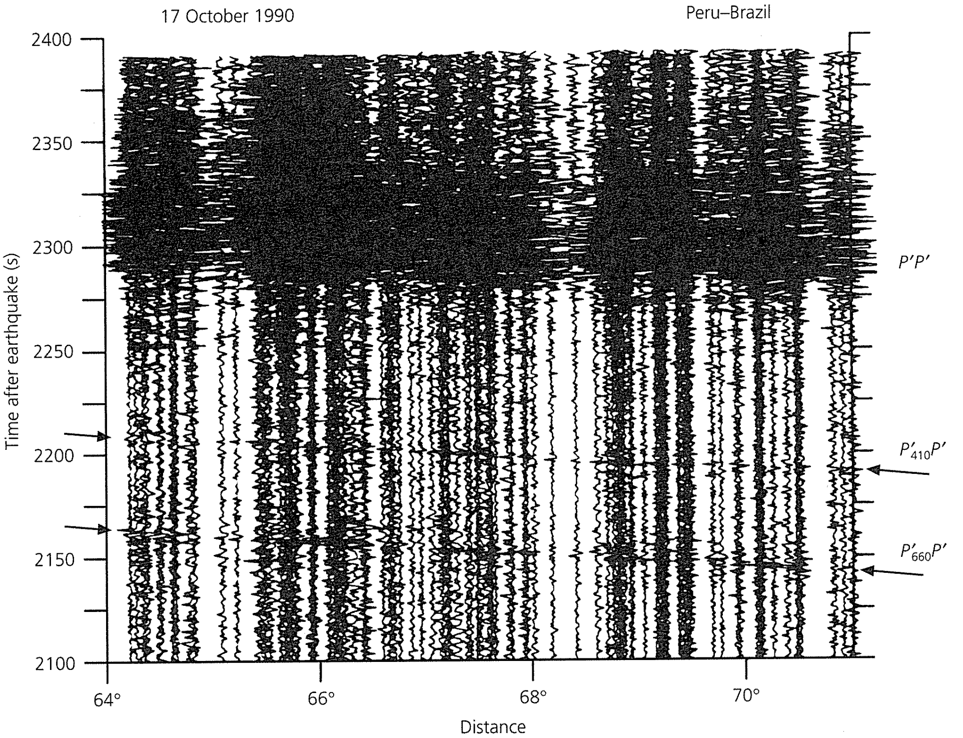


Fig. 6.6-19 Records from the short-period seismometers of California regional networks for an Oct. 17, 1990, earthquake in South America. The data reveal distinct reflections off the sharp 410 km and 660 km mantle discontinuities. The ability to examine large amount of data over a small geographical region greatly increases the resolution of earth structure. (Benz and Vidale, 1993. Reproduced with permission from *Nature*.)

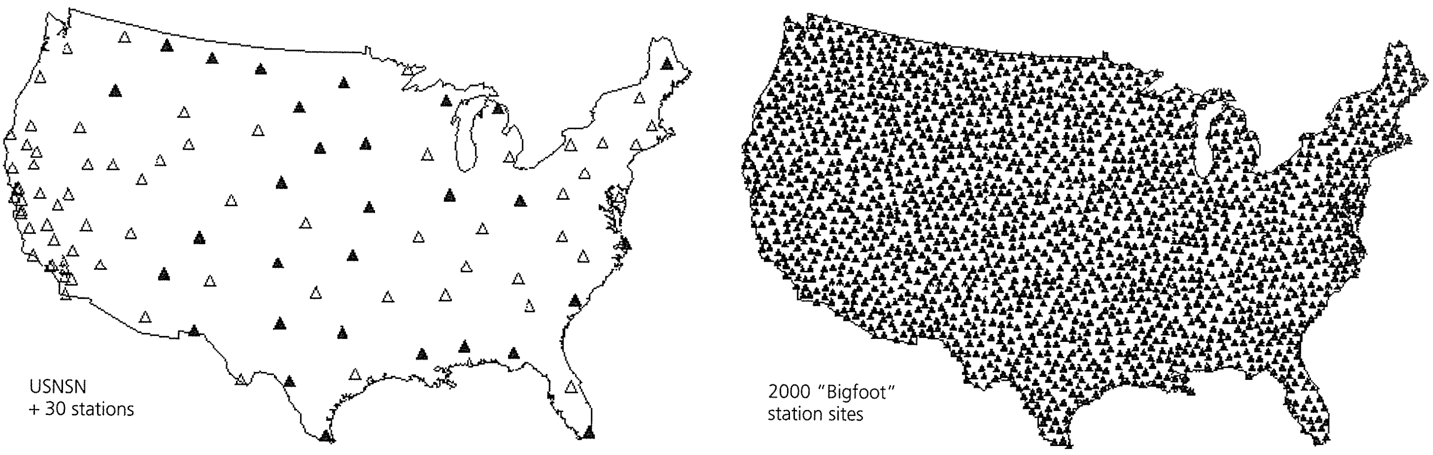


Fig. 6.6-20 Seismometer locations for the proposed USArray. *Left*: Solid triangles would be new permanent seismometers to augment the existing US National Seismic Network (open triangles). *Right*: Possible locations of 2000 sites that the moving array of 400 broadband seismometers would eventually cover. (Courtesy of P. Shearer.)

Further reading

Because of its widespread use, an excellent literature is available both for signal processing in general and for geophysical applications. These include introductory texts by Rabiner and Rader (1972), Claerbout (1976), Bracewell (1978), Robinson and Treitel (1980), Kanasewich (1981), and Hatton *et al.* (1986). Brigham (1974) discusses the FFT in detail.

Error analysis in the physical sciences is the subject of many books, including Bevington and Robinson (1992). Seismological texts, especially Aki and Richards (1980) and Lay and Wallace (1995), discuss seismological instrumentation. Scherbaum (1996) addresses seismometry, especially digital, from a signal processing viewpoint.

Problems

1. Find the coefficients analytically of the Fourier series for the functions
   (a) A step:
   $$\begin{array}{rll} f(t) = & 1 & 0 < t < 1/2 \\ & -1 & -1/2 < t < 0. \end{array}$$
   (b) A ramp: $f(t) = t$ for $-1/2 < t < 1/2$.
2. Use the formulae for the product of sine and cosine functions (Section A.2) to prove the orthogonality relations for the sine and cosine functions (Eqns 6.2.2–4).
3. Express the following complex numbers in $a + ib$ form:
   (a) $e^{i\pi}$
   (b) $4e^{i\pi/2}$
   (c) $e^{-i\pi/2}$
   (d) $3e^{i\pi/3}$
4. In the Fourier series (Eqn 6.2.1), no b_0 term is given. Why?
5. Show that
   (a) The Fourier transform is linear: if $F(\omega)$ and $G(\omega)$ are the transforms of $f(t)$ and $g(t)$, then $(aF(\omega) + bG(\omega))$ is the transform of $(af(t) + bg(t))$.
   (b) The Fourier transform of a purely real-time function has the symmetry $F(-\omega) = F^*(\omega)$.
   (c) The total energy in a Fourier transform is the same as that in the corresponding time series (Parseval's theorem):
   $$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(\omega)|^2 d\omega.$$
6. If $F(\omega)$ is the Fourier transform of $f(t)$, show that the following are also transform pairs:
   (a) $f(t-a)$ and $e^{-i\omega a}F(\omega)$,
   (b) $F(\omega - a)$ and $e^{iat}f(t)$,
   (c) df/dt and $i\omega F(\omega)$.
7. For $f(t) = \sin \omega_0 t$,
   (a) Find the Fourier transform.
   (b) Compare it to the Fourier transform of $f(t) = \cos \omega_0 t$.
   (c) Explain what operation (filter) in the frequency domain could be used to convert the Fourier transform of $\sin \omega_0 t$ to that of $\cos \omega_0 t$.
   (d) Explain how the relation between the Fourier transforms of $\sin \omega_0 t$ and $\cos \omega_0 t$ could be derived using the fact that one function is a time-shifted version of the other.
8. Show that if $f(t)$ and $F(\omega)$ are a transform pair, the inverse transform of $F(\omega)$ yields $f(t)$.
9. Use the propagation of errors relation (Eqn 6.5.18) to show how the uncertainty in the following functions of several variables depends on the variances and covariances of the variables u and v, where a and b are constants:
   (a) $z = au + bv$,
   (b) $z = auv$,
   (c) $z = au/v$,
   (d) $z = au^b$.
10. For the discrete Fourier transform and inverse discrete Fourier transform, show that:
    (a) The DFT and IDFT are linear: if $A(k)$ and $B(k)$ are the transforms of time series $a(n)$ and $b(n)$, then $\alpha A(k) + \beta B(k)$ is the transform of $\alpha a(n) + \beta b(n)$.

(b) The DFT of a real-time series has the symmetry $F(-k) = F(N-k) = F^*(k)$.

(c) If the DFT of $f(n)$ is $F(k)$, the DFT of $f(n-j)$ is $W^{kj}F(k)$, and the IDFT of $F(k-m)$ is $W^{-mn}f(n)$, where $W = e^{-2\pi i/N}$.

11. As derived in Eqn 4.3.10, the depth h of an earthquake can be estimated from the difference in arrival times δt between the direct P wave and pP, the P wave reflected from the surface, using $\delta t = (2h \cos i)/v$ where i and v are the incidence angle and velocity.

(a) Express the depth as a function of the parameters δt, v, i.

(b) Find the depth for a measured time difference of 2.7 s and assumed velocity of 6.8 km/s and incidence angle of 24°.

(c) Use the propagation of errors relation to show how the uncertainty in depth depends on the uncertainties of the three parameters.

(d) Use the results of (c) to find the uncertainty in depth corresponding to uncertainties (one standard deviation) of 0.5 s in time difference, 0.5 km/s in velocity, and 3° in incidence angle. (Remember to convert to radians.)

Computer problems

C-1. Using the Fourier series coefficients for the step function, derived in problem 1a, plot the first ten terms of the series and their sum. Also plot the sum of the first 20 and 30 terms.

C-2. Write a subroutine to prepare a time series for taking the fast Fourier transform and take it. The subroutine should call a set of separate subroutines that extend the time series to a power of 2 as required, allow for a taper of a length which you input, take the FFT using the subroutine (COOLB) provided (Box 6C-2) or another, and plot the amplitude spectrum. The subroutine should have the option to list the real and imaginary parts of the spectrum, and the amplitude and phase spectra, at each frequency.

C-3. (a) Write a subroutine to generate values of the function $\sin \frac{2\pi t}{T}$ from $t = 0$ to $t = T_{\max}$, where the time step Δt, the period T, and the total data length $T_{\max}$ are inputs.

(b) Plot this function for $\Delta t = 0.25$, T = 5, $T_{\max} = 20$.

(c) Use the results of C-2 to find the amplitude spectrum, with no tapering and with 10% and 20% tapering.

(d) Do parts (b) and (c) for $\Delta t = 0.25$, T = 8, $T_{\max} = 50$.

(e) Do parts (b) and (c) for the function

$$\sin \frac{2\pi t}{5} + (0.5) \sin \frac{2\pi t}{8},$$

with $\Delta t = 0.25$, $T_{\max} = 256$.

Box 6C-2 COOLB subroutine.[1]

```
      SUBROUTINE COOLB(NN,DATAI, SIGNI)
C CLASSIC - BUT USABLE - FFT PROGRAM
C DATAI IS DATA ARRAY, 2*NP REAL NUMBERS REPRESENTING
C NP COMPLEX POINTS, SO EACH PAIR OF POINTS ARE THE
C (REAL, IMAGINARY) PARTS OF A COMPLEX NUMBER.
C NN IS POWER OF TWO, CAN BE FOUND BY
C NN=(ALOG10(FLOAT(NP))/ALOG10(2.))+.99
C TRANSFORM DIRECTION CONTROLLED BY REAL VARIABLE
C SIGNI (SIGN OF EXPONENTIAL):-1. FORWARD, 1. TO
C INVERT.
C DIMENSIONS: IF TIME SERIES HAS TIME INCREMENT DT,
C TRANSFORM HAS DELTA FREQ=1/(2**NN*DT)
C NOTE: AFTER TAKING INVERSE FFT DIVIDE OUTPUT BY 2**NN
      INTEGER NN
      REAL SIGNI
      DIMENSION DATAI(1)
      N=2**(NN+1)
      J=1
      DO 5 I=1,N,2
      IF(I-J)1,2,2
 1    TEMPR=DATAI(J)
      TEMPI=DATAI(J+1)
      DATAI(J)=DATAI(I)
      DATAI(J+1)=DATAI(I+1)
      DATAI(I)=TEMPR
      DATAI(I+1)=TEMPI
 2    M=N/2
 3    IF(J-M)5,5,4
 4    J=J-M
      M=M/2
      IF(M-2)5,3,3
 5    J=J+M
      MMAX=2
 6    IF(MMAX-N)7,10,10
 7    ISTEP=2*MMAX
      THETA=SIGNI*6.2831831/FLOAT(MMAX)
      SINTH=SIN(THETA/2.)
      WSTPR=-2.0 *SINTH*SINTH
      WSTPI=SIN(THETA)
      WR=1.
      WI=0.
      DO 9 M=1,MMAX,2
      DO 8 I=M,N,ISTEP
      J=I+MMAX
      TEMPR=WR*DATAI(J)-WI*DATAI(J+1)
      TEMPI=WR*DATAI(J+1)+WI*DATAI(J)
      DATAI(J)=DATAI(I)-TEMPR
      DATAI(J+1)=DATAI(I+1)-TEMPI
      DATAI(I)=DATAI(I)+TEMPR
 8    DATAI(I+1)=DATA(I+1)+TEMPI
      TEMPR=WR
      WR=WR*WSTPR-WI*WSTPI+WR
 9    WI=WI*WSTPR+TEMPR*WSTPI+WI
      MMAX=ISTEP
      GO TO 6
10    RETURN
      END
```

[1] COOLB, written in 1960s vintage Fortran (note the arithmetic IF statements), has been left in original form to illustrate both the persistence of programs that work and the advantages of subsequent developments in programming practice and documentation (Section A.8.2).

C-4. (a) Write a subroutine, using the results of C-2, to use the fast Fourier transform to take a time series, filter it in the frequency domain over a specified passband, and invert the FFT, yielding a filtered time series. The subroutine should have the capability to taper in the frequency domain. This subroutine is best written as a set of subroutines.

(b) Use this routine to filter the time series in C-3e to isolate the two different frequency components.

C-5. (a) Write a subroutine, using the results of C-2 and C-4, to use the fast Fourier transform to convolve two time series.

(b) Use it on two boxcar functions of unit amplitude, one 6 s long and one 3 s long.

C-6. (a) Write a subroutine to do time domain convolution of two functions of different lengths, both sampled at a time step Δt.

(b) Use it on two boxcar functions of unit amplitude, one 6 s long and one 3 s long. Compare the results to those of C-5b.

7 Inverse Problems

Most people, if you describe a train of events to them, will tell you what the result would be. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk of reasoning backwards.

Sherlock Holmes, in *A Study in Scarlet* by Arthur Conan Doyle

7.1 Introduction

Throughout this book we have noted that seismology is largely directed at solving inverse problems dealing with earthquake sources and earth structure. We start with the end result, seismograms, and work backwards to characterize the earthquakes that generated the seismic waves and the medium through which the waves passed. To do this, we first addressed the forward problems of how features of seismic waves that are observable from seismograms, such as travel times, amplitudes, waveforms, eigenfrequencies, dispersion, and attenuation, depend on the seismic source and the medium. We have also discussed how the properties of the medium and the source, such as velocity structure and earthquake mechanisms, reflect tectonic processes within the earth. These are specific examples of the fundamental question of what we can say about the earth from seismological and other observations at its surface.

We now end our discussions by addressing some issues in solving inverse problems. Inverse problems can be posed by assuming that we understand the physics of a process which, for a set of model parameters described by a vector **m**, gives rise to a set of observed data described by the vector **d**. The data can thus be considered the result of a function, or operator, A acting on the model parameters,

$$\mathbf{d} = A(\mathbf{m}). \tag{1}$$

The forward problem, predicting the data **d** that would result from a given model described by **m**, is tractable if we understand the process. The corresponding inverse problem, finding what gave rise to a specific set of observed data, is more difficult. We assume that some physical model describes the process, and then use the data to estimate a set of model parameters that are consistent with the data. We solve the inverse problem using either mathematical inverse techniques to find **m** directly from **d**, or trial-and-error techniques that solve the forward problem repeatedly and look for the best solution. Each approach has advantages in some applications.

We have already mentioned solving inverse problems in contexts including studying the cooling of oceanic lithosphere using surface wave dispersion (Section 2.8.3), inverting travel time and amplitude data to find earth structure (Chapter 3), inverting polarity, waveform, and geodetic data to study earthquake mechanisms (Chapter 4), and using earthquake mechanisms to study plate motions and regional tectonics (Chapter 5). We have noted (Section 1.1.2) that although forward problems typically can be solved in a straightforward way, giving a unique solution, inverse problems often have no unique, exact, or "correct" solutions. Because the data are generally somewhat inconsistent due to errors, and our models simplify complex reality, no model exactly describes the data. Similarly, a range of parameters can describe the data equally well for a given model, and we have various models to choose from based on various criteria and preconceptions. Moreover, the data are often insufficient to resolve aspects of the model. We can thus only recognize and accept these limitations on the solutions.[1]

A consequence of these limitations is a trade-off between the model's *resolution*, how detailed it is, and its *stability*, or robustness. For example, inverting travel times with simple earthquake location algorithms using a laterally homogeneous velocity model shows the Wadati–Benioff zones of dipping seismicity. These results are stable, in that they do not depend

[1] This situation is summarized by the title of a paper "Interpretation of inaccurate, insufficient, and inconsistent data" (Jackson, 1972).

Table 7.1-1 Some large-scale reference models.

Model for	Observables inverted and predicted	Parameters estimated	Misfits ("anomalies") indicate
Laterally homogeneous earth structure	Travel times, eigenfrequencies	Average velocity and density versus depth	Lateral velocity variation (subduction zones, continental–ocean differences, etc.)
Relative plate motions	Rates and azimuths of plate motion	Euler vectors	Nonrigid plate behavior (plate interiors and boundary zones)
Thermal evolution of oceanic lithosphere	Variation with age in depth, heat flow, and geoid	Plate thickness, asthenospheric temperature, physical properties (e.g., α, κ, k)	Lateral thermomechanical variations (swells, etc.)

significantly on the details of the location algorithm and velocity model, but have only limited resolution for where in the slab the earthquakes occur. More detailed locations, which are more useful for relating the earthquakes to the physics of subduction, can be derived from sophisticated location algorithms using a laterally variable velocity model that better represents the slab. However, the improved resolution comes at the price of stability, in that it depends on the specific velocity model used.

The results of inverse studies can be viewed in terms of two end members. In one, we use an individual set of data to characterize a specific phenomenon, such as the location of an earthquake or the velocity structure in a specific area. In others, we describe a set of data averaged over a region or the whole earth with a simple physical model characterized by a relatively small, or *sparse*, set of parameters. Such *reference models* — the physical model with a specific set of parameters — are used to characterize large sets of data in a simple way, predict data where no observations exist, and thus identify misfits, or "anomalies," where the data deviate from the model predictions and hence the global average. We then use reference models to draw inferences about the processes that give rise to both the average situation and deviations from it. For example, body wave, surface wave, and normal mode data give average global velocity structure. This structure is used to constrain models of the average radial variations in composition and temperature, and as a reference against which velocity perturbations due to subducting slabs, continental roots, hot spots, ridges, etc. can be identified and analyzed in terms of local processes that perturb the global model. As shown in Table 7.1-1, we can view other reference models in a similar way. For example, the Euler vectors describing a plate's motion are a simple description of its behavior, and places where earthquake mechanisms differ from these predictions indicate deviations from rigid plate behavior. Similarly, simple cooling models of the oceanic lithosphere describe the average variations in depth, heat flow, and the geoid, and so give a reference model for the temperature against which other effects can be identified and modeled.

As illustrated in Fig. 1.1-8, the models are refined over time using new data and model parameterizations. Eventually, the reference model does not improve significantly. When this occurs, we are probably doing about as well as possible with this type of model. For example, as discussed in Section 3.5, laterally homogeneous global seismic velocity models have become sufficiently accurate that more attention is now directed toward the lateral variations.

In this chapter, we discuss several inverse problems to introduce some of the methods used. Because such inverse problems are crucial to seismology and the earth sciences, and also appear in other sciences, considerable attention has been directed toward them. It turns out that physically quite different problems are often described in mathematically similar ways. Our goal is to identity some common themes and approaches, rather than discuss the details. Some more sophisticated treatments are listed in the suggested reading.

7.2 Earthquake location

We first consider the classic inverse problem of locating an earthquake and finding its origin time using the arrival times of seismic waves at various stations. The velocity structure, which determines the ray paths and hence travel times, is crucial. We first regard the velocity structure as known, and then explore how it can also be estimated from the travel times.

7.2.1 Theory

Assume that an earthquake occurred at an unknown time t, at an unknown position $\mathbf{x} = (x, y, z)$, known as the *hypocenter*, or *focus* (Fig. 7.2-1). The point (x, y) on the surface above the focus is called the *epicenter*. n seismic stations at locations $\mathbf{x}_i = (x_i, y_i, z_i)$ detect the earthquake at arrival times d'_i, which depend on the origin time t and the travel time between the source and the station $T(\mathbf{x}, \mathbf{x}_i)$:

$$d'_i = T(\mathbf{x}, \mathbf{x}_i) + t. \qquad (1)$$

If the velocity structure is known, the forward problem can be written using the formulation

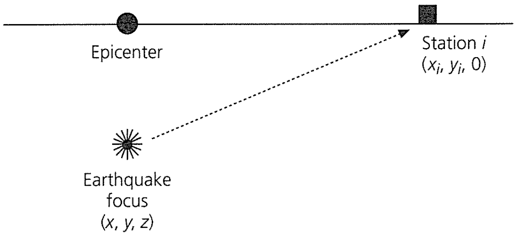


Fig. 7.2-1 Geometry for earthquake location in a homogeneous (uniform velocity) halfspace.

$$\mathbf{d} = A(\mathbf{m}), \quad \text{or} \quad d_i = A(m_j), \tag{2}$$

showing how the data vector, containing the arrival times at the stations, can be computed from an assumed model vector composed of the source location and origin time,

$$\mathbf{m} = (x, y, z, t). \tag{3}$$

The model vector consists of physically different quantities: three space coordinates and an origin time. Because the data and model are vectors, relations between them can be written in terms of either vectors ($\mathbf{d} = A(\mathbf{m})$) or their components ($d_i = A(m_j)$).

The inverse problem can be stated as: given the observed arrival times, find a model that fits them. To do this, we begin with a *starting model* $\mathbf{m}^\circ$, which is an estimate of (or guess at) a model that we hope is close to the solution we seek. The starting model predicts that we would have observed data $d_i^\circ = A(m_j^\circ)$. Unless we are lucky, these predicted data are not what were actually observed. Hence we seek changes Δm_j in the starting model

$$m_j = m_j^\circ + \Delta m_j \tag{4}$$

that will make the predicted data closer to those observed. In general, the data do not depend linearly on the model parameters, so we *linearize* the problem by expanding the data in a Taylor series about the starting model $\mathbf{m}^\circ$ and keeping only the linear term,

$$d_i \approx d_i^\circ + \sum_j \frac{\partial d_i}{\partial m_j}\bigg|_{\mathbf{m}^\circ} \Delta m_j. \tag{5}$$

This equation can be written in terms of the difference between the observed data and those predicted,

$$\Delta d_i^\circ \equiv d_i' - d_i^\circ \approx \sum_j \frac{\partial d_i}{\partial m_j}\bigg|_{\mathbf{m}^\circ} \Delta m_j^\circ. \tag{6}$$

Such relations are common in inverse problems. For simplicity, we omit the superscripts and define the partial derivative matrix as

$$G_{ij} = \frac{\partial d_i}{\partial m_j}, \tag{7}$$

so the equation becomes

$$\Delta\mathbf{d} = G\Delta\mathbf{m}, \quad \text{or} \quad \Delta d_i = \sum_j G_{ij}\Delta m_j. \tag{8}$$

Often the Δs are also suppressed, and the equation is written as $\mathbf{d} = G\mathbf{m}$. This makes the notation simpler, but can be confusing at first. In this derivation, we retain the Δs to explicitly indicate changes.

Equation 8 is a vector–matrix equation representing a system of simultaneous linear equations. To solve it, we seek a change in the model $\Delta\mathbf{m}$ that, when multiplied by the known partial derivative matrix G, gives the required change in the data $\Delta\mathbf{d}$. This is an inverse problem, in contrast to the forward problem of finding the change in the data $\Delta\mathbf{d}$ predicted by an assumed change $\Delta\mathbf{m}$ in the model. Many aspects of inverse theory deal with solving such equations under various circumstances. The earthquake location problem considered here is a simple case.

A common complexity is that we generally have arrival time observations at many (often several hundred) seismic stations, and are solving for only four model parameters. In the notation of Eqn 8, j ranges from 1 to 4, and i ranges from 1 to n, where n is much greater than 4. Because each arrival time corresponds to one equation, and each model parameter provides one unknown, G has a number of rows equal to the number of arrival time observations, and a number of columns equal to the number of model parameters. Because there are more (n) equations than unknowns (4), G has more rows than columns, so Eqn 8 looks like

$$\begin{pmatrix} \Delta d_1 \\ \Delta d_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \Delta d_n \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} & G_{13} & G_{14} \\ G_{21} & G_{22} & G_{23} & G_{24} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ G_{n1} & G_{n2} & G_{n3} & G_{n4} \end{pmatrix} \begin{pmatrix} \Delta m_1 \\ \Delta m_2 \\ \Delta m_3 \\ \Delta m_4 \end{pmatrix}. \tag{9}$$

Such *overdetermined* problems can pose difficulties. One way to see this is to recall that if n were equal to 4 the matrix G would be square (have the same number of rows and columns), so Eqn 8 could be solved by multiplication by the inverse matrix,

$$G^{-1}\Delta\mathbf{d} = G^{-1}G\Delta\mathbf{m} = \Delta\mathbf{m}, \quad \text{or}$$

$$\sum_i G_{ki}^{-1}\Delta d_i = \sum_i G_{ki}^{-1}\left(\sum_j G_{ij}\Delta m_j\right) = \Delta m_k. \tag{10}$$

If the number of arrival time observations exceeds four, this method cannot be used, because G is not square and thus does not have an inverse.[1] Our first instinct might be to use only arrival times at four stations, which would give an exact solution, and assume that the arrival times at the other stations give only extra, redundant information. In an ideal world this would be the case. In reality, the arrival times contain errors due to a variety of possible effects, including reading errors, inaccuracies in the clocks at the stations, and misidentification of the first arrivals. In addition to these errors of measurement, there are systematic errors due to the fact that the velocity structure is not perfectly known and is laterally variable. As a result, the equations are *inconsistent*: no one model can solve them exactly. Moreover, choosing four arrival times might mean selecting data poorer than those discarded. The approach taken instead is to seek the origin time and source location that "best" solve the overdetermined, inconsistent equations.

To do this, we regard the observations d_i' as having errors described by their standard deviations σ_i and find the model that minimizes the misfit,

$$\chi^2 = \sum_i \frac{1}{\sigma_i^2}\left(\Delta d_i - \sum_j G_{ij}\,\Delta m_j\right)^2, \tag{11}$$

which is the prediction error, the normalized sum of the squares of the difference between the observed arrival times and those predicted by the model. x^2, the fitting function to be minimized, weights the data by the reciprocal of their variances so that the most uncertain have the least effect. To find the best fit, we set partial derivatives of the misfit with respect to the change in model parameters Δm_k equal to zero, and use the fact that the model elements are independent, so the partial derivative of the change in one with respect to those in the others is zero,

$$\frac{\partial \Delta m_j}{\partial \Delta m_k} = \delta_{jk}. \tag{12}$$

The partial derivatives of the misfit are

$$\frac{\partial \chi^2}{\partial \Delta m_k} = 0 = 2\sum_i \frac{1}{\sigma_i^2}\left(\Delta d_i - \sum_j G_{ij}\Delta m_j\right)G_{ik}, \tag{13}$$

or

$$\sum_i \frac{1}{\sigma_i^2}\Delta d_i G_{ik} = \sum_i \frac{1}{\sigma_i^2}\left(\sum_j G_{ij}\Delta m_j\right)G_{ik}. \tag{14}$$

If the variances of the data are equal ($\sigma_i^2 = \sigma^2$), that term can be factored out, and

$$\sum_i \Delta d_i G_{ik} = \sum_i \left(\sum_j G_{ij}\Delta m_j\right)G_{ik}, \tag{15}$$

or, in matrix notation,

$$G^T\Delta\mathbf{d} = G^T G\Delta\mathbf{m}. \tag{16}$$

To see that $\sum_i \Delta d_i G_{ik} = G^T\Delta\mathbf{d}$, whereas $\sum_j G_{ij}\Delta m_j = G\Delta\mathbf{m}$, consider the dimensions.

The advantage of this form is that although the matrix G cannot be inverted, the matrix G^TG is square and can be inverted. Equation 16 thus gives $\Delta\mathbf{m}$, the standard least squares solution to a set of equations that cannot be solved exactly, because

$$\Delta\mathbf{m} = (G^TG)^{-1}G^T\Delta\mathbf{d} = G^{-g}\Delta\mathbf{d}, \quad \text{or} \quad \Delta m_j = \sum_i G_{ji}^{-g}\Delta d_i. \tag{17}$$

The operator $(G^TG)^{-1}G^T$, which acts on the data to yield the model, is called the *generalized inverse* of G, and is written as G^{-g}. It provides the "best" solution in a least squares sense, because it gives the smallest squared misfit. The generalized inverse is the analog of the inverse, but for a matrix that is not square, and hence does not have a conventional inverse. If G is square and has an inverse, then $G^{-1} = G^{-g}$. If the data errors are not equal, the least squares solution is weighted by the errors, as shown in problem 5 at the end of this chapter.

To use this method, we begin with a starting model (source location and origin time) $\mathbf{m}^\circ$ and predict the values expected for the data, $\mathbf{d}^\circ = A(\mathbf{m}^\circ)$. We then form the residual vector giving the misfit to the data, $\Delta\mathbf{d}^\circ \equiv \mathbf{d}' - \mathbf{d}^\circ$, evaluate the matrix of partial derivatives about the starting model,

$$G_{ij} = \left.\frac{\partial d_i}{\partial m_j}\right|_{\mathbf{m}^\circ}, \tag{18}$$

and use the generalized inverse (Eqn 17) to find $\Delta\mathbf{m}^\circ$, the change in the starting model that gives a better fit to the data. Thus the new model

$$\mathbf{m}^1 = \mathbf{m}^\circ + \Delta\mathbf{m}^\circ \tag{19}$$

predicts values of the data

$$\mathbf{d}^1 = A(\mathbf{m}^1) \tag{20}$$

that should be closer to the observations than the predictions of the starting model. This can be tested by computing the difference between the observations and the predicted data for the new model $\Delta\mathbf{d}^1 \equiv \mathbf{d}' - \mathbf{d}^1$, and examining the total squared misfit $\Sigma(\Delta d_i^1)^2 = \Sigma(d_i' - d_i^1)^2$. This should be less than the corresponding misfit for the starting model $\Sigma(\Delta d_i^\circ)^2$. The total squared misfit is more useful than the total misfit $\Sigma\Delta d_i$, because the latter could be small for large misfits of opposite signs.

We can often do even better. Remember that the G matrix of partial derivatives was found by expanding the function that predicts the data (travel times) about the starting model in a Taylor series, and taking the linear terms. This expansion

[1] The definition of the inverse (Section A.4.3) requires that both pre- and post-multiplication yield the identify; i.e., $A^{-1}A = AA^{-1} = I$.

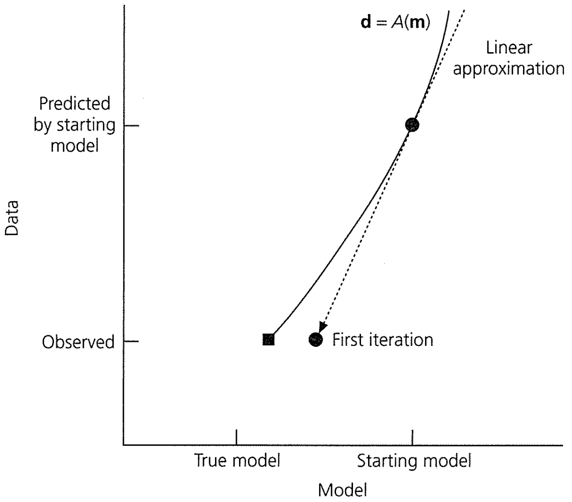


Fig. 7.2-2 Schematic illustration of the effect of linearizing about a starting model in an inverse problem. The new model is found from the difference between the observed data and that predicted for the starting model. The worse the linear approximation is, the more iterations will be needed to reach the true model.

works well if the starting model is "close" to the actual model. If this is not the case, the linear approximation may not be a good one. Figure 7.2-2 illustrates this idea schematically. The actual situation is hard to draw, because each model vector is an element in a four-dimensional (three space and one time) vector space.

As a result, the method can be iterated. Once the model has been changed, a new partial derivative matrix

$$G_{ij} = \left.\frac{\partial d_i}{\partial m_j}\right|_{\mathbf{m}^1} \tag{21}$$

is found by expanding the function that predicts the data about the *new* model. The generalized inverse method is then used to solve

$$\Delta \mathbf{d}^1 = G\Delta \mathbf{m}^1 \tag{22}$$

for a further change in the model $\Delta\mathbf{m}^1$ that reduces the remaining misfit. This process is repeated until successive iterations produce only small changes in the model, and hence in the total misfit to the data (Fig. 7.2-3).

7.2.2 *Earthquake location for a homogeneous medium*

To make these ideas less abstract, we consider the simple case of locating an earthquake in a medium of uniform velocity v. In this case the ray paths connecting an earthquake and seismic stations are straight lines. This geometry approximates a situation where the receivers are close enough to a source that the first arrivals are direct waves in a medium whose velocity does not vary significantly. Seismic waves from an earthquake that

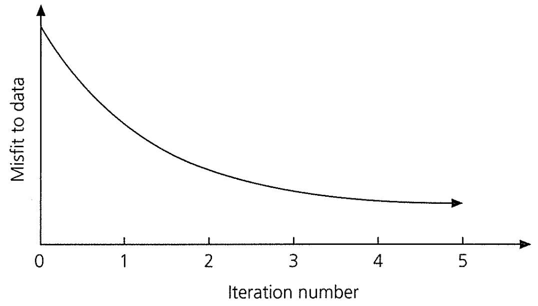


Fig. 7.2-3 Schematic illustration of the variation in misfit to the data as a function of iteration number for an inverse problem.

occurred at time t at location $\mathbf{x} = (x, y, z)$ are recorded by seismic stations at positions $\mathbf{x}_i = (x_i, y_i, z_i)$ with arrival times

$$d_i = T(\mathbf{x}, \mathbf{x}_i) + t = \frac{1}{v}[(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2]^{1/2} + t. \tag{23}$$

Although the earthquake can occur below the surface, the stations are at the surface $z_i = 0$. The travel times depend only on the distance between source and receiver, $|\mathbf{x} - \mathbf{x}_i|$.

To solve the inverse problem, we form the matrix G_{ij}. Its elements, the partial derivatives of the elements of the data vector d_i (the arrival times at each station) with respect to the model parameters m_j (the location coordinates and origin time of the earthquake) are easily found. Differentiation of the i^{th} element of the data vector is done with respect to the first element of the model vector, which is the x coordinate of the location

$$G_{i1} = \frac{\partial d_i}{\partial m_1} = \frac{\partial d_i}{\partial x} = \frac{\partial T(\mathbf{x}, \mathbf{x}_i)}{\partial x}$$
$$= \frac{(x - x_i)}{v}[(x - x_i)^2 + (y - y_i)^2 + z^2]^{-1/2}. \tag{24}$$

Similar expressions give the partial derivatives with respect to the other two space coordinates of the location. Note that these partial derivatives are functions of the spatial model parameters (x, y, z). The final partial derivative, with respect to origin time, is just

$$G_{i4} = \frac{\partial d_i}{\partial m_4} = \frac{\partial d_i}{\partial t} = 1. \tag{25}$$

Given the G matrix, the earthquake is located by choosing a starting model, forming the difference $\Delta\mathbf{d}$ between the model predictions and the observations, and solving for the change in the model $\Delta\mathbf{m}$ using the procedure in the last section.

Table 7.2-1 (*top*) illustrates a hypothetical example of locating an earthquake with ten stations located within a 100 km square. The earthquake is assumed to have occurred at time 0 seconds at the point (0, 0, 10) km. We then try to locate the

Table 7.2-1 Earthquake location example with error-free data.

Invert for location and origin time				
model evolution				
parameter	actual value	model for iteration number		
		0	1	2
x	0.0	3.0	-0.5	0.0
y	0.0	4.0	-0.6	0.0
z	10.0	20.0	10.1	10.0
origin time	0.0	2.0	0.2	0.0
station location		residual for iteration number		
		0	1	2
35.0	9.0	-2.1	-0.4	0.0
-44.0	10.0	-3.0	-0.2	0.0
-11.0	-25.0	-3.8	-0.1	0.0
23.0	-39.0	-3.0	-0.2	0.0
42.0	-27.0	-2.6	-0.3	0.0
-12.0	50.0	-2.0	-0.3	0.0
-45.0	16.0	-2.9	-0.2	0.0
5.0	-19.0	-3.7	-0.2	0.0
-1.0	-11.0	-4.1	-0.2	0.0
20.0	11.0	-2.4	-0.4	0.0
error		92.4	0.6	0.0

Invert for location, origin time, and velocity				
model evolution				
parameter	actual value	model for iteration number		
		0	1	2
x	0.0	3.0	0.2	0.0
y	0.0	4.0	0.3	0.0
z	10.0	20.0	10.2	10.0
origin time	0.0	2.0	0.7	0.0
velocity	5.0	4.0	4.9	5.0
station location		residual for iteration number		
		0	1	2
35.0	9.0	-4.0	-0.9	0.0
-44.0	10.0	-5.6	-1.0	0.0
-11.0	-25.0	-5.7	-0.9	0.0
23.0	-39.0	-5.6	-1.0	0.0
42.0	-27.0	-5.2	-1.0	0.0
-12.0	50.0	-4.6	-0.9	0.0
-45.0	16.0	-5.6	-1.0	0.0
5.0	-19.0	-5.2	-0.9	0.0
-1.0	-11.0	-5.3	-0.9	0.0
20.0	11.0	-3.8	-0.8	0.0
error		261.3	8.3	0.0

earthquake using the computed arrival times at the ten stations as "data." For a starting model, we assume the earthquake occurred at time 2 seconds at (3, 4, 20) km. As discussed in the previous section, we compute the arrival times expected at each station for a source located at the initial estimated position and time, and then form the residual, the difference between the "data" and this prediction (Eqn 6). For the starting model, the total squared misfit is 92.4 s^2.

To reduce the misfit, we form the partial derivative matrix G_{ij} evaluated at the starting model, and use the generalized inverse (Eqn 17) to solve for $\Delta\mathbf{m}^\circ$, the change in the starting model that would best fit the residuals. This change gives a source location of (−0.5, −0.6, 10.1) km and an origin time of 0.2 s. This new estimate is close to the true values. Because for a real case the true model would not be known, the new model is tested by calculating the expected arrival times, forming the residuals, and examining the total squared misfit, which is reduced to 0.6 s^2. To reduce this further, we form the partial derivative matrix evaluated at the new model and iterate again. The resulting change in the model yields the true model exactly, which fits the data perfectly.

This success is hardly surprising, because the data had no errors. We could thus have used any four data to find the model, and avoided the generalized inverse. Before turning to discuss the errors, note that the same procedure could be used to find the velocity. To do so, we regard the velocity as a fifth model parameter, and invert the data for a model vector $m = (x, y, z, t, v)$. The additional partial derivatives are

$$\frac{\partial d_i}{\partial m_5} = \frac{\partial d_i}{\partial v} = -\frac{1}{v^2}[(x-x_i)^2+(y-y_i)^2+z^2]^{1/2}. \tag{26}$$

We thus assume a velocity as part of the starting model, find the partial derivative matrix (which now has five columns), and use the generalized inverse to find the changes in the starting model. Table 7.2-1 (*bottom*) illustrates this process for the same example as before, except that we also invert for velocity.

7.2.3 *Errors*

Because earthquakes are located using arrival time data that have errors, the resulting locations and origin times have uncertainties. To assess these uncertainties, we examine how errors in the data affect the generalized inverse solution.

We characterize the errors in the data at the i^{th} station, d_i, by viewing the specific values measured as samples from a parent distribution that includes all possible $d_i^{(k)}$, $k = 1, \ldots \infty$, such that an infinite number of measurements would yield the parent distribution. In this notation, $d_i^{(k)}$ is the k^{th} sample of d_i, the arrival time at station i. Because in real applications the parent distribution for d_i is unknown, it is common to assume a Gaussian distribution with mean $\bar{d}_i$ and standard deviation σ_i, as discussed in Section 6.5. For a large number of measurements (samples) from this distribution, the mean is the average

$$\bar{d}_i = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} d_i^{(k)}, \tag{27}$$

and the "spread" of the measurements is the variance

$$\sigma_i^2 = \lim_{K\to\infty} \left(\frac{1}{K} \sum_{k=1}^{K} (d_i^{(k)} - \bar{d}_i)^2 \right). \tag{28}$$

If the Gaussian parent distribution is an appropriate choice, there is a 68% probability that any sample will fall in the range $\bar{d}_i \pm \sigma_i$, and a 95% probability that any sample will fall in the range $\bar{d}_i \pm 2\sigma_i$ (Fig. 6.5-1).

The errors at different stations are described by the variance–covariance matrix of the data

$$\sigma_{\mathrm{d}}^2 = \sigma_{d_{ij}}^2 = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} (d_i^{(k)} - \bar{d}_i)(d_j^{(k)} - \bar{d}_j). \tag{29}$$

The diagonal ($i = j$) terms are the variances for data at individual stations. The off-diagonal terms ($i \neq j$) are the covariances that describe the relation between errors at pairs of stations. If the errors are uncorrelated between two stations – for example, those due to a station clock – then how a measurement at one station differs from the mean there is unrelated to what occurs at another station, so their covariance is ideally zero. Given a finite number of real data, we expect the covariance to be small. By contrast, if the errors are correlated (for example, if one person were reading seismograms from different stations with a consistent bias), then similar deviations from the mean occur between these stations, and their covariances would be larger. Errors can also be anti-correlated, such that deviations at a station tend to occur in one direction, whereas those at another station tend the other way, yielding negative covariances. Although errors of measurement are likely to be uncorrelated, systematic errors are often correlated. For example, variations in velocity can cause systematic biases that are either correlated or anti-correlated between different stations.

The data are inverted using the generalized inverse solution

$$m_j = \sum_i G_{ji}^{-g} d_i \tag{30}$$

(here the Δs are not written). As a result, the uncertainty in a model parameter can reflect errors in *all* of the data. Thus, even if the errors in the data are uncorrelated, the resulting uncertainties in model parameters can be correlated. To see this, we write the covariances of the model parameters in terms of those for the data

$$\begin{aligned}\sigma_{\mathrm{m}}^2 = \sigma_{m_{ji}}^2 &= \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} (m_j^{(k)} - \bar{m}_j)(m_i^{(k)} - \bar{m}_i) \\ &= \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \left(\sum_p G_{jp}^{-g} (d_p^{(k)} - \bar{d}_p) \right) \left(\sum_s G_{is}^{-g} (d_s^{(k)} - \bar{d}_s) \right) \\ &= \sum_p G_{jp}^{-g} \sum_s G_{is}^{-g} \left(\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} (d_p^{(k)} - \bar{d}_p)(d_s^{(k)} - \bar{d}_s) \right) \\ &= \sum_p G_{jp}^{-g} \sum_s G_{is}^{-g} \sigma_{d_{ps}}^2 . \end{aligned} \tag{31}$$

This relation can be written in matrix form in terms of σ_{d}^2 and σ_{m}^2, the variance–covariance matrices for the data and model:

$$\sigma_{\mathrm{m}}^2 = G^{-g} \sigma_{\mathrm{d}}^2 (G^{-g})^T. \tag{32}$$

We often assume that the data errors are uncorrelated and equal, so that the data variance–covariance matrix is a constant times the identity matrix,

$$\sigma_{\mathrm{d}}^2 = \sigma^2 \delta_{ij}, \tag{33}$$

and the model variance–covariance matrix is

$$\sigma_{\mathrm{m}}^2 = \sigma^2 (G^T G)^{-1}, \tag{34}$$

as proved in problem 4.

Table 7.2-2 illustrates these ideas for the location example in the previous section. In this case, Gaussian errors with mean zero and standard deviation 0.1 s were added to the arrival times. As a result, the data are inconsistent and cannot be fit exactly by any model. The inversion thus changes the model until a good, but not perfect, fit to the data is achieved. This final model, which is no longer changing much after three

Table 7.2-2 Earthquake location example with errors.

Invert for location and origin time					
		model evolution			
parameter	actual value	model for iteration number			
		0	1	2	3
x	0.0	3.0	−0.2	0.2	0.2
y	0.0	4.0	−0.9	−0.4	−0.4
z	10.0	20.0	12.2	12.2	12.2
origin time	0.0	2.0	0.0	−0.2	−0.2
station location		residual for iteration number			
		0	1	2	3
35.0	9.0	−2.0	−0.1	0.1	0.1
−44.0	10.0	−3.0	−0.1	0.0	0.0
−11.0	−25.0	−3.8	0.0	0.1	0.1
23.0	−39.0	−3.2	−0.1	0.0	0.0
42.0	−27.0	−2.8	−0.2	−0.1	−0.1
−12.0	50.0	−2.1	−0.3	−0.1	−0.1
−45.0	16.0	−2.9	−0.1	0.0	0.0
5.0	−19.0	−3.7	−0.1	0.0	0.0
−1.0	−11.0	−4.0	−0.1	0.0	0.0
20.0	11.0	−2.5	−0.3	0.0	0.0
error		93.74	0.33	0.04	0.04
		data standard deviation		0.10	
		model variance–covariance matrix			
0.06	0.01	0.01		0.00	
0.01	0.08	−0.13		0.01	
0.01	−0.13	1.16		−0.08	
0.00	0.01	−0.08		0.01	
		model standard deviation			
x	*y*	*z*		origin time	
0.25	0.28	1.08		0.10	

iterations, is close to, but not exactly, the model used to generate the data. This simple example thus has some features of real situations.

The uncertainties in the final model are shown by the model variance–covariance matrix

$$\sigma^2_{\mathrm{m}} = \begin{pmatrix} \sigma^2_{xx} & \sigma^2_{xy} & \sigma^2_{xz} & \sigma^2_{xt} \\ \sigma^2_{yx} & \sigma^2_{yy} & \sigma^2_{yz} & \sigma^2_{yt} \\ \sigma^2_{zx} & \sigma^2_{zy} & \sigma^2_{zz} & \sigma^2_{zt} \\ \sigma^2_{tx} & \sigma^2_{ty} & \sigma^2_{tz} & \sigma^2_{tt} \end{pmatrix}. \qquad (35)$$

To see that the results seem reasonable, we compare the final inversion model, taking into account its uncertainty, to the true model. The standard deviations of each parameter are given by the square roots of the diagonal terms of the model variance–covariance matrix, so the final model (x = 0.2 ± 0.25 km, $y = -0.4 \pm 0.28$ km, $z = 12.2 \pm 1.08$ km, $t = -0.2 \pm 0.10$ s) is an acceptable representation of the true model.

The model variance–covariance matrix shows some interesting features. The variance of the depth estimate, σ^2_{zz} is larger than the corresponding terms σ^2_{xx} and σ^2_{yy}, indicating that the depth is less well constrained than the epicenter. This situation is common, and arises because all the seismometers are at the surface.[2] In some cases when the depth is poorly constrained, it is regarded as fixed, and only the epicenter and the origin are inverted for. The results of multiple inversions, each with the depth fixed at a different value, are compared to see which best fits the data. It is also possible to determine the depth from other criteria, such as the times of surface reflections (Section 4.3), and then invert with the depth fixed.

The uncertainties in the model parameter estimates are correlated, because the off-diagonal elements of the model variance–covariance matrix are nonzero. σ^2_{zt}, the covariance of the depth and origin time uncertainties, is negative, indicating a trade-off between the focal depth and the origin time. At any station, similar arrival times result if the earthquake occurred earlier (t smaller) but deeper (z larger). Similarly, σ^2_{xy}, the covariance of the x and y location uncertainties, is nonzero, so the uncertainties in these two parameters are correlated. A method often used to illustrate this is to extract the 2 × 2 submatrix

$$\begin{pmatrix} \sigma^2_{xx} & \sigma^2_{xy} \\ \sigma^2_{yx} & \sigma^2_{yy} \end{pmatrix} \qquad (36)$$

and diagonalize it by finding the eigenvalues $\lambda^{(1)}$ and $\lambda^{(2)}$, and the associated eigenvectors $(x_1^{(1)}, x_2^{(1)})$ and $(x_1^{(2)}, x_2^{(2)})$. The uncertainty in the epicenter can then be thought of as an ellipse with semi-major and semi-minor axes $\lambda^{(1)1/2}$ and $\lambda^{(2)1/2}$, oriented in a direction given by $\tan^{-1}(x_1^{(1)}/x_2^{(1)})$. In this case, the semi-major and semi-minor axes have lengths of 0.29 and 0.24 km, and the semi-major axis trends N22°E. An interesting feature of the error ellipse is that its shape and orientation depend on the $(G^TG)^{-1}$ matrix, whereas the variance of the data, σ^2_{d}, controls the size of the ellipse. Because the shape of the error ellipse depends on the geometry of the receivers, it can be examined without reference to specific data. As written, the ellipse is for a confidence level of 1σ (68%), but ellipses are sometimes also given for 2σ(95%), or 3σ(99%).

We have shown that the model variance–covariance matrix depends on the variance–covariance matrix of the data. In the example, we knew the standard deviation of the data and that the errors were uncorrelated. This information would not be available for a real experiment. However, we could estimate the standard deviation of the data from the misfit between the data and the best-fitting model, given by the sample variance s^2,

$$\sigma^2 \approx s^2 = \frac{1}{n-k}\sum_{i=1}^{n}(d_i' - d_i)^2. \qquad (37)$$

Here, d_i' are the observations, d_i are the values of the data predicted by the best-fitting model, and k is the number of model parameters determined from the data. Division by $n - k$, the number of degrees of freedom, rather than by n, the number of data, compensates for the improvement in fit resulting from the use of model parameters determined from the data. Thus, for our example, the final squared misfit is 0.4 s^2, and four parameters were determined from the data, so the sample standard deviation is $s = (0.4/(10-4))^{1/2} = 0.08$ s, a value close to the true σ, 0.1 s.

7.2.4 *Earthquake location for more complex geometries*

This formulation is not restricted to locating earthquakes in a homogeneous halfspace. Velocity variations can be incorporated in the function relating the arrival time at the i^{th} station to the origin time t and travel time $T(\mathbf{x}, \mathbf{x}_i)$,

$$d_i' = T(\mathbf{x}, \mathbf{x}_i) + t. \qquad (1)$$

For example, a model for locating local earthquakes could have a series of layers. As a result, even for a source at the surface, the travel time curve is a more complicated function of distance (Section 3.2). At close distances, the first arrival is the direct wave. At greater distances, the first arrival becomes a head wave from an interface at depth, with the relevant interface being deeper as the distance increases. The situation is similar, but more complicated, for a source at depth, because at zero distance the travel time is nonzero.

The travel time curve can be found either analytically or by tracing rays. If the receivers are on the surface at (x_i, y_i), the travel time curve $T(r, z)$ depends on the *horizontal* distance between source and receiver,

$$r_i = [(x - x_i)^2 + (y - y_i)^2]^{1/2}, \qquad (38)$$

[2] Similarly, vertical positions determined using the GPS (Section 4.5.1) by a process analogous to earthquake location are less precise than the horizontal positions.

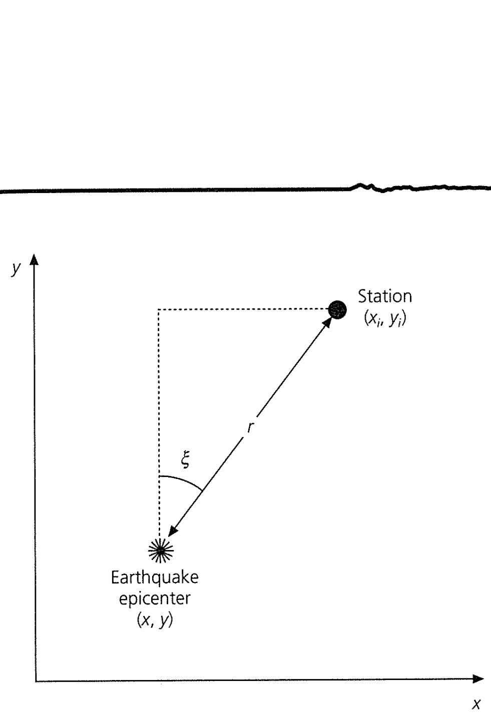


Fig. 7.2-4 Map view of the relation between an earthquake epicenter and a seismic station in Cartesian coordinates.

and the source depth z, so the arrival times are

$$d_i' = T(r_i, z) + t. \tag{39}$$

In this case, the x derivatives are found by

$$\frac{\partial d_i}{\partial x} = \frac{\partial T(r_i, z)}{\partial x} = \frac{\partial T(r_i, z)}{\partial r}\frac{\partial r_i}{\partial x} = \frac{\partial T(r_i, z)}{\partial r}\frac{(x - x_i)}{r_i}, \tag{40}$$

and similarly for the y derivatives. If ζ is the azimuth from the source to the receiver (Fig. 7.2-4),

$$(x - x_i)/r_i = -\sin \zeta_i \quad \text{and} \quad (y - y_i)/r_i = -\cos \zeta_i. \tag{41}$$

If the travel time curve is found numerically, then $T(r_i, z)$ is a set of values for various points (r, z) rather than an explicit function. The procedure for location is still the same, except that the x, y, and z partial derivatives are computed numerically. For example, if we begin by assuming that the source is at $(x^\circ, y^\circ, z^\circ)$, then the partial derivative with respect to r about

$$r_i^\circ = [(x^\circ - x_i)^2 + (y^\circ - y_i)^2]^{1/2} \tag{42}$$

is found using the tabulated travel times for points $(r_i^\circ + \delta/2, z^\circ)$ and $(r_i^\circ - \delta/2, z^\circ)$. Thus the x derivatives are found by approximating the derivative by a difference

$$\begin{aligned}\frac{\partial T(r_i, z^\circ)}{\partial x} &= \frac{\partial T(r_i, z^\circ)}{\partial r}\frac{\partial r_i}{\partial x} \\ &= \frac{T(r_i^\circ + \delta/2, z^\circ) - T(r_i^\circ - \delta/2, z^\circ)}{\delta}\frac{(x^\circ - x_i)}{r_i^\circ},\end{aligned} \tag{43}$$

and the y derivatives are found similarly. The z derivatives are found numerically by forming the difference between two depths. The inversion is then done as before.

The location of earthquakes for a spherical earth is similar. As before, we assume that velocity varies only with depth. In this case, for an earthquake at colatitude θ, longitude ϕ, focal depth z, and origin time t, we seek to estimate the model vector $\mathbf{m} = (\theta, \phi, z, t)$ from the data.

The travel time to receivers on the surface at colatitudes θ_i and longitude ϕ_i depends on the focal depth and the angular distance from the epicenter (Eqn A.7.7),

$$\cos \Delta_i = \cos \theta \cos \theta_i + \sin \theta \sin \theta_i \cos (\phi_i - \phi). \tag{44}$$

For a travel time curve $T(\Delta, z)$ the arrival times are

$$d_i = T(\Delta_i, z) + t. \tag{45}$$

Several average global travel time curves are available, as in Fig. 3.5-4. In addition, a travel time curve for a specific velocity model can be found numerically by tracing rays.

In this case, the θ derivatives are found using

$$\frac{\partial d_i}{\partial \theta} = \frac{\partial T(\Delta_i, z)}{\partial \theta} = \left.\frac{\partial T(\Delta_i, z)}{\partial \Delta}\right|_{\Delta_i} \frac{\partial \Delta_i}{\partial \theta}. \tag{46}$$

To find the last term, note that

$$\frac{\partial(\cos \Delta_i)}{\partial \theta} = \frac{\partial(\cos \Delta_i)}{\partial \Delta}\frac{\partial(\Delta_i)}{\partial \theta}, \tag{47}$$

so

$$\begin{aligned}\frac{\partial(\Delta_i)}{\partial \theta} &= \left(\frac{\partial(\cos \Delta_i)}{\partial \theta}\right) \Big/ \left(\frac{\partial(\cos \Delta_i)}{\partial \Delta}\right) \\ &= \frac{1}{\sin \Delta_i}(\sin \theta \cos \theta_i - \cos \theta \sin \theta_i \cos (\phi_i - \phi)) \\ &= \cos \zeta_i,\end{aligned} \tag{48}$$

where ζ_i is the azimuth of the i^{th} station with respect to the earthquake (Eqn. A.7.10). Thus the partial derivatives with respect to source colatitude are

$$\frac{\partial d_i}{\partial \theta} = \frac{\partial T(\Delta_i, z)}{\partial \Delta}\cos \zeta_i. \tag{49}$$

Similarly, because by the same method

$$\begin{aligned}\frac{\partial(\Delta_i)}{\partial \phi} &= \left(\frac{\partial(\cos \Delta_i)}{\partial \phi}\right) \Big/ \left(\frac{\partial(\cos \Delta_i)}{\partial \Delta}\right) \\ &= \frac{1}{\sin \Delta_i}(-\sin \theta \sin \theta_i \sin (\phi_i - \phi)) \\ &= -\sin \theta \sin \zeta_i,\end{aligned} \tag{50}$$

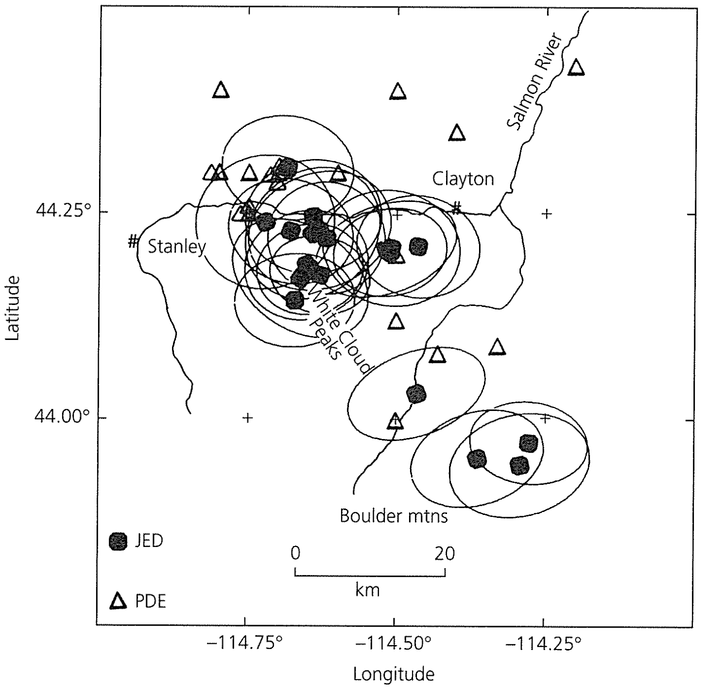


Fig. 7.2-5 Comparison of epicenters for earthquakes in central Idaho derived by a standard location program (PDE, open triangles) and from a joint epicenter determination study (JED, closed symbols). Error ellipses are shown for JED locations. The JED epicenters suggest a narrower source region than the PDE epicenters. (Dewey, 1987. © Seismological Society of America. All rights reserved.)

the partial derivatives with respect to source longitude are

$$\frac{\partial d_i}{\partial \phi} = -\frac{\partial T(\Delta_i, z)}{\partial \Delta} \sin\theta \sin\zeta_i. \tag{51}$$

The two derivatives required from the travel time table, $\partial T(\Delta_i, z)/\partial\Delta$ and $\partial T(\Delta_i, z)/\partial z$, can be approximated by forming differences between tabulated values. This approach is used to locate earthquakes all over the world using teleseismic data, often from hundreds of stations.

We can also locate earthquakes in a laterally varying structure using a numerical representation of the travel time curve. In this case, the travel times, and hence partial derivatives, depend on the actual positions of the source and the receiver, not just on the distance between them. The techniques discussed so far will work, with the modification that the travel times, and hence partial derivatives, must be computed, by tracing rays or otherwise, for each source–receiver pair. The computational effort involved is large enough that laterally homogeneous models are used whenever possible.

A number of methods are sometimes applied to improve locations derived using a laterally homogeneous model. Some treat residuals at individual stations as *station corrections* to be removed. *Master event methods* consider a particular (often the largest) earthquake in a group as the best located, and then locate a group of nearby earthquakes using a travel time correction at each station derived from the residual at each station for the master event. This procedure attempts to locate the other events more accurately with respect to the master event. *Joint hypocenter determination* methods use data from a number of nearby earthquakes, and locate them simultaneously to best fit the travel times. Figure 7.2-5 illustrates applying this technique to a group of earthquakes: the locations from a joint epicenter determination study are more closely grouped and are shifted somewhat from the epicenters for the same events found by the standard location program.

When considering earthquake location, the travel time residuals remaining once the "best" location is found are a nuisance. Following the dictum that "one person's signal is another's noise" brings us naturally to our next topic, the use of these travel time residuals to study deviations from a laterally homogeneous earth model.

7.3 Travel time tomography

In the last section we noted that travel time observations contain information about both the location and the origin time of the seismic source and the velocity structure in the region between the source and receivers. Thus, for the simple halfspace example shown, we also inverted the travel time residuals to find the best velocity. This is analogous to the way in Chapter 3 that we discussed techniques to develop layered models in which velocity varied only with depth. However, we have seen that many of the earth's most interesting processes, such as subduction, cause deviations from a laterally homogeneous velocity model. Methods have thus been developed to use seismological data to investigate laterally heterogeneous structure. For example, we have discussed using lateral variations in surface wave velocities to investigate the cooling of oceanic lithosphere (Section 2.8.3) and migration of seismic reflection data to image variable structure at depth (Section 3.3.7). In this section we introduce the concepts of *travel time tomography*, some of whose results we have seen in Sections 3.7 and 5.4. This discussion illustrates both some further general aspects of inverse problems and some specific features of inverting for earth structure.

7.3.1 *Theory*

Consider the path s of a seismic ray through a medium whose velocity v varies with position. The travel time, T, is

$$T = \int 1/v(s)\,ds = \int u(s)\,ds, \tag{1}$$

the integral of 1/velocity, the slowness, along the ray path. The ray path, in turn, is determined by the velocity distribution. Suppose now that the slowness at various points along the path

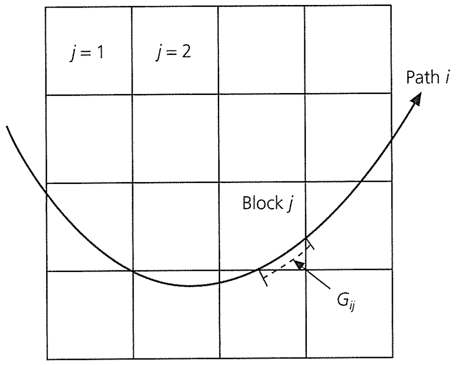


Fig. 7.3-1 Geometry of a region being studied using travel time tomography. The region is divided into blocks j, whose perturbations in velocity are to be found from the travel time along ray paths i. The velocity outside the blocks is assumed to be laterally homogeneous, so travel time perturbations with respect to the reference model are used to find the velocity perturbations within the blocks.

is perturbed by an amount $\delta u(s)$ small enough that the ray path is essentially unchanged, but the travel time changes by

$$\delta T = \int \delta u(s) ds. \tag{2}$$

We can then use the changes in travel time to study the velocity changes that caused them.

Because the travel time perturbation reflects the slowness perturbation integrated along the ray path, a single observation does not indicate how the perturbation is distributed along the path. A large localized perturbation and a smaller, but more widely distributed, one could give the same effect. To improve resolution, data from ray paths that sample the medium differently can be combined (Fig. 7.3-1). The simplest spatial distribution of the slowness perturbation divides the medium into a number of homogeneous subregions termed blocks, or cells. Thus the integral (Eqn 2) giving the travel time perturbation along the i^{th} ray path is written in discrete form

$$\Delta T_i = \sum_{j=1} G_{ij} \Delta u_j, \tag{3}$$

where G_{ij} is the distance the i^{th} ray travels in the j^{th} block, and Δu_j is the slowness perturbation in the block.

Our goal is to use the observed travel times along a number of paths through the medium to recover the slowness perturbation. Problems of this type, in which observations of properties integrated along a number of paths through the medium are used to infer the two- or three-dimensional distribution of the physical property within a medium, occur in many branches of science and are known collectively as *tomography*.[1] The two- or three-dimensional perturbation can be thought of as an image, which we seek to reconstruct from observations. The observations, one-dimensional integrals through the perturbation, are known as projections.

In travel time tomography, the inverse problem of estimating the slowness perturbation from the observed travel time perturbation has the form discussed in the last section

$$\mathbf{d} = G\mathbf{m}, \quad \text{or} \quad d_i = \sum_j G_{ij} m_j. \tag{4}$$

As before, we do not explicitly write the Δs, so the model vector $\mathbf{m}$ is the perturbation in slowness from a starting model, and the data vector $\mathbf{d}$ is the difference between the observed travel times and those predicted by the starting model. The elements of the partial derivative matrix

$$G = \frac{\partial d_i}{\partial m_j} = \frac{\partial T_i}{\partial u_j} \tag{5}$$

equal the distance the i^{th} ray travels in the j^{th} block, which is the partial derivative of the ray's travel time with respect to the slowness in the block.

The matrix G is an operator that relates model vectors and data vectors. As in the location problem, these vectors are physically different quantities with different dimensions. The model vectors have as many elements as there are blocks in the model, whereas the data vectors have a number of elements equal to the number of ray paths. Mathematically, this means that if there are r blocks in the model, any model vector is a vector in an r-dimensional model space. Similarly, if there are n travel times and thus n ray paths, any data vector is a vector in an n-dimensional data space. Because there are generally many more equations (ray paths) than unknowns (model parameters), the system of equations is *overdetermined*. Because the data contain noise, the system of equations is generally also *inconsistent*.

The inverse problem is solved by a procedure like that discussed for the location problem. For the different ray paths, the travel times and the distances traveled in each block are predicted using a starting or reference model. The starting model is generally laterally homogeneous, so the travel times are easily calculated. Travel time residuals are then computed for each ray path by subtracting the times predicted by the starting model from those observed. These travel time residuals form the data vector that is inverted using the generalized inverse to find slowness changes that predict the travel time residuals as well as possible.

To illustrate these ideas, consider a schematic experiment in which a region under a seismic array is divided into four square blocks of unit length (Fig. 7.3-2). Travel time residuals from six ray paths form the data. Four paths (1–4), which can be thought of as due to distant (teleseismic) earthquakes, traverse the model vertically. Two paths (5, 6), which can be thought of as due to local earthquakes, traverse the model horizontally.

[1] This term is Greek for "slice picture."

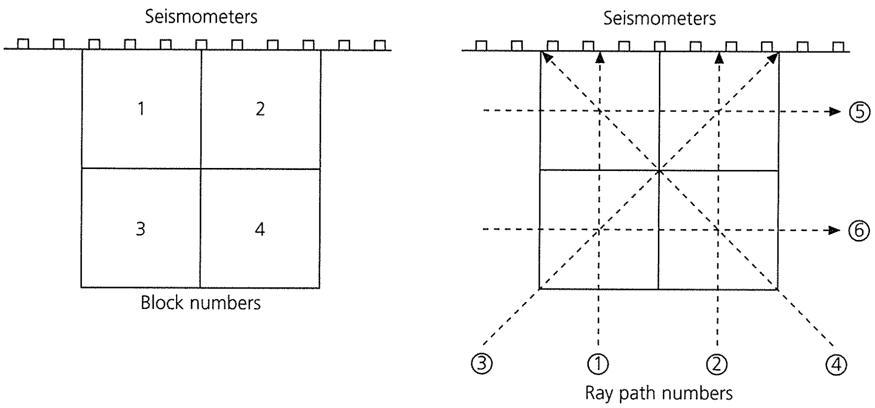


Fig. 7.3-2 Ray path and block geometry for an idealized tomographic experiment. Each block is sampled by three different ray paths.

The reference slowness model is assumed to be appropriate outside the blocks, so the entire travel time residual for each path is attributed to slowness perturbations in the blocks. Thus the problem looks like

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{pmatrix} \tag{6}$$

We encountered this problem, solving a vector–matrix equation where the matrix is not square, in the last section. As in that case, we form

$$G^T G\mathbf{m} = G^T\mathbf{d} \tag{7}$$

and invert the square matrix G^TG to form the generalized inverse solution

$$\mathbf{m}_g = (G^TG)^{-1}G^T\mathbf{d} = G^{-g}\mathbf{d}. \tag{8}$$

We next ask how $\mathbf{m}_g$, the model found by the inversion, compares to the actual slowness model that gave rise to the travel time data. To compare the two, we substitute $G\mathbf{m}$ for $\mathbf{d}$ in Eqn 8, and find that in this case

$$\mathbf{m}_g = (G^TG)^{-1}G^TG\mathbf{m} = \mathbf{m}, \tag{9}$$

so the inversion correctly resolves the true model. Naturally, if errors are present in the data, these errors propagate into the results of the inversion, as discussed previously.

7.3.2 *Generalized inverse*

An interesting situation occurs in this example if only the four teleseismic ray paths (1–4) are available. The inverse problem becomes finding the four elements of $\mathbf{m}$ from

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix}. \tag{10}$$

After multiplying by G^T, we attempt to solve this system as before, but find that the matrix G^TG has a zero determinant, so it cannot be inverted. Thus, although the system of equations (7) has four equations for four unknowns, it does not have a unique solution (Section A.4.4). It turns out that this is because the rows of G are not linearly independent. Thus the ray geometry is not adequate to fully resolve the slowness perturbations in the four blocks.

Because this situation occurs frequently in solving inverse problems, methods for dealing with it have been developed. Although a full treatment is beyond our scope, we summarize some key ideas without proof.

In the general case when G is an $n \times r$ matrix, G^TG is an $r \times r$ symmetric matrix that can be decomposed using its eigenvectors and eigenvalues (Section A.5.3)

$$G^TG = V\Lambda V^T, \tag{11}$$

where the columns of matrix V are the r eigenvectors of G^TG

$$V = \begin{pmatrix} v_1^{(1)} & \cdot & \cdot & \cdot & v_1^{(r)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ v_r^{(1)} & \cdot & \cdot & \cdot & v_r^{(r)} \end{pmatrix} \tag{12}$$

and Λ is a diagonal matrix with eigenvalues on the diagonal and zeroes elsewhere

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \lambda_r \end{pmatrix}. \tag{13}$$

Because the eigenvectors are orthogonal,

$$VV^T = V^TV = I, \quad \text{so} \quad V^T = V^{-1}. \tag{14}$$

If G^TG has an inverse,

$$(G^TG)^{-1} = (V\Lambda V^T)^{-1} = V\Lambda^{-1}V^T, \tag{15}$$

where

$$\Lambda^{-1} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & 1/\lambda_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1/\lambda_r \end{pmatrix}. \tag{16}$$

This expression shows that G^TG is singular if at least one eigenvalue is zero. In this case, the p nonzero eigenvalues are used to form the $p \times p$ diagonal matrix

$$\Lambda_p = \begin{pmatrix} \lambda_1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \lambda_p \end{pmatrix}, \tag{17}$$

and the associated eigenvectors are divided into two matrices:

$$V_p = \begin{pmatrix} v_1^{(1)} & \cdot & \cdot & \cdot & v_1^{(p)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ v_r^{(1)} & \cdot & \cdot & \cdot & v_r^{(p)} \end{pmatrix} \quad \text{and} \quad V_0 = \begin{pmatrix} v_1^{(p+1)} & \cdot & \cdot & \cdot & v_1^{(r)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ v_r^{(p+1)} & \cdot & \cdot & \cdot & v_r^{(r)} \end{pmatrix}. \tag{18}$$

V_p is the $r \times p$ matrix of the eigenvectors with nonzero eigenvalues, and V_0 is the $r \times (r - p)$ matrix of the eigenvectors with zero eigenvalues.

Similarly, the $n \times n$ matrix GG^T can be decomposed as

$$GG^T = U\Lambda U^T, \tag{19}$$

using its eigenvector matrix U. GG^T has the same p nonzero eigenvalues as G^TG, so the U matrix can be divided into U_p, the $n \times p$ matrix of the eigenvectors with nonzero eigenvalues, and U_0, the $n \times (n - p)$ matrix of the eigenvectors with zero eigenvalues. Although we do not prove it here, it is possible to decompose the matrix G using only the eigenvectors with nonzero eigenvalues:

$$G = U\Lambda V^T = U_p\Lambda_p V_p^T. \tag{20}$$

This decomposition, known as the *Lanczos decomposition*, is important, because a generalized inverse

$$G^{-p} = V_p\Lambda_p^{-1}U_p^T \tag{21}$$

that involves only the eigenvectors with nonzero eigenvalues gives an optimal solution to the inverse problem. This solution provides the best fit to the data while minimizing **m**, the change from the starting model. This is a desirable feature: for example, in the tomographic problem, we start with a laterally homogeneous model, so the best solution is that with least lateral velocity variation consistent with the data.

7.3.3 *Properties of the generalized inverse solution*

The relation between the solution to the inverse problem, the model derived from the data using

$$\mathbf{m}_p = G^{-p}\mathbf{d}, \tag{22}$$

and the "true" (although unknown) model **m**, can be found because the data are related to the "true" model by the forward problem (Eqn 4), so

$$\mathbf{m}_p = G^{-p}G\mathbf{m} = V_p\Lambda_p^{-1}U_p^TU_p\Lambda_pV_p^T\mathbf{m} = V_pV_p^T\mathbf{m}. \tag{23}$$

Thus the matrix $G^{-p}G = V_pV_p^T$ is known as the *model resolution matrix*.

The derivation used the fact that $U_p^TU_p = I$, because the columns of U_p and hence the rows of U_p^T are orthonormal eigenvectors. Similarly, $V_p^TV_p = I$. By contrast, if there are some zero eigenvalues, then $p \neq n$, $U_pU_p^T \neq I$ and $p \neq r$, $V_pV_p^T \neq I$, because the rows of U_p and V_p are no longer orthonormal eigenvectors (because the columns corresponding to the zero eigenvalues have been removed to form the V_0 and U_0 matrices).

To illustrate these ideas, consider the example in Eqn 10. The G matrix yields

$$G^TG = \begin{pmatrix} 3 & 0 & 1 & 2 \\ 0 & 3 & 2 & 1 \\ 1 & 2 & 3 & 0 \\ 2 & 1 & 0 & 3 \end{pmatrix}, \tag{24}$$

which has eigenvalues 0, 2, 4, 6, and hence is singular. The eigenvector matrices are

$$V_p = \begin{pmatrix} -0.5 & -0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ -0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 \end{pmatrix} \quad \text{and} \quad V_0 = \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}. \tag{25}$$

The model resulting from the inversion $\mathbf{m}_p$ is then related to the "true" (although unknown) model **m** by the model resolution matrix

$$\mathbf{m}_p = V_pV_p^T\mathbf{m} = \begin{pmatrix} 0.75 & -0.25 & 0.25 & 0.25 \\ -0.25 & 0.75 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.75 & -0.25 \\ 0.25 & 0.25 & -0.25 & 0.75 \end{pmatrix}\mathbf{m}. \tag{26}$$

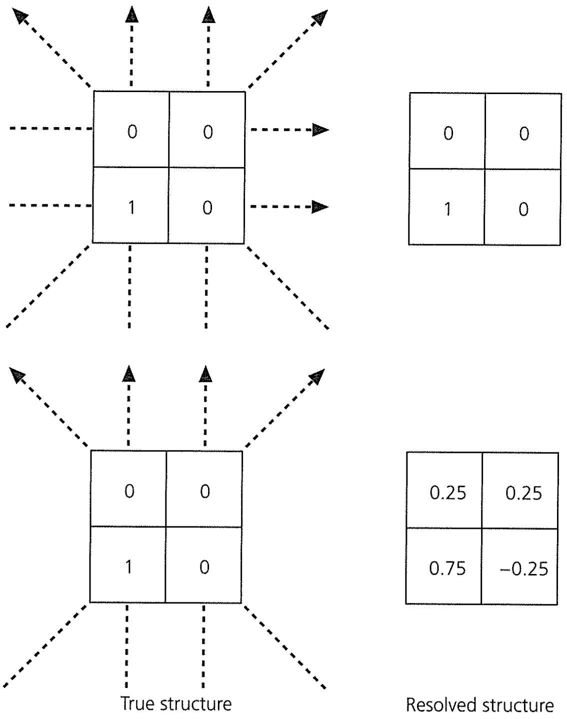


Fig. 7.3-3 Illustration of the "blurring" resulting from the tomographic experiment of Fig. 7.3-2, with incomplete ray coverage. When coverage is adequate, the true slowness perturbation (*top left*) is recovered (*top right*). When coverage is inadequate, the true slowness perturbation (*lower left*) is blurred (*lower right*), although the resulting slowness perturbations yield the correct travel time perturbation for each ray path.

The i^{th} column of the model resolution matrix shows how a unit perturbation in the i^{th} element of the true model maps into various elements of $\mathbf{m}_p$. The true model is thus "blurred" by the inversion. For example (Fig. 7.3-3), inversion of travel time data resulting from a 1% slowness perturbation in block 3 yields a model with 0.25% perturbations in blocks 1 and 2, a 0.75% perturbation in block 3, and a −0.25% perturbation in block 4. These slowness perturbations yield the correct travel time perturbations for the four paths, but because there are no horizontal paths, the solution is not exactly correct. However, most of the perturbation is correctly placed. Note that the resolved structure has a smaller net slowness perturbation than the true structure.

The relation between the resolution matrix and the model covariance matrix (Eqn 7.2.32) is interesting. The blurring illustrated by the resolution matrix results from the ray geometry and would occur even if the data contained no errors. In other words, the resolution matrix illustrates how well the inverse problem could be solved for perfect data. Because the data usually contain errors, the uncertainty in the model, given by the model covariance matrix, reflects errors induced in the model by both the ray geometry and the data errors.

Because the resolution matrix shows how a perturbation in any block is resolved by the inversion, it can be used to find how well the inversion can recover an arbitrary slowness anomaly. Thus the ray geometry, which gives the G and hence V matrices their form, controls the resolution. Note that in the first example, in which all six ray paths are used, Eqn 9 shows that the model from the inversion was the true model. In this case the resolution matrix is the identity matrix.

To see how the lack of resolution in the four-ray case arises, consider what would occur if G^TG had no zero eigenvalues and could be inverted. Then, by Eqns 21 and 22, the model derived from the data would be

$$\mathbf{m}_p = V\Lambda_p^{-1}U_p^T\mathbf{d}, \tag{27}$$

because $V_p = V$. The model is thus a linear combination of the columns of V, or the eigenvectors of G^TG. Because there are r (in this case four) linearly independent eigenvectors, and the model vector has r elements, the eigenvectors span the r-dimensional model space. Thus any vector in the model space is a possible model.

If instead, as in this case, some of the eigenvalues are zero, the eigenvectors associated with them are excluded from the V_p matrix. The model

$$\mathbf{m}_p = V_p\Lambda_p^{-1}U_p^T\mathbf{d} \tag{28}$$

is then a linear combination of only the columns of V_p, the eigenvectors associated with the nonzero eigenvalues. In this case, there are $r-p$ (here three) rather than r linearly independent eigenvectors. Hence not all possible vectors in the model space can be constructed. The model resulting from the inversion contains no linear combinations of the eigenvectors associated with the zero eigenvalues.

To illustrate this idea, consider the four-ray case where the eigenvector associated with the zero eigenvalue is (from Eqn 25)

$$\mathbf{v} = (0.5, 0.5, -0.5, -0.5)^T. \tag{29}$$

This vector corresponds to equal slowness perturbations in blocks 1 and 2 and equal perturbations of opposite sign in blocks 3 and 4. Physically, this means changing the slowness everywhere in the upper layer by some amount, and making the opposite change in the lower layer. Because all four teleseismic rays have equal path lengths in the upper and lower layers, their travel times are unaffected, so travel time data cannot resolve any such change.

Another way to see this is to consider Eqn 7 and note that if $\mathbf{v}$ is an eigenvector whose eigenvalue is zero,

$$(G^TG)\mathbf{v} = 0, \tag{30}$$

so that even if the model contains a linear combination of such eigenvectors, they have no effect on the problem. The zero eigenvectors thus limit the resolution of the model. Because any linear combination of these eigenvectors has no effect, the model resulting from the inversion is not unique. It is possible to prove that the generalized inverse G^{-p} finds a "best" model with no contribution from these eigenvectors. Mathematically, the resulting model is restricted to the V_p space and has no components in the V_0 space. As a result, this model is the minimum possible solution consistent with the data. In this application, the minimum model gives the least lateral perturbation in slowness consistent with the travel time data. Philosophically, this is an attractive approach.

The six-ray case, by contrast, had no zero eigenvalues. Because one ray traveled only in the upper layer and another traveled only in the lower layer, a change in the slowness in either layer would affect the travel times. This ray geometry avoids the ambiguity of the four-ray case, so the model is fully resolved. There is no V_0 space, so $V = V_p$, G^TG can be inverted, and the solution is found using the generalized inverse G^{-g} (Eqn 8). To see how this is related to the generalized inverse G^{-p}, we use the Lanczos decomposition (Eqn 20) to expand G:

$$G^TG = (V\Lambda_p U_p^T)(U_p\Lambda_p V^T) = V\Lambda_p^2 V^T, \qquad (31)$$

$$(G^TG)^{-1} = V\Lambda_p^{-2} V^T, \qquad (32)$$

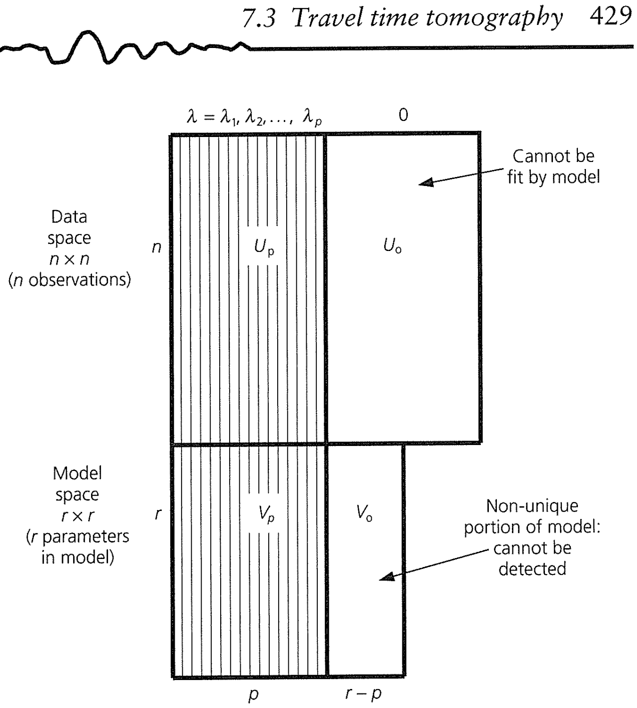
where the matrix products $\Lambda_p^2 = \Lambda_p\Lambda_p$ and $\Lambda_p^{-2} = \Lambda_p^{-1}\Lambda_p^{-1}$. Thus, if G^TG can be inverted, the generalized inverse

$$G^{-g} = (G^TG)^{-1}G^T = (V\Lambda_p^{-2}V^T)(V\Lambda_p U_p^T) = V\Lambda_p^{-1}U_p^T = G^{-p}. \qquad (33)$$

Hence G^{-p} is the general form of the generalized inverse, and G^{-g} is the special form that applies if G^TG can be inverted. The later form, G^{-g}, is easier to compute because it does not require the eigenvector decomposition. Fortunately, it can often be used in applications such as earthquake location.

The eigenvector decomposition also divides the data space into two portions, U_p and U_0, reflecting the nonzero and zero eigenvalues. Data vectors in the U_0 space, linear combinations of the eigenvectors whose eigenvalues are zero, cannot be generated by the operator G for any model. For example, in the six-ray case there cannot be six linearly independent observations because the model has only four parameters. Thus two of the six eigenvectors of the 6×6 matrix GG^T must have zero eigenvalues. These eigenvectors represent travel time observations that should be impossible, given the geometry of the experiment. If the data contained some linear combinations of these eigenvectors, perhaps due to noise in the data, the inversion process could never generate a model capable of matching them.

Figure 7.3-4 summarizes these ideas: the operator G and its generalized inverse G^{-p} relate the model and data spaces. Portions of these spaces are not "illuminated." Any part of the model in the V_0 portion of the model space has no effect on the data, and thus cannot be detected. Thus, if V_0 space exists, the model found by solving the inverse problem is not unique. This situation can only be improved by additional types of data, such as a new set of ray paths in the tomographic example (Fig. 7.3-3).[2] Similarly, any part of the data in the U_0 portion of the data space cannot be described by any possible model.



λ = λ₁, λ₂,..., λₚ
0
Cannot be fit by model
Data space n × n (n observations)
n
Uₚ
U₀
Model space r × r (r parameters in model)
r
Vₚ
V₀
Non-unique portion of model: cannot be detected
p
r − p

Fig. 7.3-4 Schematic illustration of the relation between the model and data spaces for the inverse problem $\mathbf{d} = G\mathbf{m}$. The observed data $\mathbf{d}$ form a vector in the n-dimensional data space, the model $\mathbf{m}$ sought is a vector in the r-dimensional model space, and the known partial derivative matrix G has dimensions $n \times r$. Matrix U, whose columns are the eigenvectors of the matrix GG^T, can be decomposed into U_p, the matrix of the p eigenvectors with nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$, and U_0, the matrix of the eigenvectors with zero eigenvalues. Similarly, the matrix V, whose columns are the eigenvectors of the matrix G^TG, can be decomposed into V_p, the matrix of the eigenvectors with nonzero eigenvalues, and V_0, the matrix of the eigenvectors with zero eigenvalues. (After Lanczos, 1961.)

Thus, if a U_0 space exists, the model found by solving the inverse problem is not an exact solution.

7.3.4 *Variants of the solution*

A number of variants of the least squares solution that we have developed using earthquake location and tomography are also used in these and other inverse problems.

One variant arises from the fact that although the eigenvector decomposition gives insights, it may not be the best approach in some real applications. First, it involves significant computations when the matrices are large. Second, it associates difficulties with the eigenvalues that are zero, whereas in real problems complications and noisy data are more likely to yield small, but nonzero, eigenvalues. These small eigenvalues cause the sort of difficulties that occur formally for zero eigenvalues. To see this, note that in Eqn 27 the model is derived by multiplying the data by the matrix Λ^{-1}, which contains the reciprocals of the eigenvalues. Thus the small eigenvalues, representing the worst-constrained features of the data and model spaces, can have large effects on the solution. For example, we noted in Section 4.4.7 that using the generalized inverse to estimate the moment tensor gives good estimates of components on which seismograms depend strongly, but

[2] As Sherlock Holmes says in *The Copper Beeches*, "I have devised seven separate explanations, each of which would cover the facts so far as we know them. But which of these is correct can only be determined by fresh information."

poorer ones for components on which the seismogram depends weakly.

This issue can be addressed in several ways. One is to exclude small eigenvalues from the inversion. Another, which avoids the eigenvector decomposition, is to modify the function used to measure the misfit between the data predicted by the model and those observed (Eqn 7.2.11) to

$$\chi^2 = \sum_i \frac{1}{\sigma_i^2}\left(\Delta d_i - \sum_j G_{ij}\Delta m_j\right)^2 + \varepsilon^2 \sum_j (\Delta m_j)^2. \qquad (34)$$

This function is the sum of the net misfit and the change in length of the model vector, weighted by ε^2. Hence minimizing it is a compromise between the best fit to the data and the least change from the starting model. The resulting solution, written with the Δs suppressed,

$$\mathbf{m} = (G^T G + \varepsilon^2 I)^{-1} G^T \mathbf{d}, \qquad (35)$$

is called the *damped least squares* solution. If ε is zero, we have the best-fit solution (Eqn 7.2.17), whereas larger values of ε reduce or damp the change in the starting model by accepting a poorer fit to the data. The damping parameter ε is chosen empirically to yield a solution that seems plausible, and thus of necessity reflects our ideas about the solution sought, because damping the poorly constrained and undesired changes in the model also damps the better constrained and desired changes.

Another common situation is that we want some data to have greater effect on the solution, usually because we consider them to be better known. We thus incorporate a data-weighting matrix W_d into the solution. The simplest is to weight by $W_d = (\sigma_{\mathrm{d}}^2)^{-1}$, the inverse of the variance–covariance matrix of the data, so the data with the smallest uncertainties have the greatest effect. Problem 5 shows that this *weighted least squares* solution is

$$\mathbf{m} = (G^T W_d G)^{-1} G^T W_d \mathbf{d}. \qquad (36)$$

We may also want to have the model change smoothly, such that each element varies only slightly with respect to its neighbors. For instance, if the model were a continuous function of one variable, we measure the smoothness, or *flatness*, f, of the changes by forming

$$\mathbf{f} = \begin{pmatrix} -1 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ \cdot \\ \cdot \\ m_r \end{pmatrix} = F\mathbf{m}, \qquad (37)$$

where F is the *flatness matrix*, which is a numerical approximation to the derivative at the edges of each element. The overall flatness of the solution is then

$$\mathbf{f}^T\mathbf{f} = \mathbf{m}^T F^T F \mathbf{m} = \mathbf{m}^T W_m \mathbf{m}, \qquad (38)$$

so the matrix $W_m = F^T F$ is a weighting matrix for the model. For more complicated model geometries, F is changed appropriately.

We can combine the model and data weighting in a *weighted damped least squares* inversion, which yields the solution

$$\mathbf{m} = (G^T W_d G + \varepsilon^2 W_m)^{-1} G^T W_d \mathbf{d}. \qquad (39)$$

As noted earlier, the damping parameter ε is chosen empirically. If we do not weight the data and model, the weighting matrices W_m and W_d are identity matrices, and Eqn 39 is just the simple damped least squares solution (Eqn 35).

An example of such an inversion was shown for *P*-wave velocities at the base of the mantle in Fig. 3.5-17. A grid of 660 nodes that were roughly equally spaced were used to represent the base of the mantle. The damping factor, $\varepsilon = 1.2$, was a compromise between the best fit, which minimizes the prediction error, and minimizing the undetermined part of the solution. Because each node is surrounded by 5 or 6 nodes that are roughly equidistant, the rows of the model flatness matrix F were chosen with the diagonal term equal to -1 and the terms of the nearest N neighbors equal to $1/N$ (with $N = 5$ or 6). The data were weighted empirically so that the diagonal elements of the W_d matrix ranked the quality of the observations from 9 (excellent) through 4 (good) to 1 (poor). These choices again bear out that we have various ways of solving inverse problems, so the solution we develop depends on choices about the data we use and the model we seek, based on our ideas about what seems reasonable. Hence our solutions are in part objective and in part subjective, and different approaches yield different solutions.

7.3.5 Examples

Studies using travel time tomography yield interesting results for various areas. For example, Fig. 7.3-5 (*top*) shows the model geometry used in a study of the upper mantle in the region including Central Europe, the Mediterranean, and the Middle East. The model contains nine layers, each divided into 1040 1° by 1° blocks. The layer thickness increases with depth from 33 km at the top to 130 km at a depth of 670 km. The data consist of approximately half a million travel times from about 25,000 earthquakes, recorded at stations both within the model region and at distances to 90°.

The data used are travel time anomalies relative to the Jeffreys–Bullen values, which can result from earthquake mislocations as well as variations in seismic velocities. The location and origin time of the earthquakes were thus also inverted for, so the number of unknowns reflects both the number of blocks (9360) and four times the number of earthquakes used. To reduce these numbers, procedures were used to combine data from nearby earthquakes and from stations close to each other. The problem to be solved thus involves approximately 300,000 equations for 20,000 unknowns.

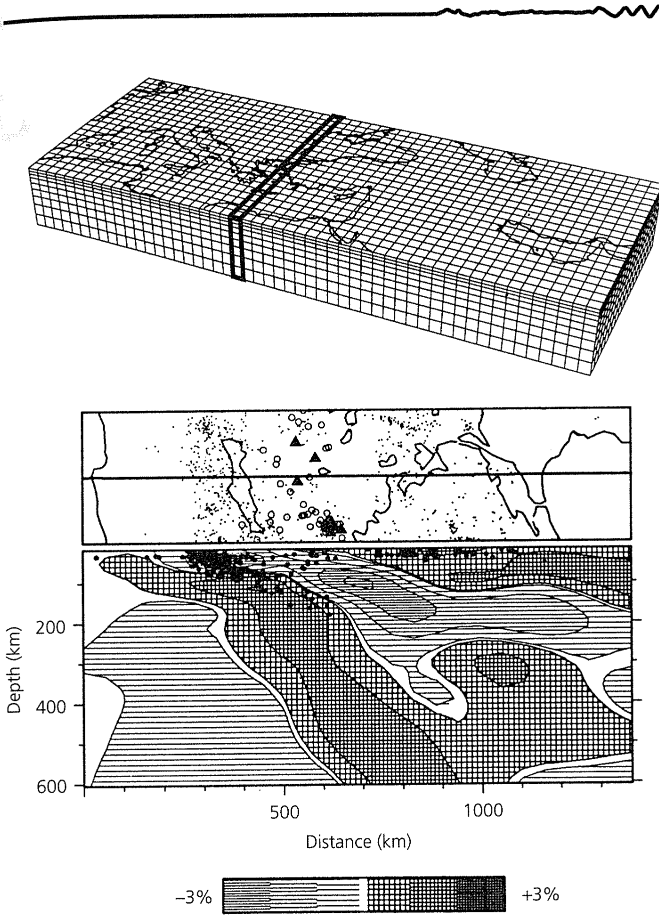


Fig. 7.3-5 *Top*: Block model for a travel time tomographic study of the upper mantle in the region including Central Europe, the Mediterranean, and the Middle East. The heavy line indicates the location of the cross-section shown below. (Spakman and Nolet, 1988, with kind permission from Kluwer Academic Publishers.) *Bottom*: Cross-section through the block model across the Hellenic trench region, showing P-wave velocity perturbations with respect to the JB model. (Spakman *et al.*, 1988. *Geophys. Res. Lett.*, *15*, 60–3, copyright by the American Geophysical Union.)

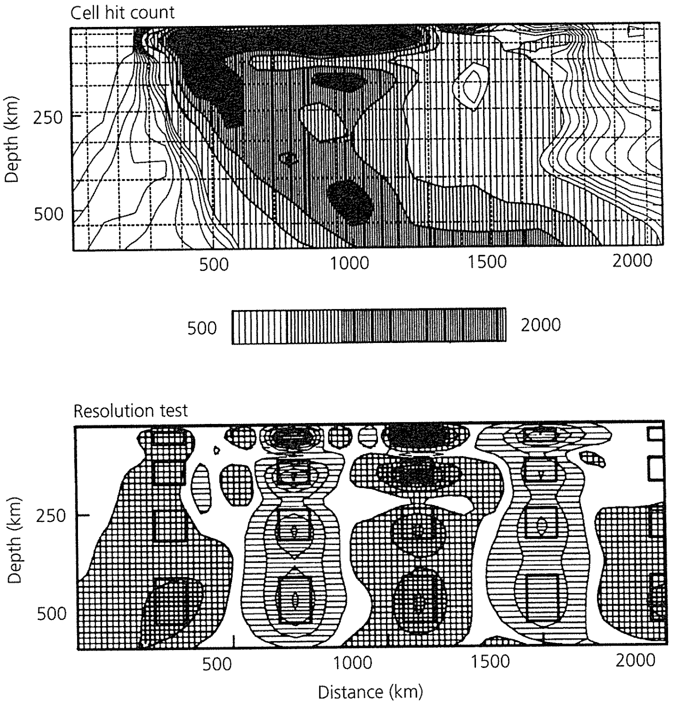


Fig. 7.3-6 Analysis of the tomographic image in Fig. 7.3-5 (*bottom*). *Top*: Hit count plot, showing the number of times each block is sampled. Black regions indicate the best-sampled blocks (hit counts in excess of 2000). *Bottom*: Resolution test using synthetic velocity anomalies. Travel times are generated for a model with 5% velocity perturbations, of alternating sign, in each of the blocks marked by heavy lines. How well the perturbations are recovered illustrates how much the image is blurred. (Spakman and Nolet, 1988, with kind permission from Kluwer Academic Publishers.)

Solving matrix equations of this size poses major difficulties. The matrices are so large (in this case 6×10^9 elements) that they are difficult to store in a computer and operate on. As a result, numerical methods are used, some of which allow only a single row of the matrix to be manipulated at any time. The properties of these algorithms and methods of improving the resulting image form an active research area.

The resulting three-dimensional velocity model can be shown as either cross-sections or map views at various depths. Figure 7.3-5 (*bottom*) shows a cross-section across the Hellenic trench region, where the African plate subducts beneath Crete and the Aegean basin (Fig. 5.6-8). The tomographic image shows velocity anomalies in percent of the velocity predicted for that depth by the JB model. A planar high-velocity (positive) anomaly, presumably the cold downgoing slab, dips NW from the trench and extends to depths well below the deepest earthquakes (dots). Above the slab, a low-velocity (negative) region occurs, presumably due to flow behind the arc. Such observations are valuable for modeling the subduction history and dynamics.

Because tomographic images are solutions to an inverse problem, they are neither unique nor exact. Hence it is important to assess which features in the image are likely to be geologically real, and which are more likely to be artifacts of the inversion. As we have seen, an important factor is how well parts of the model are sampled by the ray paths. Figure 7.3-6 (*top*) shows a *hit count* plot for the section of Fig. 7.3-5 (*bottom*), showing the number of ray paths that sample each block. The better-illuminated regions should be better resolved than poorly sampled regions. Additional insight comes from analyzing how a perturbation in one model block is blurred by the inversion into nearby blocks. This information, given by the resolution matrix (Eqn 23), can also be found by placing a perturbation in one block, computing the forward problem, and inverting the result. Because this would be time consuming for such a large model, perturbations were placed in various blocks, and the combined resolution was estimated by computing synthetic travel time data and inverting it. Figure 7.3-6

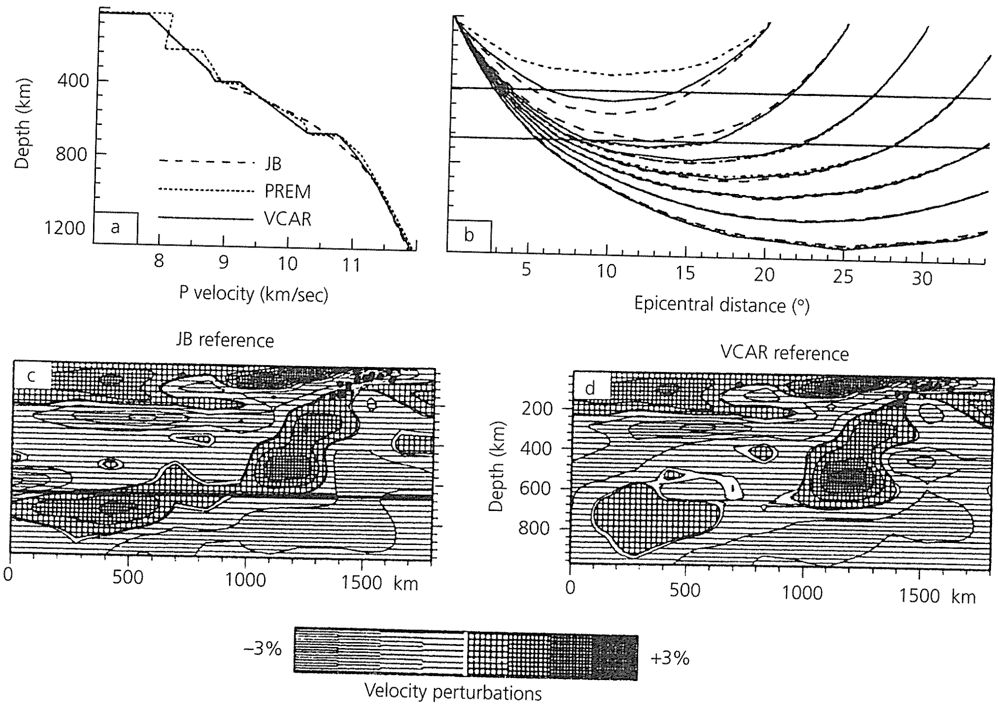


Fig. 7.3-7 Illustration of the effects of the reference model in travel time tomography. Velocity structure (a) and ray paths (b) for global reference models JB and PREM and a local reference model VCAR. The differences (c, d) between the tomographic images reflect differences between the reference models near 600 km depth. (van der Hilst and Spakman, 1989. *Geophs. Res. Lett.*, *16*, 1093–6, copyright by the American Geophysical Union.)

(*bottom*) illustrates this method for a 5% velocity contrast whose sign alternates between columns. If resolution were perfect, the image would be reconstructed exactly: each anomaly would be confined to the original block (heavy line). Due to the ray geometry, the anomalies "blur", but are still concentrated in the correct locations. Comparison with the hit counts shows that better-sampled regions, such as the second column from the left, are better resolved than poorly sampled regions like the lower left column. The reconstructed image is further degraded when the effects of noise in the data are simulated. Even in this case, the inversion results locate the perturbed blocks reasonably well and retrieve the sign of the perturbation. These tests suggest strongly that the high-velocity slab in the image is real.

Typically, the major features of tomographic inversions seem likely to be real, but assessing how much of the detailed structure is real is more difficult. For example, Fig. 5.4-7 showed the results of a numerical experiment to see how well a tomographic study would reconstruct the image of a theoretical subducting slab. It turned out that the general shape of the slab was resolved, but was blurred by artifacts implying velocity anomalies that are not present in the original model. In this case these artifacts, generally of low amplitude, caused the slab to appear to broaden, shallow in dip, or flatten out. The extent to which these artifacts appear depends on ray geometry, so the image could be improved by using upgoing as well as downgoing rays.

Another important factor in tomographic images is the reference model with respect to which the velocity anomalies are shown. In examining images, it is natural to focus on the lateral variations. However, because these variations are with respect to a starting model, which is usually laterally homogeneous, the resulting images depend on the starting model. Figure 7.3-7 shows an example for the Lesser Antilles. The ray paths predicted by the global JB and PREM reference models differ somewhat from those predicted by a model VCAR developed for this region. As a result, tomographic images relative to the JB and VCAR models differ. Although both show the high-velocity North American plate subducting westward beneath the Caribbean, the JB image implies that the slab flattens at the 660 km discontinuity, whereas this suggestion is much less in the VCAR image. The flattening in the JB image results from the fact that the inversion yields "streaks" of velocities relative to JB that are lower than those observed above 660 km, and higher than those observed below 660 km. This effect arises because, compared to VCAR, the JB model predicts higher velocity above 660 km, and lower velocity below. Thus a bias in the reference model can produce spurious lateral heterogeneity. Similar reference model artifacts, in which a common state seems abnormal due to the standard used, appear in various inverse problems and other situations.[3] However, the choice of reference model is subjective, so making a choice requires recognizing its consequences. For example, a global velocity reference model that excluded subducting slabs would be slower than the actual global average, whereas one including slabs would predict slow anomalies elsewhere.

[3] 90% of motorists are said to consider themselves above-average drivers, and "all children are above average" in the mythical town of Lake Wobegon in the radio show *Prairie Home Companion*.

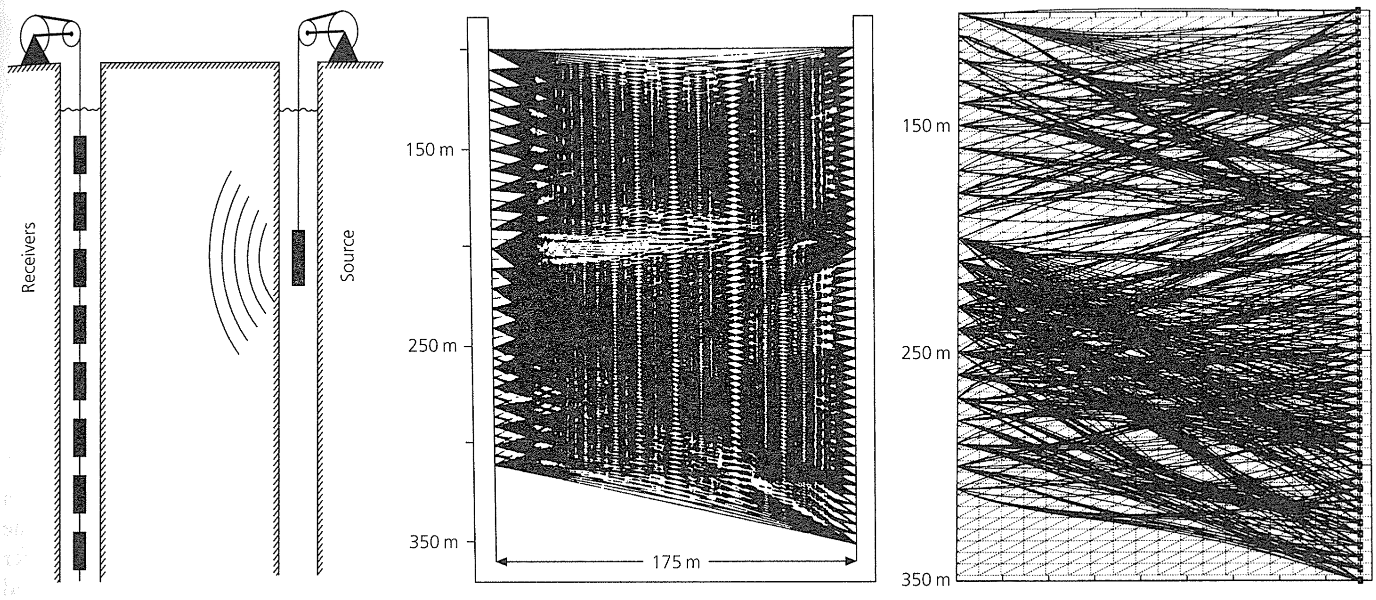


Fig. 7.3-8 An example of cross-borehole tomography in Manitoba, Canada. *Left*: Travel times are recorded from a source at different depths in one borehole to receivers in the other. The experiment is then reversed, yielding dense ray path coverage. *Center*: Straight ray paths computed for the laterally homogeneous starting model. *Right*: Ray paths for the laterally varying model found from the inversion. (Wong *et al.*, 1987.)

In addition to ray geometry and reference model artifacts, it is worth noting that tomographic images can also be affected by something as simple as the contouring scheme used. Sometimes when features are not robust aspects of the image, their tectonic interpretation depends in part on preconceptions, much like the ink-blot tests used by psychologists. Thus, despite the power and value of tomographic images, it is important to bear their limitations in mind.

Tomography is also used in other seismological applications. One important use, providing detailed near-surface images, is illustrated by Fig. 7.3-8 showing tomography between two boreholes. The source and receivers were moved to generate dense coverage with many crossing ray paths. The travel time observations were then inverted for velocity structure. In this experiment, the ray paths were recomputed for the perturbed model and used to compute travel times for later iterations. The differences between the initial and perturbed ray paths show the advantages of recomputing the ray paths for each successive model, a process called *nonlinear tomography*. This updating ensures that the ray paths, and hence predicted travel time anomalies, are consistent with the velocity structure being found. However, for practical reasons it is common to conduct linearized tomography using ray paths from the starting model even as the model is perturbed, and to assume that the resulting errors are small.

It is interesting to compare travel time tomography to the surface wave tomography discussed in Section 2.8.3, where the average surface wave velocity along multiple paths through oceanic lithosphere of various ages is used to infer the velocity structure for each age range. The approach is to find the phase or group wave velocity as a function of frequency for each age range, and then infer the variation in the medium velocity with depth from the dispersion curve giving the variation in apparent velocity as a function of frequency. Hence this is tomography in the lateral direction, and dispersion analysis vertically. We will see in the next section that dispersion analysis is an example of methods that infer earth structure using functions that sample the structure at depth in different ways.

Tomographic methods can be used for waveforms as well as travel times. As noted earlier — for example, in Fig. 3.7-7 — waveforms sample earth structure over broader regions than travel times, which, in the limit, correspond to sampling along narrow geometric rays. Figure 7.3-9 shows some results from global tomography in which velocity perturbations were inferred by fitting both waveforms from 27,000 long-period seismograms and 14,000 travel times. The seismograms include body wave records (from the *P* or *PKP* arrival to the start of the surface waves) and "mantle wave" records, which are low-pass filtered seismograms about 4.5 hours in length. The travel time data include both absolute shear wave arrival times and differential (*SS–S* and *ScS–S*) times. Rather than inverting for the velocity perturbations in blocks, the velocity perturbation was described by a series of orthogonal functions, and the inversion was for the coefficients of the functions. The lateral structure was described by spherical harmonics (Section 2.9.3), and the vertical structure was modeled using Chebyshev polynomials.

In addition, we saw in Section 3.7 that amplitude tomography can infer attenuation variations along the ray paths. Amplitude tomography is similar to medical tomography,[4] in which the image indicates the degree to which X-rays are

[4] The medical term "CAT scan" is for computed axial tomography.

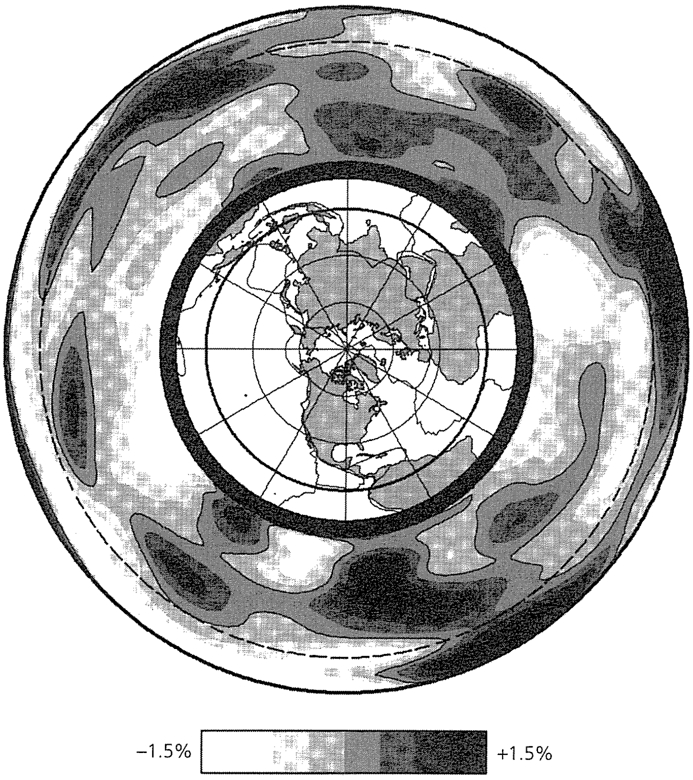


Fig. 7.3-9 Tomographic image of shear wave velocities along a great circle slice through the Equator, obtained by inversion of both waveforms and travel times. (Su *et al.*, 1994. *J. Geophys. Res.*, *99*, 6945–80, copyright by the American Geophysical Union.)

absorbed in different portions of the subject. Medical tomography has the advantages that the subject can be uniformly illuminated from all sides, and that the internal structure is both well understood and subject to later, direct observation.

7.4 Stratified earth structure

Quantities that can be determined using seismological data are often the integrals of a physical property of the earth. For example, the travel time is the integral of slowness along a ray path. As discussed in the last section, although a single travel time gives only the average slowness along the ray path, travel times for different ray paths can be combined to find the spatial distribution of slowness.

A common such problem is finding earth structure for laterally homogeneous or stratified earth models, in which physical properties are assumed to vary only with depth. Frequently, an observable quantity d_i can be expressed as the integral over the radius of a physical property $m(r)$,

$$d_i = \int_0^a G_i(r)m(r)dr, \tag{1}$$

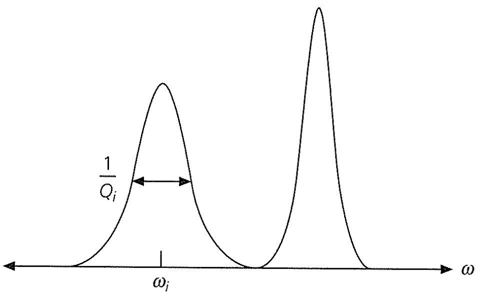


Fig. 7.4-1 Schematic amplitude spectrum of a seismogram, showing the observations used to invert normal mode data for eath structure. Each mode peak is described by a width proportional to Q_i^{-1}, which describes its attenuation, and an eigenfrequency ω_i.

where $G_i(r)$ is a known function of depth called a kernel. Given a set of d_i with different kernels, each of which samples the distribution of $m(r)$ differently, the inverse problem is to infer $m(r)$. Although the relation between the observed quantity and earth structure is sometimes less intuitive than for travel time and slowness, the problems can be formulated in a similar way.

We encountered this idea in discussing Love wave dispersion in Section 2.7.4. The apparent phase velocity along the free surface varies as a function of period, because waves of different period sample the velocity at depth differently. Hence this variation can be used to study the velocity at depth.

7.4.1 *Earth structure from normal modes*

The concepts of inverting observations for the structure of a stratified medium can be illustrated using normal modes (Sections 2.9 and 3.7). The displacement field of the i^{th} mode excited by an earthquake can be written

$$\mathbf{u}_i(t) = \mathrm{C}_i(t)\exp(-\omega_i t/2Q_i). \tag{2}$$

The mode's eigenfrequency ω_i and quality factor Q_i, which describes the attenuation, and thus the width of the peak, can be found from the Fourier transform of the seismogram (Fig. 7.4-1). Because ω_i and Q_i depend on the variation with depth of the seismic velocities, density, and attenuation, these observations can be used to study earth structure.

To do this, we begin with an earth model described by $\alpha(r)$, $\beta(r)$, and $\rho(r)$ and find the eigenfrequencies of the different modes, ω_i. This calculation also gives the partial derivative functions

$$\frac{\partial\omega_i}{\partial\alpha}(r),\quad \frac{\partial\omega_i}{\partial\beta}(r),\quad \frac{\partial\omega_i}{\partial\rho}(r), \tag{3}$$

showing how a mode's eigenfrequency changes if the velocity or density at a given depth is perturbed. The total change in the eigenfrequency is the integral over the radius of the perturbations in the earth model:

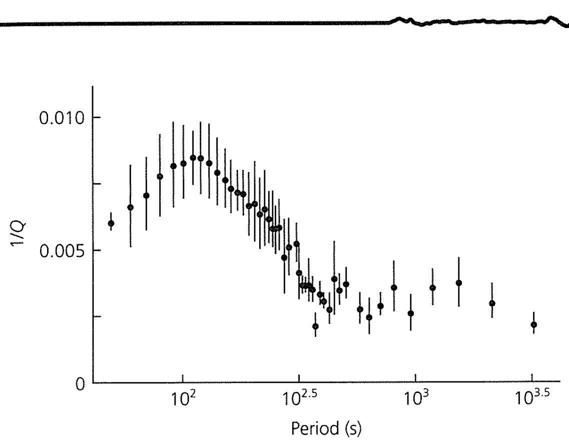


Fig. 7.4-2 Observed attenuation for fundamental spheroidal modes ${}_0S_2$–${}_0S_{191}$. The variation in Q^{-1} with period reflects the depth variation of $q^{-1}(r)$. (Stein *et al.*, 1981. *Anelasticity in the Earth*, 39–53, copyright by the American Geophysical Union.)

$$\Delta\omega_i = \int_0^a \left[\frac{\partial\omega_i}{\partial\alpha}(r)\Delta\alpha(r) + \frac{\partial\omega_i}{\partial\beta}(r)\Delta\beta(r) + \frac{\partial\omega_i}{\partial\rho}(r)\Delta\rho(r) \right] dr. \qquad (4)$$

Thus the difference between a measured eigenfrequency and that predicted by an earth model can be inverted to find the perturbation in the model required to fit the data. Although a single mode observation gives only the average over depth of the required perturbation, a set of modes gives more information, because the partial derivatives vary between modes.

We illustrate the method using the corresponding inverse problem for attenuation, which has a simple linear form. If attenuation within the earth is described by the function $q(r)$, the quality factor for the i^{th} mode is

$$Q_i^{-1} = \int_0^a G_i(r) q^{-1}(r) dr, \qquad (5)$$

where the kernels $G_i(r)$ are derived from the partial derivatives (Eqn 4), using the formulation of the quality factor as an imaginary part of the frequency that is related to an imaginary part of the velocity (Section 3.7.6). Although the symbol Q is commonly used for both the modes' quality factor and the attenuation as a function of depth, using $q(r)$ for the latter emphasizes the distinction. The problem is written using the reciprocals $q^{-1}(r)$ and Q_i^{-1}, so higher attenuation (larger loss of seismic energy) corresponds to larger values.

Figure 7.4-2 shows measured values of the attenuation of fundamental spheroidal modes, which for periods less than a few hundred seconds correspond to fundamental mode Rayleigh waves. The attenuation is low for the longest-period modes, rises to its highest values at periods slightly above 100 seconds, and then decreases again for the shortest periods (about 50 seconds) shown. This variation occurs because the kernels differ between modes (Fig. 7.4-3). Because Q^{-1} for a mode is the integral of the attenuation weighted by the kernel, the shape of the kernel with depth illustrates a mode's sensitivity to attenuation at various depths. Long-period modes are most sensitive in the lower mantle, periods near 100 seconds sample the low-velocity zone heavily, and periods near 50 seconds are most sensitive to structure in the "lid" region above the low-velocity

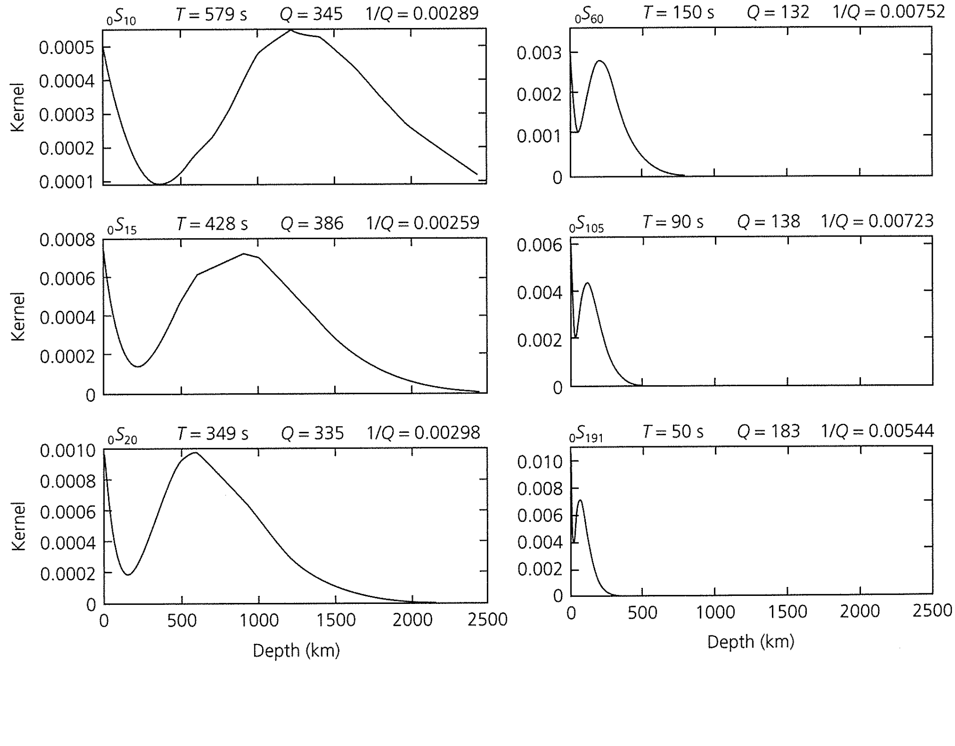


Fig. 7.4-3 Attenuation kernels for various modes, illustrating the different depth sampling. Attenuation values are for the third model in Fig. 7.4-5. (Stein *et al.*, 1981. *Anelasticity in the Earth*, 39–53, copyright by the American Geophysical Union.)

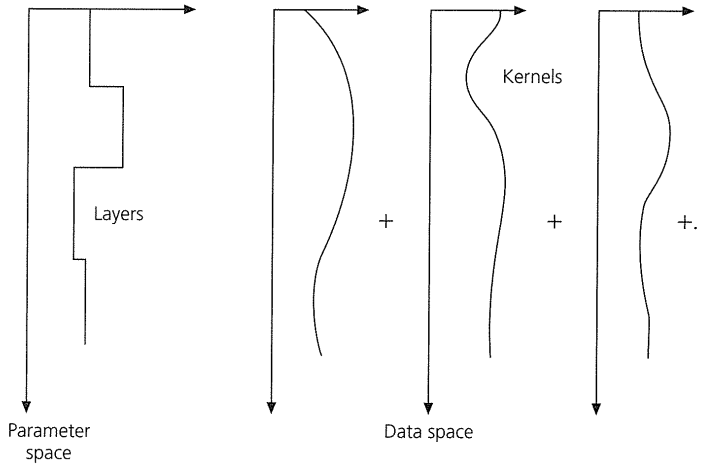


Fig. 7.4-4 Schematic illustration of the model parameterizations for two types of inversion methods. In parameter space inversions, the model is divided into layers; in data space inversions the model is treated as a weighted sum of the kernels.

zone. Q^{-1} is a smooth function of the period, because the kernels of fundamental modes with similar periods are similar.

The inverse problem is to use the observed mode attenuation Q_i^{-1} and the known kernels $G_i(r)$ to infer the function $q^{-1}(r)$ describing the variation of attenuation with depth in the earth that best fits the data. This problem can be approached in several ways, two of which we discuss briefly.

7.4.2 *Parameter and data space inversions*

The most direct approach, *parameter space inversion*, is to regard the unknown model $q^{-1}(r)$ as constant in a set of layers (Fig. 7.4-4, *left*), such that in the j^{th} layer

$$q^{-1}(r) = q_j^{-1}, \quad r_j \le r \le r_{j+1}. \tag{6}$$

The inverse problem is then converted from an integral to a matrix equation

$$Q_i^{-1} = \int_0^a G_i(r) \sum_j q_j^{-1} dr = \sum_j A_{ij} q_j^{-1}, \tag{7}$$

where the matrix elements are

$$A_{ij} = \int_{r_j}^{r_{j+1}} G_i(r) dr. \tag{8}$$

The observations are inverted for the value of the parameter q_j^{-1} in each layer.

By choosing a smaller number of layers than mode observations, we obtain an overdetermined system of equations. As before, the generalized inverse gives the "best" solution in a least squares sense. The concepts developed previously are useful for assessing the solution. Columns of the resolution matrix, called resolving kernels, indicate how well the value in the corresponding layer could be determined independently of those in the other layers if the data had no errors. This uncertainty results from the inverse problem itself, and reflects the best resolution possible, given the available kernels, analogous to the resolution matrix (Eqn 7.3.23) in the tomographic example. It is also useful to consider the model covariance matrix, which indicates the uncertainty in the model due to both the nature of the inverse problem and the errors in the observations. Often a weighted average over a number of layers is the best resolution obtainable, analogous to the blurring in travel time tomography.

Parameter space inversion has a few unattractive features. First, the layers in which attenuation is treated as constant must be chosen in advance. This choice might not be a meaningful one. Second, parametrizing the model as constant in these layers yields a model with "steps" at layer boundaries. These steps may be quite unphysical; in many cases our intuition (admittedly sometimes a poor guide) suggests that physical properties should vary smoothly with depth.

In an alternative formulation, *data space inversion*, the unknown model describing attenuation as a function of depth is expanded not into constant layers, but as a weighted sum of the kernels themselves (Fig. 7.4-4, *right*),

$$q^{-1}(r) = \sum_j v_j G_j(r). \tag{9}$$

The inverse problem is then

$$Q_i^{-1} = \int_0^a G_i(r) \sum_j v_j G_j(r) dr = \sum_j A_{ij} v_j, \tag{10}$$

where the matrix elements are

$$A_{ij} = \int_0^a G_i(r) G_j(r) dr. \tag{11}$$

The model is found by inverting for the expansion coefficients v_j.

Data space inversion is less intuitive than parameter space inversion, but has the attractive features that the resulting model is a smooth function of depth, and need not be parametrized in depth in advance. Moreover, it is in some sense "natural" to use the kernels as basis functions for the model, because the observations sample the model along these kernels. However, these solutions often seem too smooth for our instincts, just as the parameter space solutions often seem too jagged. We often both expect changes in properties near certain

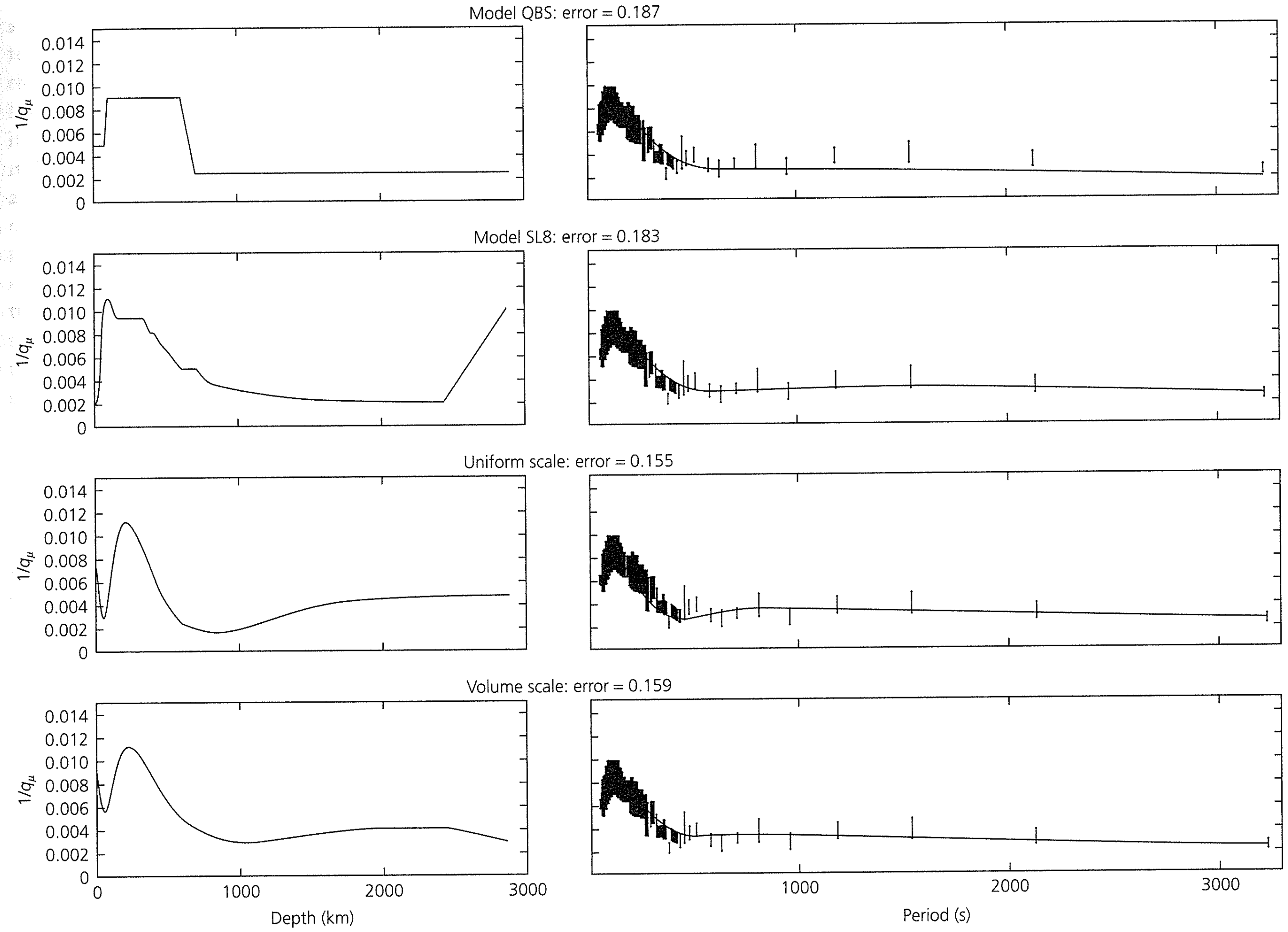


Fig. 7.4-5 Comparison of various attenuation models. Despite the differences, all reproduce the general features of the data in Fig. 7.4-2, as shown in the right hand panels. (Stein *et al.*, 1981. *Anelasticity in the Earth*, 39–53, copyright by the American Geophysical Union.)

depths and are reluctant to force them into the solution. This dilemma is an example of the general issue of deciding how much we want the inversion solution to reflect our preconceptions, some of which may be correct, especially when derived from other data, and some of which may be incorrect. We can choose to focus on what the data require, what the data permit, or a combination of the two.

These issues are illustrated in Fig. 7.4-5, which shows several models for attenuation as a function of depth, all generally consistent with the data in Fig. 7.4-2. Model SL8 was derived by parameter space inversion, whereas the others were derived from data space inversion. The lower two models were derived by inverting the data in Fig. 7.4-2 with different misfit functions, whereas the upper two were derived from different data. Although the models differ, all have low attenuation in the lower mantle, high attenuation in the upper mantle associated with the low-velocity zone, and moderate attenuation in the "lid" above the low-velocity zone. The models illustrate the range of acceptable solutions. For example, the high attenuation zone at the base of the mantle in model SL8 is permissible, and thus survives if included in the starting model, but is not required by the data. This ambiguity results from the fact that the data have little resolution for structure at this depth, as shown by the kernels in Fig. 7.4-3.

7.4.3 *Features of the solutions*

The inverse problem for attenuation (Eqn 5) has a simple form, because each mode's quality factor depends linearly on $q^{-1}(r)$, so the observations can be inverted directly for the attenuation structure. If this is not the case, we linearize about a starting model (Section 7.2.1), so the change in a datum depends linearly on the change in model parameters

$$\Delta d_i = \int_0^a G_i(r)\Delta m(r)dr. \tag{12}$$

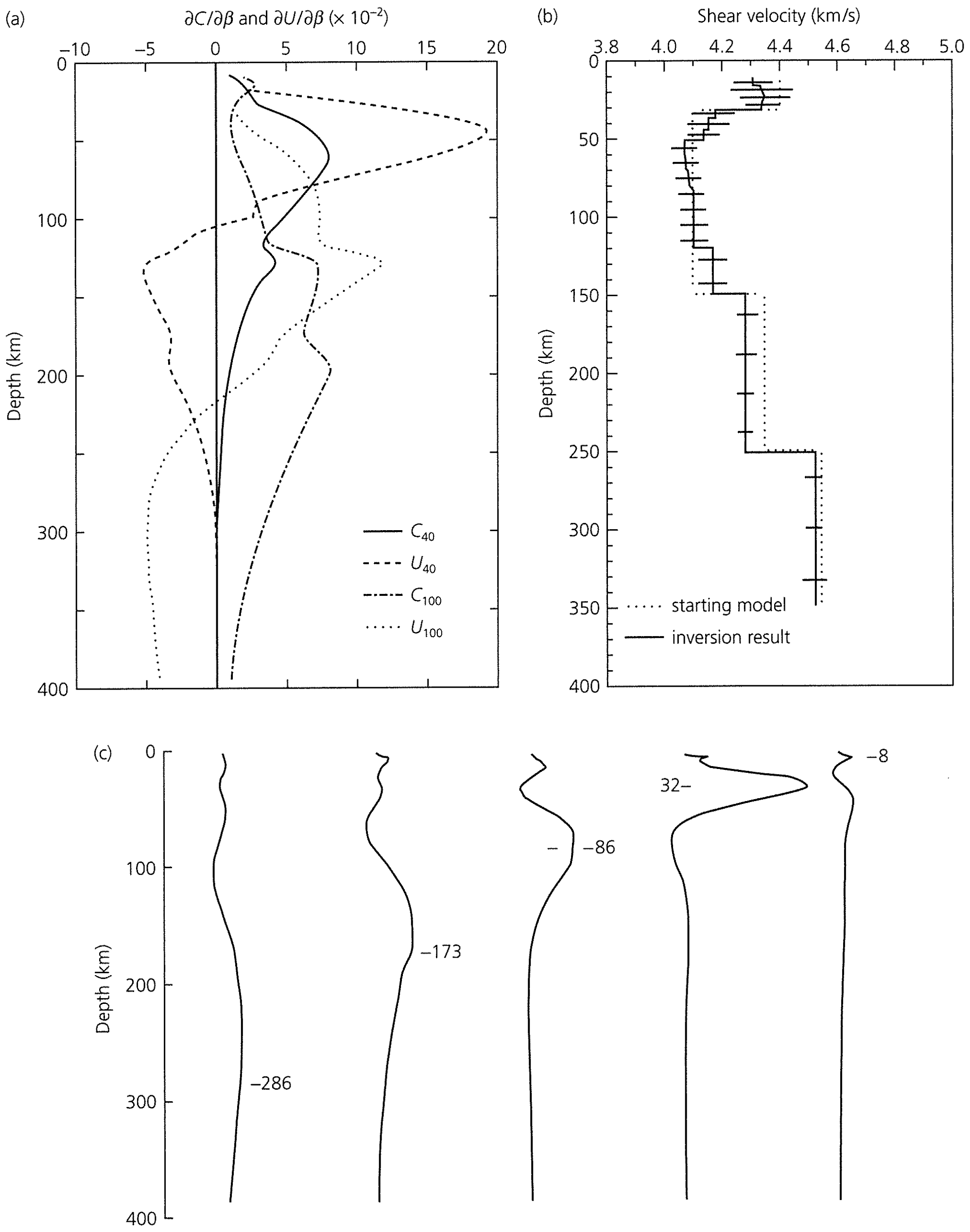


Fig. 7.4-6 Inversion of Rayleigh wave phase and group velocity measurements for shear wave velocity structure beneath the Pacific. (a): Phase and group velocity partial derivatives at 40 and 100 s periods. (b): Starting (dotted line) model and final model derived by parameter space inversion. Horizontal lines indicate the model standard deviation in each layer. (c): Resolving kernels for various depths. The number and horizontal line indicate the depth for each kernel. (Yu and Mitchell, 1979.)

Figure 7.4-6 illustrates a parameter space inversion for vertical shear velocity structure from Rayleigh waves. Using the partial derivatives

$$\frac{\partial C(T)}{\partial \beta}(r), \quad \frac{\partial U(T)}{\partial \beta}(r), \qquad (13)$$

which show how the phase and group velocities at a particular period change in response to a shear velocity perturbation at each depth, the starting model is modified to fit the observed dispersion. The resolving kernels that illustrate the vertical "smearing" are largest at the depth for which they are computed, but have nonzero amplitudes at other depths. The best resolution occurs when the kernel is sharply peaked at the desired depth.

As we noted earlier, the generalized inverse solution yields the minimum change in the model that best produces a desired change in the data. Hence the final model is as close to the starting model as possible. Features of a model derived by linearized inversion can thus depend on the starting model. For example, in a parameter space inversion, a layer whose value in the starting model is assumed to differ significantly from adjacent layers will often retain this feature in the solution. One way to avoid this is to start off with a model whose properties are uniform with depth. In other cases, data not included in the inversion can be used to find a starting model more appropriate than a uniform one. Another approach is to do inversions with different starting models and compare the resulting solutions. If the solutions differ, they are likely local minima of the misfit function (Eqn 7.2.11) that the inversion minimized, whereas if the different starting models yield the same solution, it is more likely to be the global minimum that we seek. Yet another approach is to search numerically for the minimum in the model space by varying the model parameters. Such "brute force" approaches, in which we solve the inverse problem by solving the forward problem many times, are attractive when the number of model parameters is small, because they avoid the issue of linearizing about the starting model and show the trade-offs between various parameters. For example, Fig. 5.3-8 showed the trade-off between plate thickness and basal temperature in inverting oceanic depth and heat flow data for thermal structure.

Parameter space and data space inversions can be carried out using more sophisticated variations. For example, parameter space inversion can be smoothed to reduce the jumps at layer boundaries. Data space inversion can be formulated in terms of a set of orthogonal kernels, rather than the actual kernels, which are often quite similar to each other. This approach expands the model in the simplest possible way with the minimum number of parameters. In addition, the model can be constrained to fit the data only within the error bars, rather than attempt to fit the mean value of each datum.

Due to the structure of inverse problems and the range of possible techniques available, various solutions can generally be derived for a set of seismological observations. As a result, inverse problems remain an important research area. The choices, ambiguities, and trade-offs in the solutions of these problems are sometimes key features of the solution. Attempts to explain these issues can be frustrating to nonseismologists, as illustrated by the joke that in response to the question "How much is 2 + 2," an engineer replies "3.9999," a geologist replies, "Somewhere in the mid-single digits," and a geophysicist replies, "How much do you want it to be?"

7.5 Inverting for plate motions

We end our discussion of inverse problems with the issue of determining the Euler vectors that describe relative plate motions. As we have noted, these Euler vectors are derived in part from earthquake focal mechanisms, and are then used as a reference model to predict the directions and rates of plate motions for applications including estimating earthquake recurrence, slip partitioning, and the fractions of seismic and aseismic slip at plate boundaries.

7.5.1 *Method*

The forward problem (Section 5.2.1) is that at any point **r** along their boundary, the linear velocity of plate j with respect to plate i is

$$\mathbf{v}_{ji} = \omega_{ji} \times \mathbf{r}, \tag{1}$$

where ω_{ji} is the relative angular velocity, or Euler vector. Hence the rate and direction of plate motion are given by the north–south and east–west components of **v**,

$$\text{rate} = |\mathbf{v}| = \sqrt{(v^{\text{NS}})^2 + (v^{\text{EW}})^2},$$
$$\text{azimuth} = 90° - \tan^{-1}[(v^{\text{NS}})/(v^{\text{EW}})]. \tag{2}$$

The corresponding inverse problem is to find a model, or set of Euler vectors, that best predicts the observed motions. Because Euler vectors can be added, assuming that the plates are rigid, m plates are specified by $m - 1$ Euler vectors, and thus their $3(m - 1)$ components. Hence we use a data vector **d** composed of rates and azimuths to estimate the model vector **m** composed of the Euler vector components. Both the model and data vectors consist of physically different quantities: the model vector is made up of Euler pole latitudes, longitudes, and rotation rates

$$\mathbf{m} = (\theta_1, \theta_2, \ldots \theta_{m-1}, \phi_1, \phi_2, \ldots \phi_{m-1}, |\omega_1|, |\omega_2|, \ldots |\omega_{m-1}|), \tag{3}$$

whereas the data vector contains rates and azimuths

$$\mathbf{d} = (r_1, r_2, \ldots r_k, az_1, az_2, \ldots az_{n-k}). \tag{4}$$

As written, the inverse problem is not linear because the data are complicated functions of the model parameters. Thus, as in the previous examples, we linearize about a starting model by forming the partial derivative matrix

$$G_{ij} = \frac{\partial d_i}{\partial m_j}, \tag{5}$$

showing how a change in the j^{th} model parameter affects the prediction of the i^{th} datum. The derivatives are found by differentiating the expressions for v^{NS} and v^{EW} (Eqn 5.2.7). We then have the usual equation

$$\Delta\mathbf{d} = G\Delta\mathbf{m}, \quad \text{or} \quad \Delta d_i = \sum_j G_{ij}\Delta m_j, \tag{6}$$

relating the changes in the data and the model. The system is usually overdetermined, because we generally have data at many sites and solve for only a few plate model parameters. For example, the NUVEL-1 model has 12 plates whose motions were estimated from 1122 data (Fig. 1.1-9). We thus use the weighted least squares solution

$$\Delta\mathbf{m} = (G^T W_d G)^{-1} G^T W_d \Delta\mathbf{d}, \tag{7}$$

where the variance–covariance matrix of the data, $W_d = (\sigma_d^2)^{-1}$, contains our estimates of the uncertainty in rates from magnetic anomalies and the uncertainties in directions associated with estimating transform azimuths and determining earthquake slip vectors. The weighted solution is needed because the uncertainties have different dimensions and vary between data points.

Thus uncertainties in the estimated Euler vectors are given by the model variance–covariance matrix

$$\sigma^2_{\mathrm{m}} = (G^T W_d G)^{-1}. \tag{8}$$

Uncertainties associated with the Euler poles are often shown by error ellipses analogous to those for earthquake locations, whereas those for the rates are quoted separately. Alternatively, we can view the pole and rate uncertainties as forming a three-dimensional ellipsoid. Hence two Euler vectors are distinct if their error ellipsoids do not overlap. As we have seen, conventional global plate motion studies using magnetic anomalies, transforms, and earthquake slip vectors yield solutions similar to those obtained by using the same formulation to invert the rates and azimuths of plate motions determined by space-based geodesy (Section 5.2.3). This agreement is gratifying, given that the conventional solutions combine data from magnetic anomalies averaged over millions of years, the azimuths of transform faults that formed over long times, and the slip vectors of earthquakes, whereas the space-geodetic solutions based on data spanning only a few years have different uncertainties.

7.5.2 *Testing the results with χ^2 and* F*-ratio tests*

Given a model derived by inversion, the natural question is, how good is it? This issue is a specific case of the general one of testing how well a model fits data, which is discussed in statistics texts. For our purposes we focus on two issues and note some results without proof.

One common way to test how well a model fits data uses the misfit function χ^2 that we minimized to derive the least squares solution (Eqn 7.2.11). We write it as

$$\chi^2 = \sum_i \frac{(d_i - d_i^m)^2}{\sigma_i^2}, \tag{9}$$

where d_i^m are the data predicted by the model, d_i are the data observed, and σ_i are their uncertainties. Lower values of χ^2 correspond to better fits. However, because a model derived from these data is bound to fit better than one derived without them, we examine the *reduced chi square*

$$\chi^2_\nu = \chi^2/\nu \tag{10}$$

where the parameter ν, known as the number of *degrees of freedom*, equals $n - p$ where n is the number of data and p is the number of model parameters estimated in the inversion.

If the model is a good fit to the data and our estimates of the uncertainties are reasonable, then we expect χ^2_ν to be around 1.

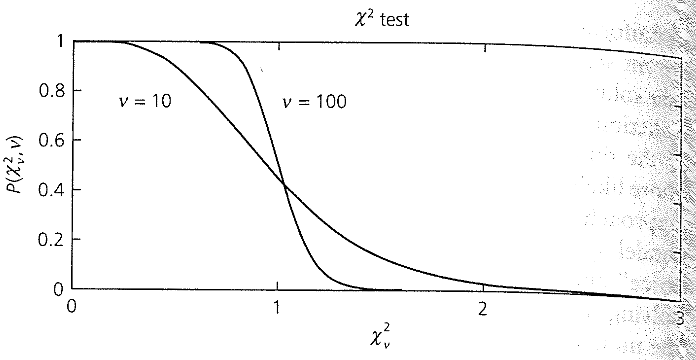


Fig. 7.5-1 Cumulative probability distribution $P(\chi^2_\nu, \nu)$, giving the probability of observing χ^2_ν above a certain value, plotted for 10 and 100 degrees of freedom. The more the degrees of freedom, the more likely χ^2_ν is to be near 1, and the less likely much higher or lower values are.

Statistically, this means that there is a reasonable possibility that the observed data are samples from a parent distribution described by the model, given the random uncertainties of measurement. However, if χ^2_ν is much larger than 1, it is unlikely that the data are samples from this distribution. This issue is addressed using the cumulative probability distribution $P(\chi^2_\nu, \nu)$ given by statistical tables or mathematical software that gives the probability of observing χ^2_ν above a certain value (Fig. 7.5-1). In other words, this test asks what the probability is that such a high value would be observed purely by chance due to the uncertainties of measurement. The more the degrees of freedom, the less likely a high value is. For example, the chance of observing χ^2_ν greater than 1.5 is about 13% for $\nu = 10$, but less than 1% for $\nu = 100$. Thus, the more data we have, the more the degrees of freedom, and closer to 1 we expect χ^2 to be. This test does not tell specifically whether the data observed are samples from the distribution predicted by the model, but gives instead some insight into the probability. If χ^2_ν is too large, there is likely to be something wrong.

One possibility is that the model does not include some crucial factors. For example, a plate motion model may not include an important plate boundary, and so does not describe the data well. In this case, the misfit is greater than expected from considering only random uncertainties of measurement, because systematic errors are also present. Similarly, the misfit to travel time in an earthquake location includes both errors of measurement and the effects of velocity structure like lateral heterogeneity. We sometimes rescale the uncertainties to make $\chi^2_\nu = 1$, which lets us assign confidence limits using χ^2_ν. This rescaling does not address the causes of the misfit, but implicitly lumps the systematic errors in with the errors of measurement. To do better requires improving the model.

Conversely, if χ^2_ν is too small, Fig. 7.5-1 indicates that something is also likely to be wrong. For example, for $\nu = 10$, there is only about a 2% chance of observing χ^2_ν less than 0.3, and the probability is less for more degrees of freedom. This is because the data are unlikely to be fit that well, given errors of measurement. About one-third (100 – 68%) of the data should be misfit by at least 1σ, and about 5% should be outside the 2σ range. Hence a low χ_ν^2 value, which we might view as showing an excellent fit, is more likely to imply that the uncertainties in the data have been overestimated, and have thus made χ_ν^2 appear too small. For example, χ_ν^2 for the NUVEL-1 model is 0.24, whereas it is expected to lie with 95% probability between 0.93 and 1.07. This effect is also seen for other plate motion models, suggesting that the assigned data uncertainties are more like 95% (2σ) confidence limits than one standard deviation. If so, the uncertainties in the model are correspondingly less than implied by the model variance–covariance matrix. Thus the χ^2 test formalizes the adage that if something seems too good to be true, it probably is.[1]

A second issue is whether the number of model parameters is appropriate. As discussed in Chapter 5, there are often several possible plate boundary geometries for an area. Naturally, more plates can describe plate motions in an area better because the model has more parameters. Thus we ask whether the improved fit shown by a lower value of χ_ν^2 is more than expected purely by chance due to the additional parameters. For example, a set of data in the x–y plane are always better fit by a higher-order polynomial, such as a quadratic versus a straight line.

This issue can be addressed using the *F-ratio* test, which gives insight into whether a set of data are significantly better fit by a model with more parameters. The idea is that if a set of n data are fit by two models, one with r parameters ($n - r$ degrees of freedom) and a second with p parameters ($n - p$ degrees of freedom) with p greater than r, the second model should fit the data better, and $\chi^2(p)$ should be less than $\chi^2(r)$. To test if the reduction in χ^2 is greater than would be expected simply because additional model parameters are added, we form the statistic

$$F = \frac{[\chi^2(r) - \chi^2(p)]/(p - r)}{\chi^2(p)/(n - p)}. \tag{11}$$

Statistical tables or mathematical software give the probability $P_F(F, \nu_1, \nu_2)$ of observing an F value greater than that observed for a random sample with $\nu_1 = (p - r)$ and $\nu_2 = (n - p)$. Thus, for example, if P_F is 0.01, there is only a 1% chance that the improved fit of the model with more parameters is due purely to chance. Because this test depends on the ratio of χ^2, it is not affected if the uncertainties are consistently over- or under-estimated.

We can use F to test whether the fit to n relative motion data of a model with $p + 1$ plates is significantly better than that of one with p plates. The p plate model has $3(p - 1)$ parameters ($n - 3p + 3$ degrees of freedom), whereas the $p + 1$ plate model has $3p$ parameters ($n - 3p$ degrees of freedom). Thus

$$F = \frac{[\chi^2(p \text{ plates}) - \chi^2(p + 1 \text{ plates})]/3}{\chi^2(p + 1 \text{ plates})/(n - 3p)} \tag{12}$$

is tested using $P_F(F, \nu_1, \nu_2)$ with $\nu_1 = 3$ and $\nu_2 = (n - 3p)$. If the risk that the improved fit would occur by chance is small, perhaps less than 1%, then we treat the additional plate as distinct. Conversely, if the improved fit is likely to result simply from the additional parameters, the data do not strongly indicate the presence of an additional plate. For example, such tests show that although the boundary between them is indistinct, North and South America should be treated as separate plates. This approach is used to investigate complicated regions where the plate geometry is unclear, such as near Japan and in the Indian Ocean. Similarly, we can investigate regions of intraplate deformation to see whether there is resolvable motion.

In many applications these or other statistical tests can be used to examine how well a model fits the data and to gain insight into whether the model is too simple (underparametrized) to explain the data or more complicated (overparametrized) than is required by the data. For example, we can examine cases when adding more layers to a velocity model significantly improves the fit to travel time data, when a more complex earthquake source model fits seismograms significantly better, or when a more complex model of earthquake recurrence describes an earthquake history better. In these applications the statistical tests address only the data used, so a more complex model may be justified based on other data, even if it is not required by the data tested. Moreover, we often suspect that the earth is more complicated than we would like when using simple statistical models. In particular, we often have little a priori knowledge of how to estimate the random and systematic errors. Even so, it is worth subjecting models to tests and seeing how well the data support our beliefs. This testing is a key part of the cycle (Fig. 1.1-8) by which models are refined using new data and model parameterizations.

Further reading

Many discussions of inverse theory, including ours, are based on Lanczos (1961). Applications in the earth sciences, especially seismology, are discussed in texts and reviews including Parker (1977), Aki and Richards (1980), and Menke (1984). Treatments of tomographic methods in seismology are given by Nolet (1987), Thurber and Aki (1987), Spakman and Nolet (1988), Humphreys and Clayton (1988), and Romanowicz (1991). Inversion for the properties of stratified media is reviewed by Wiggins (1972).

Tests for goodness of fit are discussed in statistical texts such as Bevington and Robinson (1992) and Freedman *et al.* (1991); the latter treats the issue of Mendel's results. Chase (1972) and Minster *et al.* (1974) present the inverse problem for plate motions; the latter gives the partial derivatives. Stein and Gordon (1984) and DeMets *et al.* (1990) discuss applications of the *F*-ratio test to plate motions and intraplate deformation.

[1] This approach has been used to argue that Mendel's famous results in 1865 that established the science of genetics are so good — the probability of observing them is 0.004% — that they are suspect. Similarly, instructors have used χ^2 tests to show that students' results reported in laboratory classes are so good that they are unlikely to have actually been obtained.

Problems

1. Show the following matrix identities:
   (a) For an arbitrary (not square) matrix A, the matrices A^TA and AA^T are symmetric.
   (b) For an arbitrary (not square) matrix B and a symmetric matrix A, $(B^TAB)^T = B^TAB$.
   (c) For square matrices A and B such that $(AB)^{-1}$ exists, $(AB)^{-1} = B^{-1}A^{-1}$.
2. Show that if a square matrix G has an inverse, the inverse and generalized inverse are identical.
3. Show that if the variance–covariance matrix of the data is diagonal, $\sigma^2_{\mathrm{d}} = \sigma^2_{ij}\delta_{ij}$ (with no summation implied), its inverse is another diagonal matrix $W_d = \delta_{ij}/\sigma^2_{ij}$. (Also with no summation implied.)
4. Show that the model variance–covariance matrix (Eqn 7.2.32) $\sigma^2_{\mathrm{m}} = G^{-g}\sigma^2_{\mathrm{d}}(G^{-g})^T$ reduces to $\sigma^2_{\mathrm{m}} = \sigma^2(G^TG)^{-1}$ when the data errors are uncorrelated and equal, so the data variance–covariance matrix is a constant times the identity matrix, $\sigma^2_{\mathrm{d}} = \sigma^2\delta_{ij}$.
5. Show that if the data errors are uncorrelated but not equal, such that the data variance–covariance matrix of the data is the diagonal matrix $\sigma^2_{\mathrm{d}} = \sigma^2_{ij}\delta_{ij}$ with inverse W_{d} (problem 3):
   (a) The least squares criterion (Eqn 7.2.14) for the inverse problem gives rise to the weighted least squares solution $\Delta\mathbf{m} = (G^TW_dG)^{-1}G^TW_d\Delta\mathbf{d}$.
   (b) The model variance–covariance matrix is $\sigma^2_{\mathrm{m}} = (G^TW_dG)^{-1}$.
6. For a halfspace with uniform (and known) velocities α and β:
   (a) Show how the location problem can be formulated to use both P-wave and S-wave first arrival times as data. Write the data vector, model vector, and partial derivatives. How do these differ from the case for P waves alone?
   (b) Show how the location problem can be formulated to use only the *difference* between P-wave and S-wave first arrival times as data. Write the data vector, model vector, and partial derivatives. How do these differ from the case for P waves alone? How might you apply this method if only the P velocity were known? Under what conditions might this method be useful?
7. For the idealized tomographic experiment in Figure 7.3-2:
   (a) Show how one row of the G matrix in Eqn 7.3.10 can be derived from the others, such that the four teleseismic ray paths are not linearly independent. Give a physical interpretation of this result.
   (b) Find four rows of the G matrix in Eqn 7.3.6 that are linearly independent, and give a physical interpretation of this result.

Computer problems

C-1. Write a subroutine to find the generalized inverse $G^{-g} = (G^TG)^{-1}G^T$ of an $(n \times r)$ matrix G, using a matrix inversion subroutine. As a test, check that the solution satisfies the criterion that for a square matrix G that has an inverse, the inverse and generalized inverse are identical.

C-2. For a homogeneous halfspace with P-wave velocity α:
   (a) Write a subroutine to compute the distance and travel time between two points (x, y, z) and (x_i, y_i, z_i). Test this for some simple cases.
   (b) Use the result of (a) to write a program that reads an earthquake location, origin time, and medium velocity and the locations of n seismic stations, and finds the first arrival time at each station.
   (c) Write a subroutine using the result of (a) to compute the partial derivatives of the first arrival time at a station with respect to changes in the model parameters (location, origin time, and medium velocity).
   (d) Modify the result of (b) to compute arrival times for a starting model (assumed location, origin time, and medium velocity), and then locate the earthquake by inverting these synthetic data to find the best-fitting model. The result of C-1 should be useful. Have the program iterate until the model change between iterations is less than a parameter you set. The program should have the option to invert for velocity or hold velocity fixed at an assumed starting value.

C-3. Test the location program with a set of station locations, a "real" origin time and location, and an incorrect starting model. The program should retrieve the "real" model. Once this works for error-free data, add some errors to the travel times, either by using your computer's random number function or by simply choosing some numbers. Invert for the best-fitting model, and see how the result of the inversion changes as the errors become a larger fraction of the travel times. How do the results depend on whether the velocity is held fixed or inverted for?

C-4. Compute and compare χ^2 and χ^2_ν for C-3 for cases in which you inverted for velocity and in which the velocity is fixed at an incorrect value. Using the F-ratio test, does the improved fit due to inverting for velocity seem significant?

Appendix: Mathematical and Computational Background

If you wish to learn about nature, to appreciate nature, it is necessary to understand the language she speaks in. She offers her information only in one form; we are not so unhumble as to demand that she change before we pay attention.

Richard Feynman, *The Character of Physical Law* (1982)

A.1 Introduction

The study of seismology follows a pattern characteristic of many scientific disciplines. We first identify phenomena that we seek to understand, such as the propagation of seismic waves through the solid earth. We then consider the physics of the simplest relevant case, such as the propagation of a wave of a single frequency through a uniform material, formulate the problem mathematically, and derive a solution. From this solution, we build up mathematical solutions to more complex problems, each of which is ideally a better approximation to the complexities of the real earth. Although the simpler problems can be solved analytically, eventually the complexities require numerical techniques.

We thus rely on a set of mathematical techniques often used in physical problems. Experience suggests that although many readers are familiar with most of the mathematics required in this book, a review is often helpful. This appendix briefly summarizes a broad range of material. The first sections treat a variety of mathematical topics. The final section reviews some concepts relevant to the use of computers for scientific calculations.

In using these mathematical techniques, it is worth bearing in mind that we are invoking the special power of mathematics to deal with physical problems. This power is that if a physical problem is posed correctly in mathematical terms, then applying mathematical techniques to this formulation yields quite different, and often apparently unrelated, statements that also correctly describe the physical world. For example, in Section 2.4 we used the equations of elasticity and applied vector calculus to derive the properties of seismic waves that we observe. Similarly, in Section 2.5 we derived an observed physical relation, Snell's law, starting from three different physical formulations. Conversely, we have seen that different physical phenomena can be described using similar mathematical approaches and so have some deep similarities. Although in hindsight such successes may not seem surprising, because many of the mathematical methods we use were developed to solve such physical problems, they illustrate the intimate connection between sciences like seismology and mathematics.[1]

A.2 Complex numbers

In several of our applications, notably in describing propagating waves and their frequency content, complex numbers are helpful. We thus briefly review some of their properties.

The complex number $z = a + ib$, where $i = \sqrt{-1}$, has a real part, a, and an imaginary part, b. These relations are sometimes written $a = \mathrm{Re}\,(z)$ and $b = \mathrm{Im}\,(z)$. Complex numbers are typically plotted in the complex plane with their real parts on the x_1 axis and their imaginary parts on the x_2 axis (Fig. A.2-1). Alternatively, a complex number can be written in *polar coordinate* form as

$$z = a + ib = re^{i\theta} = r(\cos\theta + i\sin\theta). \qquad (1)$$

[1] Most seismologists are more conservative than Paul Dirac, a leader in the development of quantum physics, who invented the delta function. Dirac regarded mathematical beauty as a guiding principle, stating that "it is more important to have beauty in one's equations than to have them fit experiment."

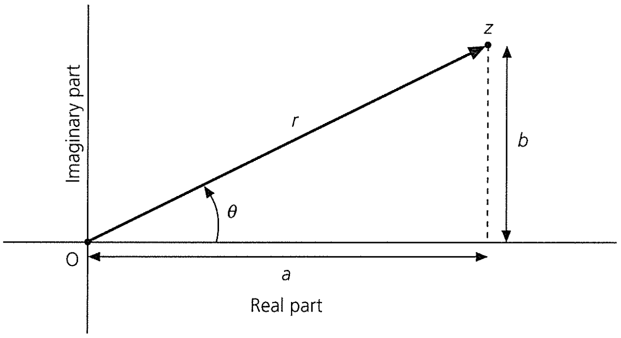


Fig. A.2-1 A number in the complex plane can be represented in terms of its real and imaginary parts, $z=a+ib$, or in polar form $z=re^{i\theta}$.

The *polar coordinates*, the magnitude r and the phase angle θ, can be expressed in terms of the real and imaginary parts as

$$r=\sqrt{a^2+b^2}, \quad \theta=\tan^{-1}(b/a). \tag{2}$$

and, conversely,

$$a=r\cos\theta, \quad b=r\sin\theta. \tag{3}$$

To describe complex numbers in all four quadrants of the complex plane, θ ranges from 0 to 2π. Because the inverse tangent is periodic with period π, the signs of the real and imaginary parts are used to obtain the correct phase.

Complex numbers are equal when they have the same real and imaginary parts. Two complex numbers in $(a+ib)$ form are added by adding the real parts and the imaginary parts:

$$(a_1+ib_1)+(a_2+ib_2)=(a_1+a_2)+i(b_1+b_2). \tag{4}$$

Complex numbers can be multiplied either in the $(a+ib)$ form:

$$(a_1+ib_1)(a_2+ib_2)=(a_1a_2-b_1b_2)+i(a_1b_2+b_1a_2), \tag{5}$$

or in the magnitude and phase form:

$$r_1e^{i\theta_1}r_2e^{i\theta_2}=r_1r_2e^{i(\theta_1+\theta_2)}. \tag{6}$$

The conjugate of a complex number z, z^*, has the same real part and an imaginary part of opposite sign. Because

$$\begin{aligned} z^* &= a-ib=r\cos\theta-ir\sin\theta \\ &= r\cos(-\theta)+ir\sin(-\theta)=re^{-i\theta}, \end{aligned} \tag{7}$$

the conjugate has the same magnitude but the opposite phase. Hence the square of the magnitude of a complex number can be found by multiplication by the complex conjugate,

$$|z|^2=zz^*=(a+ib)(a-ib)=(a^2+b^2)=re^{i\theta}re^{-i\theta}=r^2. \tag{8}$$

By combining

$$e^{i\theta}=\cos\theta+i\sin\theta \quad \text{and} \quad e^{-i\theta}=\cos\theta-i\sin\theta \tag{9}$$

we obtain the definitions of the sine and cosine functions in terms of complex exponentials

$$\cos\theta=(e^{i\theta}+e^{-i\theta})/2 \quad \text{and} \quad \sin\theta=(e^{i\theta}-e^{-i\theta})/2i. \tag{10}$$

These relations yield formulae for the trigonometric functions of the sum of the angles because

$$e^{i(\theta_1+\theta_2)}=\cos(\theta_1+\theta_2)+i\sin(\theta_1+\theta_2) \tag{11}$$

and, by Eqn 6,

$$\begin{aligned} e^{i(\theta_1+\theta_2)} &= e^{i\theta_1}e^{i\theta_2}=(\cos\theta_1+i\sin\theta_1)(\cos\theta_2+i\sin\theta_2) \\ &= (\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2) \\ &\quad +i(\sin\theta_1\cos\theta_2+\cos\theta_1\sin\theta_2), \end{aligned} \tag{12}$$

so we can equate the real and imaginary parts and find

$$\cos(\theta_1+\theta_2)=\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2 \tag{13}$$

and

$$\sin(\theta_1+\theta_2)=\sin\theta_1\cos\theta_2+\cos\theta_1\sin\theta_2. \tag{14}$$

These expressions are symmetric in θ_1 and θ_2, as expected. The corresponding relations for the trigonometric functions of the difference of two angles are found by making θ_2 negative. Setting $\theta_1=\theta_2$ gives expressions for $\cos(2\theta)$ and $\sin(2\theta)$.

The relations for the product of trigonometric functions of two angles can also be found using complex exponentials

$$\begin{aligned} \cos\theta_1\cos\theta_2 &= \frac{(e^{i\theta_1}+e^{-i\theta_1})}{2}\frac{(e^{i\theta_2}+e^{-i\theta_2})}{2} \\ &= \frac{1}{4}[(e^{i(\theta_1+\theta_2)}+e^{-i(\theta_1+\theta_2)})+(e^{i(\theta_1-\theta_2)}+e^{-i(\theta_1-\theta_2)})] \\ &= \frac{1}{2}[\cos(\theta_1+\theta_2)+\cos(\theta_1-\theta_2)] \end{aligned} \tag{15}$$

and, similarly,

$$\begin{aligned} \sin\theta_1\sin\theta_2 &= \frac{(e^{i\theta_1}-e^{-i\theta_1})}{2i}\frac{(e^{i\theta_2}-e^{-i\theta_2})}{2i} \\ &= \frac{1}{4}[(e^{i(\theta_1-\theta_2)}+e^{-i(\theta_1-\theta_2)})-(e^{i(\theta_1+\theta_2)}+e^{-i(\theta_1+\theta_2)})] \\ &= \frac{1}{2}[\cos(\theta_1-\theta_2)-\cos(\theta_1+\theta_2)]. \end{aligned} \tag{16}$$

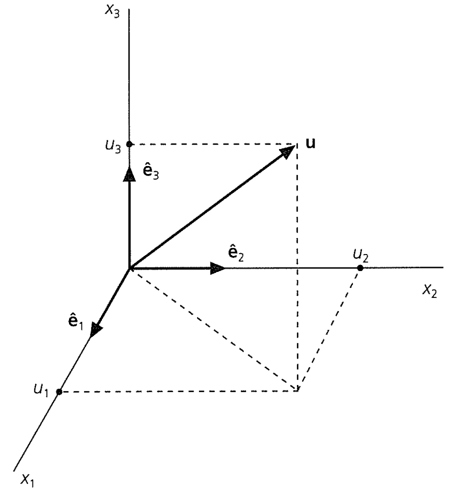


Fig. A.3-1 A vector **u** is expressed by the Cartesian unit basis vectors and its components: $\mathbf{u} = u_1\hat{e}_1 + u_2\hat{e}_2 + u_3\hat{e}_3$.

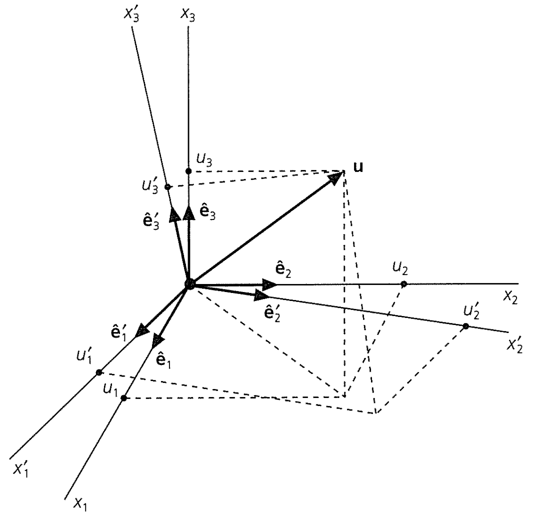


Fig. A.3-2 A vector **u** is described in each of two orthogonal coordinate systems by the Cartesian unit basis vectors of the coordinate system and the components of the vector in the coordinate system: $\mathbf{u} = u_1\hat{e}_1 + u_2\hat{e}_2 + u_3\hat{e}_3 = u'_1\hat{e}'_1 + u'_2\hat{e}'_2 + u'_3\hat{e}'_3$. Although the components differ between coordinate systems, the vector remains the same.

A.3 Scalars and vectors

A.3.1 Definitions

In seismology, we deal with several types of physical quantities. The simplest, *scalars*, are numbers describing a physical property at a given point that is independent of the coordinate system used to identify the point. Temperature, pressure, mass, and density are familiar examples. Mathematically, if a point is described in one coordinate system by (x_1, x_2, x_3) and in a second by (x'_1, x'_2, x'_3), the value of a scalar function ϕ in the first coordinate system equals that of the corresponding scalar function in the second

$$\phi(x_1, x_2, x_3) = \phi'(x'_1, x'_2, x'_3). \tag{1}$$

The distance between two points is a scalar because although the coordinates of the points depend on the coordinate system, the distance does not.

Vectors are more complicated entities that have magnitude and direction. In seismology, the most common vector is the motion, or *displacement*, of a piece of material within the earth due to the passage of a seismic wave. Vectors transform between different coordinate systems in a specific way. Thus, if the horizontal ground motion is recorded with seismometers oriented northeast–southwest and northwest–southeast, the north–south and east–west components of the displacement can be found using the properties of vectors. We will see that although the components depend on the coordinate system, the magnitude and direction of the vector remain the same.

Consider the familiar Cartesian coordinate system (Fig. A.3-1) with three mutually perpendicular (orthogonal) coordinate axes. There are two standard notations for these coordinates and axes: either the x_1, x_2, and x_3, or the x, y, and z axes. Each notation has advantages. The x_1, x_2, x_3 notation is more convenient for some derivations, and the x, y, z notation is sometimes clearer in physical problems. We use the x_1, x_2, and x_3 notation in this appendix, and use whichever notation seems more convenient in other discussions.

A point in this coordinate system is described by its x_1, x_2, and x_3 coordinates. Because a vector can be defined by a line from the origin (0, 0, 0) to the point (u_1, u_2, u_3), the three numbers u_1, u_2, and u_3 are the *components* of the vector **u**. A vector is denoted either by boldface type or by a set of its components

$$\mathbf{u} = (u_1, u_2, u_3) = (u_x, u_y, u_z). \tag{2}$$

A Cartesian coordinate system is described by three orthogonal unit basis vectors, $\hat{e}_1$, $\hat{e}_2$, and $\hat{e}_3$, along the x_1, x_2, and x_3 coordinate axes:

$$\hat{e}_1 = (1, 0, 0) \quad \hat{e}_2 = (0, 1, 0) \quad \hat{e}_3 = (0, 0, 1). \tag{3}$$

The caret, or "hat" superscript, indicates a *unit vector*, whose length is 1. The vector **u** is formed from its components and the basis vectors

$$\mathbf{u} = u_1\hat{e}_1 + u_2\hat{e}_2 + u_3\hat{e}_3 = (u_1, u_2, u_3). \tag{4}$$

Now, consider a second Cartesian coordinate system with the same origin and different axes x'_1, x'_2, and x'_3, along which unit basis vectors $\hat{e}'_1$, $\hat{e}'_2$, and $\hat{e}'_3$ are defined (Fig. A.3-2). In this coordinate system the components of **u** are different,

$$\mathbf{u} = u'_1\hat{e}'_1 + u'_2\hat{e}'_2 + u'_3\hat{e}'_3 = (u'_1, u'_2, u'_3). \tag{5}$$

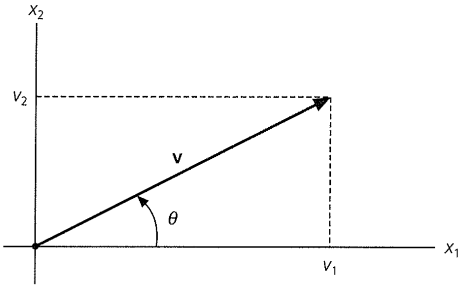


Fig. A.3-3 A vector in two dimensions making an angle θ with the x_1 axis.

Thus the *same* physical vector is represented in a different coordinate system, described by a different set of basis vectors, using different components. The essential idea is that the vector remains the same, or invariant, regardless of the coordinate system, although the numerical values of its components change. Physical laws, like Newton's law stating that the force vector equals the product of the mass and the acceleration vector (the second derivative with respect to time of the displacement vector), are written in vector form because the physical phenomenon does not depend on the coordinate system used to describe it.

The length or *magnitude* of a vector, $|\mathbf{u}|$, is a scalar, and thus the same in different coordinate systems. By the Pythagorean theorem, the length is

$$|\mathbf{u}| = (u_1^2 + u_2^2 + u_3^2)^{1/2} = (u_1'^2 + u_2'^2 + u_3'^2)^{1/2}. \tag{6}$$

The zero vector, **0**, all of whose components are zero in any coordinate system, has zero magnitude.

A vector is specified in either Cartesian coordinates by its components or in polar coordinates by its magnitude and direction. For example, in a two-dimensional (x_1, x_2) coordinate system (Fig. A.3-3), the vector **v** can be written in terms of its components

$$\mathbf{v} = (v_1, v_2) \tag{7}$$

or its magnitude

$$|\mathbf{v}| = (v_1^2 + v_2^2)^{1/2} \tag{8}$$

and direction, given by the angle θ that **v** makes with the x_1 direction

$$\theta = \tan^{-1}(v_2/v_1). \tag{9}$$

Just as $|\mathbf{v}|$ and θ are given by the components, so the components are given by $|\mathbf{v}|$ and θ

$$v_1 = |\mathbf{v}| \cos\theta \quad \text{and} \quad v_2 = |\mathbf{v}| \sin\theta. \tag{10}$$

By analogy, a vector in three dimensions is specified by either its three components or its magnitude and the angles it forms with two of the coordinate axes. It is worth noting that the mathematical convention of defining angles counterclockwise from x_1 differs from the geographical convention of defining angles clockwise from North (x_2), so conversions are often needed.

A.3.2 *Elementary vector operations*

The simplest vector operation is multiplication of a vector by a scalar

$$\alpha\mathbf{u} = (\alpha u_1, \alpha u_2, \alpha u_3). \tag{11}$$

For example, in two dimensions,

$$\alpha\mathbf{v} = (\alpha v_1, \alpha v_2) \tag{12}$$

yields a vector with magnitude

$$((\alpha v_1)^2 + (\alpha v_2)^2)^{1/2} = |\alpha| (v_1^2 + v_2^2)^{1/2} = |\alpha| |\mathbf{v}| \tag{13}$$

whose direction is given by

$$\tan\theta = \alpha v_2/\alpha v_1 = v_2/v_1. \tag{14}$$

Multiplication by a positive scalar thus changes the magnitude of a vector but preserves its direction. Similarly, multiplication by a negative scalar changes the magnitude of a vector and reverses its direction. $\hat{\mathbf{u}}$, a unit vector in the direction of **u** is formed by dividing **u** by its magnitude

$$\hat{\mathbf{u}} = \mathbf{u}/|\mathbf{u}|. \tag{15}$$

The sum of two vectors is another vector whose components are the sums of the corresponding components, so if

$$\mathbf{a} = a_1\hat{e}_1 + a_2\hat{e}_2 + a_3\hat{e}_3 \quad \text{and} \quad \mathbf{b} = b_1\hat{e}_1 + b_2\hat{e}_2 + b_3\hat{e}_3,$$
$$\mathbf{a} + \mathbf{b} = (a_1 + b_1)\hat{e}_1 + (a_2 + b_2)\hat{e}_2 + (a_3 + b_3)\hat{e}_3 = \mathbf{b} + \mathbf{a}. \tag{16}$$

Addition can be done graphically (Fig. A.3-4) by shifting one vector, while preserving its orientation, so that its "tail" is at the "head" of the other, and forming the vector sum. For example, the total force vector acting on an object is the vector sum of the individual force vectors. Equation 16 and Fig. A.3-4 show that vector addition is commutative; it does not matter in which order the vectors are added.

A.3.3 *Scalar products*

There are two methods of multiplying vectors. The first, the *scalar product* (also called the dot product or inner product), yields a scalar:

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3 = |\mathbf{a}| |\mathbf{b}| \cos\theta, \tag{17}$$

where θ is the angle between two vectors. To see that the two definitions of the scalar product are equivalent, consider a two-

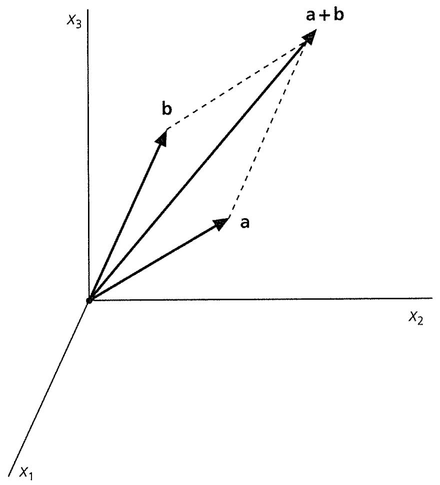


Fig. A.3-4 Addition of vectors a and **b**. The addition can be done analytically, by adding components, or graphically. Vector addition is commutative, as the order of addition is irrelevant.

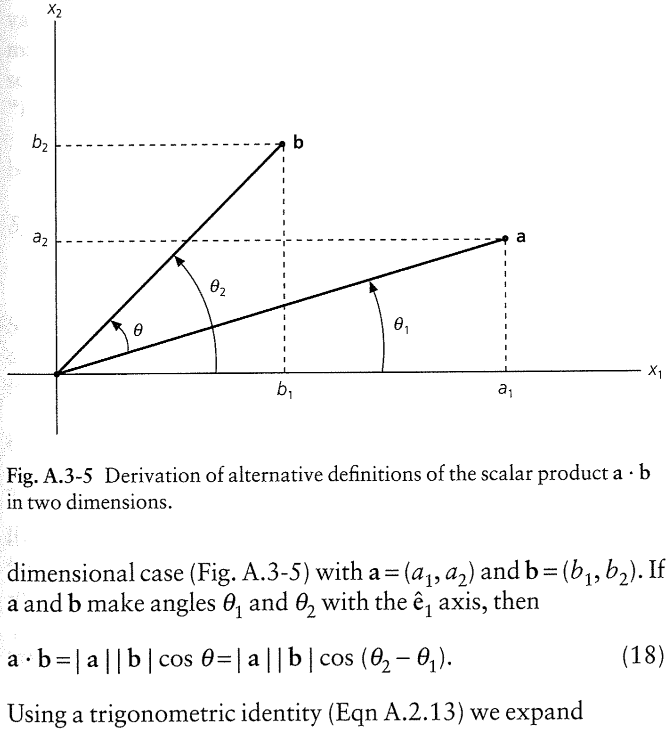


Fig. A.3-5 Derivation of alternative definitions of the scalar product **a** · **b** in two dimensions.

dimensional case (Fig. A.3-5) with $\mathbf{a}=(a_1, a_2)$ and $\mathbf{b}=(b_1, b_2)$. If **a** and **b** make angles θ_1 and θ_2 with the $\hat{e}_1$ axis, then

$$\mathbf{a}\cdot\mathbf{b}=|\mathbf{a}||\mathbf{b}|\cos\theta=|\mathbf{a}||\mathbf{b}|\cos(\theta_2-\theta_1). \tag{18}$$

Using a trigonometric identity (Eqn A.2.13) we expand

$$\cos\theta=\cos(\theta_2-\theta_1)=\cos\theta_2\cos\theta_1+\sin\theta_2\sin\theta_1. \tag{19}$$

Because

$$\cos\theta_1=a_1/(a_1^2+a_2^2)^{1/2} \quad \text{and} \quad \sin\theta_1=a_2/(a_1^2+a_2^2)^{1/2}, \tag{20}$$

and similar definitions hold for θ_2 and **b**, substitutions for the angles in Eqn 18 show that

$$|\mathbf{a}||\mathbf{b}|\cos\theta=\frac{|\mathbf{a}||\mathbf{b}|(a_1b_1+a_2b_2)}{(a_1^2+a_2^2)^{1/2}(b_1^2+b_2^2)^{1/2}}=a_1b_1+a_2b_2. \tag{21}$$

Equation 17 shows several features of the scalar product:

- The scalar product commutes: $\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a}$.
- The scalar product of two perpendicular vectors is zero, because cos 90° = 0.
- The scalar product of a vector with itself is its magnitude squared:

$$\mathbf{a}\cdot\mathbf{a}=a_1a_1+a_2a_2+a_3a_3=|\mathbf{a}|^2. \tag{22}$$

The definition of the scalar product is generalized for vectors with complex components. To see why, note that for a vector $\mathbf{a}=(i, 1, 0)$, where $i=\sqrt{-1}$, Eqn 22 would give a squared magnitude of zero. Because we would like only the zero vector, all of whose elements are zero, to have zero magnitude, Eqn 17 is generalized to

$$\mathbf{a}\cdot\mathbf{b}=a_1^*b_1+a_2^*b_2+a_3^*b_3 \tag{23}$$

where * indicates the complex conjugate. Thus the definition of the squared magnitude (Eqn 22) becomes

$$\mathbf{a}\cdot\mathbf{a}=a_1^*a_1+a_2^*a_2+a_3^*a_3=|\mathbf{a}|^2. \tag{24}$$

For example, the squared magnitude of $|(i, 1, 0)|^2=(i)(-i)+(1)(1)=2$. These complex definitions reduce to the familiar cases, (Eqns 17 and 22), for vectors with real components.

The relations between the unit basis vectors for a Cartesian coordinate system, $\hat{e}_1$, $\hat{e}_2$, and $\hat{e}_3$, are easily stated using their scalar products. Because each is perpendicular to the other two, the scalar product of any two different ones is zero,

$$\hat{e}_1\cdot\hat{e}_2=\hat{e}_1\cdot\hat{e}_3=\hat{e}_2\cdot\hat{e}_3=0, \tag{25}$$

and the scalar product of each with itself is its squared magnitude

$$\hat{e}_1\cdot\hat{e}_1=\hat{e}_2\cdot\hat{e}_2=\hat{e}_3\cdot\hat{e}_3=1. \tag{26}$$

The unit basis set of vectors is *orthonormal*; each is orthogonal (perpendicular) to the others and normalized to unit magnitude.

The *projection*, or component of a vector in a direction given by a unit vector, is the scalar product of a vector with the unit vector. Using this idea, a component of a vector can be found from its projection on the unit basis vector along the corresponding axis. Thus the x_1 component of **u** is

$$\mathbf{u}\cdot\hat{e}_1=(u_1\hat{e}_1+u_2\hat{e}_2+u_3\hat{e}_3)\cdot\hat{e}_1=u_1, \tag{27}$$

with the other components defined similarly.

A.3.4 Vector products

A second form of multiplication, the *vector* or *cross* product, forms a third vector from two vectors by

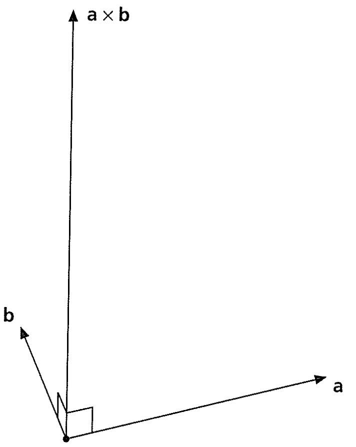


Fig. A.3-6 Illustration of the right-hand rule giving the orientation of the vector product a × b.

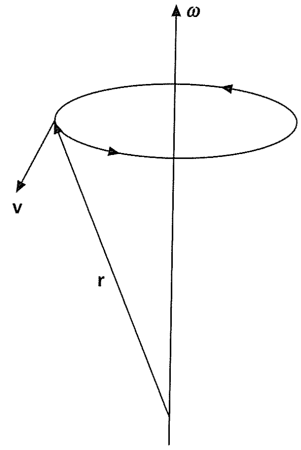


Fig. A.3-7 The vector product $\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}$ describes a rotation.

$$\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2)\hat{\mathbf{e}}_1 + (a_3 b_1 - a_1 b_3)\hat{\mathbf{e}}_2 + (a_1 b_2 - a_2 b_1)\hat{\mathbf{e}}_3, \tag{28}$$

which can be written as the determinant

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}. \tag{29}$$

The vector product of two vectors is perpendicular to both vectors. For example, if **a** and **b** are in the x_1–x_2 plane, $a_3 = b_3 = 0$, and by Eqn 28, the vector product has only an $\hat{\mathbf{e}}_3$ component. This can be shown in general by evaluating $\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 0$. Geometrically, the direction of the vector product is found by a "right-hand rule" (Fig. A.3-6): if the fingers of a right hand rotate from **a** to **b**, the thumb points in the direction **a** × **b**. The magnitude of the cross product is

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin \theta, \tag{30}$$

where θ is the angle between the two vectors. The cross product is zero for parallel vectors because sin 0° = 0, so the cross product of a vector with itself is zero.

The vector product often appears in connection with rotations, such as those used to describe the motion of lithospheric plates (Section 5.2). For example, if an object located at a position **r** undergoes a rotation, its linear velocity **v** is given by

$$\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}, \tag{31}$$

where ω is the rotation vector, which is oriented along the axis of rotation, with a magnitude $|\omega|$ that is the angular velocity (Fig. A.3-7). Similarly, the vector product is used to define the torque, which gives the rate of change of angular momentum. A force **F**, acting at a point **r**, gives a torque

$$\tau = \mathbf{r} \times \mathbf{F}. \tag{32}$$

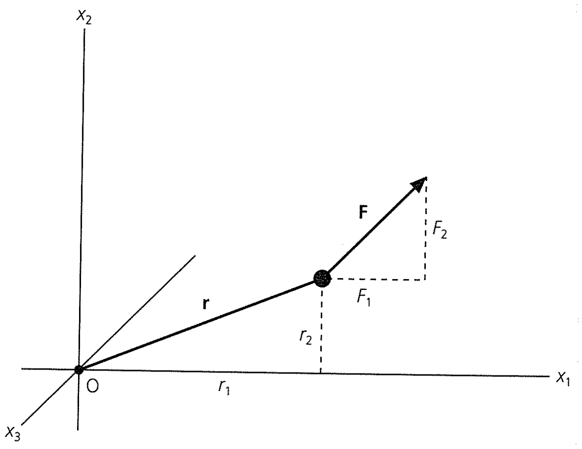


Fig. A.3-8 The x_3 component of the vector product $\tau = \mathbf{r} \times \mathbf{F}$ gives the torque, $r_1 F_2 - r_2 F_1$ about the x_3 axis. In this case $r_1 F_2$ is greater than $r_2 F_1$, so counterclockwise rotation about the x_3 axis occurs.

For example, the torque about the x_3 axis is $\tau_3 = (r_1 F_2 - r_2 F_1)$, so each component of the force contributes a counterclockwise torque equal to the component times its lever arm, the perpendicular distance of the point from that axis (Fig. A.3-8).

Some useful identities, whose proofs are left as problems, are

$$\begin{aligned} &\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} \\ &\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \\ &\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) \\ &\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b}). \end{aligned} \tag{33}$$

A.3.5 *Index notation*

Vector equations, such as the definition of the cross product, can be cumbersome when written in terms of the components. Simplification can be obtained using *index notation*, whereby an index assuming all possible values replaces the subscripts indicating coordinate axes. For example, the vector $\mathbf{u} = (u_1, u_2, u_3)$ is written u_i, where i can be 1, 2, or 3. In this notation, the scalar product is

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3 = \sum_{i=1}^{3} a_i b_i . \tag{34}$$

Because the sum over all coordinates appears frequently, the *Einstein summation convention* is often used, whereby an index repeated twice implies a summation over that index, and the summation sign is not explicitly written. Hence the scalar product of two real vectors is written

$$\mathbf{a} \cdot \mathbf{b} = a_i b_i, \tag{35}$$

using implied summation over the repeated index i. Similarly, the square of the magnitude of a real vector is

$$|\mathbf{u}|^2 = u_i u_i. \tag{36}$$

A repeated index is called a "dummy" index, like a dummy variable of integration, because it is used only within the summation. The form of the expression indicates that $u_i u_i$ is a scalar; because the repeated index is summed, no index remains "free." By contrast, u_i is a vector, because there is a free index.

Index notation is further simplified by introducing two symbols, δ_{ij} and ε_{ijk}. The *Kronecker delta*, δ_{ij}, is defined

$$\begin{aligned} \delta_{ij} &= 0 \quad \text{if } i \neq j, \\ &= 1 \quad \text{if } i = j. \end{aligned} \tag{37}$$

So, for example, $\delta_{11} = 1$, but $\delta_{12} = 0$. Using the Kronecker delta symbol, the relations between the Cartesian basis vectors (Eqns 25, 26) can be written compactly as

$$\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j = \delta_{ij}. \tag{38}$$

The Kronecker delta, a function of two discrete variables i and j, is analogous to the Dirac delta function which is a function of a continuous variable (Section 6.2.5).

The *permutation symbol*, ε_{ijk}, is defined as

$$\begin{aligned} \varepsilon_{ijk} &= 0 \quad \text{if any of the indices are the same,} \\ &= 1 \quad \text{if } i, j, k \text{ are in order, i.e., } (1,2,3), (2,3,1), \text{ or } (3,1,2) \\ &= -1 \quad \text{if } i, j, k \text{ are out of order, i.e., } (2,1,3), (3,2,1), (1,3,2). \end{aligned} \tag{39}$$

Cases where the indices are in order are known as even, or cyclic, permutations of the indices; those in which the indices are out of order are known as odd permutations. Because of the symmetries in the definition, $\varepsilon_{ijk} = \varepsilon_{jki} = \varepsilon_{kij}$. A useful relation, whose proof is left for the problems, is

$$\varepsilon_{ijk}\varepsilon_{ist} = \delta_{js}\delta_{kt} - \delta_{jt}\delta_{ks}. \tag{40}$$

Using index notation, the definition of the vector product (Eqn 28) becomes

$$(\mathbf{a} \times \mathbf{b})_i = \sum_{j=1}^{3}\sum_{k=1}^{3} \varepsilon_{ijk} a_j b_k = \varepsilon_{ijk} a_j b_k, \tag{41}$$

where the last form uses the summation convention. The notation shows that the cross product yields a vector because only one index, i, remains free after the repeated indices j and k are summed. To see that the index notation gives the correct definition, we expand the $i = 2$ component as

$$\begin{aligned} (\mathbf{a} \times \mathbf{b})_2 &= \varepsilon_{211}a_1b_1 + \varepsilon_{212}a_1b_2 + \varepsilon_{213}a_1b_3 + \varepsilon_{221}a_2b_1 + \varepsilon_{222}a_2b_2 \\ &\quad + \varepsilon_{223}a_2b_3 + \varepsilon_{231}a_3b_1 + \varepsilon_{232}a_3b_2 + \varepsilon_{233}a_3b_3 \\ &= (a_3b_1 - a_1b_3), \end{aligned} \tag{42}$$

because the only nonzero ε_{ijk} terms are $\varepsilon_{213} = -1$ and $\varepsilon_{231} = 1$.

Index notation points out an interesting feature of the vector product. Because $a_ib_i = b_ia_i$, the scalar product commutes. By contrast, the properties of the permutation symbol show that

$$\mathbf{a} \times \mathbf{b} = \varepsilon_{ijk} a_j b_k = -\varepsilon_{ijk} b_j a_k = -\mathbf{b} \times \mathbf{a}, \tag{43}$$

so the order matters for the vector product.

Although index notation seems unnatural at first, it does more than simply shorten expressions. The notation explicitly indicates what operations must be performed, and thus makes them easier to evaluate. For example, suppose we seek to show that the cross product of a vector with itself is zero. In contrast to $(\mathbf{a} \times \mathbf{a})$, the notation $\varepsilon_{ijk}a_ja_k$ shows how the cross product should be evaluated. Because a_ja_k is symmetric in the indices j and k, the permutation symbol makes the terms involving any pair of j and k sum to zero. We will see that index notation makes the complicated expressions that we encounter in studying stress and strain easier to evaluate.

A.3.6 Vector spaces

These concepts for vectors can be generalized in several ways. In three dimensions any vector is a weighted combination of three basis vectors. The usual choice of basis vectors along coordinate axes is for simplicity. We could choose any three mutually orthogonal vectors, which need not be of unit length, to be the basis vectors. To see this, remember that a physical vector does not depend on the coordinate system.

Moreover, the idea of vectors in two- or three-dimensional space can be generalized to spaces with a larger number of dimensions. For example, given unit vectors

$$\begin{aligned} &\hat{\mathbf{e}}_1 = (1,0,0,0,0), \quad \hat{\mathbf{e}}_2 = (0,1,0,0,0), \quad \hat{\mathbf{e}}_3 = (0,0,1,0,0), \\ &\hat{\mathbf{e}}_4 = (0,0,0,1,0), \quad \hat{\mathbf{e}}_5 = (0,0,0,0,1), \end{aligned} \tag{44}$$

a vector $\mathbf{u}$ can be formed from the basis vectors and components

$$\mathbf{u} = u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3 + u_4\hat{\mathbf{e}}_4 + u_5\hat{\mathbf{e}}_5 = (u_1, u_2, u_3, u_4, u_5). \tag{45}$$

This vector is defined in a five-dimensional space, with five axes each orthogonal to the others, because their scalar products are zero. Although this is difficult to visualize (or draw), the mathematics carries through directly from the three-dimensional case. N mutually orthogonal vectors thus provide a basis for an N-dimensional space.

These ideas are formalized in terms of vectors in a general *linear vector space*. For our purposes, a vector space is a collection of vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, satisfying several criteria:

- The sum of any two vectors in the space is also in the space.
- Vector addition commutes: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
- Vector addition is associative: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
- There exists a unique vector $\mathbf{0}$ such that for all $\mathbf{x}$, $\mathbf{x} = \mathbf{x} + \mathbf{0}$.
- There exists a unique vector $-\mathbf{x}$ such that for all $\mathbf{x}$, $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
- Scalar multiplication is associative: $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$.
- Scalar multiplication is distributive: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ and $(\alpha + \beta)\mathbf{x} = (\alpha\mathbf{x} + \beta\mathbf{x})$.

A point worth considering is the number of independent vectors in a vector space. Given N vectors $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N$ in a linear vector space, a weighted sum $\sum \alpha_i \mathbf{x}^i$ is called a *linear combination.* The N vectors are *linearly independent* if

$$\sum_{i=1}^{N} \alpha_i \mathbf{x}^i = 0 \text{ only when all } \alpha_i = 0, \tag{46}$$

so that no vector can be expressed as a combination of the others. Otherwise, the vectors are *linearly dependent*, and one can be expressed as a linear combination of the others.

This idea corresponds to that of basis vectors. If N basis vectors are mutually orthogonal, they are linearly independent. Because any vector in an N-dimensional space is a linear combination of N linearly independent basis vectors, the basis vectors *span* the space. Thus the dimension of a vector space is the number of linearly independent vectors within it. For example, we cannot find four linearly independent vectors in three dimensions.

Though vector spaces sound abstract, they are useful in seismology. For example, in Chapter 2 we represent travelling waves by normal modes, which are orthogonal basis vectors in a vector space, so any wave is a weighted sum of them. The modes of a string (Section 2.2.5) form a Fourier series (Chapter 6), in which a function is expanded into sine and cosine functions that are the basis vectors of a vector space. A similar approach is also used for the modes of the spherical earth (Section 2.9). Vector space ideas are also used in inverting seismological observations to study earth structure (Chapter 7).

A.4 Matrix algebra

A.4.1 *Definitions*

Matrix algebra is a powerful tool often used to study systems of equations. As a result, it appears in seismological applications, including stresses and strains, locating earthquakes, and seismic tomography. We thus review some basic ideas, often stating results without proof and leaving proofs for the problems. Further discussion of these topics can be found in linear algebra texts.

Given a matrix A with m rows and n columns, called an $m \times n$ matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{1}$$

and a second matrix B, also with m rows and n columns, matrix addition is defined by

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{pmatrix}. \tag{2}$$

The usual convention is to indicate matrices with capital letters and their elements with lower-case ones.

Matrix multiplication is defined such that for a matrix A that is $m \times n$ and a matrix B that is $n \times r$, the ij^{th} element of the $m \times r$ product matrix $C = AB$ is defined by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{ik} b_{kj}. \tag{3}$$

The ij^{th} element of C is the scalar product of the i^{th} row of A and the j^{th} column of B. As a result, for matrix multiplication the two matrices need not have the same number of rows and columns, but must have the number of columns in the first matrix equal to the number of rows in the second. Often the numbers of rows and columns in the two matrices allow multiplication in only one order. Thus, in the example above, A "premultiplies" B, or B "postmultiplies" A. A convenient way to remember this is that the number of columns in the first matrix must equal the number of rows in the second, but this dimension does not appear in the product. In the case of $AB = C$, written schematically, we have $[m \times n][n \times r] = [m \times r]$. Hence, in the final form in Eqn 3, the summation convention shows that k is summed out, leaving i and j as free indices, so c_{ij} is a matrix element. Furthermore, even if both AB and BA are allowed, the two products are generally not equal, so matrix multiplication is not commutative.

The *identity matrix*, I, is a square matrix (one with the same number of rows and columns) whose diagonal elements are equal to 1 while all other elements are 0:

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}. \tag{4}$$

The identity matrix has the property that for any square matrix A,

$$AI = IA = A. \tag{5}$$

The *transpose* of a matrix A, A^T, is derived by placing the rows of A into the columns of A^T, so for $C = A^T$,

$$c_{ij} = a_{ji}. \tag{6}$$

The transpose has the properties that for matrices A and B,

$$(A+B)^T = A^T + B^T \quad \text{and} \quad (AB)^T = B^T A^T. \tag{7}$$

With these definitions, vector operations can be expressed using matrix algebra, by treating vectors as matrices with one column. For example, premultiplication of a vector by a matrix yields another vector, $\mathbf{y} = A\mathbf{x}$, such that

$$y_i = \sum_j a_{ij} x_j \quad \text{or} \quad y_i = a_{ij} x_j, \tag{8}$$

where the second form uses the summation convention. Each component y_i is the scalar product of the i^{th} row of A with $\mathbf{x}$. Similarly, the scalar product of two vectors is given by the matrix product

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = \sum_i a_i b_i = a_i b_i. \tag{9}$$

Thus the scalar product of two vectors yields a scalar, because a $1 \times m$ matrix times an $m \times 1$ matrix is a 1×1 matrix, or single value. The squared magnitude of a real vector can be written as

$$|\,\mathbf{u}\,|^2 = \mathbf{u} \cdot \mathbf{u} = \mathbf{u}^T \mathbf{u} = \sum_i u_i u_i = u_i u_i. \tag{10}$$

For vectors with complex components, the scalar product (Eqn A.3.23) is

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{*T} \mathbf{b} = \sum_i a_i^* b_i = a_i^* b_i. \tag{11}$$

This brings us to a minor point of notation. In linear algebra, as in the last few equations, it is common to treat vectors as column vectors represented by $n \times 1$ matrices with n rows and one column

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{pmatrix}, \tag{12}$$

whose transposes are row vectors (one row, n columns) like

$$\mathbf{u}^T = (u_1, u_2, \dots u_n). \tag{13}$$

Nonetheless, to save space, we sometimes write

$$\mathbf{u} = (u_1, u_2, \dots u_n), \tag{14}$$

while treating $\mathbf{u}$ as a column vector when required. Strictly speaking, we should call the row vector $\mathbf{u}^T$.

We often encounter matrices that are *symmetric*, or equal their transposes,

$$A = A^T, \quad a_{ij} = a_{ji}. \tag{15}$$

For a matrix A with complex elements, the conjugate matrix A^* is formed by taking the conjugate of each element, and the transpose is generalized to the *adjoint* matrix $A^+ = A^{*T}$, which is the complex conjugate of A^T. Note that if the elements of A are real, $A^+ = A^T$. A matrix A is *Hermitian* if it equals its adjoint

$$A = A^+, \quad a_{ij} = a_{ji}^*. \tag{16}$$

If A is real, "Hermitian" and "symmetric" are equivalent.

A.4.2 *Determinant*

A useful entity is the *determinant* of a matrix, written det A, or $|\,A\,|$. For an $n \times n$ matrix,

$$\det \mathrm{A} = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \dots \sum_{j_n=1}^{n} s(j_1, j_2, \dots j_n) a_{1j_1} a_{2j_2} \dots a_{nj_n}. \tag{17}$$

This complicated sum over n indices, $j_1, j_2, \dots j_n$, uses a generalized form of the permutation symbol

$$s(j_1, j_2, \dots j_n) = \operatorname{sgn} \prod_{1 \le p < q \le n} (j_q - j_p). \tag{18}$$

The sgn function is one times the sign of its argument, so that it equals 1 if its argument is positive, −1 if its argument is negative, and 0 if its argument is zero. For $n = 3$,

$$s(j_1, j_2, j_3) = \operatorname{sgn}\,[(j_2 - j_1)(j_3 - j_1)(j_3 - j_2)], \tag{19}$$

so that, for example,

$$s(1,2,3)=1, \quad s(2,1,3)=-1, \quad s(1,1,3)=0. \tag{20}$$

Because $s(j_1, j_2, j_3)$ suppresses terms with two equal indices, and assigns others a sign depending on the order of the indices, it is the same as the permutation symbol, $\varepsilon_{j_1 j_2 j_3}$ (Eqn A.3.39).

The definition of the determinant gives the familiar result for $n=2$:

$$\begin{aligned} |A| = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} &= \sum_{j_1=1}^{2}\sum_{j_2=1}^{2} s(j_1, j_2) a_{1j_1} a_{2j_2} \\ &= s(1,1)a_{11}a_{21} + s(1,2)a_{11}a_{22} + s(2,1)a_{12}a_{21} + s(2,2)a_{12}a_{22} \\ &= a_{11}a_{22} - a_{12}a_{21}, \end{aligned} \tag{21}$$

because $s(1, 1) = s(2, 2) = 0$, $s(1, 2) = 1$, and $s(2, 1) = -1$. For a matrix with only one element, the determinant equals the matrix element.

Among the properties of determinants that we will find useful in solving systems of equations are:

- The determinant of a matrix equals that of its transpose, $|A| = |A^T|$.
- If two rows or columns of a matrix are interchanged, the determinant has the same absolute value but changes sign.
- If one row (or column) is multiplied by a constant, the determinant is multiplied by that constant.
- If a multiple of one row (or column) is added to another row (or column), the determinant is unchanged.
- If two rows or columns of a matrix are the same, the determinant is zero.

Proving these properties is left for the problems.

A.4.3 *Inverse*

For an $n \times n$ square matrix A, the *inverse* matrix A^{-1} is defined such that multiplication by the inverse gives the identity matrix

$$A^{-1}A = AA^{-1} = I. \tag{22}$$

A^{-1} can be written in terms of the *cofactor matrix*, C, whose elements

$$c_{ij} = (-1)^{i+j} |A_{ij}| \tag{23}$$

are formed from the determinants of A_{ij}, an $(n-1) \times (n-1)$ square matrix formed by deleting the i^{th} row and j^{th} column from A. If $|A|$ is not zero,

$$A^{-1} = C^T / |A|. \tag{24}$$

For the familiar $n=2$ case, see problem 7.

A matrix whose determinant is zero does not have an inverse, and is called *singular*. Because the determinant of a matrix with two equal rows or columns is zero, such a matrix is singular. More generally, a matrix is singular if a row or column is a linear combination of the others.

The inverse of the matrix product AB, if AB is nonsingular, obeys

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{25}$$

A matrix A whose transpose equals its inverse,

$$A^{-1} = A^T, \tag{26}$$

is called *orthogonal*. By extension, a matrix A with complex elements is *unitary* if its adjoint and inverse are equal

$$A^{-1} = A^{+}. \tag{27}$$

A.4.4 *Systems of linear equations*

A vector–matrix representation is often used for systems of linear equations. In this formulation, a system of m equations for n unknown variables x_i,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 \ldots + a_{2n}x_n &= b_2 \\ &\ldots \\ a_{m1}x_1 + a_{m2}x_2 \ldots + a_{mn}x_n &= b_m \end{aligned} \tag{28}$$

is written in the form

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{or} \quad A\mathbf{x} = \mathbf{b}, \tag{29}$$

by defining the matrix of coefficients and column vectors for the unknowns and right-hand side,

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix}. \tag{30}$$

The coefficient matrix A is $m \times n$, because there is one row for each equation, and one column for each unknown.

The $A\mathbf{x} = \mathbf{b}$ form illustrates that whether a system of equations can be solved depends on the matrix A. A system of equations is called *homogeneous* in the special case that $\mathbf{b} = \mathbf{0}$, and *inhomogeneous* for all other cases in which $\mathbf{b} \neq \mathbf{0}$. We consider here only systems where the number of unknowns and equations are equal, so the coefficient matrix A is square. If A possesses an inverse, both sides can be premultiplied by A^{-1}, and

$$A^{-1}A\mathbf{x} = A^{-1}\mathbf{b} = I\mathbf{x} = \mathbf{x} \tag{31}$$

yields a unique solution vector **x**. For inhomogeneous systems, computing A^{-1} provides a straightforward manner of solving for the unknown variables x_i. For homogeneous systems of equations, the equation shows that $\mathbf{x} = 0$ if A^{-1} exists. Thus, for a homogeneous system to have a nonzero or *nontrivial* solution, A must be singular. This occurs if the determinant of A is zero, implying that some of the rows (or columns) of A are not linearly independent. If a nontrivial solution of the homogeneous system exists, any constant times that solution is also a solution.

If the coefficient matrix is singular, the corresponding inhomogeneous system of equations does not have unique solutions, and may have none. The existence of A^{-1} and the solvability of the equations thus depend on whether the rows and columns of A are linearly independent. For example, if the rows are linearly dependent, there are fewer independent equations than unknowns and difficulties result, as discussed in the context of inverse problems (Chapter 7).

A.4.5 *Solving systems of equations on a computer*

Standard methods exist to solve linear equations on a computer. Consider the basic problem

$$A\mathbf{x} = \mathbf{b} \tag{32}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

in which we solve for **x**, given A and **b**. If A were a *triangular* matrix T, with zeroes below the diagonal, it would be easy to solve the system

$$T\mathbf{x} = \mathbf{d} \tag{33}$$

$$\begin{pmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{23} \\ 0 & 0 & t_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$$

by starting with the simplest (bottom) equation, solving for x_3, and solving the other equations in succession to find x_2 and then x_1. In other words, the solution

$$x_3 = d_3/t_{33} \tag{34}$$

can be substituted into the middle equation to find

$$x_2 = (d_2 - t_{23}x_3)/t_{22}. \tag{35}$$

Then, by substituting x_3 and x_2 into the first equation,

$$x_1 = (d_1 - t_{13}x_3 - t_{12}x_2)/t_{11}. \tag{36}$$

The importance of this idea is that an arbitrary matrix can be triangularized. Consider that the solution of the system of equations is not changed by any of the following *elementary row operations*:

(i) Rearranging the equations, which corresponds to interchanging rows in the **b** vector and matrix, i.e.,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_3 \\ b_2 \end{pmatrix}. \tag{37}$$

The solution is unchanged because the order of the equations is arbitrary.

(ii) Multiplying an equation by a constant c, which corresponds to multiplying a row of A and the corresponding element of **b** by a constant, i.e.,

$$\begin{pmatrix} ca_{11} & ca_{12} & ca_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} cb_1 \\ b_2 \\ b_3 \end{pmatrix}. \tag{38}$$

(iii) Adding two equations, which corresponds to adding a multiple of one row to another, i.e.,

$$\begin{pmatrix} ca_{11} + a_{21} & ca_{12} + a_{22} & ca_{13} + a_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} cb_1 + b_2 \\ b_2 \\ b_3 \end{pmatrix}. \tag{39}$$

Thus if the system $A\mathbf{x} = \mathbf{b}$ is transformed into $T\mathbf{x} = \mathbf{d}$ using elementary row operations, the two systems of equations have the same solutions **x**. This provides a fast method of solving the system: combine A and **b** into a single *augmented matrix*

$$(A, \mathbf{b}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{pmatrix} \tag{40}$$

and triangularize the augmented matrix to obtain

$$(T, \mathbf{d}) = \begin{pmatrix} t_{11} & t_{12} & t_{13} & d_1 \\ 0 & t_{22} & t_{23} & d_2 \\ 0 & 0 & t_{33} & d_3 \end{pmatrix}, \tag{41}$$

which represents a set of equations easily solved for **x** by the method in Eqns 34–6.

The matrix is triangularized using the following method column by column:

- Find the element of maximum absolute value in the column on or below the diagonal.
- If this "pivot" element is below the diagonal, interchange rows to get it on the diagonal.

- Subtract multiples of the pivot row from rows below it to get zeroes below the diagonal.

The pivoting, though not absolutely necessary, avoids possible numerical difficulties. Note that once a column is zeroed below the diagonal, we do not have to think about it any more.

For an illustration of this method, called *Gaussian elimination with partial pivoting*, consider solving the system of equations

$$\begin{aligned} x_1+x_2&=5,\\ 4x_1+x_2+x_3&=4,\\ 2x_1+2x_2+2x_3&=3. \end{aligned} \tag{42}$$

This can be expressed in matrix form as

$$\begin{pmatrix} 1 & 1 & 0 \\ 4 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix}, \tag{43}$$

and solved by triangularizing the augmented matrix

$$\begin{pmatrix} 1 & 1 & 0 & 5 \\ 4 & 1 & 1 & 4 \\ 2 & 2 & 2 & 3 \end{pmatrix}. \tag{44}$$

To get zeroes below the diagonal in the first column, we first move 4, the element with the largest absolute value in the first column, to the diagonal by interchanging rows

$$\begin{pmatrix} 4 & 1 & 1 & 4 \\ 1 & 1 & 0 & 5 \\ 2 & 2 & 2 & 3 \end{pmatrix}. \tag{45}$$

We then subtract 1/4 times the first row from second, and 1/2 times the first row from third, leaving

$$\begin{pmatrix} 4 & 1 & 1 & 4 \\ 0 & 0.75 & -0.25 & 4 \\ 0 & 1.5 & 1.5 & 1 \end{pmatrix}. \tag{46}$$

Next, to zero the elements below the diagonal in the second column, we interchange rows to get the pivot for this column, 1.5, on the diagonal:

$$\begin{pmatrix} 4 & 1 & 1 & 4 \\ 0 & 1.5 & 1.5 & 1 \\ 0 & 0.75 & -0.25 & 4 \end{pmatrix} \tag{47}$$

and subtract $0.75/1.5=0.5$ times the second row from the third

$$\begin{pmatrix} 4 & 1 & 1 & 4 \\ 0 & 1.5 & 1.5 & 1 \\ 0 & 0 & -1 & 3.5 \end{pmatrix} \tag{48}$$

to complete the triangularization. We then solve the equations for $\mathbf{x}$, beginning with the bottom one, as in Eqns 34–6.

A similar procedure can be used to invert a matrix. This method uses the idea that two vector–matrix equations

$$A\mathbf{x}=\mathbf{b} \quad \text{and} \quad A\mathbf{y}=\mathbf{c} \tag{49}$$

can be combined into one by forming an augmented matrix from each pair of vectors,

$$X=(\mathbf{x},\mathbf{y}), \quad B=(\mathbf{b},\mathbf{c}), \tag{50}$$

and writing the matrix equation

$$AX=B. \tag{51}$$

Because $\mathbf{x}$, the solution to $A\mathbf{x}=\mathbf{b}$, is not changed by elementary row operations on the augmented matrix $(A, \mathbf{b})$, the corresponding solution to $AX = B$ is unaffected by elementary row operations on the augmented matrix (A, B).

To apply this to matrix inversion, consider a special case

$$AX=I, \tag{52}$$

whose solution $X = A^{-1}$ is the inverse of the $n \times n$ matrix A. X is unaffected by elementary row operations that convert the augmented matrix

$$(A,I)=\begin{pmatrix} a_{11} & \cdot & \cdot & a_{1n} & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \cdot & \cdot & a_{nn} & 0 & \cdot & \cdot & 1 \end{pmatrix} \tag{53}$$

to one whose left side is the identity

$$(I,B)=\begin{pmatrix} 1 & \cdot & \cdot & 0 & b_{11} & \cdot & \cdot & b_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 1 & b_{n1} & \cdot & \cdot & b_{nn} \end{pmatrix}, \tag{54}$$

so the corresponding equation

$$IX=B \tag{55}$$

shows that the right side of the matrix gives $B=X=A^{-1}$, the inverse of A. The sequence of operations used to diagonalize the left (A) side of the augmented matrix (A, I) are similar to those that triangularize a matrix.

A.5 Vector transformations

In seismology, we often apply two types of transformations to vectors. In the first, the same vector is expressed in two

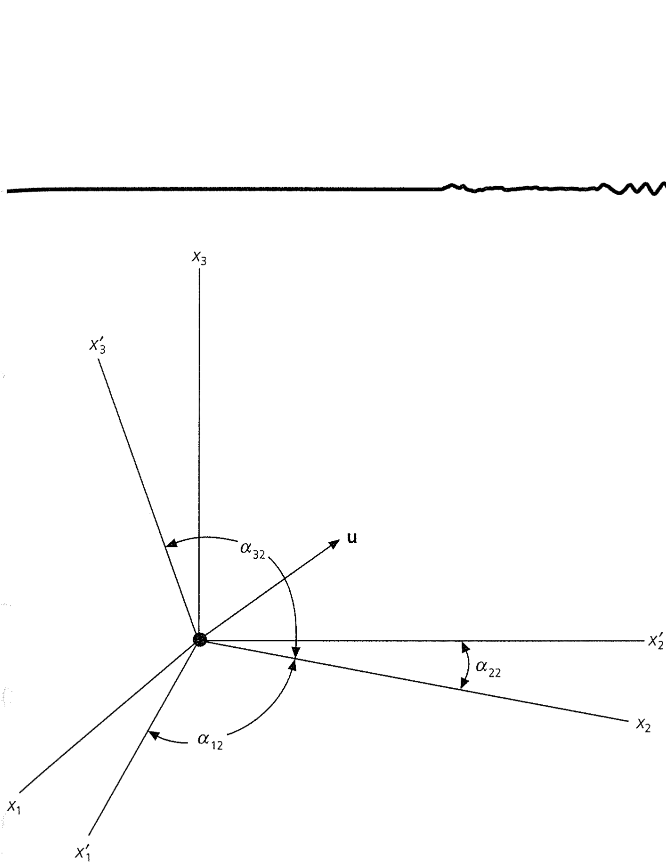


Fig. A.5-1 The relation between two orthogonal coordinate systems with the same origin is described by the angles α_{ij} between the two sets of axes.

different coordinate systems. In the second, some operation converts a vector to another vector expressed in the same coordinate system. In this section we summarize these transformations and their differences.

A.5.1 Coordinate transformations

We have seen that vectors remain the same regardless of the coordinate system in which they are defined, although their components differ between coordinate systems. Thus vectors can be defined in one coordinate system (for example, one oriented along an earthquake fault plane) and reexpressed in another (such as a geographic coordinate system). This property is very useful for solving problems and gives valuable insight into the nature of vectors.

To define the relation between vector components and coordinate systems, consider two orthogonal Cartesian coordinate systems (Fig. A.5-1). Because the origins are the same, one coordinate system can be obtained by rotating the other through three angles. The relation between the two sets of unit basis vectors, $\hat{e}_1, \hat{e}_2, \hat{e}_3$ and $\hat{e}_1', \hat{e}_2', \hat{e}_3'$, is given by their scalar products, called *direction cosines*,

$$\hat{e}_i' \cdot \hat{e}_j = \cos \alpha_{ij} = a_{ij}, \tag{1}$$

where the angles α_{ij} are the angles between the two sets of axes.

A vector can be expressed in terms of its components in the two coordinate systems

$$\mathbf{u} = u_1\hat{e}_1 + u_2\hat{e}_2 + u_3\hat{e}_3 = u_1'\hat{e}_1' + u_2'\hat{e}_2' + u_3'\hat{e}_3'. \tag{2}$$

Given the components u_i in the unprimed system, the components u_i' in the primed system are found by taking the scalar products of the vector with the basis vectors of the primed system:

$$\begin{aligned} u_1' &= \hat{e}_1' \cdot \mathbf{u} = (\hat{e}_1' \cdot \hat{e}_1)u_1 + (\hat{e}_1' \cdot \hat{e}_2)u_2 + (\hat{e}_1' \cdot \hat{e}_3)u_3 \\ &= a_{11}u_1 + a_{12}u_2 + a_{13}u_3, \\ u_2' &= \hat{e}_2' \cdot \mathbf{u} = a_{21}u_1 + a_{22}u_2 + a_{23}u_3, \\ u_3' &= \hat{e}_3' \cdot \mathbf{u} = a_{31}u_1 + a_{32}u_2 + a_{33}u_3. \end{aligned} \tag{3}$$

These can be written as a matrix equation

$$\mathbf{u}' = A\mathbf{u}, \quad \text{or} \quad \begin{pmatrix} u_1' \\ u_2' \\ u_3' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \tag{4}$$

where A is the matrix that transforms a vector from the unprimed to the primed system. Note that this is not a relation between two different vectors **u** and **u′** — it is a relationship between the *components* of the *same* vector in two coordinate systems. It turns out that the matrix A uniquely describes the transformation between these coordinate systems.

For example, a unit basis vector for the unprimed system

$$\hat{e}_1 = 1\hat{e}_1 + 0\hat{e}_2 + 0\hat{e}_3 = (1, 0, 0) \tag{5}$$

has components in the primed system given by

$$\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{6}$$

and so is written

$$a_{11}\hat{e}_1' + a_{21}\hat{e}_2' + a_{31}\hat{e}_3' = (a_{11}, a_{21}, a_{31}) \tag{7}$$

in the primed system. The last expression is just the first column of A. Similarly, the components of $\hat{e}_2$ and $\hat{e}_3$ in the primed system are the second and third columns of A, respectively. Thus the columns of the transformation matrix A are the basis vectors of the unprimed system written in terms of their components in the primed system.

For example, consider rotating a Cartesian coordinate system by θ counterclockwise about the $\hat{e}_3$ axis, so that the only rotation occurs in the $\hat{e}_1$–$\hat{e}_2$ plane. The $\hat{e}_3$ axis is also the $\hat{e}_3'$ axis (Fig. A.5-2). The elements of the transformation matrix are found by evaluating the scalar products of the basis vectors $a_{ij} = \hat{e}_i' \cdot \hat{e}_j$, so

$$\begin{aligned} &a_{11} = \hat{e}_1' \cdot \hat{e}_1 = \cos\theta, && a_{12} = \hat{e}_1' \cdot \hat{e}_2 = \cos(90° - \theta) = \sin\theta, \\ &a_{22} = \hat{e}_2' \cdot \hat{e}_2 = \cos\theta, && a_{21} = \hat{e}_2' \cdot \hat{e}_1 = \cos(90° + \theta) = -\sin\theta, \\ &a_{33} = \hat{e}_3' \cdot \hat{e}_3 = 1, && a_{13} = a_{23} = a_{31} = a_{32} = 0, \end{aligned} \tag{8}$$

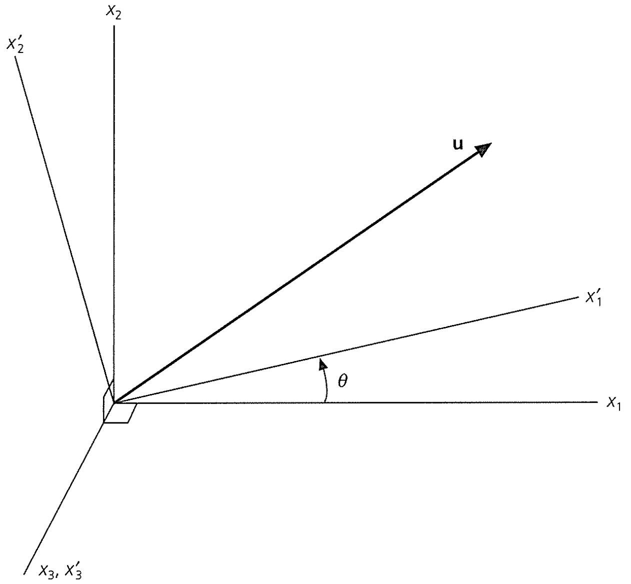


Fig. A.5-2 The relation between the axes of two orthogonal coordinate systems differing by a rotation θ in the x_1–x_2 plane.

and the components of a vector in the two systems are related by

$$\begin{pmatrix} u_1' \\ u_2' \\ u_3' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}. \tag{9}$$

Thus the $\hat{e}_1$ and $\hat{e}_1'$, and the $\hat{e}_2$ and $\hat{e}_2'$ components differ, whereas the $\hat{e}_3$ and $\hat{e}_3'$ components are the same. To check this, consider the case where $\theta = 90°$. As expected, (1, 0, 0) in the unprimed system becomes (0, –1, 0) in the primed system, and (0, 1, 0) in the unprimed system becomes (1, 0, 0) in the primed system, while (0, 0, 1) in the unprimed system remains (0, 0, 1) in the primed system.

Seismologists often use such a geometry. Because the ground motion is a vector, seismometers are generally oriented to record its components in the east–west, north–south, and up–down directions. This decomposition is less useful than decomposing ground motion into its *radial* and *transverse* components, those along and perpendicular to the great circle connecting the earthquake and seismometer. The vertical component is useful as is, so a rotation about the vertical by the angle between East and the great circle connecting the earthquake and seismometer converts the E–W and N–S components into the new representation. The relevant angle, the *back azimuth* to the source from the receiver, is discussed in Section A.7.2.

We can also reverse the transformation. By analogy to Eqn 3, the components in the unprimed system can be found from those in the primed system as

$$u_1 = \hat{e}_1 \cdot \mathbf{u}' = (\hat{e}_1 \cdot \hat{e}_1')u_1' + (\hat{e}_1 \cdot \hat{e}_2')u_2' + (\hat{e}_1 \cdot \hat{e}_3')u_3'$$
$$= a_{11}u_1' + a_{21}u_2' + a_{31}u_3',$$

$$u_2 = \hat{e}_2 \cdot \mathbf{u}' = a_{12}u_1' + a_{22}u_2' + a_{32}u_3',$$
$$u_3 = \hat{e}_3 \cdot \mathbf{u}' = a_{13}u_1' + a_{23}u_2' + a_{33}u_3'. \tag{10}$$

Combining these to express the reverse transformation in vector–matrix form,

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \begin{pmatrix} u_1' \\ u_2' \\ u_3' \end{pmatrix}, \tag{11}$$

shows that the reverse transformation matrix is just the transpose of the transformation matrix A

$$\mathbf{u} = A^T\mathbf{u}'. \tag{12}$$

Hence a unit basis vector in the primed system

$$\hat{e}_1' = 1\hat{e}_1' + 0\hat{e}_2' + 0\hat{e}_3' \tag{13}$$

becomes, by the matrix transformation,

$$a_{11}\hat{e}_1 + a_{12}\hat{e}_2 + a_{13}\hat{e}_3 \tag{14}$$

in the unprimed system. This is the first row of A, so the rows of A are the primed basis vectors expressed in the unprimed coordinates. This is natural because the transformations are related by the matrix transpose.

Alternatively, the reverse transformation can be found directly by starting with $\mathbf{u}' = A\mathbf{u}$ and multiplying both sides by the inverse matrix

$$A^{-1}\mathbf{u}' = A^{-1}A\mathbf{u} = \mathbf{I}\mathbf{u} = \mathbf{u}. \tag{15}$$

Comparison with Eqn 12 shows that the inverse of the transformation matrix equals its transpose, so the transformation matrix is an orthogonal matrix. This seems reasonable because the columns of A, which represent orthogonal basis vectors, are orthogonal. Similarly, the rows of A are orthogonal. As a result, such coordinate transformations are called *orthogonal transformations*. An important feature of orthogonal transformations, whose proof is left as a homework problem, is that they preserve the length of vectors.

The transformation relations, Eqns 4 and 12, provide a mathematical definition of a vector. Any vector must transform between coordinate systems in this way. A set of three entities defined at points in space (for example, temperature, pressure, and density) that does not obey the transformation equations is not a vector.

A.5.2 Eigenvalues and eigenvectors

The product of an arbitrary $n \times n$ matrix A and an arbitrary n-component vector $\mathbf{x}$

$$\mathbf{y} = A\mathbf{x} \tag{16}$$

is also a vector in n dimensions. This is not the same as coordinate transformation; the vector $\mathbf{x}$ is transformed into another distinct vector, with both vectors expressed in the same coordinate system.

A physically important class of transformations convert a vector into one parallel to the original vector, so that

$$A\mathbf{x} = \lambda\mathbf{x}, \tag{17}$$

where A is a matrix, and λ is a scalar. The only effect of the transformation is that the length of $\mathbf{x}$ changes by a factor of λ. For a given A, it is useful to know which vectors $\mathbf{x}$ and scalars λ satisfy this equation.

In three dimensions, the case most commonly encountered, Eqn 17 can be written

$$(A - \lambda I)\mathbf{x} = 0$$

$$\begin{pmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{18}$$

This is a homogeneous system of linear equations, so nontrivial solutions exist only if the matrix $(A - \lambda I)$ is singular. We thus seek values of λ such that the determinant

$$|(A - \lambda I)| = \det \begin{pmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{pmatrix} = 0. \tag{19}$$

Evaluating the determinant gives the *characteristic polynomial*

$$\lambda^3 - I_1\lambda^2 + I_2\lambda - I_3 = 0, \tag{20}$$

which depends on three constants called the *invariants* of A:

$$I_1 = a_{11} + a_{22} + a_{33},$$

$$I_2 = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} + \det\begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix},$$

$$I_3 = \det A. \tag{21}$$

I_1, the first invariant, or *trace*, of A, is the sum of the diagonal elements of A. The invariants of a matrix have significance for stresses, strains, and earthquake moment tensors, because they are not changed by orthogonal transformations.

The characteristic polynomial is a cubic equation in λ with three roots, or *eigenvalues*, λ_m for which the determinant $|A - \lambda I|$ is zero. For each eigenvalue there is an associated nontrivial *eigenvector*, $\mathbf{x}^{(m)}$, satisfying

$$A\mathbf{x}^{(m)} = \lambda_m\mathbf{x}^{(m)}. \tag{22}$$

The components of the eigenvector, $x_1^{(m)}$, $x_2^{(m)}$, $x_3^{(m)}$, are found by solving

$$\begin{pmatrix} a_{11} - \lambda_m & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda_m & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda_m \end{pmatrix} \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ x_3^{(m)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{23}$$

Each eigenvalue and its associated eigenvector form a pair satisfying Eqn 22. In general, an eigenvalue and the eigenvector associated with a different eigenvalue will not satisfy the equation.

For example, the eigenvalues of

$$A = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix} \tag{24}$$

are found by solving the characteristic polynomial

$$\lambda^3 - 8\lambda^2 + 19\lambda - 12 = 0, \tag{25}$$

whose roots are $\lambda_1 = 4$, $\lambda_2 = 3$, $\lambda_3 = 1$. Next, the equations

$$\begin{pmatrix} 3 - \lambda_m & -1 & 0 \\ -1 & 2 - \lambda_m & -1 \\ 0 & -1 & 3 - \lambda_m \end{pmatrix} \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ x_3^{(m)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{26}$$

are solved for each eigenvalue to yield the associated eigenvector. Thus for $\lambda_3 = 1$,

$$\begin{aligned} 2x_1^{(3)} - x_2^{(3)} &= 0, \\ -x_1^{(3)} + x_2^{(3)} - x_3^{(3)} &= 0, \\ -x_2^{(3)} + 2x_3^{(3)} &= 0. \end{aligned} \tag{27}$$

All three unknowns cannot be found uniquely, because these are homogeneous equations. We thus set $x_1^{(3)}$ equal to 1 and find the other two unknowns, $x_2^{(3)} = 2$, $x_3^{(3)} = 1$. Similarly, the other eigenvectors are found by substituting λ_2 and λ_1 in Eqn 26, so

$$\mathbf{x}^{(3)} = (1, 2, 1), \quad \mathbf{x}^{(2)} = (1, 0, -1), \quad \mathbf{x}^{(1)} = (1, -1, 1). \tag{28}$$

Because the eigenvectors are solutions to a set of homogeneous equations, any multiple of an eigenvector is also an eigenvector. The eigenvectors thus determine a direction in space, but the magnitude of the vector is arbitrary. Often the eigenvectors are normalized to unit magnitude. The set we have found can be written as

$$\mathbf{x}^{(1)} = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3}), \quad \mathbf{x}^{(2)} = (1/\sqrt{2}, 0, -1/\sqrt{2}),$$

$$\mathbf{x}^{(3)} = (1/\sqrt{6}, 2/\sqrt{6}, 1/\sqrt{6}). \tag{29}$$

Sometimes complications arise, as for the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{30}$$

with eigenvalues 1, 1, and 0. Using the method given above to find the eigenvector for $\lambda_3 = 0$ by setting $x_1^{(3)} = 1$ yields no solution. Setting $x_2^{(3)} = 1$, however, yields a correct solution for the eigenvector, (0, 1, 0). Because this has no $\hat{e}_1$ component, we could not have set $x_1^{(3)} = 1$ and found the other components.

This example illustrates a complication that arises for a *degenerate*, or repeated, eigenvalue: e.g., $\lambda_1 = \lambda_2 = 1$. In this case, the eigenvalue corresponds not to an eigenvector but to an entire plane, and any vector contained within it is an eigenvector. Two eigenvectors spanning this plane can be found by finding the eigenvector of the nondegenerate eigenvalue, and then choosing two independent vectors orthogonal to it. Because the eigenvector for the nondegenerate eigenvalue is (0, 1, 0), two possible orthogonal eigenvectors for the degenerate eigenvalue are (1, 0, 0) and (0, 0, 1).

A.5.3 Symmetric matrix eigenvalues, eigenvectors, diagonalization, and decomposition

The eigenvalues and eigenvectors of a symmetric matrix have interesting properties. An $n \times n$ matrix H has a characteristic polynomial of degree n, each of whose n roots is an eigenvalue. Consider two eigenvalues and their associated eigenvectors

$$H\mathbf{x}^{(i)} = \lambda_i \mathbf{x}^{(i)}, \quad H\mathbf{x}^{(j)} = \lambda_j \mathbf{x}^{(j)}. \tag{31}$$

Multiplication of the first equation by $\mathbf{x}^{(j)T}$ (the transpose of $\mathbf{x}^{(j)}$) and the second equation by $\mathbf{x}^{(i)T}$ yields

$$\mathbf{x}^{(j)T} H\mathbf{x}^{(i)} = \lambda_i \mathbf{x}^{(j)T}\mathbf{x}^{(i)}, \quad \mathbf{x}^{(i)T} H\mathbf{x}^{(j)} = \lambda_j \mathbf{x}^{(i)T}\mathbf{x}^{(j)}. \tag{32}$$

Transposing both sides of the second part of Eqn 32 and subtracting it from the first gives

$$\mathbf{x}^{(j)T} H\mathbf{x}^{(i)} - \mathbf{x}^{(j)T} H^T\mathbf{x}^{(i)} = (\lambda_i - \lambda_j)\mathbf{x}^{(j)T}\mathbf{x}^{(i)}. \tag{33}$$

Because H is symmetric, it equals its transpose, $H = H^T$, so the left-hand side is zero

$$0 = (\lambda_i - \lambda_j)\mathbf{x}^{(j)T}\mathbf{x}^{(i)}. \tag{34}$$

Thus, if $i \neq j$ and the two eigenvalues are different, their associated eigenvectors must be orthogonal so that their scalar product $\mathbf{x}^{(j)T}\mathbf{x}^{(i)}$ is zero. Thus, for a symmetric matrix, eigenvectors associated with distinct eigenvalues are orthogonal.

This result lets us diagonalize a symmetric matrix. To illustrate this for a 3×3 case, consider a matrix U whose columns are the eigenvectors of the symmetric matrix H

$$U = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & x_1^{(3)} \\ x_2^{(1)} & x_2^{(2)} & x_2^{(3)} \\ x_3^{(1)} & x_3^{(2)} & x_3^{(3)} \end{pmatrix}. \tag{35}$$

If the eigenvalues of H are distinct, the eigenvectors of H, and hence the columns of the eigenvector matrix, are orthogonal, so U is an orthogonal matrix satisfying $U^{-1} = U^T$.

The entire set of eigenvalue–eigenvector pairs, each of which satisfy $H\mathbf{x}^{(i)} = \lambda_i \mathbf{x}^{(i)}$, can be written as the matrix equation

$$HU = U\Lambda, \tag{36}$$

where Λ is the diagonal matrix with eigenvalues on the diagonal

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}. \tag{37}$$

Premultiplying both sides of Eqn 36 by the inverse of the eigenvector matrix yields

$$U^{-1}HU = U^T HU = \Lambda, \tag{38}$$

which shows how the eigenvector matrix can be used to diagonalize a symmetric matrix. This result can also be stated as

$$H = U\Lambda U^T, \tag{39}$$

which illustrates how a symmetric matrix can be decomposed into a diagonal eigenvalue matrix and the orthogonal eigenvector matrix. Similar results apply for complex Hermitian matrices.

We will see that if a matrix contains the components of vectors expressed in a coordinate system, the physical problem under discussion can be simplified by diagonalizing the matrix. This corresponds to rewriting the problem in its "natural" coordinate system, whose basis set is the eigenvectors, an idea used in discussing stresses in the earth (Section 2.3.4) and the seismic moment tensor (Section 4.4.5).

A.6 Vector calculus

A.6.1 Scalar and vector fields

Many phenomena in seismology depend on how physical quantities vary in space. Some, like density or temperature, are *scalar fields*, scalar valued functions of the position vector $\mathbf{x}$ denoted by expressions like $\phi(\mathbf{x})$ or $\phi(x_1, x_2, x_3)$. Similarly, a vector that varies in space is described by a *vector field*. For example, seismic waves are described by the variation in the displacement vector

$$\mathbf{u}(\mathbf{x}) = \mathbf{u}(x_1, x_2, x_3)$$
$$= u_1(x_1, x_2, x_3)\hat{e}_1 + u_2(x_1, x_2, x_3)\hat{e}_2 + u_3(x_1, x_2, x_3)\hat{e}_3 \quad (1)$$

as a function of position, and result in turn from forces derived from spatial derivatives of the stress tensor.

Spatial variations of scalar, vector, or tensor fields are described using the vector differential operator "del", $\mathbf{\nabla}$,

$$\mathbf{\nabla} = \left(\hat{e}_1 \frac{\partial}{\partial x_1}, \hat{e}_2 \frac{\partial}{\partial x_2}, \hat{e}_3 \frac{\partial}{\partial x_3}\right). \quad (2)$$

This operator has the form of a vector, but has meaning only when applied to a scalar, vector, or tensor field. We first review uses of the $\mathbf{\nabla}$ operator in Cartesian coordinates, and in the next section discuss the more complicated forms for spherical coordinates.

A.6.2 Gradient

The simplest application of the $\mathbf{\nabla}$ operator is the *gradient*, a vector field formed from the spatial derivatives of a scalar field. If $\phi(\mathbf{x})$ is a scalar function of position, the gradient is defined by

$$\text{grad } \phi(\mathbf{x}) = \mathbf{\nabla}\phi(\mathbf{x}) = \frac{\partial\phi(\mathbf{x})}{\partial x_1}\hat{e}_1 + \frac{\partial\phi(\mathbf{x})}{\partial x_2}\hat{e}_2 + \frac{\partial\phi(\mathbf{x})}{\partial x_3}\hat{e}_3, \quad (3)$$

where $\partial\phi(\mathbf{x})/\partial x_1$ is the partial derivative of $\phi(x_1, x_2, x_3)$ with respect to x_1, for x_2 and x_3 held constant. The gradient is a vector field whose components equal the partial derivative with respect to the corresponding coordinate.

Expressions like Eqns 1 and 3 can be written more compactly if the dependences on position are not written explicitly, i.e.,

$$\mathbf{\nabla}\phi = \frac{\partial\phi}{\partial x_1}\hat{e}_1 + \frac{\partial\phi}{\partial x_2}\hat{e}_2 + \frac{\partial\phi}{\partial x_3}\hat{e}_3. \quad (4)$$

In this notation, it is implicit that ϕ, its derivatives, and hence the gradient, vary with position.

For example, the elevation $\phi(x_1, x_2)$ is a scalar field describing the topography as a function of position in a two-dimensional region. This is often plotted using topographic contours (Fig. A.6-1), curves along which ϕ is constant. At any point, $\partial\phi/\partial x_1$ is the slope in the x_1 direction, and $\partial\phi/\partial x_2$ is the slope in the x_2 direction.

The gradient can be used to find the slope in any direction. The projection of a vector in a given direction is the scalar product of the vector and the unit normal vector in that direction, $\hat{\mathbf{n}} = (n_1, n_2)$. Thus the scalar product of the gradient with the normal vector,

$$\hat{\mathbf{n}} \cdot \mathbf{\nabla}\phi = n_1\frac{\partial\phi}{\partial x_1} + n_2\frac{\partial\phi}{\partial x_2}, \quad (5)$$

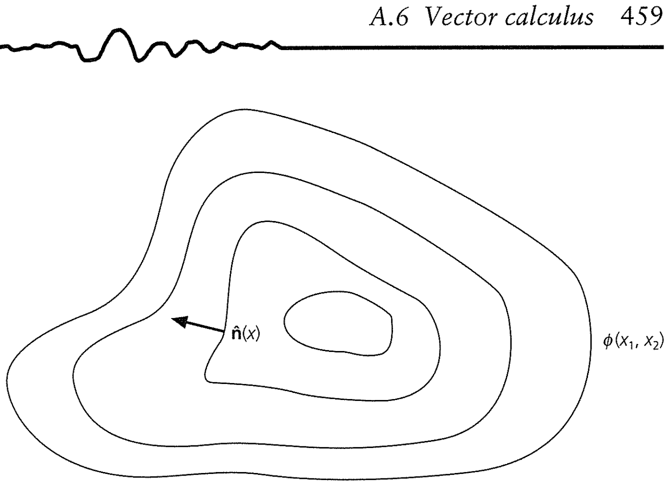


Fig. A.6-1 A scalar field demonstrating the concept of a gradient. If $\phi(x_1, x_2)$ gives the elevation, the gradient can be used to find the slope in the $\hat{\mathbf{n}}$ direction at a point (x_1, x_2).

gives the *directional derivative* in the $\hat{\mathbf{n}}$ direction. Because both $\hat{\mathbf{n}}$ and $\mathbf{\nabla}\phi$ are functions of position, the directional derivative varies in space. At any point, the maximum value of the scalar product occurs for $\hat{\mathbf{n}}$ parallel to the gradient, so the gradient points in the direction of the steepest slope along which ϕ changes most rapidly. The scalar product is zero when $\hat{\mathbf{n}}$ is perpendicular to the gradient, so the gradient is perpendicular to curves of constant ϕ. These concepts are also used in three dimensions.

In index notation, the gradient is written as

$$(\mathbf{\nabla}\phi)_i = \frac{\partial\phi}{\partial x_i} = \phi_{,i}, \quad (6)$$

where the last form uses a common (if sometimes confusing) notation in which differentiation is indicated by a comma. The notation, with one free index, shows that the gradient is a vector. By contrast, the directional derivative, written as

$$\hat{\mathbf{n}} \cdot \mathbf{\nabla}\phi = n_i\frac{\partial\phi}{\partial x_i} = n_i\phi_{,i}, \quad (7)$$

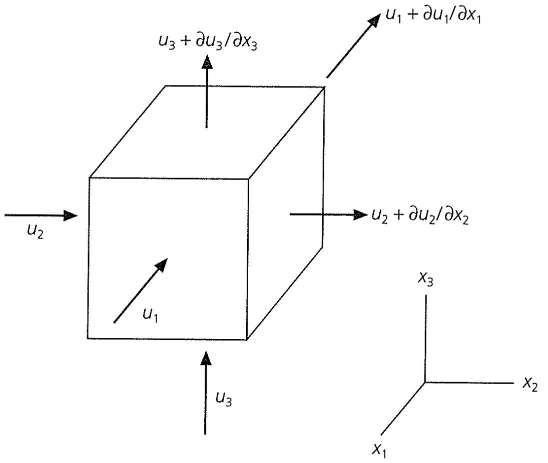
has an implied sum over i and is a scalar.

Often, the gradients of quantities are important physically because an effect depends on spatial variations of a field. For example, the flow of heat depends on the gradient of the temperature field (Sections 5.3.2, 5.4.1), and the gradient of the pressure field in the atmosphere is important for the weather.

A.6.3 Divergence

A related operation that describes the spatial variation of a vector field is the *divergence*. The divergence of a vector field $\mathbf{u}(\mathbf{x})$ is given by the scalar product of the $\mathbf{\nabla}$ operator with $\mathbf{u}(\mathbf{x})$ as

$$\text{div}\,\mathbf{u} = \mathbf{\nabla} \cdot \mathbf{u} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}, \quad (8)$$



$u_1 + \partial u_1/\partial x_1$
$u_3 + \partial u_3/\partial x_3$
u_2
$u_2 + \partial u_2/\partial x_2$
u_1
x_3
x_2
u_3
x_1

Fig. A.6-2 The divergence, formed from the differences between the flow into one face of a volume and the flow out of the opposite face, gives the net flow through a unit volume.

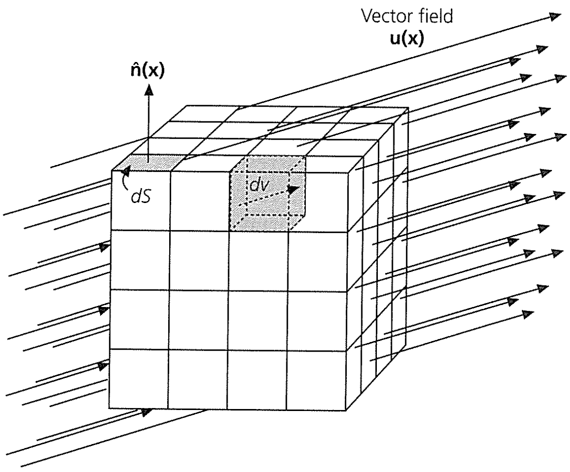


Fig. A.6-3 Geometry for the divergence theorem: $\hat{\mathbf{n}}(\mathbf{x})$ is a unit vector pointing outward at the point $\mathbf{x}$ from an element dS of the surface S that encloses a volume dV.

which yields a scalar field because the vector components and their derivatives are functions of position.

The divergence frequently arises in conservation equations. For example, if $\mathbf{u}(\mathbf{x})$ is the velocity as a function of position in a fluid, $\nabla \cdot \mathbf{u}(\mathbf{x})$ gives the net outflow of material per unit time from a unit volume at position $\mathbf{x}$ (Fig. A.6-2). To see this, note that, to first order, the net flow in the x_2 direction is the difference between the flow out the far side, $u_2 + \partial u_2/\partial x_2$, and that into the near side, u_2, given as

$$u_2 + \frac{\partial u_2}{\partial x_2} - u_2 = \frac{\partial u_2}{\partial x_2}. \tag{9}$$

Adding similar terms for the net flow in the x_1 and x_3 directions gives the divergence (Eqn 8). If the divergence is positive, there is a net outward flow, whereas a negative divergence indicates a net inflow.

This idea can be applied to any vector field $\mathbf{u}(\mathbf{x})$. Consider the problem of finding the net outflow from a region with volume V and surface S. If $\hat{\mathbf{n}}(\mathbf{x})$ is the unit normal vector pointing outward at a point $\mathbf{x}$ on the surface (Fig. A.6-3), the scalar product $\hat{\mathbf{n}}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x})$ gives the outward *flux* per unit area at that point. Integrating the flux over the surface then gives the total flux. Another way to compute the total flux is to integrate the divergence over the volume. These two methods give the same flux, so

$$\int_S \hat{\mathbf{n}} \cdot \mathbf{u} dS = \int_V \nabla \cdot \mathbf{u} dV. \tag{10}$$

This relation, *Gauss's theorem*, or the *divergence theorem*, says that what accumulates inside a volume is determined by the integral over its surface of what goes out. If we think of the volume as many adjacent cells, the flow out of one cell is the flow into an adjacent cell, which cancels to zero. Only flow in or out of the volume's surface is not canceled out in this way. Written in full, $\int dV$ is a triple integral over the volume, and $\int dS$ is a double integral over the surface.

In index notation, using the summation convention, the divergence is written

$$\nabla \cdot \mathbf{u} = \frac{\partial u_i}{\partial x_i} = u_{i,i}, \tag{11}$$

which is a scalar because no free index remains. Gauss's theorem is written

$$\int_S u_i n_i dS = \int_V \frac{\partial u_i}{\partial x_i} dV, \tag{12}$$

or, using the comma notation for derivatives,

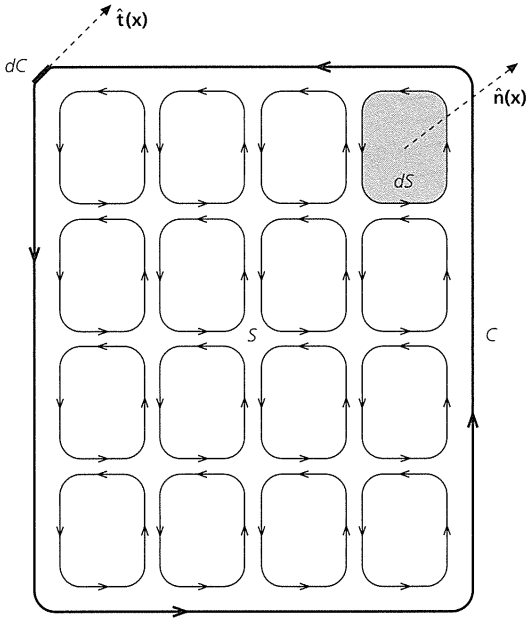
$$\int_S u_i n_i dS = \int_V u_{i,i} dV. \tag{13}$$

As before, it is implicit in the notation that the field $\mathbf{u}$, its derivatives, and the normal vector $\hat{\mathbf{n}}$ vary with position.

A.6.4 Curl

The *curl* operator, the cross product of the ∇ operator with a vector field, yields another vector field

$$\nabla \times \mathbf{u} = \hat{e}_1\left(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}\right) + \hat{e}_2\left(\frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}\right) + \hat{e}_3\left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right). \tag{14}$$



$\hat{\mathbf{t}}(\mathbf{x})$
dC
$\hat{\mathbf{n}}(\mathbf{x})$
dS
S
C

Fig. A.6-4 Geometry for Stokes' theorem: $\hat{\mathbf{n}}(\mathbf{x})$ is a unit vector pointing outward at the point $\mathbf{x}$ from an element dS of the surface S. dC is an element of the curve C bounding S, with tangent $\hat{\mathbf{t}}(\mathbf{x})$.

This can be written as a determinant

$$\nabla \times \mathbf{u} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \dfrac{\partial}{\partial x_1} & \dfrac{\partial}{\partial x_2} & \dfrac{\partial}{\partial x_3} \\ u_1 & u_2 & u_3 \end{pmatrix}, \tag{15}$$

or, using index notation, in a compact form as

$$\nabla \times \mathbf{u} = \varepsilon_{ijk} \frac{\partial u_k}{\partial x_j} = \varepsilon_{ijk} u_{k,j}. \tag{16}$$

Some physical insight into the curl comes from *Stokes' theorem*, which relates the integral of the curl of a vector field over a surface S to the line integral around a curve C bounding S (Fig. A.6-4) as

$$\int_C \mathbf{u} \cdot \hat{\mathbf{t}} dC = \int_S (\nabla \times \mathbf{u}) \cdot \hat{\mathbf{n}} dS. \tag{17}$$

Here dS is an element of surface area with normal $\hat{\mathbf{n}}(\mathbf{x})$, and dC is an element of the curve with tangent $\hat{\mathbf{t}}(\mathbf{x})$. Analogous to the case of Gauss's theorem applied to a volume, we can think of the surface as composed of infinitesimal tiles, each with a line integral of $\mathbf{u} \cdot \hat{\mathbf{t}}$ around it. The border of each tile is shared with another tile, but, because the line integral, or *circulation*, is computed in a counterclockwise manner, the integrals along this border are the same but of opposite sign for the two tiles, and therefore cancel. The segments of the line integrals cancel between all the tiles except those on the outer border that have no adjacent circulation to cancel them.

If the line integral is nonzero, the vector field has a net rotation along the curve, so the integral of its curl over the surface is nonzero. The curl of a vector field shows where rotations arise. A common application is describing the velocity field of a moving fluid. The upper portion of Fig. A.6-5 shows streamlines, lines parallel to the velocity vector at any point, for a viscous fluid flowing past a circular object. The velocity is zero at the object, and increases with distance away from it. The flow is symmetric on the bottom of the object. The lower portion of the figure shows contours of the curl of the velocity field with larger values, indicating greater rotations, close to the object.

Two useful identities, whose proofs are left for the problems, are that the curl of a gradient and the divergence of a curl are zero:

$$\nabla \cdot (\nabla \times \mathbf{u}) = 0 \tag{18}$$

$$\nabla \times (\nabla \phi) = 0. \tag{19}$$

Equation 19 can be used with Stokes' theorem to show that for a vector field written as the gradient of a scalar, the curl, and hence circulation around an arbitrary curve, are zero. This idea is used in mechanics to prove that a conservative force (one that can be written as the gradient of a potential) has a line integral that is independent of path, because its circulation around any path is zero. These relations give insight into seismic waves, because P waves have no curl and S waves have no divergence (Section 2.4.1).

A.6.5 Laplacian

The *Laplacian* operator is formed by taking the divergence of the gradient of a scalar field, which yields a scalar field

$$\nabla^2 \phi = \nabla \cdot \nabla \phi = \frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2} + \frac{\partial^2 \phi}{\partial x_3^2} = \phi_{,ii}, \tag{20}$$

where the last form uses index notation and the summation convention. By analogy, the Laplacian of a vector field is a vector field whose components in Cartesian coordinates are the Laplacians of the original vector components,

$$\nabla^2 \mathbf{u} = (\nabla^2 u_1, \nabla^2 u_2, \nabla^2 u_3). \tag{21}$$

For example, the $\hat{\mathbf{e}}_1$ component of $\nabla^2 \mathbf{u}$ is

$$\frac{\partial^2 u_1}{\partial x_1^2} + \frac{\partial^2 u_1}{\partial x_2^2} + \frac{\partial^2 u_1}{\partial x_3^2}. \tag{22}$$

In Cartesian coordinates, the Laplacian of a vector satisfies

$$\nabla^2 \mathbf{u} = \nabla(\nabla \cdot \mathbf{u}) - \nabla \times (\nabla \times \mathbf{u}), \tag{23}$$

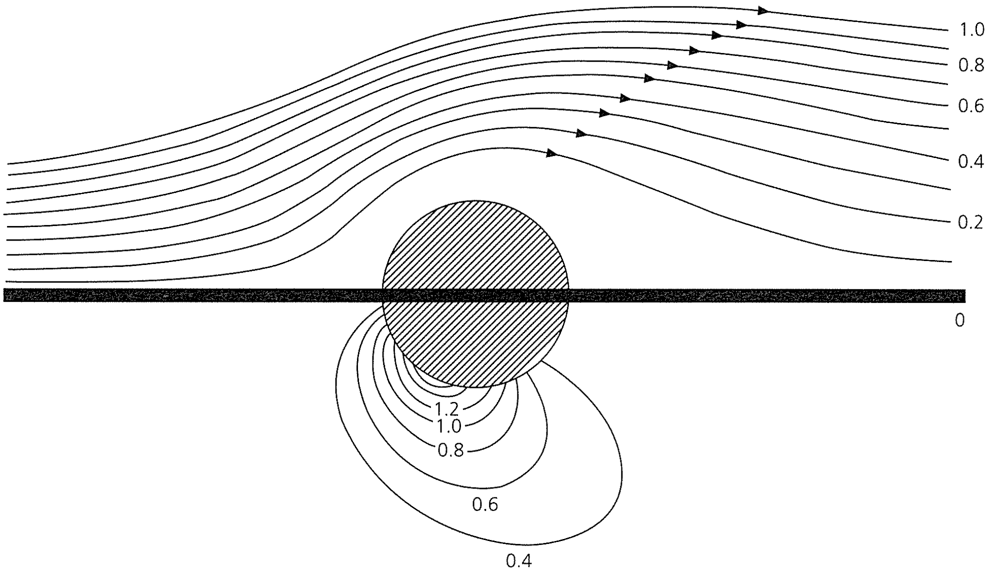


Fig. A.6-5 *Top*: streamlines showing the velocity of fluid flow around an object. Numbers on streamlines show the magnitude of the velocity. *Bottom*: contours of the curl for this velocity field. The curl is greatest near the sphere, where the fluid flow lines are the most curved. (After Batchelor, 1967. Reprinted with the permission of Cambridge University Press.)

an obscure-looking relation that is useful in deriving the existence of *P* and *S* waves.

A.7 Spherical coordinates

The vector operations discussed so far were performed in Cartesian coordinates, in which the unit basis vectors $(\hat{e}_1, \hat{e}_2, \hat{e}_3)$ point in the same direction everywhere. There are, however, situations in which non-Cartesian coordinate systems without these nice properties are useful. In particular, *spherical* coordinates often simplify the solution of problems with a high degree of symmetry about a point.

A.7.1 The spherical coordinate system

In a spherical coordinate system, a point defined by a position vector $\mathbf{x}$ is described by its radial distance from the origin, $r = |\mathbf{x}|$, and two angles. θ is the *colatitude*, or angle between $\mathbf{x}$ and the x_3 axis, and ϕ, the *longitude*, is measured in the x_1–x_2 plane. Often the *latitude*, $90° - \theta$, is used instead of the colatitude. Spherical coordinates are often used in seismology because the earth is approximately spherically symmetric, varying with depth much more than laterally. Thus properties like velocity and density are often approximated as functions only of r, independent of θ and ϕ.

Figure A.7-1 shows the relations between rectangular and spherical coordinates. If the vector $\mathbf{x}$ is written as

$$\mathbf{x} = x_1\hat{e}_1 + x_2\hat{e}_2 + x_3\hat{e}_3, \tag{1}$$

then its components in rectangular coordinates (x_1, x_2, x_3) are described by spherical coordinates as

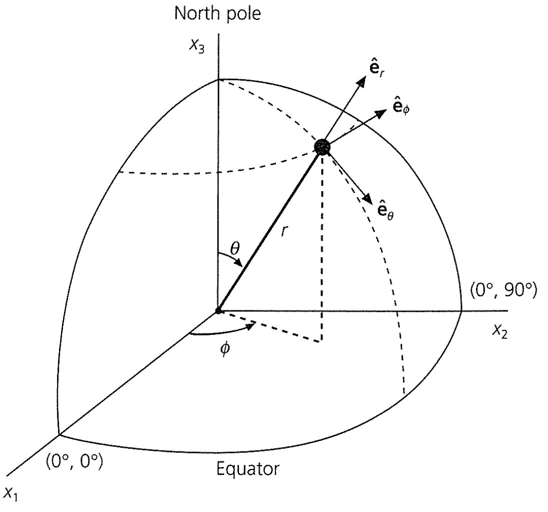


Fig. A.7-1 Relations between spherical (r, θ, ϕ) and Cartesian coordinates (x_1, x_2, x_3). (After Marion, 1970. From *Classical Dynamics of Particles and Systems*, 2nd edn, copyright 1970 by Academic Press, reproduced by permission of the publisher.)

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r\sin\theta\cos\phi \\ r\sin\theta\sin\phi \\ r\cos\theta \end{pmatrix}. \tag{2}$$

Conversely, the spherical coordinates r, θ, and ϕ can be written as

$$r = (x_1^2 + x_2^2 + x_3^2)^{1/2}, \quad \theta = \cos^{-1}(x_3/r), \quad \phi = \tan^{-1}(x_2/x_1). \tag{3}$$

In the equatorial $(x_1$–$x_2)$ plane, $\theta = 90°$, $\cos\theta = 0$, $\sin\theta = 1$, so $x_1 = r\cos\phi$, $x_2 = r\sin\phi$, and $x_3 = 0$. This is the same as the polar

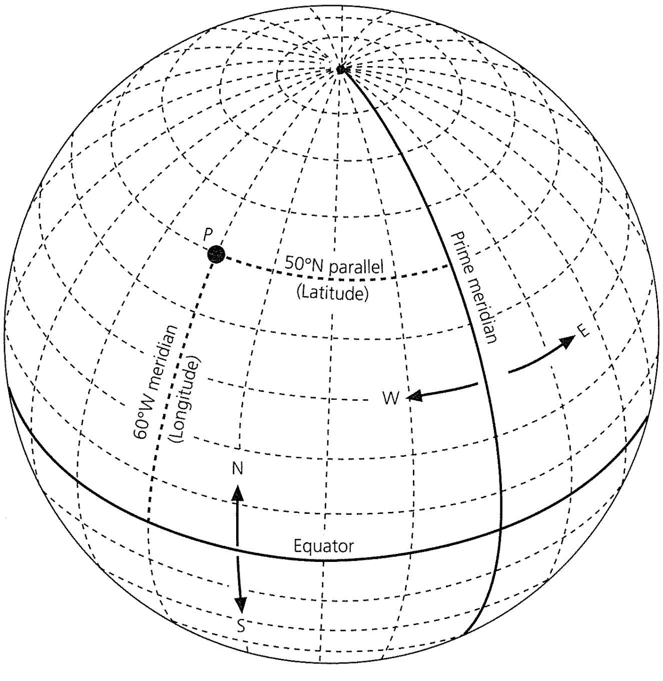


Fig. A.7-2 Geometry of the latitude and longitude system used to locate points on the earth's surface. A point P at 50°N, 60°W ($\theta = 40°$, $\phi = -60°$) is shown. (After Strahler, 1969.)

coordinate system described in Section A.3.1. Along the x_3 axis we have $\theta = 0°$, so $x_1 = x_2 = 0$, and $x_3 = r$. Any of these expressions written in terms of colatitude θ can be converted to latitude $\lambda = 90° - \theta$, using $\cos\theta = \sin\lambda$ and $\sin\theta = \cos\lambda$.

This coordinate system is the familiar one (Fig. A.7-2) used to locate points within the earth or on its surface, $r = a$. For this purpose, the origin is placed at the center of the earth, and the x_3 axis is defined by a line from the center of the earth through the north pole. The intersections of planes containing the x_3 axis with the earth's surface define *meridians*, lines of constant longitude. The x_1 axis intersects the equator at the *prime meridian*, on which ϕ is defined as zero, which has been chosen to run through Greenwich, England. The intersection of planes perpendicular to the x_3 axis with the earth's surface define *parallels*, lines of constant colatitude or latitude. Meridians are a special case of *great circles*, lines on the surface defined by the intersection of a plane through the origin with the surface of the spherical earth. Parallels are a special case of *small circles*, which are lines on the surface defined by the intersection of the surface of the spherical earth with a plane normal to a radius vector.

These conventions allow the colatitude θ ($0° \le \theta < 180°$) and longitude ϕ ($0° \le \phi < 360°$) to define a unique point on the earth's surface. Often locations are described in terms of latitudes north and south of the equator, and longitudes east and west of Greenwich. North and south latitudes correspond, respectively, to colatitudes less than or greater than 90°. Because ϕ measures longitude east of the prime meridian, west longitudes correspond to values of ϕ less than 0° or greater than 180°. Thus a point at (10°S, 110°W) has $\theta = 90° + 10° = 100°$, and $\phi = -110° = 360° - 110° = 250°$.

At any point, unit spherical basis vectors ($\hat{e}_r$, $\hat{e}_\theta$, $\hat{e}_\phi$) can be defined in the direction of increasing r, θ, and ϕ. $\hat{e}_r$ points away from the origin, and gives the upward vertical direction. $\hat{e}_\theta$ points south, and $\hat{e}_\phi$ points east. These two are sometimes written in terms of north- and east-pointing unit vectors, $\hat{e}_{NS} = -\hat{e}_\theta$ and $\hat{e}_{EW} = \hat{e}_\phi$.

An important feature of the unit spherical basis vectors is that at different points they are oriented differently with respect to the Cartesian axes. The Cartesian unit basis vectors ($\hat{e}_1$, $\hat{e}_2$, $\hat{e}_3$) point in the same direction everywhere. By contrast, for example, $\hat{e}_r$ points in the $\hat{e}_3$ direction at the north pole, and in the $-\hat{e}_3$ direction at the south pole. This effect is described by the Cartesian ($\hat{e}_1$, $\hat{e}_2$, $\hat{e}_3$) components of the unit spherical basis vectors, at a point with colatitude θ and longitude ϕ:

$$\hat{e}_\phi = \begin{pmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{pmatrix}, \quad \hat{e}_\theta = \begin{pmatrix} \cos\theta\cos\phi \\ \cos\theta\sin\phi \\ -\sin\theta \end{pmatrix}, \quad \hat{e}_r = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix}. \qquad (4)$$

The dependence on the colatitude and longitude describes how the orientation with respect to the Cartesian axes changes.

At any point, the spherical basis vectors ($\hat{e}_r$, $\hat{e}_\theta$, $\hat{e}_\phi$) form an orthonormal set. For problems whose spatial extent is small enough that the curvature of the earth can be ignored, these basis vectors provide a useful local coordinate system.

A.7.2 *Distance and azimuth*

Spherical coordinates are especially useful in describing the geographic relation between two points on the earth's surface. A common application is to find the distance between points and the direction of the great circle arc joining them. A great circle arc is the shortest path between points on a sphere, so if seismic velocity varies only with depth, the fastest path along the surface is the great circle arc, and the fastest paths through the interior are in the plane of the great circle and the center of the earth. Because velocities vary laterally by only a few percent throughout most of the earth (and imperceptibly in the liquid outer core), this is a good approximation for most seismic applications. The source-to-receiver distance is often given in terms of the angle Δ subtended at the center of the earth by the great circle arc between the two points (Fig. A.7-3). If Δ is expressed in radians, then the length s (in km) of the arc along the earth's surface is $R\Delta$, where R is the earth's radius (≈ 6371 km). If Δ is expressed in degrees, $s = R\Delta\pi/180$, so one degree of arc equals 111.2 km.

Consider the great circle arc connecting an earthquake whose epicenter is at (θ_E, ϕ_E) and a seismic station at (θ_S, ϕ_S). Seismic waves that traveled along the great circle arc (or in the plane of this arc and the center of the earth) left the earthquake in a direction given by the *azimuth* angle ζ measured clockwise from the local direction of north at the epicenter to the great

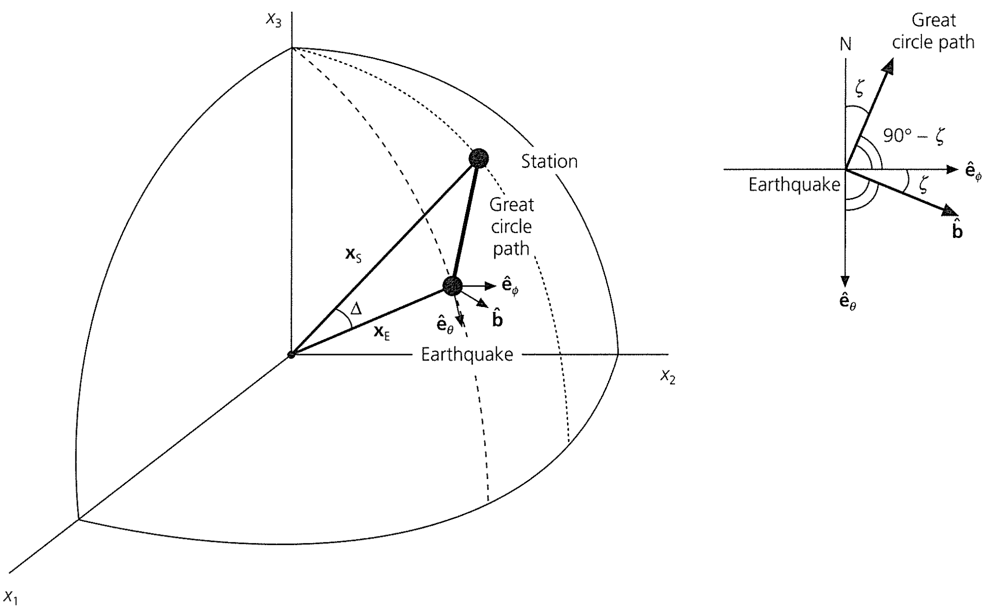


Fig. A.7-3 Geometry of the great circle path between an earthquake epicenter and seismic station (*left*), showing the convention for defining the azimuth, ζ (*right*).

circle arc. These waves arrive at the seismometer from a direction described by the *back azimuth* angle ζ' measured clockwise from the local direction of north at the seismometer to the great circle arc. To find these quantities, the Cartesian components of the position vectors for the earthquake and the station are written, using Eqn 2:

$$\mathbf{x}_E = \begin{pmatrix} R \sin\theta_E \cos\phi_E \\ R \sin\theta_E \sin\phi_E \\ R \cos\phi_E \end{pmatrix} \quad \mathbf{x}_S = \begin{pmatrix} R \sin\theta_S \cos\phi_S \\ R \sin\theta_S \sin\phi_S \\ R \cos\theta_S \end{pmatrix}. \qquad (5)$$

The distance Δ, the angle between $\mathbf{x}_S$ and $\mathbf{x}_E$, is given by the scalar product

$$\mathbf{x}_S \cdot \mathbf{x}_E = R^2 \cos\Delta, \qquad (6)$$

so

$$\Delta = \cos^{-1}\left[\cos\theta_E \cos\theta_S + \sin\theta_E \sin\theta_S \cos(\phi_S - \phi_E)\right]. \qquad (7)$$

This formula defines Δ uniquely between 0 and 180°. This shorter portion of the great circle is called the *minor arc* connecting the two points; the longer portion, known as the *major arc*, is $(360° - \Delta)$ degrees long.

To compute the azimuth from the earthquake to the station, consider $\hat{\mathbf{b}}$, a unit vector normal to the great circle in the local horizontal plane at $\mathbf{x}_E$, which is written using the vector product of the position vectors

$$\mathbf{x}_S \times \mathbf{x}_E = \hat{\mathbf{b}} R^2 \sin\Delta. \qquad (8)$$

Evaluation of the vector product gives

$$\hat{\mathbf{b}} = \frac{1}{\sin\Delta} \begin{pmatrix} \sin\theta_S \cos\theta_E \sin\phi_S - \sin\theta_E \cos\theta_S \sin\phi_E \\ \cos\theta_S \sin\theta_E \cos\phi_E - \cos\theta_E \sin\theta_S \cos\phi_S \\ \sin\theta_S \sin\theta_E \sin(\phi_E - \phi_S) \end{pmatrix}. \qquad (9)$$

The azimuth angle ζ, measured clockwise from north, is then given (Fig. A.7-3) by

$$\cos\zeta = \hat{\mathbf{b}} \cdot \hat{\mathbf{e}}_\phi = \frac{1}{\sin\Delta}\left(\cos\theta_S \sin\theta_E - \sin\theta_S \cos\theta_E \cos(\phi_S - \phi_E)\right) \qquad (10)$$

and

$$\sin\zeta = \hat{\mathbf{b}} \cdot \hat{\mathbf{e}}_\theta = \frac{1}{\sin\Delta} \sin\theta_S \sin(\phi_S - \phi_E). \qquad (11)$$

Use of both $\sin\zeta$ and $\cos\zeta$ makes the angle ζ unambiguous $(0° \le \zeta < 360°)$. The azimuth from an earthquake to a receiver is useful, because earthquakes radiate more energy in some directions than in others (Chapter 4), so measurements at different azimuths yield information about the source.

The back azimuth ζ', obtained by reversing the indices E and S in Eqns 10 and 11, shows the direction from which seismic energy arrives at a seismometer. Seismometers typically record the north–south and east–west components of horizontal ground motion. Using the back azimuth, these observations can be converted into *radial* (along the great circle path) and *transverse* (perpendicular to the great circle path) components by a vector transformation (Eqn A.5.9). This distinction is made because waves appearing on these components propagated differently (Section 2.4). The azimuth and back azimuth

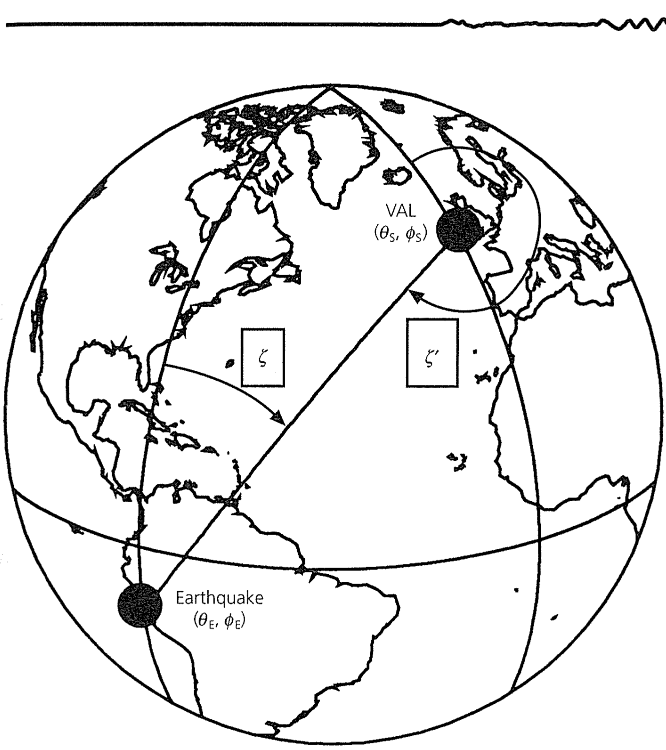


Fig. A.7-4 Geometry of the great circle path for an earthquake in the Peru trench recorded at station VAL (Valentia, Ireland). The azimuth, ζ, and back azimuth, ζ', are not simply related, due to the sphericity of the earth.

angles are measured clockwise from north, a geographic convention which contrasts with the mathematical one of measuring angles counterclockwise from the x_1 direction. Figure A.7-4 illustrates this geometry for an earthquake in the Peru trench ($\theta_E = 102°$, $\phi_E = -78°$) recorded at station VAL (Valentia, Ireland; $\theta_S = 38°$, $\phi_S = -10.25°$). The resulting distances and azimuths are $\Delta = 86°$, $\zeta = 35°$, $\zeta' = 245°$.[1]

This analysis assumes that the earth is perfectly spherical. In fact, the earth is flattened by its rotation into a shape close to an oblate ellipsoid, so the radius varies with colatitude approximately as

$$r(\theta) = R_e(1 - f\cos^2\theta), \tag{12}$$

where R_e is the equatorial radius, 6378 km. The flattening factor f is approximately 3.35×10^{-3}, or about 1/298, so the polar radius R_p is 6357 km. An average radius can be defined as the radius of a sphere with the same volume as the earth, if it were a perfect ellipsoid. Because the volume of an ellipsoidal earth would be $(4/3)\pi R_e^2 R_p$, and a sphere of radius R has volume $(4/3)\pi R^3$, the average radius is 6371 km. For certain applications the ellipticity is included in precise distance calculations.

A.7.3 Choice of axes

Spherical coordinates are also used with axes different from the geographic ones. Because the physics of a problem does not depend on the choice of coordinates, a set of coordinates that simplifies the relevant expressions is used. For example, in earthquake source studies, the x_3 axis can be chosen to go from the center of the earth to the location of the earthquake. The prime meridian, and hence x_1, axis can be selected so that the fault is oriented in the direction $\phi = 0$. These axes simplify the description of the seismic waves radiated by an earthquake, because the distance Δ from the source is now the colatitude. Moreover, the radiation pattern generally has a high degree of symmetry about the fault, so simple functions of ϕ appear. By contrast, the radiation pattern need have no symmetry about the North pole and Greenwich meridian, so a description in those coordinates would usually be more complicated.

Fortunately, a coordinate system referred to the earthquake location does not make describing the propagation of waves from the source any more difficult. Because earth structure varies primarily with depth, the spherical symmetry about the center of the earth is independent of the axis orientation chosen. The geographical convention in which the earth rotates about the x_3 axis is helpful for navigation. In most seismological applications, however, the north direction has no particular significance because the propagation of seismic waves is essentially unaffected by the earth's rotation. The choice of a prime meridian is arbitrary; in the early nineteenth century some American maps had it through Washington DC, and some French maps had it through Paris.

A.7.4 Vector operators in spherical coordinates

Because at a point on the sphere the unit spherical basis vectors are oriented up, south, and east, the basis vectors at different locations are generally not parallel. This makes the vector differential operators more complicated, because these operators involve taking spatial derivatives of vectors. In Cartesian coordinates the unit basis vectors are not affected by this differentiation because they do not change orientation, so only derivatives of the components need be taken. In spherical coordinates, because a vector **u** is

$$\mathbf{u} = u_r\hat{e}_r + u_\theta\hat{e}_\theta + u_\phi\hat{e}_\phi, \tag{13}$$

differential operators acting on **u** must incorporate the derivatives of the basis vectors. Thus, in spherical coordinates, for a scalar field ψ and a vector field **u**:

$$\text{grad }\psi = \hat{e}_r\frac{\partial\psi}{\partial r} + \hat{e}_\theta\frac{1}{r}\frac{\partial\psi}{\partial\theta} + \hat{e}_\phi\frac{1}{r\sin\theta}\frac{\partial\psi}{\partial\phi} \tag{14}$$

$$\text{div }\mathbf{u} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2u_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, u_\theta) + \frac{1}{r\sin\theta}\frac{\partial u_\phi}{\partial\phi} \tag{15}$$

[1] These distance–azimuth equations also have nonseismological applications because ships and aircraft follow the shortest (great circle) paths between two points when possible.

$$\text{curl}\,\mathbf{u} = \hat{e}_r \frac{1}{r\sin\theta}\left(\frac{\partial}{\partial\theta}(\sin\theta\, u_\phi) - \frac{\partial u_\theta}{\partial\phi}\right)$$
$$+ \hat{e}_\theta \frac{1}{r\sin\theta}\left(\frac{\partial u_r}{\partial\phi} - \sin\theta\frac{\partial}{\partial r}(ru_\phi)\right)$$
$$+ \hat{e}_\phi \frac{1}{r}\left(\frac{\partial}{\partial r}(ru_\theta) - \frac{\partial u_r}{\partial\theta}\right) \qquad (16)$$

$$\nabla^2\psi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)$$
$$+ \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}. \qquad (17)$$

These expressions are used when we discuss spherical waves in Section 2.4 and the earth's normal modes in Section 2.9.

A final point worth noting is that the elements of volume and surface used in integrals are different in spherical coordinates from rectangular coordinates. In spherical coordinates (Fig. A.7-5) there are several scale factors, so an element of surface area is

$$dS = r^2 \sin\theta\, d\theta d\phi, \qquad (18)$$

and an element of volume is

$$dV = r^2 \sin\theta dr\, d\theta d\phi. \qquad (19)$$

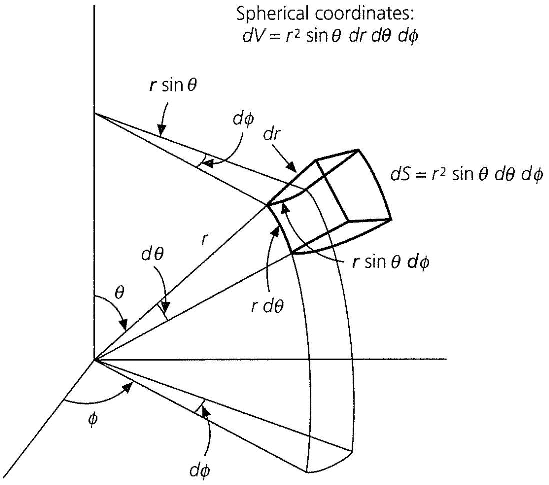


Fig. A.7-5 Definition of the element of volume in spherical coordinates. Unlike the case of Cartesian coordinates, the volume element in spherical coordinates in not a cube. (Marion, 1970. From *Classical Dynamics of Particles and Systems*, 2nd edn, copyright 1970 by Academic Press, reproduced by permission of the publisher.)

A.8 Scientific programming

Most seismological applications require computers, and these requirements, especially in exploration applications with very large data volumes, have spurred the development of computer software and hardware. Some remarks about the use of computers in seismology thus seem appropriate.

Computer usage in seismology includes several broad and overlapping categories:

- Computers are often used in data acquisition and recording systems.
- Data are initially displayed and manipulated using computers.
- Subsequent analysis is frequently done using computers. For example, seismograms can be filtered to enhance certain frequencies or combined to better resolve certain features.
- Theoretical, or *synthetic*, seismograms are often computed for a range of the parameters under study and compared to data to find the best fit.
- Computers are used to *invert* seismological data to determine the parameters of a model which best matches the data.
- Computer modeling is often used to draw geological inferences from seismological observations. For example, seismic velocity data are compared to the predictions of models for the velocity of rock as a function of composition, temperature, and pressure.

These applications often require *scientific programming*, a programming style used for essentially mathematical applications. Some problems in this book also require scientific programming. Although programming is a matter of personal style, this section discusses several points that may be helpful. The suggested reading provides some starting points for readers interested in pursuing these topics further.

A.8.1 Example: synthetic seismogram calculation

Consider a program to compute a synthetic seismogram for waves in a one-dimensional constant-velocity medium, a mathematically idealized string that illustrates features of wave behavior. The program is based on $u(x, t)$, the displacement as a function of position x and time t. The displacement is zero at the fixed ends of the string, $x = 0$ and $x = L$, between which waves travel at speed v. As in Section 2.2.5, the displacement can be written as the sum of the normal modes of the string, each of which is a standing wave with n half wavelengths along the string,

$$u_n(x, t) = \sin(n\pi x/L)\cos(\omega_n t), \qquad (1)$$

and vibrates at a characteristic frequency, or *eigenfrequency*,

$$\omega_n = n\pi v/L. \qquad (2)$$

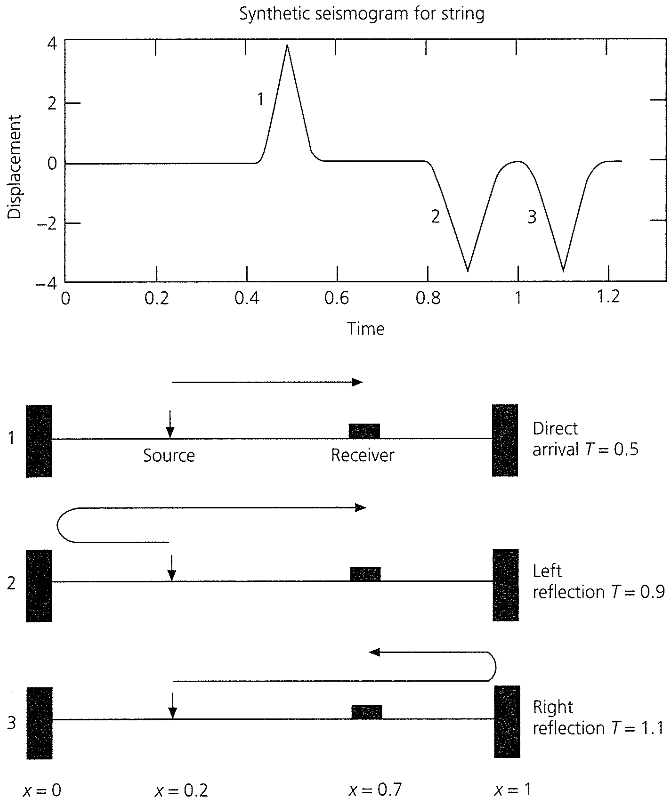


Fig. A.8-1 *Top*: Synthetic seismogram for a string showing the direct wave arrival (1) and reflections (2, 3) from both ends. *Bottom*: Geometry showing source and receiver positions, and the times of the direct and reflected arrivals.

If a source at position x_s generates a pulse at time zero with duration τ, the propagating waves are described by a weighted sum of the modes

$$u(x, t) = \sum_{n=1}^{\infty} \sin(n\pi x/L) \sin(n\pi x_s/L) \cos(\omega_n t) \exp[-(\omega_n \tau)^2/4]. \tag{3}$$

Given the displacement $u(x, t)$ for any position and time, a seismogram ("stringogram") giving the displacement versus time at a receiver position x_r is $u(x_r, t)$. Alternatively, a "snapshot" of the displacement everywhere on the string at time t_0 is $u(x, t_0)$.

Consider a program to evaluate a synthetic seismogram using this sum. For simplicity, we use a string of length 1 m[1] with a wave speed 1 m/s, a source at $x_s = 0.2$ m and a receiver at $x_r = 0.7$ m. To approximate the infinite sum, the program adds up 200 modes. The seismogram (Fig. A.8-1, *top*) is calculated at 50 time steps, covering 1.25 s. This program is written in Fortran, a language that is especially suitable for scientific programming and is therefore commonly used in seismology (and thus in this book). The program could be also written in other languages, but the general points would still apply.

```
C SYNTHETIC SEISMOGRAM FOR HOMOGENEOUS STRING
C DISPLACEMENT U AS FUNCTION OF TIME T
C CALCULATED BY NORMAL MODE SUMMATION
      DIMENSION U(200)
      PI = 3.1415927
C
C PARAMETERS (NORMALLY WOULD COME FROM INPUT)
C STRING LENGTH (M)
      ALNGTH = 1.0
C VELOCITY (M/S)
      C = 1.0
C NUMBER OF MODES
      NMODE = 200
C SOURCE POSITION (M)
      XSRC = 0.2
C RECEIVER POSITION (M)
      XRCVR = 0.7
C SEISMOGRAM TIME DURATION (S)
      TDURAT = 1.25
C NUMBER TIME STEPS
      NTSTEP = 50
C TIME STEP (S)
      DT = TDURAT/NTSTEP
C SOURCE SHAPE TERM
      TAU = .02
C
C LIST PARAMETERS
      WRITE (6,3000)
3000  FORMAT('SYNTHETIC SEISMOGRAM FOR STRING')
      WRITE (6,3001) NMODE
3001  FORMAT('NUMBER OF MODES', I6)
      WRITE (6,3002) ALNGTH, C
3002  FORMAT ('LENGTH (M)' F7.3, 'VELOCITY,
     X (M/S)', F7.3)
      WRITE (6,3003) XSRC, XRCVR
3003  FORMAT ('POSITION (M): SOURCE', F7.3,
     X 'RECEIVER', F7.3)
      WRITE (6,3004) TDURAT, NTSTEP
3004  FORMAT ('SEISMOGRAM DURATION (S)', F7.3,
     X I6, 'TIME STEPS')
      WRITE (6,3005) TAU
3005  FORMAT ('SOURCE SHAPE TERM', F7.3)
C
C INITIALIZE DISPLACEMENT
        DO 5 I = 1, NTSTEP
      U(I) = 0.0
5     CONTINUE
C
C OUTER LOOP OVER MODES
      DO 10 N = 1, NMODE
        ANPIAL = N*PI/ALNGTH
```

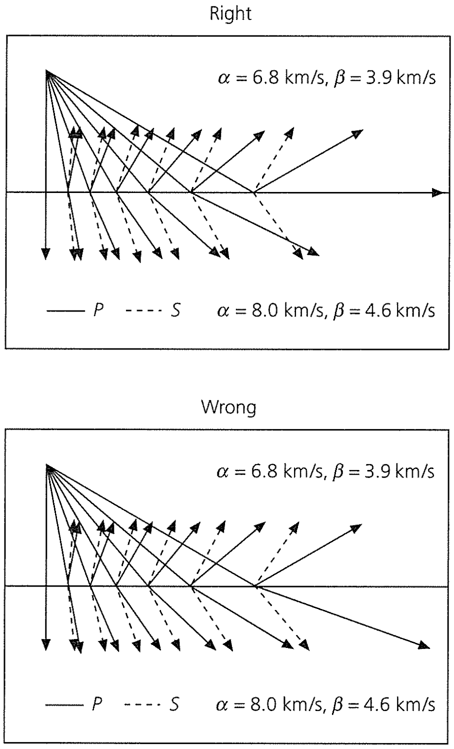
[1] It is easy to use arbitrary values on a computer; we could also use 1 km or 1 furlong. Finding a physical 1 km string is another matter . . .

```
C SPACE TERMS: SOURCE AND RECEIVER
        SXS = SIN(ANPIAL*XSRC)
        SXR = SIN(ANPIAL*XRCVR)
C MODE FREQUENCY
        WN = N*PI*C/ALNGTH
C TIME INDEPENDENT TERMS
        DMP = (TAU*WN)**2
        SCALE = EXP(-DMP/4.)
        SPACE = SXS*SXR*SCALE
C
C INNER LOOP OVER TIME STEPS
        DO 15 J = 1, NTSTEP
          T = DT*(J - 1)
          CWT = COS(WN*T)
C COMPUTE DISPLACEMENT
          U(J) = U(J) + CWT*SPACE
15      CONTINUE
10    CONTINUE
C
C OUTPUT SEISMOGRAM FOR LATER PLOTTING
      WRITE (6, 3101)(U(J), J = 1, NTSTEP)
3101  FORMAT (7F10.4)
      STOP
      END
```

This example brings out several points:

- *Is the answer correct?* Two different types of error occur in scientific programs. First, the *program* may be wrong. In this case, the mathematical formulation correctly describes the physical problem, but the program incorrectly implements this formulation. This is the usual situation, in which "bugs" are identified and corrected. Second, the *formulation* may be wrong, so the program correctly implements an incorrect mathematical model. This could occur because of a mathematical error, like an attempt to sum a divergent series, or a physical error, such as an equation that does not correctly describe waves on a string. An incorrect formulation is particularly disturbing because it cannot be detected by checking the program. For example, Fig. A.8-2 shows two computer simulations for waves bending as they pass from one medium into another with higher velocities. Figure A.8-2 (*top*) uses the correct formulation of Snell's law (Section 2.5), whereas Fig. A.8-2 (*bottom*) looks equally convincing but is wrong because the equation which the program illustrates is incorrect.

Programmers check for both types of errors by choosing cases for which the results can be predicted analytically and comparing the results to those of the program. Several tests are easily done for the string. The wave following the shortest (direct) path appears at the expected time, 0.5 s (Fig. A.8-1, *bottom*), because the source and the receiver are 0.5 m apart. The next two arrivals, reflections from the ends of the string, also occur at the expected times. Moreover, these arrivals have polarities opposite that of the initial pulse, as should occur (Section 2.2.3) upon reflection at the string's fixed ends. The program can also be checked for different string lengths, speeds, and source and receiver positions. Similarly, in addition to synthetic seismograms, displacements along the string at fixed times could be computed. Such tests are important, because if the mathematical model is not appropriate for the physical situation, then time spent debugging, documenting, and optimizing the program is wasted.


Right

Wrong
α = 6.8 km/s, β = 3.9 km/s
P
S
α = 8.0 km/s, β = 4.6 km/s
α = 6.8 km/s, β = 3.9 km/s
P
S
α = 8.0 km/s, β = 4.6 km/s

Fig. A.8-2 Demonstration of the danger that a program accurately computes an incorrect mathematical formulation. *Top*: A correct simulation of wave refraction using Snell's law, $\sin i_1/v_1 = \sin i_2/v_2$. *Bottom*: The same simulation using a wrong formula for Snell's law, $i_1/v_2 = i_2/v_2$.

- *The program is reasonably comprehensible.* Several features help clarify the program. The program's purpose and method are stated. Variable names somewhat resemble those in the equation: "SXS" is $\sin x_s$, and so on. Comments identify the functions of portions of the program.

- *The program uses loops and arrays.* The seismogram is described by the array U(J), and its values at successive times are calculated by looping. Using an array, rather than discrete variables UT1, UT2, etc., makes the program clearer, closer to the mathematical formulation, and simplifies output. The loop structure also makes the program clearer and allows the number of time steps to be changed simply by changing the parameter NTSTEP. Similarly, the number of modes is easily changed.

- *The output is labeled.* The seismogram was placed in an output file for later plotting. The parameters used to compute the seismogram are included, so examination of the output

```
C OUTER LOOP OVER MODES
        DO 10 N = 1, NMODE

            terms for each mode
            that do not depend on time

C INNER LOOP OVER TIME STEPS
            DO 15 J = 1, NTSTEP

            terms that depend on time

C COMPUTE DISPLACEMENT

15          CONTINUE
10      CONTINUE
```

Fig. A.8-3 Structure of the loops for the string synthetic seismogram calculation.

shows how it was computed. This helps avoid the common situation where, given a large collection of computer output, cases are rerun because it is unclear what parameters were used. Moreover, subsequent "improved" versions of the program can be checked to see whether they give the same results.

- *The program is somewhat efficient.* Some thought is generally put into *optimizing* scientific programs to make them run rapidly. The program could find the displacement by looping over time and summing all the modes at each time step. However, consideration of the equation shows that three terms, sin $(n\pi x/l)$, sin $(n\pi x_s/l)$, and exp $[-(\omega_n\tau)^2/4]$ are evaluated only once for each mode, whereas only cos $(\omega_n t)$ is evaluated for each time step. It is thus more efficient to loop over the modes and evaluate each at all times (Fig. A.8-3). Because the outer (mode) loop is executed 200 times, whereas the inner (time) loop is executed $200 \times 50 = 10{,}000$ times, the inner loop should be as efficient as possible. The program would run more slowly if the loops were reversed. The difference, though not significant for this calculation, might be significant for much larger numbers of time steps and modes.

Further improvements could be made to fully optimize the program. Optimization is not an end in itself, because the programmer's time and the intelligibility of the program are also important. Programmers typically try to write reasonably optimized programs without making them impossible to understand and debug. Once fully tested, a program that will be used heavily may be worth further optimization if the computer time savings justify the effort required. There is no point in "getting the wrong answer as fast as possible."[2] Certain computers, such as those using parallel processors, may require specialized optimization.

A.8.2 *Programming style*

The style in which programs are written can make them easier to develop, debug and use. A few suggestions, though not absolute rules, may be useful.

- *Document the program.* Computer programs can be almost useless without adequate documentation. Stonehenge has been described as "the world's largest undocumented computer system."[3] Failure to document is often justified by the assumption that the program will not be used again. This rationalization is self-fulfilling, because even the author may find an undocumented program difficult to reuse once the details are forgotten.

Documentation should state the program's goals and method. The input and output variables, their units, and how they are defined should be listed. Implicit assumptions and restrictions are worth noting. Comments should identify major portions of the program and describe their functions.

Documentation is best written when writing a program because it can aid in debugging. Moreover, once a program is fully written, it is harder to remember how it works. Documentation included in the program is less prone to be lost than that written separately.

Finally, documentation helps scientists exchange programs and work in collaboration. This can be useful, except in the apocryphal cases of programmers writing gigantic undocumented programs to maximize their job security.

- *Use modular programming.* Large programs can generally be divided into smaller subroutines or functions, which can be used like the functions (e.g., sine, square root) supplied by many computer languages. Each subroutine can be tested separately and then used in various programs. Subroutines can handle applications that frequently recur, such as reading or plotting data or carrying out a mathematical operation. This approach saves the time needed to write and debug portions of a program similar to one already available. Moreover, the overall structure of a program containing a set of calls to subroutines is generally easier to understand, because many complexities are isolated into subroutines.

- *Make programs comprehensible.* It is helpful to be able to understand programs once written. Clear documentation and modular programming help. In addition, it should be easy to tell what portions will be executed under which circumstances. For this purpose, portions of a program should be executed sequentially, rather than jumping backwards and forwards within a program.

Similarly, the statements themselves can be written clearly. The use of mnemonic variable names and natural groupings of variables can help. For example, it is somewhat unclear that

X = 0.23873*A / (Y*Y*Y)

gives the average density X of a planet with mass A and radius Y, whereas

RHO = AMASS / ((4.0/3.0) * PI * (RADIUS**3))

[2] Kernighan and Plauger (1978).

[3] Brooks (1975).

is clearer. For clarity, the latter expression is more verbose than required, has π previously defined, and is slightly less efficient.

- *Don't be clever.* Sometimes the shortest, "cleverest" way of programming something can be the worst. In addition to giving rise to lack of clarity, some shortcuts make it difficult to transfer programs between computers. This is especially true of programs that exploit specific properties of an individual computer or compiler, such as local variants of a standard programming language.

- *Keep a perspective on precision.* The program calculates and manipulates numbers that, at least in theory, correspond to physical entities. It is worth keeping track of the precision associated with the data and other quantities, and of that required to compute the desired results.

- *Organize programs and data.* Related programs and the associated files can be grouped into directories which include files listing and explaining the directory's contents. Data files can be organized similarly. Often seismograms, for example, go through multiple processing stages carried out by different programs. A common practice is to use specific types of file names to indicate various intermediate stages. In addition, the data files begin with *headers*, information identifying the data and recording the operations applied to it. The headers and file names should be updated by the programs themselves, rather than "by hand" at each stage. The output, whether text or graphic, should contain the parameters required to replicate the result. This can be especially important for interactive data processing because input files are not kept.

A.8.3 *Representation of numbers*

Several simple concepts about numerical calculations on a computer are worth bearing in mind. One is the consequences of the way in which numbers are represented and manipulated. Because computers use binary (base 2) arithmetic, numbers are written as sets of *bits*, single binary digits, grouped into *words*. Some general ideas about these representations can be illustrated without going into the schemes used by various computers.

Integers are represented by their binary equivalent. Thus 46 (decimal) is 101110, because

$$46 = 1\times 2^5 + 0\times 2^4 + 1\times 2^3 + 1\times 2^2 + 1\times 2^1 + 0\times 2^0.$$

Many computers represent integers by 16- or 32-bit words. The word length governs the range of possible integers. For example, using 16 bits, one of which indicates the sign, the largest positive integer that can be represented is

$$111\ 1111\ 1111\ 1111\ (\text{binary}) = 2^{15} - 1 = 32{,}767.$$

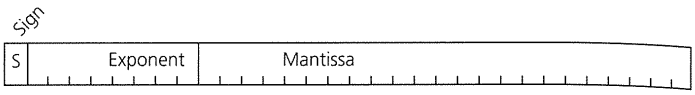


Fig. A.8-4 Representation of a floating point number using 32 bits.

Because a greater range is needed for scientific computation, floating point numbers are used:

$$\text{number} = (\text{mantissa}) \times 2^{\text{exponent}}.$$

Floating point numbers can accommodate fractions, with digits to the right of the binary point representing negative powers of two, just as digits to the left of the point represent positive powers of two. For example,

$$\begin{aligned}46.625\ (\text{decimal}) &= 1\times 2^5 + 0\times 2^4 + 1\times 2^3 + 1\times 2^2 + 1\times 2^1\\ &\quad + 0\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}\\ &= 101{,}110.101\ (\text{binary}) = 0.101110101\times 2^6.\end{aligned}$$

To represent binary floating point numbers on a computer, a certain number of bits are assigned to the mantissa and the exponent. Figure A.8-4 shows one way in which a single precision floating point number might be represented by a 32-bit word. One bit is reserved for the sign of the mantissa, 8 bits are used for the exponent including its sign, and the remaining 23 bits contain the mantissa. The number of bits available for the exponent determines the *range* of the floating point numbers. Because $2^8 = 256$, the exponent can represent numbers between approximately 2^{127} and 2^{-128} or approximately 10^{38} to 10^{-39}. The number of bits in the mantissa determines the *precision* or number of significant digits. Because 2^{-23} is approximately 10^{-7}, the maximum number of significant decimal digits is about seven. Further precision can be obtained using *double precision* numbers with additional bits for the mantissa. The precise values of the range and the precision depend on details of the implementation.

The range and precision in use are worth bearing in mind because computers do not always issue "overflow" or "underflow" warnings. The computer may assign a value, such as the largest floating point number, and proceed. It can be frustrating to find that the peculiar answers produced by a program result from numbers outside the computer's range.

A related malady is *round-off error*, the loss of computational precision due to the limited number of significant digits. To illustrate the concept, suppose that a computer used six bits for the mantissa. The decimal addition

$$0.65625 + 0.96875 = 1.625$$

would, in binary, be

$$0.10101 + 0.11111 = 1.10100,$$

which, because no precision was lost, equals the exact answer. Now, consider the decimal addition

$$5.25 + 0.96875 = 6.21875,$$

which, in binary, becomes

$$0.101010 \times 2^3 + 0.111110 \times 2^0.$$

To carry out the binary addition, because the numbers have different exponents, the mantissa of the smaller number is shifted to produce a common exponent. If some of the bits representing the smaller number are lost, inaccuracy may result. For example, in this case,

$$\begin{aligned} 0.101010 \times 2^3 + 0.000111 \times 2^3 &= 0.110001 \times 2^3 \\ &= 6.125 \text{ (decimal)}. \end{aligned}$$

The precision available on a computer is generally adequate to avoid significant round-off error. Nonetheless, it is a potential problem to keep in mind, especially in long calculations or in those such as a series sum where the answer is the difference between large numbers.

A.8.4 *A few pitfalls*

Difficulties often can be avoided by considering how various statements in the program will be executed. This is especially the case when using compilers that provide little error checking and few helpful warning and error messages. The computer, following its explicit rules, may yield results differing from those expected. The foibles here are for Fortran, but similar ones often appear in other computer languages.

- *Statement execution*. Problems often stem from the distinction between integers and floating point numbers. For example, if I and J are integer variables,

J = 5
I = 1/J

yields zero, because integer division yields an integer. This problem is not cured by setting the result equal to a floating point variable, or performing a floating point operation on the integer result:

X = 1/J
Z = 1.0*(1/J)

yield zero, because division is done as an integer operation, and the result (0) is converted to floating point (0.0). On the other hand, most compilers give 0.2 as the result of

X = 1.0/J,

although a conservative policy is to explicitly convert the integer to floating point

X = 1.0/FLOAT(J).

A second class of problems can result from the order in which operations are performed. For example, it may be unclear whether

−1.0**2

should be interpreted as $(-1.0)^2 = 1.0$ or $-(1.0)^2 = -1.0$. Although the computer language rules are explicit, it may be wise to use parentheses, e.g.,

(−1.0)**2

to ensure that operations are carried out as desired. The additional parentheses can also make the program more comprehensible.

- *Subroutines*. Subroutines are heavily used in writing scientific programs. As a result, problems can result while using computer languages like Fortran in which what appear to be arguments passed to a subroutine are actually the locations in memory of these arguments.

A common error is exemplified by the following program

```
CALL SUB(1.0)
X = 1.0
WRITE (6,*) 'X = ', X
STOP
END

SUBROUTINE SUB(Y)
Y = 5.0
RETURN
END
```

which, when executed, yields “X = 5.0.” Because Y, a parameter in the subroutine definition, was set equal to 5.0, the value of the corresponding parameter in the subroutine call, “1.0” has been redefined as 5.0. This situation, which sometimes underlies inexplicable behavior by programs, can be avoided by not passing numerical values of an argument explicitly to a subroutine if the argument will be redefined. For example, had the first statements been

```
Z = 1.0
CALL SUB(Z)
```

the variable Z would equal 5.0, but “1.0” would not be affected.

Other errors occur when either the type or number of arguments in a call to a subroutine do not match those in its definition. For example, calling a subroutine with an integer variable may yield unexpected results if the definition is in terms of a real variable.

- *Arrays*. Scientific computing often involves dealing with *arrays*, groups of data addressed by their indices. For example, a seismogram giving a single component (e.g., vertical) of ground motion can be written as an array (U(1), U(2) . . .) of displacement versus time. Similarly, a seismogram giving all three (vertical, north–south, east–west) ground motion components can be written as a two-dimensional array

```
U(1, 1), U(1, 2), U(1, 3), U(1, 4) . . .
U(2, 1), U(2, 2), U(2, 3), U(2, 4) . . .
U(3, 1), U(3, 2), U(3, 3), U(3, 4) . . .
```

whose first index gives the component, and second index indicates the time.

Arrays are defined initially by statements giving their dimensions, i.e.,

```
DIMENSION A(N, M)
```

or

```
REAL A(N, M).
```

Typically, the computer selects a memory location for the first element in A and reserves N × M successive locations. Similarly, N × M × R locations are reserved for a three-dimensional array dimensioned (N, M, R). In Fortran, regardless of the number of dimensions, an array is stored as one-dimensional with the first index varying the most rapidly, then the second, and so on. In other words, if A is dimensioned (2, 3), the storage order is

```
A(1,1), A(2,1), A(1,2), A(2,2), A(1,3), A(2,3).
```

For two-dimensional arrays, this can be thought of as storing the array by columns. An individual array element is found by calculating its location relative to that of the first element. Thus, for an array dimensioned (N, M), with element (1, 1) at location 1, element (I, J) is found at location

```
1 + (I - 1) + (J - 1) × N.
```

Several computational difficulties can arise in dealing with arrays. A common set of errors involve being "off by one," either by starting or ending on the wrong element. This is especially easy because some computer languages (e.g., Fortran) start with the first element in an array being "1," whereas others (e.g., "C") start with the first array element as "0." Thus one needs to make sure that the array elements correspond to the expected variable values, such as seismic record times. Often, when an array index is computed by the program, an error yields an index outside the bounds dimensioned for the array. Because many compilers do not check for such errors unless specifically requested, a statement like

```
A(9) = 4.0
```

will usually be executed even for an array dimensioned

```
DIMENSION A(5).
```

Typically, the computer places 4.0 in whatever is 8 locations in memory beyond A(1). This location may contain some other variable, or a portion of the program itself. Often the program continues until it requires the contents of the overwritten location, at which point several things may occur. At best, the program "crashes"; at worst, it continues the calculation with erroneous values that propagate. Array element out-of-bounds problems are among the most common and most frustrating difficulties in scientific programming. When a compiler provides array bounds checking, it is worth using.

The nature of array storage can also lead to inefficient programs. On many computers, data which are actually on disk can be treated as resident in memory, and are automatically "swapped" into physical memory when needed. For efficiency, large adjacent regions of the disk are often swapped into physical memory together. Efficient programs minimize swapping by making the most possible use of data that reside in physical memory. By contrast, inefficient programs can produce "thrashing," a situation in which much of the computer's time is spent swapping rather than computing.

For example, consider[4]

```
DIMENSION A(1000, 1000)
DO 10 I = 1,1000
DO 10 J = 1,1000
10 A(I, J) = I + J
```

Because the elements of A are stored in column order, A(1, 1) and A(1, 2) are a thousand locations apart. It would be more efficient to reverse the loops

```
10 A(J, I) = I + J
```

so that adjacent locations (A(1, 1), A(2, 1) . . .) were used successively.

- *Uninitialized variables*. Problems frequently result from *uninitialized variables*: those used in calculation without their values being set. A common example, summing an array

```
DO 10 I = 1, N
10 SUM = SUM + A(I)
```

can give strange results unless the compiler initializes SUM as zero. Because this is not always the case, it is thus wise to explicitly initialize, e.g.,

```
SUM = 0.0
```

before executing the loop. Proper initialization also helps to ensure that programs do not give different results on different computers.

- *The computer may be wrong*. Although most problems result from programming errors, a very small fraction of the time the error may be the computer's. Compilers have been known to contain "bugs" in common routines such as square root, tangent, or complex arithmetic. This tempting explanation for the failure of a long and intricate program can generally be rejected unless a test program that carries out only the suspect operation yields the wrong answer.

A.8.5 Some philosophical points

To close our discussion, a few general thoughts are worth considering. Historically, computers were considered a scarce and valuable resource. Currently, as computer power increases and costs fall, it is increasingly practical to carry out investi-

[4] Hatton (1983c).

gations numerically. One example is the change, both in exploration and in global seismology, from earth models whose properties vary only with depth, to three-dimensional models that are evaluated numerically.

The role of analytic solutions is also changing. In addition to the traditional goal of providing exact solutions to simplified problems, analytic solutions provide test cases for numerical solutions of more complex problems. Analytic solutions can also yield the insight needed to evaluate numerical results.

Along with the increase in the complexity of problems that can be solved computationally comes an increase in the volume of output. Fortunately, a parallel development has been the increasing role of graphic output, often in color. The proverb "A picture is worth a thousand words" may be unduly conservative in this context. A thousand words on a computer might be 32,000 bits; graphic output often makes it possible to visualize data with millions of bits.

Finally, software such as spreadsheets or programs with sophisticated general mathematical capabilities often eliminates the need to write programs for a specific application. In this book, we do not assume that such software will be used for the problems, although many could be done this way. We think that programming without using such software gives a deeper understanding of the underlying principles. Hence, in educational applications, we strongly favor programming, even if in non-educational applications ease of use may favor sophisticated software.

Further reading

Many texts cover portions of the mathematical material summarized here. Feynman (1982) discusses general issues of the relations between mathematics and science. Butkov (1968) and Menke and Abbott (1990) provide introductions to many of these topics. Fung (1969), Hay (1953), Jeffreys and Jeffreys (1950), and Marion (1970) treat vectors, vector transformations, and vector differential operators. Applied linear algebra texts such as Franklin (1968) and Noble (1969) deal with the range of the subject including numerical methods.

Articles by Hatton (1983a–d, 1984a,b, 1985) provide a broad and witty introduction to computer science for geophysicists. Eckhouse and Morris (1979) and Sloan (1980) cover topics in computer software, including the representation of numbers and arithmetic operations. Kernighan and Plauger (1976, 1978) discuss topics in programming style. Brooks (1975) treats issues in the development and organization of computer software. Numerical analysis texts like Froberg (1969) cover round-off and other sources of error in numerical computations. Harkrider (1988) gives an entertaining anecdotal account of early (*c.*1960) computer usage in seismology.

The application of spherical geometry to the paths between an earthquake and a receiver, including the effects of the earth's ellipticity are discussed by Ben-Menahem and Singh (1981) and Bullen and Bolt (1985). The theory of the earth's shape is treated by Cook (1973) and Jeffreys (1976).

Problems

1. Find the angle between the vectors (1, 4, 2) and (2, 3, 1).
2. Show, using index notation, that for the three-dimensional vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$:
   (a) $\mathbf{a} \times \mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$.
   (b) $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin\theta$, where θ is the angle between the two vectors.
   (c) $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$.
   (d) $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}$.
   (e) $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$.
   (f) $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$.
3. Show that for arbitrary matrices A, B, and C:
   (a) $(AB)^T = B^T A^T$.
   (b) $(ABC)^T = C^T B^T A^T$.
4. Prove the following properties of determinants for the case of a 2×2 matrix:
   (a) The determinant of a matrix equals the determinant of its transpose.
   (b) If two rows or columns of a matrix are interchanged, the determinant has the same absolute value, but its sign changes.
   (c) If a multiple of one row (or column) of a matrix is added to another row (or column), the determinant is unchanged.
   (d) If two rows or columns of a matrix are the same, the determinant is zero.
5. Express the determinant of a 3×3 matrix using the definition in Eqn A.4.17.
6. Prove that if A has an inverse, the two solutions $\mathbf{x}$ and $\mathbf{y}$ satisfying $A\mathbf{x} = \mathbf{b}$ and $A\mathbf{y} = \mathbf{b}$ are equal.
7. Find the inverse of the matrix
   $$\begin{pmatrix} 1 & 2 \\ 5 & 4 \end{pmatrix}$$
   both by the cofactor method and by row operations. Check that the solution is in fact the inverse.
8. Show that the inverse of a 2×2 matrix A is given by
   $$A^{-1} = \frac{1}{|A|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$
9. Show that A, the transformation matrix for a rotation about the $\hat{\mathbf{e}}_3$ axis (Eqn A.5.9) satisfies $A^T A = I$ and is thus orthogonal.
10. Prove that the magnitude of a vector is preserved by an orthogonal transformation.
11. Expand the determinant that give the eigenvalues of a 3×3 matrix (Eqn A.5.19) and verify that the invariants (Eqn A.5.21) are the coefficients of the characteristic polynomial.
12. Prove the following vector identities using index notation:
    (a) For any vector field $\mathbf{u}(\mathbf{x})$, $\nabla \cdot (\nabla \times \mathbf{u}) = 0$.
    (b) For any scalar function $\phi(\mathbf{x})$, $\nabla \times \nabla\phi = 0$.
13. For the vector field $\mathbf{u}(x, y, z) = (3x^2y^2 + z, 2x^3y + 2y, x)$, find:
    (a) $\nabla \cdot \mathbf{u}$.
    (b) $\nabla \times \mathbf{u}$.
    (c) $\nabla^2 \mathbf{u}$.
    (d) A scalar field $\phi(x, y, z)$ such that $\mathbf{u} = \nabla\phi$.

14. Use index notation to show that the Laplacian in Cartesian coordinates of any vector field $\mathbf{u}(\mathbf{x})$ satisfies

$$\nabla^2\mathbf{u} = \nabla(\nabla \cdot \mathbf{u}) - \nabla \times \nabla \times \mathbf{u}.$$

15. Show that at any point in a spherical coordinate system, the spherical basis vectors $(\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_\phi)$ form an orthonormal set.
16. Use Eqn A.7.6 to derive the angular distance Δ between the locations of an earthquake and a seismic station as given in Eqn A.7.7.

Computer problems

The solutions may be useful for other problems in this and other chapters.

C-1. Find the largest integer your computer allows by starting with "2," "2 × 2," "2 × 2 × 2," and doing successive multiplication by 2. What happens when you exceed this number? Do the same for floating point numbers using "10.0" instead of "2" in both single and double precision. Does double precision allow larger floating point numbers?

C-2. Find when your computer starts to show round-off error by starting with "10.0" and doing successive multiplications by 10.0. At each step, add 1.0 to the result and subtract the two numbers. When does the difference become zero? Do the same in double precision.

C-3. Write subroutines to do the following operations on an input vector in three dimensions:
(a) Find the magnitude of a vector.
(b) Find the sum of two vectors.
(c) Find the scalar product of two vectors.
(d) Find the vector product of two vectors.
Your subroutines should include comment lines explaining the purpose of the routine and the various inputs and outputs.

C-4. Write a subroutine using the necessary subroutines from problem C-3 to find the angle between two vectors.

C-5. Use the solutions to problems C-3 and C-4 to find the magnitude, sum, scalar product, and vector product of the vectors (1, 4, 2) and (2, 3, 1), and the angle between the two vectors.

C-6. (a) Write a subroutine to multiply an $n \times m$ matrix by an m-element vector.
(b) Write a subroutine to multiply an $n \times m$ matrix by an $m \times r$ matrix.
(c) Write a subroutine to find the determinant of a 3×3 matrix.

C-7. (a) Write a subroutine that uses Gaussian elimination with partial pivoting to solve the system of equations $A\mathbf{x} = \mathbf{b}$. The routine should take an arbitrary 3×3 matrix A and 3-element vector $\mathbf{b}$ as inputs. The program should test the solution by multiplying $A\mathbf{x}$ and subtracting $\mathbf{b}$ from the result. The subroutines from C-6 may be helpful.
(b) Use the subroutine to solve

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix}.$$

C-8. (a) Write functions that return the values of the δ_{ij} and ε_{ijk} symbols given the indices as arguments. Test the functions and show that they give the correct values.
(b) Write a program that uses these two functions to prove the identity

$$\varepsilon_{ijk}\varepsilon_{ist} = \delta_{js}\delta_{kt} - \delta_{jt}\delta_{ks}$$

by testing all possible combinations of indices.

C-9. (a) Write a subroutine to invert a 3×3 matrix using elementary row operations. The subroutine should first check to see if the matrix is singular. It should test the result by multiplying by the original matrix.
(b) Use this routine to invert

$$\begin{pmatrix} 1 & -1 & -1 \\ 3 & -1 & 2 \\ 2 & 2 & 3 \end{pmatrix}.$$

C-10. (a) Write a program to solve a 3×3 system of equations $A\mathbf{x} = \mathbf{b}$ using the matrix inversion routine from the previous problem. The program should test the solution by multiplying $A\mathbf{x}$ and subtracting $\mathbf{b}$ from the result. The subroutines from C-6 may be helpful.
(b) Use the program to solve the system of equations in C-7.

C-11. (a) Write a subroutine to find the roots of a general cubic equation using the method given below.[1]

A cubic equation $y^3 + py^2 + qy + r = 0$ may be converted to

$$x^3 + ax + b = 0$$

by defining

$$y = x - p/3, \quad a = (3q - p^2)/3, \quad b = (2p^3 - 9pq + 27r)/27.$$

If p, q, and r are real, the quantity

$$c = b^2/4 + a^3/27$$

characterizes the roots: if $c > 0$, there is one real root and two conjugate imaginary roots; if $c = 0$, there are three real roots, of which two are equal; and if $c < 0$, there are three real and unequal roots. Using

$$A = (-b/2 + c^{1/2})^{1/3}, \quad B = (-b/2 - c^{1/2})^{1/3},$$

the values of x given by

$$x = A + B, \quad [-(A + B) + (A - B)\sqrt{-3}]/2,$$
$$-[(A + B) + (A - B)\sqrt{-3}]/2$$

are the roots.

The subroutine requires complex arithmetic and should test the roots by substituting back into the equation.
(b) Use the result to solve

$$y^3 - 8y^2 + 19y - 12 = 0.$$

C-12. (a) Write a subroutine to find the eigenvalues and eigenvectors of a real, symmetric 3×3 matrix, using the results of C-11. The program should check that the eigenvectors and eigenvalues satisfy their definition. Be careful to avoid dividing by zero.
(b) Use this subroutine to find the eigenvalues and eigenvectors of

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}.$$

[1] Beyer (1984).

C-13. (a) Write a program that accepts the latitude and longitude of two points on the earth's surface and finds the angular distance and distance along the earth's surface between them, and the azimuth and back azimuth.

(b) Use your program to find the distances and azimuths between:

(i) Cairo, Illinois (37°N, 89°W) and Cairo, Egypt (30°N, 32°E).

(ii) Berlin, New Hampshire (44.5°N, 71.5°W) and Berlin, Germany (52.5°N, 13.5°E).

(iii) Montevideo, Minnesota (45°N, 95.5°W) and Montevideo, Uruguay (35°S, 56°W).

(iv) Mexico, Maine (44.5°N, 70.5°W) and Mexico City, Mexico (19°N, 99°W).

References

Agnew, D. C., B. Berger, R. Buland, W. Farrell, and F. Gilbert (1976) International deployment of accelerometers: A network for very long period seismology, *Eos. Trans. Am. Geophys. Un.*, *57*, 180–8.

Agnew, D. C., *et al.* (1988) *Probabilities of Large Earthquakes Occurring in California on the San Andreas Fault*, US Geol. Survey, Open-File Rep.

Ahrens, T. J. (ed.) (1995a) *Global Earth Physics: a handbook of physical constants*, Am. Geophys. Un., Washington, DC.

Ahrens, T. J. (ed.) (1995b) *Mineral Physics and Crystallography: a handbook of physical constants*, Am. Geophys. Un., Washington, DC.

Ahrens, T. J. (ed.) (1995c) *Rock Physics and Phase Relations: a handbook of physical constants*, Am. Geophys. Un., Washington, DC.

Aki, K. (1980) Presidential address to the Seismological Society of America, *Bull. Seism. Soc. Am.*, *70*, 1969–76.

Aki, K., and P. G. Richards (1980) *Quantitative Seismology: theory and methods*, W. H. Freeman, San Francisco.

Al-eqabi, G. I., K. Koper, and M. E. Wysession (2001) Source characterization of Nevada test site explosions and western United States earthquakes using Lg waves, with implications for regional discrimination, *Bull. Seism. Soc. Am.*, *91*, 140–53.

Alexander, D. (1993) *Natural Disasters*, Chapman and Hall, New York.

Ambraseys, N. (1989) Studies begin on Armenian quake, *Eos Trans. Am. Geophys. Un.*, *70*, 17.

Anderson, D. L. (1989) *Theory of the Earth*, Blackwell, Oxford.

Ando, M. (1975) Source mechanisms and tectonic significance of historical earthquakes along the Nankai Trough, Japan, *Tectonophysics*, *27*, 119–40.

Ando, M., Y. Ishikawa, and F. Yamazaki (1983) Shear wave polarization anisotropy in the upper mantle beneath Honshu, Japan, *J. Geophys. Res.*, *88*, 5850–64.

Argus, D. F., R. G. Gordon, C. DeMets, and S. Stein (1989) Closure of the Africa-Eurasia-North America plate motion circuit and tectonics of the Gloria fault, *J. Geophys. Res.*, *94*, 5585–602.

Astiz, L., P. S. Earle, and P. Shearer (1996) Global stacking of broadband seismograms, *Seism. Res. Lett.*, *67*, 8–18.

Atkinson, G. M., and I. Beresnev (1997) Don't call it stress drop, *Seism. Res. Lett.*, *68*, 3–4.

Atkinson, G. M., and D. M. Boore (1995) Ground-motion relations for Eastern North America, *Bull. Seism. Soc. Am.*, *85*, 17–30.

Babuska, V., and M. Cara (1991) *Seismic Anisotropy in the Earth*, Kluwer Academic Publishers, Boston.

Baker, B. B., and E. T. Copson (1950) *The Mathematical Theory of Huygens' Principle*, Clarendon Press, Oxford.

Batchelor, G. (1967) *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge.

Bath, M., and A. J. Berkhout (1984) *Mathematical Aspects of Seismology*, Geophysical Press, London.

Bebout, G. E., D. W. Scholl, S. H. Kirby, and J. P. Platt (1996) *Subduction: top to bottom*; Geophysical Monograph 96, Am. Geophys. Un., Washington, DC.

Benioff, H. (1955) Seismic evidence for crustal structure and tectonic activity, in A. Poldervaart (ed.), *Crust of the Earth*, Geol. Soc. Amer. Spec. Pap. 62, pp. 61–74.

Ben-Menahem, A., and S. J. Singh (1981) *Seismic Waves and Sources*, Springer-Verlag, New York.

Bennett, R. A., J. L. Davis, and B. P. Wernicke (1999) Present-day pattern of Cordilleran deformation in the Western United States, *Geology*, *27*, 371–4.

Bent, A. (1995) A complex double-couple source mechanism for the M_s 7.2 1929 Grand Banks earthquake, *Bull. Seism. Soc. Am.*, *85*, 1003–20.

Benz, H. M., and J. E. Vidale (1993) Sharpness of upper-mantle discontinuities determined from high-frequency reflections, *Nature*, *365*, 147–50.

Bevington, P. R., and D. K. Robinson (1992) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.

Beyer, W. H. (1984) *CRC Standard Mathematical Tables*, CRC Press, Boca Raton, FL.

Bina, C. R. (1997) Patterns of deep seismicity reflect buoyancy stresses due to phase transitions, *Geophys. Res. Lett.*, *24*, 3301–4.

Bina, C. R., and M. Liu (1995) A note on the sensitivity of mantle convection models to composition-dependent phase relations, *Geophys. Res. Lett.*, *22*, 2565–8.

Bina, C. R., and B. J. Wood (1987) The olivine-spinel transitions: Experimental and thermodynamic constraints and implications for the nature of the 400 km seismic discontinuity, *J. Geophys. Res.*, *92*, 4853–66.

Birch, F. (1952) Elasticity and constitution of the Earth's interior, *J. Geophys. Res.*, *57*, 227–86.

Birch, F. (1954) The earth's mantle: Elasticity and constitution, *Trans. Am. Geophys. Un.*, *35*, 79–85.

Birch, F. (1968) On the possibility of large changes in the earth's volume, *Phys. Earth Planet. Inter.*, *1*, 141–7.

Bland, D. R. (1988) *Wave Theory and Applications*, Oxford University Press, New York.

Bodine, J. H., M. S. Steckler, and A. B. Watts (1981) Observations of flexure and the rheology of the oceanic lithosphere, *J. Geophys. Res.*, *86*, 3695–707.

Boehler, R. (1996) Melting temperature of the Earth's mantle and core: Earth's thermal structure, *Ann. Rev. Earth Planet. Sci.*, *24*, 15–40.

Bolt, B. A. (1976) *Nuclear Explosions and Earthquakes: the parted veil*, W. H. Freeman, San Francisco.

Bolt, B. A. (1982) *Inside the Earth*, W. H. Freeman, San Francisco.

Bolt, B. A. (1999) *Earthquakes*, 4th edn, W. H. Freeman, San Francisco.

Bonini, W. E., and R. R. Bonini (1979) Andrija Mohorovičić: Seventy years ago an earthquake shook Zagreb, *Eos Trans. Am. Geophys. Un.*, *60*, 699–701.

Boore, D. M. (1977) Strong-motion recordings of the California earthquake of April 18, 1906, *Bull. Seism. Soc. Am.*, *67*, 561–77.

Boschi, E., G. Ekstrom, and A. Morelli (eds) (1996) *Seismic Modelling of Earth Structure*, Editrice Compositori, Rome.

Bott, M. H. P. (1982) *The Interior of the Earth: its structure, constitution and evolution*, Elsevier Science Publishing Co., Inc., New York.

Bott, M. H. P., A. P. Holder, R. E. Long, and A. L. Lucas (1970) Crustal structure beneath the granites of south-west England, in G. Newall and N. Rast (eds), *Mechanism of Igneous Intrusion*, Geol. J. Special Issue, 2, pp. 93–102.

Brace, W. F., and J. D. Byerlee (1970) California earthquakes: Why only shallow focus?, *Science*, *168*, 1573–5.

Brace, W. F., and D. L. Kohlstedt (1980) Limits on lithospheric stress imposed by laboratory experiments, *J. Geophys. Res.*, *85*, 6248–52.

Bracewell, R. (1978) *The Fourier Transform and its Applications*, McGraw-Hill, New York.

Braile, L. W., and C. S. Chiang (1986) The continental Mohorovičić discontinuity: Results from near vertical and wide angle seismic reflection studies, in M. Barazangi and L. Brown (eds), *Reflection Seismology: a global perspective*, Geodynamics Series, 13, Am. Geophys. Un., Washington, DC, pp. 257–72.

Braile, L. W., and R. B. Smith (1975) Guide to the interpretation of crustal refraction profiles, *Geophys. J. R. Astron. Soc.*, *40*, 145.

Braile, L. W., W. J. Hinze, R. G. Keller, E. G. Lidiak, and J. L. Sexton (1986) Tectonic development of the New Madrid rift complex, Mississippi embayment, North America, *Tectonophysics*, *131*, 1–21.

Braile, L. W., W. J. Hinze, R. R. B. von Frese, and G. Randy Keller (1989) Seismic properties of the crust and uppermost mantle of the conterminous United States and Canada, in L. C. Pakiser and W. D. Mooney (eds), *Geophysical Framework of the Continental United States*, Geol. Soc. Amer. Mem. 172, Boulder, CO., pp. 655–79.

Bray, J. D. (1995) Geotechnical earthquake engineering, in W. F. Chen (ed.), *The Civil Engineering Handbook*, CRC Press, Boca Raton, FL.

Brennan, B. J., and D. E. Smylie (1981) Linear viscoelasticity and dispersion in seismic wave propagation, *Rev. Geophys.*, *19*, 233–46.

Brigham, E. O. (1974) *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, NJ.

Brooks, F. P. (1975) *The Mythical Man-Month*, Addison-Wesley, Reading, MA.

Brown, G. C., and A. E. Mussett (1993) *The Inaccessible Earth*, Chapman and Hall, London.

Brumbaugh, D. (1999) *Earthquakes: science and society*, Prentice-Hall, Upper Saddle River, NJ.

Brune, J. N., W. M. Ewing, and J. T. F. Kuo (1961) Group and phase velocities for Rayleigh waves of period greater than 380 seconds, *Science*, *133*, 757–8.

Bullen, K. E. (1975) *The Earth's Density*, Chapman and Hall, London.

Bullen, K. E., and B. A. Bolt (1985) *An Introduction to the Theory of Seismology*, 4th edn, Cambridge University Press, Cambridge.

Bürgmann, R., P. A. Rosen, and E. J. Fielding (2000) Synthetic aperture radar interferometry to measure earth's surface topography and its deformation, *Ann. Rev. Earth Planet. Sci.*, *28*, 169–209.

Butkov, E. (1968) *Mathematical Physics*, Addison-Wesley, Reading, MA.

Byerlee, J. D. (1978) Friction of rocks, *Pure Appl. Geophys.*, *116*, 615–26.

Capon, J. (1969) Investigation of long-period noise at the large aperture seismic array, *J. Geophys. Res.*, *74*, 3182–94.

Chase, C. G. (1972) The *n*-plate problem of plate tectonics, *Geophys. J. R. Astron. Soc.*, *29*, 117–22.

Chase, C. G. (1978) Plate kinematics: The Americas, East Africa, and the rest of the world, *Earth Planet. Sci. Lett.*, *37*, 355–68.

Chave, A. D. (1979) Lithospheric structure of the Walvis Ridge from Rayleigh wave dispersion, *J. Geophys. Res.*, *84*, 6840–8.

Chen, W.-P., and P. Molnar (1983) Focal depths of intracontinental and intraplate earthquakes and their implications for the thermal and mechanical properties of the lithosphere, *J. Geophys. Res.*, *88*, 4183–214.

Chinnery, M. A. (1961) The deformation of the ground around surface faults, *Bull. Seism. Soc. Am.*, *51*, 355–72.

Chopra, A. K. (1995) *Dynamics of Structures: theory and applications to earthquake engineering*, Prentice-Hall, Upper Saddle River, NJ.

Choy, G., and P. G. Richards (1975) Pulse distortion and Hilbert transformation in multiply reflected and refracted body waves, *Bull. Seism. Soc. Am.*, *65*, 55–70.

Christensen, U. R. (1995) Effects of phase transitions on mantle convection, *Ann. Rev. Earth Planet. Sci.*, *23*, 65–87.

Chu, D., and R. G. Gordon (1998) Current plate motions across the Red Sea, *Geophys. J. Int.*, *135*, 313–28.

Chu, D., and R. G. Gordon (1999) Evidence for motion between Nubia and Somalia along the Southwest Indian ridge, *Nature*, *398*, 64–7.

Chung, W.-Y., and H. Kanamori (1980) Variation of seismic source parameters and stress drops within a descending slab and its implications in plate mechanics, *Phys. Earth Planet. Inter.*, *23*, 134–59.

Claerbout, J. F. (1976) *Fundamentals of Geophysical Data Processing*, McGraw-Hill, New York.

Claerbout, J. F. (1985) *Imaging the Earth's Interior*, Blackwell, Oxford.

Cloetingh, S., and R. Wortel (1985) Regional stress field of the Indian plate, *Geophys. Res. Lett.*, *12*, 77–80.

Coburn, A. W., and R. J. S. Spence (1992) *Earthquake Protection*, Wiley, New York.

Cook, A. H. (1973) *Physics of the Earth and Planets*, Wiley, New York.

Cox, A. (1973) *Plate Tectonics and Geomagnetic Reversals*, W. H. Freeman, San Francisco.

Cox, A., and R. B. Hart (1986) *Plate Tectonics: how it works*, Blackwell, Palo Alto, CA.

Creager, K. C. (1992) Anisotropy of the inner core from differential travel times of the phases PKP and PKIKP, *Nature*, *356*, 309–14.

Crossley, D. J. (ed.) (1997) *Earth's Deep Interior*, Gordon and Breach, Amsterdam.

Dahlen, F. A., and J. Tromp (1998) *Theoretical Global Seismology*, Princeton University Press, Princeton, NJ.

Davies, G. F. (1999) *Dynamic Earth: plates, plumes and mantle convection*, Cambridge University Press, Cambridge.

Davis, P., D. Jackson, and Y. Kagan (1989) The longer it has been since the last earthquake, the longer the expected time till the next, *Bull. Seism. Soc. Am.*, *79*, 1439–56.

DeMets, C., R. G. Gordon, D. F. Argus, and S. Stein (1990) Current plate motions, *Geophys. J. Int.*, *101*, 425–78.

DeMets, C., R. G. Gordon, D. F. Argus, and S. Stein (1994) Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motion, *Geophys. Res. Lett.*, *21*, 2191–4.

Dewey, J. W. (1987) Instrumental seismicity of Central Idaho, *Bull. Seism. Soc. Am.*, *77*, 819–36.

Diebold, J. B., and P. L. Stoffa (1981) The travel time equation, tau-p mapping and inversion of common midpoint data, *Geophysics*, *46*, 238–54.

Dixon, T. H. (1991) An introduction to the global positioning system and some geological applications, *Rev. Geophys.*, *29*, 249–76.

Dobrin, M. B. (1976) *Introduction to Geophysical Prospecting*, McGraw-Hill, New York.

Dobrin, M. B., and C. H. Savit (1988) *Introduction to Geophysical Prospecting*, 4th edn, McGraw-Hill, New York.

Douglas, A., J. A. Hudson, and R. G. Pearce (1988) Directivity and the Doppler effect, *Bull. Seism. Soc. Am.*, *78*, 1367–72.

Doyle, H. (1995) *Seismology*, John Wiley & Sons, Chichester.

Dziewonski, A. M., and D. L. Anderson (1981) Preliminary reference Earth model, *Phys. Earth Planet. Inter.*, *25*, 297–356.

Dziewonski, A. M., T.-A. Chou, and J. H. Woodhouse (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. Geophys. Res.*, *86*, 2825–52.

Eakins, P. R. (1987) Faults and faulting, in C. K. Seyfert (ed.), *Encyclopedia of Structural Geology and Plate Tectonics*, Van Nostrand Reinhold, New York, pp. 228–39.

Eaton, J. P., D. H. Richter, and W. U. Ault (1961) The tsunami of May 23, 1960 on the island of Hawaii, *Bull. Seism. Soc. Am.*, *51*, 135–57.

Eckhouse, R. E., and L. R. Morris (1979) *Minicomputer Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Ekeland, I. (1993) *The Broken Dice*, University of Chicago Press, Chicago.

Engeln, J. F., and S. Stein (1984) Tectonics of the Easter plate, *Earth Planet. Sci. Lett.*, *68*, 259–70.

Engeln, J. F., D. A. Wiens, and S. Stein (1986) Mechanisms and depths of Atlantic transform earthquakes, *J. Geophys. Res.*, *91*, 548–77.

Engeln, J. F., S. Stein, J. Werner, and R. Gordon (1988) Microplate and shear zone models for oceanic spreading center reorganizations, *J. Geophys. Res.*, *93*, 2839–56.

England, P., and J. Jackson (1989) Active deformation of the continents, *Ann. Rev. Earth Planet. Sci.*, *17*, 197–226.

Evans, B., and T.-F. Wong (1992) *Fault Mechanics and Transport Properties of Rocks*, Academic Press, San Diego.

Evans, R. (1997) Assessment of schemes for earthquake prediction, *Geophys. J. Int.*, *131*, 413–20.

Ewing, W. M., W. S. Jardetsky, and F. Press (1957) *Elastic Waves in Layered Media*, McGraw-Hill, New York.

Few, A. A. (1980) Thunder, in *Atmospheric Phenomena*, W. H. Freeman, San Francisco, pp. 111–21.

Feynman, R. P. (1982) *The Character of Physical Law*, MIT Press, Cambridge MA.

Feynman, R. P. (1988) *What Do You Care What Other People Think?*, W. W. Norton, New York.

Feynman, R. P., R. B. Leighton, and M. Sands (1963) *The Feynman Lectures on Physics*, Addison-Wesley, Reading, MA.

Finlayson, D. M., J. H. Leven, and K. D. Wake-Dyster (1989) Large-scale lenticles in the lower crust under an intra-continental basin in eastern Australia, in R. F. Mereu, S. Mueller and D. M. Fountain (eds), *Properties and Processes of Earth's Lower Crust*, IUGG 6, Am. Geophys. Un., Washington, DC, pp. 1–16.

Fischman, J. (1992) Falling into the gap, *Discover*, Oct., 58–63.

Flesch, L. M., W. E. Holt, A. J. Haines, and B. Shen-Tu (2000) Dynamics of the Pacific-North American plate boundary zone in the western United States, *Science*, *287*, 834–6.

Forsyth, D. W. (1975) The early structural evolution and anisotropy of the oceanic upper mantle, *Geophys. J. R. Astron. Soc.*, *43*, 103–62.

Forsyth, D. W., and S. Uyeda (1975) On the relative importance of the driving forces of plate motion, *Geophys. J. R. Astron. Soc.*, *43*, 162–200.

Forsyth, D. W., *et al.* (1998) Imaging the deep seismic structure beneath a mid-ocean ridge: The MELT experiment, *Science*, *280*, 1215–18.

Fouch, M. J., K. M. Fischer, E. M. Parmentier, M. E. Wysession, and T. J. Clarke (2000) Shear wave splitting, continental keels, and patterns of mantle flow, *J. Geophys. Res.*, *105*, 6255–76.

Foulger, G. R. *et al.* (2001) Seismic tomography shows that upwelling beneath Iceland is confined to the upper mantle, *Geophys. J. Int.*, 146, 504–30.

Fountain, D. M., and N. I. Christensen (1989) Composition of the crust and upper mantle: a review, in L. C. Pakiser and W. D. Mooney (eds), *Geophysical Framework of the Continental United States*, Geol. Soc. Amer. Mem. 172, Boulder, Co, pp. 711–41.

Fowler, C. M. R. (1990) *The Solid Earth: an introduction to global geophysics*, Cambridge University Press, Cambridge.

Frankel, A., C. Mueller, T. Barnhard, D. Perkins, E. Leyendecker, N. Dickman, S. Hanson, and M. Hopper (1996) *National Seismic Hazard Maps Documentation*, US Geol. Survey, Open-File Rep. 96–532, US Government Printing Office, Washington, DC.

Franklin, J. N. (1968) *Matrix Theory*, Prentice-Hall, Englewood Cliffs, NJ.

Freedman, D., R. Pisani, R. Purves, and A. Adhikari (1991) *Statistics*, W. W. Norton, New York.

French, A. P. (1971) *Vibrations and Waves*, W. W. Norton, New York.

Freymueller, J. T., S. C. Cohen, and H. J. Fletcher (2000) Spatial variations in present-day deformation, Kenai Peninsula, Alaska, and their implications, *J. Geophys. Res.*, *105*, 8079–101.

Froberg, C. E. (1969) *Introduction to Numerical Analysis*, 2nd edn, Addison-Wesley, Reading, MA.

Frohlich, C. (1989) The nature of deep focus earthquakes, *Ann. Rev. Earth Planet. Sci.*, *17*, 227–54.

Fung, Y. C. (1965) *Foundations of Solid Mechanics*, Prentice-Hall, Englewood Cliffs, NJ.

Fung, Y. C. (1969) *A First Course in Continuum Mechanics*, Prentice-Hall, Englewood Cliffs, NJ.

Garnero, E. (2000) Heterogeneity of the lowermost mantle, *Ann. Rev. Earth Planet. Sci.*, *28*, 509–37.

Geller, R. J. (1976) Scaling relations for earthquake source parameters and magnitudes, *Bull. Seism. Soc. Am.*, *66*, 1501–23.

Geller, R. J. (1997) Earthquake prediction: A critical review, *Geophys. J. Int.*, *131*, 425–50.

Geller, R. J., and H. Kanamori (1977) Magnitudes of great shallow earthquakes from 1904 to 1952, *Bull. Seism. Soc. Am.*, *67*, 587–98.

Geller, R. J., and S. Stein (1977) Split free oscillation amplitudes for the 1960 Chilean and 1964 Alaskan earthquakes, *Bull. Seism. Soc. Am.*, *67*, 651–60.

Geller, R. J., and S. Stein (1978) Normal modes of a laterally heterogeneous body: A one dimensional example, *Bull. Seism. Soc. Am.*, *68*, 103–16.

Geller, R. J., D. D. Jackson, Y. Kagan, and F. Mulargia (1997) Earthquakes cannot be predicted, *Science*, *275*, 1616–17.

Gere, J. M., and H. C. Shah (1984) *Terra Non Firma: understanding and preparing for earthquakes*, W. H. Freeman, New York.

Geschwind, C.-H. (2001) *California Earthquakes: science, risk, and the politics of hazard mitigation*, Johns Hopkins University Press, Baltimore.

Gibson, R. L., and A. R. Levander (1988) Lower crustal reflectivity patterns in wide-angle seismic recordings, *Geophys. Res. Lett.*, *15*, 617–20.

Gledhill, K., and D. Gubbins (1996) SKS splitting and the seismic anisotropy of the mantle beneath the Hikurangi subduction zone, New Zealand, *Phys. Earth Planet. Inter.*, *95*, 227–36.

Gordon, R. G. (1998) The plate tectonic approximation: Plate non-rigidity, diffuse plate boundaries, and global reconstructions, *Ann. Rev. Earth Planet. Sci.*, *26*, 615–42.

Gordon, R. G., and S. Stein (1992) Global tectonics and space geodesy, *Science*, *256*, 333–42.

Green, H. W., II, and H. Houston (1995) The mechanics of deep earthquakes, *Ann. Rev. Earth Planet. Sci.*, *23*, 169–213.

Gregersen, S., and P. Basham (1989) *Earthquakes at North-Atlantic Passive Margins: neotectonics and post-glacial rebound*, Kluwer, Dordrecht.

Griffiths, D. H., and R. F. King (1981) *Applied Geophysics for Geologists and Engineers; the elements of geophysical prospecting*, Pergamon, Oxford.

Gubbins, D. (1990) *Seismology and Plate Tectonics*, Cambridge University Press, Cambridge.

Gurnis, M., M. E. Wysession, E. Knittle, and B. Buffett (eds) (1998) *The Core–Mantle Boundary Region*, Am. Geophys. Un., Washington, DC.

Gutenberg, Beno (1959) *Physics of the Earth's Interior*, Academic Press, London.

Hale, L. D., and G. A. Thompson (1982) The seismic reflection character of the continental Mohorovicic discontinuity, *J. Geophys. Res.*, *87*, 4625–35.

Hanks, T. C. (1997) Imperfect science: Uncertainty, diversity, and experts, *Eos Trans. Am. Geophys. Un.*, *78*, 369–77.

Hanks, T. C., and C. A. Cornell (1994) Probabilistic seismic hazard analysis: A beginner's guide, in *Proceedings of the Fifth Symposium on Current Issues Related to Nuclear Power Plant Structures, Equipment, and Piping*, I/1–1 to I/1–17, North Carolina State University, Raleigh, NC.

Hanks, T. C., and H. Kanamori (1979) A moment magnitude scale, *J. Geophys. Res.*, *84*, 2348–50.

Hannay, J. H. (1986) Intensity fluctuations from a one-dimensional random wavefront, in B. J. Uscinski (ed.), *Wave Propagation and Scattering*, Oxford University Press, Oxford, pp. 37–48.

Harkrider, D. G. (1988) The early years of computational seismology at Caltech, *Bull. Seism. Soc. Am.*, *78*, 2105–9.

Hasegawa, A., N. Umino, and A. Takagi (1978) Double-planed structure of the deep seismic zone in the north-eastern Japan arc, *Tectonophysics*, *47*, 43–58.

Hasegawa, H. S., and H. Kanamori (1987) Source mechanism of the magnitude 7.2 Grand Banks earthquake of November 1929; double couple or submarine landslide?, *Bull. Seism. Soc. Am.*, *77*, 1984–2004.

Hatton, L. (1983a) Computer science for geophysicists, part I: Elements of a seismic data processing system, *First Break*, *1*, June, 18–24.

Hatton, L. (1983b) Computer science for geophysicists, part II: Seismic computer system architecture, *First Break*, *1*, Sept., 18–22.

Hatton, L. (1983c) Computer science for geophysicists, part III: Operating systems, I/O and the interrupt, *First Break*, *1*, Oct., 13–19.

Hatton, L. (1983d) Computer science for geophysicists, part IV: The user-interface, *First Break*, *1*, Nov., 18–23.

Hatton, L. (1984a) Computer science for geophysicists, part V: Databases and expert systems, *First Break*, *2*, Jan., 9–15.

Hatton, L. (1984b) Computer science for geophysicists, part VI: Communications and networks, *First Break*, *2*, Sept., 9–17.

Hatton, L. (1985) Computer science for geophysicists, part VII: Form and structure in programming, *First Break*, *3*, April, 9–19.

Hatton, L., M. H. Worthington, and J. Makin (1986) *Seismic Data Processing*, Blackwell, Oxford.

Hay, G. E. (1953) *Vector and Tensor Analysis*, Dover, New York.

Heaton, T. H., and S. H. Hartzell (1988) Earthquake ground motions, *Ann. Rev. Earth Planet. Sci.*, *16*, 121–45.

Hedlin, M. A., P. M. Shearer, and P. S. Earle (1997) Seismic evidence for small-scale heterogeneity throughout the Earth's mantle, *Nature*, *387*, 145–50.

Helffrich, G., S. Stein, and B. Wood (1989) Subduction zone thermal structure and mineralogy and their relation to seismic wave reflections and conversions at the slab/mantle interface, *J. Geophys. Res.*, *94*, 753–63.

Helmberger, D. V., and L. J. Burdick (1979) Synthetic seismograms, *Ann. Rev. Earth Planet. Sci.*, *7*, 417–42.

Henrion, M., and B. Fischhoff (1986) Assessing uncertainty in physical constants, *Am. J. Phys.*, *54*, 791–8.

Hernandez, B., F. Cotton, M. Campillo, and D. Massonet (1997) A comparison between short-term (coseismic) and long-term (1 year) slip for the Landers earthquake: Measurements from strong motion and SAR interferometry, *Geophys. Res. Lett.*, *24*, 1579–82.

Hill, D. P. (1998) Science, geologic hazards, and the public in a large, restless caldera, *Seism. Res. Lett.*, *69*, 400–2.

Hindle, D., J. Kley, E. Klosko, S. Stein, T. Dixon, and E. Norabuena (2002) Consistency of geologic and geodetic displacements during Andean orogenesis, *Geophys. Res. Lett.*, *29(7)*, 10.1029/2001GL013757, 2002.

Hough, S., J. G. Armbruster, L. Seeber, and J. F. Hough (2000) On the Modified Mercalli intensities and magnitudes of the 1811/1812 New Madrid, central United States, earthquakes, *J. Geophys. Res.*, *105*, 23,839–64.

Howell, B. F., Jr. (1985) On the effect of too small a data base on earthquake frequency diagrams, *Bull. Seism. Soc. Am.*, *75*, 1205–7.

Huang, P. Y., and S. C. Solomon (1988) Centroid depths of mid-ocean ridge earthquakes: Dependence on spreading rate, *J. Geophys. Res.*, *93*, 13,445–77.

Hubbard, W. (1984) *Planetary Interiors*, Van Nostrand, New York.

Hudnut, K., *et al.* (1996) Coseismic displacements of the 1994 Northridge, California, earthquake, *Bull. Seism. Soc. Am.*, *86*, S19–36.

Hudson, J. A. (1980) *The Excitation and Propagation of Elastic Waves*, Cambridge University Press, Cambridge.

Humphreys, E., and R. Clayton (1988) Adaptation of back projection tomography to seismic travel time problems, *J. Geophys. Res.*, *93*, 1073–86.

Huygens, C. (1962 [1690]) *Treatise on Light*, trans. S. P. Thompson, Dover, New York.

Igarashi, G., S. Saeki, N. Takahata, K. Sumikawa, S. Tasaka, Y. Sasaki, M. Takahashi, and Y. Sano (1995) Ground-water radon anomaly before the Kobe earthquake in Japan, *Science*, *269*, 60–1.

Isacks, B., and M. Barazangi (1977) Geometry of Benioff zones: Lateral segmentation and downwards bending of the subducted lithosphere, in M. Talwani and W. C. Pitman, III (eds), *Island Arcs, Deep Sea Trenches and Back Arc Basins*, Maurice Ewing Ser., 1, Am. Geophys. Un., Washington, DC, pp. 99–114.

Jackson, D. D. (1972) Interpretation of inaccurate, insufficient, and inconsistent data, *Geophys. J. R. Astron. Soc.*, *28*, 97–109.

Jackson, D. D., and Y. Y. Kagan (1993) Reply, *J. Geophys. Res.*, *98*, 9919–20.

Jackson, I. (1993) Progress in the experimental study of seismic wave attenuation, *Ann. Rev. Earth Planet. Sci.*, *21*, 375–406.

Jackson, J., and D. McKenzie (1988) The relationship between plate motions and seismic moment tensors, and the rates of active deformation in the Mediterranean and Middle East, *Geophys. J. R. Astron. Soc.*, *93*, 45–73.

Jacobs, J. A. (1987) *The Earth's Core*, 2nd edn, Academic Press, London.

Jaeger, J. C. (1970) *Elasticity, Fracture and Flow, with Engineering and Geological Applications*, 3rd edn, Barnes & Noble, New York.

Jaeger, J. C., and N. G. W. Cook (1976) *Fundametals of Rock Mechanics*, Chapman and Hall, London.

Jarchow, C. M., and G. A. Thompson (1989) The nature of the Mohorovičić discontinuity, *Ann. Rev. Earth Planet. Sci.*, *17*, 475–506.

Jarosch, H., and E. Aboodi (1970) Towards a unified notation of source parameters, *Geophys. J. R. Astron. Soc.*, *21*, 513–29.

Jeanloz, R. (1990) The nature of the earth's core, *Ann. Rev. Earth Planet. Sci.*, *18*, 357–86.

Jeffreys, H. (1976) *The Earth: its origin, history, and physical constitution*, 6th edn, Cambridge University Press, Cambridge.

Jeffreys, H., and K. E. Bullen (1940) *Seismological Tables*, British Association Seismological Committee, London.

Jeffreys, H., and B. S. Jeffreys (1950) *Methods of Mathematical Physics*, Cambridge University Press, Cambridge.

Jost, M. L., and R. B. Hermann (1989) A student's guide to and review of moment tensors, *Seism. Res. Lett.*, *60*, 37–57.

Julian, B. R., and S. A. Sipkin (1985) Earthquake processes in the Long Valley Caldera area, California, *J. Geophys. Res.*, *90*, 11,155–70.

Kagan, Y. Y., and D. D. Jackson (1991) Seismic gap hypothesis: Ten years after, *J. Geophys. Res.*, *96*, 21,419–31.

Kanamori, H. (1970a) Synthesis of long-period surface waves and its application to earthquake source studies — Kurile Islands earthquake of October 13, 1963, *J. Geophys. Res.*, *75*, 5011–27.

Kanamori, H. (1970b) The Alaska earthquake of 1964: Radiation of long-period surface waves and source mechanism, *J. Geophys. Res.*, *75*, 5029–40.

Kanamori, H. (1977a) The energy release in great earthquakes, *J. Geophys. Res.*, *82*, 2981–7.

Kanamori, H. (1977b) Seismic and aseismic slip along subduction zones and their tectonic implications, in M. Talwani and W. C. Pitman, III (eds), *Island Arcs, Deep Sea Trenches and Back Arc Basins*, Maurice Ewing Ser., 1, Am. Geophys. Un., Washington, DC, pp. 163–74.

Kanamori, H. (1978) Quantification of earthquakes, *Nature*, *271*, 411–14.

Kanamori, H. (1986) Rupture process of subduction-zone earthquakes, *Ann. Rev. Earth Planet. Sci.*, *14*, 293–322.

Kanamori, H. (1988) Importance of historical seismograms for geophysical research, in W. H. K. Lee, H. Meyers and K. Shimizaki (eds), *Historical Seismograms and Earthquakes of the World*, Academic Press, San Diego, pp. 16–36.

Kanamori, H. (1994) Mechanics of earthquakes, *Ann. Rev. Earth Planet. Sci.*, *22*, 207–37.

Kanamori, H., and K. Abe (1968) Digital processing of surface waves and structure of island arcs, *J. Phys. Earth*, *16*, 137–40.

Kanamori, H., and D. L. Anderson (1975) Theoretical basis of some empirical relations in seismology, *Bull. Seism. Soc. Am.*, *65*, 1073–95.

Kanamori, H., and D. L. Anderson (1977) Importance of physical dispersion in surface wave and free oscillation problems: Review, *Rev. Geophys. Space Phys.*, *15*, 105–12.

Kanamori, H., and E. Boschi (1983) *Earthquakes: observation, theory, and interpretation*, Proc. Int. Sch. Phys. "Enrico Fermi", Course 85, North Holland, Amsterdam.

Kanamori, H., and J. J. Cipar (1974) Focal process of the great Chilean earthquake May 22, 1960, *Phys. Earth Planet. Inter.*, *9*, 128–36.

Kanamori, H., and J. W. Given (1981) Use of long-period surface waves for rapid determination of earthquake-source parameters, *Phys. Earth Planet. Inter.*, *27*, 8–31.

Kanamori, H., and J. W. Given (1982) Analysis of long-period seismic waves excited by the May 18, 1980 eruption of Mount St. Helens – a terrestrial monopole?, *J. Geophys. Res.*, *87*, 5422–32.

Kanamori, H., and G. S. Stewart (1976) Mode of the strain release along the Gibbs Fracture Zone, Mid-Atlantic Ridge, *Phys. Earth Planet. Inter.*, *11*, 312–32.

Kanamori, H., E. Hauksson, and T. H. Heaton (1997) Real-time seismology and earthquake hazard mitigation, *Nature*, *390*, 461–4.

Kanasewich, E. R. (1981) *Time Sequence Analysis in Geophysics*, University of Alberta Press, Edmonton.

Karato, S., and H. A. Spetzler (1990) Defect microdynamics in minerals and solid state mechanisms of seismic wave attenuation and velocity dispersion in the mantle, *Rev. Geophys.*, *28*, 399–421.

Kaula, W. M. (1975) The seven ages of a planet, *Icarus*, *26*, 1–15.

Kearey, P., and M. Brooks (1984) *An Introduction to Geophysical Exploration*, Blackwell, Oxford.

Kearey, P., and F. Vine (1990) *Global Tectonics*, Blackwell, Oxford.

Keller, E., and N. Pinter (1996) *Active Tectonics: earthquakes, uplift, and the landscape*, Prentice-Hall, Upper Saddle River, NJ.

Kendall, J. M., and P. G. Silver (1996) Constraints from seismic anisotropy on the nature of the lowermost mantle, *Nature*, *381*, 409–12.

Kennett, B. L. N. (1977) Towards a more detailed seismic picture of the oceanic crust and mantle, *Mar. Geophys. Res.*, *3*, 7–42.

Kennett, B. L. N. (1983) *Seismic Wave Propagation in Stratified Media*, Cambridge University Press, Cambridge.

Kennett, B. L. N., and E. R. Engdahl (1991) Traveltimes for global earthquake location and phase identification, *Geophys. J. Int.*, *105*, 429–65.

Kennett, B. L. N., E. R. Engdahl, and R. Buland (1995) Constraints on seismic velocities in the Earth from travel times, *Geophys. J. Int.*, *122*, 108–24.

Kernighan, B. W., and P. J. Plauger (1976) *Software Tools*, Addison-Wesley, Reading, MA.

Kernighan, B. W., and P. J. Plauger (1978) *The Elements of Programming Style*, McGraw-Hill, New York.

Kikuchi, M., and H. Kanamori (1991) Inversion of complex body waves – III, *Bull. Seism. Soc. Am.*, *81*, 2335–50.

Kirby, S. H. (1980) Tectonic stresses in the lithosphere: Constraints provided by the experimental deformation of rocks, *J. Geophys. Res.*, *85*, 6353–63.

Kirby, S. H. (1983) Rheology of the lithosphere, *Rev. Geophys. Space Phys.*, *21*, 1458–87.

Kirby, S. H., and A. K. Kronenberg (1987) Rheology of the lithosphere: Selected topics, *Rev. Geophys.*, *25*, 1219–44.

Kirby, S. H., E. R. Engdahl, and R. Denlinger (1996a) Intermediate-depth intraslab earthquakes and arc volcanism as physical expressions of crustal and uppermost mantle metamorphism in subducting slabs, in G. E. Bebout, D. W. Scholl, S. H. Kirby and J. P. Platt (eds), *Subduction: Top to Bottom*, Am. Geophys. Un., Washington, DC, pp. 195–214.

Kirby, S. H., S. Stein, E. A. Okal, and D. C. Rubie (1996b) Metastable phase transformations and deep earthquakes in subducting oceanic lithosphere, *Rev. Geophys.*, *34*, 261–306.

Klein, C., and C. Hurlbut, Jr. (1985) *Manual of Mineralogy*, John Wiley & Sons, Inc., New York.

Klein, M. V., and T. E. Furtak (1986) *Optics*, 2nd edn, John Wiley & Sons, Inc., New York.

Klosko, E., J. DeLaughter, and S. Stein (2000) Technology in introductory geophysics: The high-low mix, *Comp. Geosci.*, *26*, 693–8.

Kovach, R. L. (1995) *Earth's Fury: an introduction to natural hazards and disasters*, Prentice-Hall, Englewood Cliffs, NJ.

Krinitzsky, E. L., J. P. Gould, and P. H. Edinger (1993) *Fundamentals of Earthquake Resistant Construction*, John Wiley & Sons, New York.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Kuo, B. Y., D. W. Forsyth, and M. W. Wysession (1987) Lateral heterogeneity and azimuthal anisotropy in the North Atlantic determined from SS-S differential travel times, *J. Geophys. Res.*, *92*, 6421–36.

Lachenbruch, A. H., and J. H. Sass (1988) The stress-heat flow paradox and thermal results from Cajon Pass, *Geophys. Res. Lett.*, *15*, 981–4.

Lambeck, K. (1988) *Geophysical Geodesy: the slow deformations of the earth*, Clarendon Press, Oxford.

Lanczos, C. (1961) *Linear Differential Operators*, Van Nostrand, London.

Langston, C. A. (1978) Moments, corner frequencies, and the free surface, *J. Geophys. Res.*, *83*, 3422–6.

Lapwood, E. R., and T. Usami (1981) *Free Oscillations of the Earth*, Cambridge University Press, Cambridge.

Larson, K., R. Bürgmann, R. Bilham, and J. T. Freymueller (1999) Kinematics of the India-Eurasia collision zone from GPS measurements, *J. Geophys. Res.*, *104*, 1077–94.

Lay, T. (1992) Nuclear testing and seismology, in *The Encyclopedia of Earth System Science*, vol. 3, Academic Press, New York, pp. 333–51.

Lay, T. (1994) *Structure and Fate of Subducting Slabs*, Academic Press, New York.

Lay, T., and T. C. Wallace (1995) *Modern Global Seismology*, Academic Press, New York.

Lewis, B. T. R. (1978) Evolution of ocean crust seismic velocities, *Ann. Rev. Earth Planet. Sci.*, *6*, 377–404.

Liu, H.-P., D. L. Anderson, and H. Kanamori (1976) Velocity dispersion due to anelasticity: Implications for seismology and mantle composition, *Geophys. J. R. Astron. Soc.*, *47*, 41–58.

Liu, M., Y. Zhu, S. Stein, Y. Yang, and J. Engeln (2000) Crustal shortening in the Andes: Why do GPS rates differ from geological rates?, *Geophys. Res. Lett.*, *18*, 3005–8.

Lomnitz, C. (1989) Comment on "temporal and magnitude dependence in earthquake recurrence models" by C. A. Cornell and S. R. Winterstein, *Bull. Seism. Soc. Am.*, *79*, 1662.

Lomnitz, C. (1994) *Fundamentals of Earthquake Prediction*, Wiley, New York.

Lorenz, E. (1993) *The Essence of Chaos*, University of Washington Press, Seattle.

Lowrie, W. (1997) *Fundamentals of Geophysics*, Cambridge University Press, Cambridge.

Lundgren, P. R., and D. Giardini (1994) Isolated deep earthquakes and the fate of subduction in the mantle, *J. Geophys. Res.*, *99*, 15,833–42.

Madariaga, R. I. (1972) Toroidal free oscillations of the laterally heterogeneous Earth, *Geophys. J. R. Astron. Soc.*, *27*, 81–100.

Main, I. G. (1978) *Vibrations and Waves in Physics*, Cambridge University Press, Cambridge.

Main, I. (1996) Statistical physics, seismogenesis, and seismic hazard, *Rev. Geophys.*, *34*, 433–62.

Malvern, L. E. (1969) *Introduction to the Mechanics of a Continuous Medium*, Prentice-Hall, Englewood Cliffs, NJ.

Marion, J. B. (1970) *Classical Dynamics of Particles and Systems*, 2nd edn, Academic Press, New York.

Marone, C. (1998) Laboratory-derived friction laws and their application to seismic faulting, *Ann. Rev. Earth Planet. Sci.*, *26*, 643–96.

Mavko, G. M. (1981) Mechanics of motion on major faults, *Ann. Rev. Earth Planet. Sci.*, *9*, 81–111.

McCann, W. R., S. P. Nishenko, L. R. Sykes, and J. Krause (1979) Seismic gaps and plate tectonics: Seismic potential for major plate boundaries, *Pure Appl. Geophys.*, *117*, 1082–147.

McClusky, S., *et al.* (2000) Global positioning system constraints on plate kinematics and dynamics in the eastern Mediterranean and Caucasus, *J. Geophys. Res.*, *105*, 5695–719.

McElhinny, M. W. (ed.) (1979) *The Earth, Its Origin, Structure and Evolution*, Academic Press Inc., New York.

McKenzie, D. P. (1969) Speculations on the consequences and causes of plate motions, *Geophys. J. R. Astron. Soc.*, *18*, 1–32.

McKenzie, D. P., and F. M. Richter (1978) Simple plate models of mantle convection, *J. Geophys.*, *44*, 441–71.

Medawar, P. (1979) *Advice to a Young Scientist*, Basic Books, New York.

Meissner, R. (1986) *The Continental Crust*, Academic Press, Inc., San Diego.

Melchior, P. (1986) *The Physics of the Earth's Core*, Pergamon Press, Oxford.

Meltzer, A. S., A. R. Levander, and W. D. Mooney (1987) Upper crustal structure in the Livermore valley and vicinity, *Bull. Seism. Soc. Am.*, *77*, 1655–73.

Menard, H. W. (1986) *The Ocean of Truth: a personal history of global tectonics*, Princeton Series in Geology and Paleontology, ed. A. G. Fischer, Princeton University Press, Princeton, NJ.

Mendiguren, J. A. (1973) High resolution spectroscopy of the Earth's free oscillations knowing the earthquake source mechanism, *Science*, *179*, 179–80.

Menke, W. (1984) *Geophysical Data Analysis: discrete inverse theory*, Academic Press, Inc., Orlando, FL.

Menke, W., and D. Abbott (1990) *Geophysical Theory*, Columbia University Press, New York.

Michael, A. J., and R. J. Geller (1984) Linear moment tensor inversion for shallow thrust earthquakes combining first motion and surface wave data, *J. Geophys. Res.*, *89*, 1889–97.

Michaels, A., D. Malmquist, A. Knap, and A. Close (1997) Climate science and insurance risk, *Nature*, *389*, 225–7.

Minster, J. B., and T. H. Jordan (1978) Present-day plate motions, *J. Geophys. Res.*, *83*, 5331–54.

Minster, J. B., T. H. Jordan, P. Molnar, and E. Haines (1974) Numerical modeling of instantaneous plate tectonics, *Geophys. J. R. Astron. Soc.*, *36*, 541–76.

Mitchell, B. J. (1995) Anelastic structure and evolution of the continental crust and upper mantle from seismic surface wave attenuation, *Rev. Geophys.*, *33*, 441–62.

Mitchell, B. J., J. Xie, and S. Baqer (1997) *Lg Excitation, Attenuation, and Source Spectral Scaling in Central and Eastern North America*, Report to the Nuclear Regulatory Commission, NUREG/CR-6563, Washington, DC.

Molnar, P. (1988) Continental tectonics in the aftermath of plate tectonics, *Nature*, *335*, 131–7.

Mooney, W. D., and C. S. Weaver (1989) Regional crustal structure and tectonics of the Pacific coastal states: California, Oregon, and Washington, in L. C. Pakiser and W. D. Mooney (eds), *Geophysical Framework of the Continental United States*, Geol. Soc. Amer. Mem. 172, Boulder, Co, pp. 129–61.

Mooney, W. D., G. Laske, and T. G. Masters (1998) CRUST 5.1; a global crustal model at 5 degrees X5 degrees, *J. Geophys. Res.*, *103*, 727–47.

Moores, E. M., and R. J. Twiss (1995) *Tectonics*, W. H. Freeman, New York.

Morris, G. B., R. W. Raitt, and G. G. Shor, Jr. (1969) Velocity anisotropy and delay-time maps of the mantle near Hawaii, *J. Geophys. Res.*, *74*, 4300–16.

Morse, P. M., and H. Feshbach (1953) *Methods of Theoretical Physics*, McGraw-Hill, New York.

Nakamura, Y. (1983) Seismic velocity structure of the moon's upper mantle, *J. Geophys. Res.*, *88*, 677–86.

Nataf, H. C. (2000) Seismic imaging of mantle plumes, *Ann. Rev. Earth Planet. Sci.*, *28*, 391–417.

Nettles, M., and G. Ekström (1998) Faulting mechanism of anomalous earthquakes near Bardarbunga Volcano, Iceland, *J. Geophys. Res.*, *103*, 17,973–83.

Newman, A., S. Stein, J. Weber, J. Engeln, A. Mao, and T. Dixon (1999) Slow deformation and lower seismic hazard at the New Madrid Seismic Zone, *Science*, *284*, 619–21.

Newman, A., J. Schneider, S. Stein, and A. Mendez (2001) Uncertainties in seismic hazard maps for the New Madrid Seismic Zone and implications for seismic hazard communication, *Seism. Res. Lett.*, *72*, 653–67.

Ni, J., and M. Barazangi (1984) Seismotectonics of the Himalayan continental collision zone: geometry of the underthrusting Indian plate beneath the Himalayas, *J. Geophys. Res.*, *89*, 1147–64.

Nicolas, A. (1995) *The Mid-Oceanic Ridges*, Springer-Verlag, Berlin.

Nishenko, S. P., and L. R. Sykes (1993) Comment on "Seismic gap hypothesis: ten years after" by Kagan and Jackson, *J. Geophys. Res.*, *98*, 9909–16.

Nishimura, C., and D. Forsyth (1989) The anisotropic structure of the upper mantle in the Pacific, *Geophys. J. R. Astron. Soc.*, *96*, 203–26.

Noble, B. (1969) *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ.

Nolet, G. (1987) *Seismic Tomography*, D. Riedel, Dordrecht.

Norabuena, E., L. Leffler-Griffin, A. Mao, T. Dixon, S. Stein, I. S. Sacks, L. Ocala, and M. Ellis (1998) Space geodetic observations of Nazca-South America convergence along the Central Andes, *Science*, *279*, 358–62.

Officer, C. B. (1958) *Introduction to the Theory of Sound Transmission, with Application to the Ocean*, McGraw-Hill, New York.

Okada, Y. (1985) Surface deformation due to shear and tensile faults in a half-space, *Bull. Seism. Soc. Am.*, *75*, 1135–54.

Okal, E. A. (1992) A student's guide to teleseismic body wave amplitudes, *Seism. Res. Lett.*, *63*, 169–80.

Okal, E. A., and R. J. Geller (1979) On the observability of isotropic seismic sources: The July 31, 1970 Colombian earthquake, *Phys. Earth Planet. Inter.*, *18*, 176–96.

Okal, E. A., and B. Romanowicz (1994) On the variation of b-values with earthquake size, *Phys. Earth Planet. Inter.*, *87*, 55–76.

Oliver, J., and B. Isacks (1967) Deep earthquake zones, anomalous structures in the upper mantle, and the lithosphere, *J. Geophys. Res.*, *72*, 4259–75.

Owens, T. J., S. R. Taylor, and G. Zandt (1987) Crustal structure at regional seismic test network stations determined from the inversion of broadband teleseismic P waveforms, *Bull. Seism. Soc. Am.*, *77*, 631–62.

Pacheco, J., L. R. Sykes, and C. H. Scholz (1993) Nature of seismic coupling along simple plate boundaries of the subduction type, *J. Geophys. Res.*, *98*, 14,133–59.

Pakiser, L. C., and W. D. Mooney (eds.) (1989) *Geophysical Framework of the Continental United States*, Geol. Soc. Amer. Mem. 172, Boulder, Co.

Parker, R. L. (1977) Understanding inverse theory, *Ann. Rev. Earth Planet. Sci.*, *5*, 35–64.

Parsons, B., and F. M. Richter (1980) A relation between the driving force and geoid anomaly associated with mid-ocean ridges, *Earth Planet. Sci. Lett.*, *51*, 445–50.

Parsons, B., and J. G. Sclater (1977) An analysis of the variation of ocean floor bathymetry and heat flow with age, *J. Geophys. Res.*, *82*, 803–27.

Pearce, R. G. (1977) Fault plane solutions using the relative amplitudes of P and pP, *Geophys. J.*, *50*, 381–94.

Pearce, R. G. (1980) Fault plane solutions using the relative amplitudes of P and surface reflections: further studies, *Geophys. J.*, *60*, 459–87.

Peltier, W. R. (ed.) (1989) *Mantle Convection*, Gordon and Breach, New York.

Pho, H.-T., and L. Behe (1972) Extended distances and angles of incidence of P waves, *Bull. Seism. Soc. Am.*, *62*, 885–902.

Poirier, J.-P. (2000) *Introduction to the Physics of the Earth's Interior*, 2nd edn, Cambridge University Press, Cambridge.

Press, F., and R. Siever (1982) *Earth*, 3rd edn, W. H. Freeman, San Francisco.

Rabiner, L. R., and C. M. Rader (1972) *Digital Signal Processing*, IEEE Press, New York.

Ragan, D. M. (1968) *Structural Geology*, Wiley, New York.

Ranalli, G. (1987) *Rheology of the Earth*, Allen and Unwin, Boston.

Rebollar, C. J., L. Quintanar, J. Yamamoto, and A. Uribe (1999) Source process of the Chiapas, Mexico, intermediate-depth earthquake, *Bull. Seism. Soc. Am.*, *89*, 348–58.

Reiter, L. (1990) *Earthquake Hazard Analysis*, Columbia University Press, New York.

Reynolds, J. M. (1997) *An Introduction to Applied and Environmental Geophysics*, John Wiley & Sons, Chichester.

Rial, J. A., and V. F. Cormier (1980) Seismic waves at the epicenter's antipode, *J. Geophys. Res.*, *85*, 2661–8.

Richards, P. G., and J. Zavales (1990) Seismic discrimination of nuclear explosions, *Ann. Rev. Earth Planet. Sci.*, *18*, 257–86.

Richardson, W. P., S. Stein, C. Stein, and M. T. Zuber (1995) Geoid data and the thermal structure of the oceanic lithosphere, *Geophys. Res. Lett.*, *22*, 1913–16.

Richter, C. F. (1958) *Elementary Seismology*, W. H. Freeman, San Francisco.

Ringwood, A. E. (1975) *Composition and Petrology of the Earth's Mantle*, McGraw-Hill, New York.

Ringwood, A. E. (1979) Composition and origin of the Earth, in M. W. McElhinny (ed.), *The Earth, Its Origin, Structure and Evolution*, Academic Press Inc., New York, pp. 1–58.

Robbins, J. W., D. E. Smith, and C. Ma (1993) Horizontal crustal deformation and large scale plate motions inferred from space geodetic techniques, in D. E. Smith and D. L. Turcotte (eds), *Contributions of Space Geodesy to Geodynamics: crustal dynamics*, Geodynamics Series 23, Am. Geophys. Un., Washington, DC, pp. 21–36.

Robinson, E. A. (1983) *Migration of Geophysical Data*, International Human Resources Development Corp., Boston.

Robinson, E. A., and S. Treitel (1980) *Geophysical Signal Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

Roeloffs, E. A., and J. Langbein (1994) The earthquake prediction experiment at Parkfield, California, *Rev. Geophys.*, *32*, 315–36.

Romanowicz, B. (1991) Seismic tomography of the earth's mantle, *Ann. Rev. Earth Planet. Sci.*, *19*, 77–99.

Romanowicz, B. (1992) Strike-slip earthquakes on quasi-vertical transcurrent faults: Inferences for general scaling relations, *Geophys. Res. Lett.*, *19*, 481–4.

Romanowicz, B. (1995) A global tomographic model of shear attenuation in the upper mantle, *J. Geophys. Res.*, *100*, 12,375–94.

Romanowicz, B. (1998) Attenuation tomography of the Earth's mantle: A review of current status, *Pure App. Geophys.*, *153*, 257–72.

Romanowicz, B., and P. Guillemant (1984) An experiment in the retrieval of depth and source mechanism of large earthquakes using very long period Rayleigh-wave data, *Bull. Seism. Soc. Am.*, *74*, 417–37.

Rosendahl, B. R. (1987) Architecture of continental rifts with special reference to East Africa, *Ann. Rev. Earth Planet. Sci.*, *15*, 445–503.

Roth, E. G., D. A. Wiens, L. M. Dorman, J. Hildebrand, and S. C. Webb (1999) Seismic attenuation tomography of the Tonga-Fiji region using phase pair methods, *J. Geophys. Res.*, *104*, 4795–809.

Ruff, L., and H. Kanamori (1980) Seismicity and the subduction process, *Phys. Earth Planet. Inter.*, *23*, 240–52.

Sadigh, K., C.-Y. Chang, J. A. Egan, F. Makdisi, and R. R. Youngs (1997) Attenuation relationships for shallow crustal earthquakes based on California strong motion data, *Seism. Res. Lett.*, *68*, 180–9.

Sangree, J. B., and J. M. Widmier (1979) Interpretation of depositional facies from seismic data, *Geophysics*, *44*, 131–60.

Sarewitz, D., and R. Pielke, Jr. (2000) Breaking the global-warming gridlock, *Atlantic Monthly*, July, 56–64.

Sarewitz, D., R. Pielke, Jr., and R. Byerly, Jr. (2000) *Prediction: science, decision making, and the future of nature*, Island Press, Washington, DC.

Sato, H., and M. C. Fehler (1998) *Seismic Wave Propagation and Scattering in the Heterogeneous Earth*, Springer-Verlag, New York.

Savage, J. C. (1983) A dislocation model of strain accumulation and release at a subduction zone, *J. Geophys. Res.*, *88*, 4984–96.

Savage, J. C. (1991) Criticism of some forecasts of the national earthquake prediction council, *Bull. Seism. Soc. Am.*, *81*, 862–81.

Savage, J. C. (1993) The Parkfield prediction fallacy, *Bull. Seism. Soc. Am.*, *83*, 1–6.

Scherbaum, F. (1996) *Of Poles and Zeros*, Kluwer, Dordrecht.

Schneider, W. A. (1971) Developments in seismic data processing and analysis (1968–1970), *Geophysics*, *36*, 1043–73.

Scholz, C. H. (1990) *The Mechanics of Earthquakes and Faulting*, Cambridge University Press, Cambridge.

Segall, P., and J. Davis (1997) GPS applications for geodynamics and earthquake studies, *Ann. Rev. Earth Planet. Sci.*, *25*, 301–36.

Shearer, P. M. (1994) Imaging Earth's seismic response at long periods, *Eos Trans. Am. Geophys. Un.*, *75*, 449, 451, 452.

Shearer, P. M. (1996) Transition zone velocity gradients and the 520-km discontinuity, *J. Geophys. Res.*, *101*, 3053–66.

Shearer, P. M. (1999) *Introduction to Seismology*, Cambridge University Press, Cambridge.

Shedlock, K., D. Giardini, G. Grunthal, and P. Zhang (2000) The GSHAP global seismic hazard map, *Seism. Res. Lett.*, *71*, 679–86.

Sheriff, R. E., and L. P. Geldart (1982) *Exploration Seismology*, Cambridge University Press, Cambridge.

Shimazaki, K., and T. Nakata (1980) Time-predictable recurrence model for large earthquakes, *Geophys. Res. Lett.*, *7*, 279–82.

Sieh, K., and S. LeVay (1998) *The Earth in Turmoil: earthquakes, volcanos, and their impact on humankind*, W. H. Freeman, New York.

Sieh, K., M. Stuiver, and D. Brillinger (1989) A more precise chronology of earthquakes produced by the San Andreas fault in southern California, *J. Geophys. Res.*, *94*, 603–24.

Silver, P. G. (1996) Seismic anisotropy beneath the continents: Probing the depths of geology, *Ann. Rev. Earth Planet. Sci.*, *24*, 385–432.

Silver, P. G., R. W. Carlson, and P. Olson (1988) Deep slabs, geochemical heterogeneity, and the large-scale structure of mantle convection, *Ann. Rev. Earth Planet. Sci.*, *16*, 477–541.

Simon, R. B. (1981) *Earthquake Interpretations*, William Kaufmann, Inc., Los Altos, CA.

Sipkin, S. A., and T. H. Jordan (1979) Frequency dependence of Q_{ScS}, *Bull. Seism. Soc. Am.*, *69*, 1055–79.

Sleep, N. H. (1990) Hotspots and mantle plumes: Some phenomenology, *J. Geophys. Res.*, *95*, 6715–36.

Sleep, N. H. (1992) Hotspots and mantle plumes, *Ann. Rev. Earth Planet. Sci.*, *20*, 19–43.

Sleep, N. H., and K. Fujita (1997) *Principles of Geophysics*, Blackwell, Malden, MA.

Sleep, N. H., and B. R. Rosendahl (1979) Topography and tectonics of mid-oceanic ridge axes, *J. Geophys. Res.*, *84*, 6831–9.

Sloan, M. E. (1980) *Introduction to Minicomputers and Microcomputers*, Addison-Wesley, Reading, MA.

Smith, D. E., and D. L. Turcotte (1993) *Contributions of Space Geodesy to Geodynamics*, Geodynamics Ser. 23, Am. Geophys. Un., Washington, DC.

Smith, R. B., and L. W. Braile (1994) The Yellowstone hotspot, *J. Volcan. Geotherm. Res.*, *61*, 121–87.

Smithson, S. B. (1989) Contrasting types of lower crust, in R. F. Mereu, S. Mueller and D. M. Fountain (eds), *Properties and Processes of Earth's Lower Crust*, IUGG 6, Am. Geophys. Un., Washington, DC, pp. 53–63.

Snelson, C. M., T. J. Henstock, G. R. Keller, K. C. Miller, and A. Levander (1998) Crustal and uppermost mantle structure along the Deep Probe seismic profile, *Rocky Mountain Geology*, *33*, 181–98.

Snieder, R. K. (2001) *A Guided Tour of Mathematical Physics*, Cambridge University Press, Cambridge.

Snoke, J. A., I. S. Sacks, and D. E. James (1979) Subduction beneath western South America: Evidence from converted phases, *Geophys. J. R. Astron. Soc.*, *59*, 219–25.

Solomon, S. C., and N. C. Burr (1979) The relationship of source parameters of ridge-crest and transform earthquakes to the thermal structure of oceanic lithosphere, *Tectonophysics*, *55*, 107–26.

Solomon, S. C., and D. R. Toomey (1992) The structure of mid-ocean ridges, *Ann. Rev. Earth Planet. Sci.*, *20*, 329–64.

Song, X., and D. V. Helmberger (1993) Anisotropy of Earth's inner core, *Geophys. Res. Lett.*, *20*, 2591–4.

Spakman, W., and G. Nolet (1988) Imaging algorithms, accuracy and resolution in delay time tomography, in N. J. Vlaar, G. Nolet, M. J. R. Wortel and S. A. P. L. Cloetingh (eds), *Mathematical Geophysics*, Reidel, Dordrecht, pp. 155–87.

Spakman, W., N. J. Vlaar, and M. J. R. Wortel (1988) The Hellenic subduction zone: A tomographic image and its geodynamic implications, *Geophys. Res. Lett.*, *15*, 60–3.

Spakman, W., S. Stein, R. van der Hilst, and R. Wortel (1989) Resolution experiments for NW Pacific subduction zone tomography, *Geophys. Res. Lett.*, *16*, 1097–110.

Spudich, P., and J. Orcutt (1980) A new look at the seismic velocity structure of the oceanic crust, *Rev. Geophys. Space Phys.*, *18*, 627–45.

Stacey, F. D. (1992) *Physics of the Earth*, 3rd edn, Brookfield Press, Kenmore, Brisbane.

Stein, C. A., and S. Stein (1992) A model for the global variation in oceanic depth and heat flow with lithospheric age, *Nature*, *359*, 123–9.

Stein, C. A. and S. Stein (1993) Constraints on Pacific midplate swells from global depth-age and heat flow-age models, in M. Pringle, W. Sager, W. Sliter, and S. Stein (eds), *The Mesozoic Pacific*, Geophysical Monog. 77, Am. Geophys. Un., Washington, DC, pp. 53–76.

Stein, C. A., S. Stein, and A. M. Pelayo (1995) Heat flow and hydrothermal circulation, in S. Humphris, L. Mullineaux, R. Zierenberg and R. Thomson (eds), *Seafloor Hydrothermal Systems, Physical, Chemical, Biological, and Geological Interactions*, Geophys. Mono. 91, Am. Geophys. Un., Washington, DC, pp. 425–45.

Stein, R. S. (1999) The role of stress transfer in earthquake occurrence, *Nature*, *402*, 605–9.

Stein, R. S., and R. S. Yeats (1989) Hidden earthquakes, *Sci. Am.*, *260*, 48–57.

Stein, R. S., G. C. P. King, and J. Lin (1994) Stress triggering of the 1994 M = 6.7 Northridge, California, earthquake by its predecessors, *Science*, *265*, 1432–5.

Stein, S. (1978) An earthquake swarm on the Chagos-Laccadive Ridge and its tectonic implications. *Geophys. J. R. Astron. Soc.*, *55*, 577–88.

Stein, S. (1992) Seismic gaps and grizzly bears, *Nature*, *356*, 387–8.

Stein, S. (1993) Space geodesy and plate motions, in D. E. Smith and D. L. Turcotte (eds), *Contributions of Space Geodesy to Geodynamics: crustal dynamics*, Geodynamics Series 23, Am. Geophys. Un., Washington, DC, pp. 5–20.

Stein, S., and R. J. Geller (1978) Time-domain observation and synthesis of split spheroidal and torsional free oscillations of the 1960 Chilean earthquake: Preliminary results, *Bull. Seism. Soc. Am.*, *68*, 325–32.

Stein, S., and R. G. Gordon (1984) Statistical tests of additional plate boundaries from plate motion inversions, *Earth Planet. Sci. Lett.*, *69*, 401–12.

Stein, S., and E. R. Klosko (2002) Earthquake mechanisms and plate tectonics, in R. A. Meyers (ed.), *The Encyclopedia of Physical Science and Technology*, Academic Press, San Diego.

Stein, S., and G. C. Kroeger (1980) Estimating earthquake source parameters from seismological data, in S. Nemat-Nasser (ed.), *Solid Earth Geophysics and Geotechnology*, AMD Symp. Ser., 42, Amer. Soc. Mech. Engin., New York, pp. 61–71.

Stein, S., and A. Pelayo (1991) Seismological constraints on stress in the oceanic lithosphere, *Phil. Trans. R. Soc. London Ser. A*, *337*, 53–72.

Stein, S., and D. C. Rubie (1999) Deep earthquakes in real slabs, *Science*, *286*, 909–10.

Stein, S., and C. A. Stein (1996) Thermo-mechanical evolution of oceanic lithosphere: Implications for the subduction process and deep earthquakes, in G. E. Bebout, D. W. Scholl, S. H. Kirby and J. P. Platt (eds), *Subduction: top to bottom*, Am. Geophys. Un., Washington, DC, pp. 1–17.

Stein, S., and D. Wiens (1986) Depth determination for shallow teleseismic earthquakes: Methods and results, *Rev. Geophys. Space Phys.*, *24*, 806–32.

Stein, S., and D. F. Woods (1989) Seismicity: Midocean ridge, in D. E. James (ed.), *The Encyclopedia of Solid Earth Geophysics*, Van Nostrand Reinhold, New York, pp. 1050–4.

Stein, S., N. H. Sleep, R. J. Geller, S. C. Wang, and G. C. Kroeger (1979) Earthquakes along the passive margin of eastern Canada, *Geophys. Res. Lett.*, *6*, 537–40.

Stein, S., J. M. Mills, Jr., and R. J. Geller (1981) Q^{-1} models from data space inversion of fundamental spheroidal mode attenuation measurements, in Stacey *et al.* (eds), *Anelasticity in the Earth*, Geodynamics Series, 4, Am. Geophys. Un., Washington, DC, pp. 39–53.

Stein, S., J. F. Engeln, C. DeMets, R. G. Gordon, D. Woods, P. Lundgren, D. Agrus, C. Stein, and D. A. Wiens (1986) The Nazca–South America convergence rate and the recurrence of the great 1960 Chilean earthquake, *Geophys. Res. Lett.*, *13*, 713–16.

Stein, S., S. Cloetingh, N. Sleep, and R. Wortel (1989) Passive margin earthquakes, stresses, and rheology, in S. Gregerson and P. Basham (eds), *Earthquakes at North-Atlantic Passive Margins: neotectonics and postglacial rebound*, Kluwer, Dordrecht, pp. 231–60.

Stixrude, L. (1998) Elastic constants and anisotropy of $MgSiO_3$ perovskite, periclase, and SiO_2 at high pressure, in M. Gurnis, M. E. Wysession, E. Knittle and B. Buffett (eds), *The Core-Mantle Boundary Region*, Am. Geophys. Un., Washington, DC, pp. 83–96.

Stixrude, L., and R. E. Cohen (1995) High-pressure elasticity of iron and anisotropy of Earth's inner core, *Science*, *267*, 1972–5.

Strahler, A. N. (1969) *Physical Geography*, John Wiley, New York.

Su, W., R. L. Woodward, and A. M. Dziewonski (1994) Degree 12 model of shear velocity heterogeneity in the mantle, *J. Geophys. Res.*, *99*, 6945–80.

Sykes, L. R., and D. M. Davis (1987) The yields of Soviet strategic weapons, *Sci. Am.*, *256*, 29–37.

Sykes, L. R., and S. P. Nishenko (1984) Probabilities of occurrence of large plate rupturing earthquakes for the San Andreas, San Jacinto, and Imperial Faults, California, 1983–2003, *J. Geophys. Res.*, *89*, 5905–27.

Sykes, L. R., B. E. Shaw, and C. H. Scholz (1999) Rethinking earthquake prediction, *Pure Appl. Geophys.*, *155*, 207–32.

Talandier, J., and E. A. Okal (1979) Human perception of T waves: The June 22, 1977 Tonga earthquake felt on Tahiti, *Bull. Seism. Soc. Am.*, *69*, 1475–86.

Taner, M. T., and F. Kohler (1969) Velocity spectra — digital computer derivation and application of velocity spectra, *Geophysics*, *34*, 859–81.

Tappin, D., *et al.* (1999) Sediment slump likely caused Papua New Guinea tsunami, *Eos Trans. Am. Geophys. Un.*, *80*, 329–34.

Tapponnier, P., G. Peltzer, A. Le Dain, R. Armijo, and P. Cobbold (1982) Propagating extrusion tectonics in Asia: New insights from simple experiments with plasticine, *Geology*, *10*, 611–16.

Tatham, R. (1989) Tau-p filtering, in P. L. Stoffa (ed.), *Tau-p, a Plane Wave Approach to the Analysis of Seismic Data*, Kluwer, Dordrecht, pp. 35–70.

Telford, W. M., L. P. Geldart, R. E. Sheriff, and D. A. Keys (1976) *Applied Geophysics*, Cambridge University Press, Cambridge.

Thatcher, W. (1983) Nonlinear strain buildup and the earthquake cycle on San Andreas fault, *J. Geophys. Res.*, *88*, 5893–902.

Thio, H. K., and H. Kanamori (1996) Source complexity of the 1994 Northridge, California, earthquake and its relation to aftershock mechanisms, *Bull. Seism. Soc. Am.*, *86*, S84–92.

Thurber, C. H., and K. Aki (1987) Three dimensional seismic imaging, *Ann. Rev. Earth Planet. Sci.*, *15*, 115–39.

Toomey, D. R., S. C. Solomon, and G. M. Purdy (1988) Microearthquakes beneath the median valley of the Mid-Atlantic Ridge near 23°N: Tomography and tectonics, *J. Geophys. Res.*, *93*, 9093–112.

Torge, W. (1991) *Geodesy*, de Gruyter, Berlin.

Triep, E. G., and L. R. Sykes (1997) Frequency of occurrence of moderate to great earthquakes in intracontinental regions: Implications for changes in stress, earthquake prediction, and hazard assessments, *J. Geophys. Res.*, *102*, 9923–48.

Tromp, J. (1993) Support for anisotropy of the Earth's inner core from free oscillations, *Nature*, *366*, 678–81.

Tsai, Y.-B., and K. Aki (1970) Precise focal depth determination from amplitude spectra of surface waves, *J. Geophys. Res.*, *75*, 5729–43.

Tse, S. T., and J. R. Rice (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties, *J. Geophys. Res.*, *91*, 9452–72.

Turcotte, D. L. (1991) Earthquake prediction, *Ann. Rev. Earth Planet. Sci.*, *19*, 263–82.

Turcotte, D. L. (1992) *Fractals and Chaos in Geology and Geophysics*, Cambridge University Press, Cambridge.

Turcotte, D. L., and G. Schubert (1982) *Geodynamics: applications of continuum physics to geological problems*, John Wiley, New York.

Udias, A. (1999) *Principles of Seismology*, Cambridge University Press, Cambridge.

Usselman, T. N. (1975) Experimental approach to the state of the core, I: The liquidus relations of the Fe-rich portion of the Fe-Ni-S system from 30 to 100 kb, *Am. J. Sci.*, *275*, 291–303.

Uyeda, S. (1978) *The New View of the Earth*, W. H. Freeman, San Francisco.

van der Hilst, R. D., and W. Spakman (1989) Importance of the reference model in linearized tomography and images of subduction below the Caribbean plate, *Geophys. Res. Lett.*, *16*, 1093–6.

van der Hilst, R. D., S. Widiyantoro, K. C. Creager, and T. J. McSweeney (1998) Deep subduction and aspherical variations in P-wavespeed at the base of earth's mantle, in M. Gurnis, M. E. Wysession, E. Knittle, and B. Buffett (eds), *The Core-Mantle Boundary Region*, Am. Geophys. Un., Washington, DC, pp. 5–20.

van der Lee, S., and G. Nolet (1997) Upper mantle S velocity structure of North America, *J. Geophys. Res.*, *102*, 22,815–38.

Vassiliou, M. S. (1984) The state of stress in subducting slabs as revealed by earthquakes analysed by moment tensor inversion, *Earth Planet. Sci. Lett.*, *69*, 195–202.

Vassiliou, M. S., B. H. Hager, and A. Raefsky (1984) The distribution of earthquakes with depth and stress in subducting slabs, *J. Geodynam.*, *1*, 11–28.

Vera, E. E., J. C. Mutter, P. Buhl, J. A. Orcutt, A. J. Harding, M. E. Kappus, R. S. Detrick, and T. M. Brocher (1990) The structure of 0–0.2-m.y.-old oceanic crust at 9°N on the East Pacific Rise from expanded spread profiles, *J. Geophys. Res.*, *95*, 15,529–56.

Verhoogen, J. (1980) *Energetics of the Earth*, National Academy of Sciences, Washington, DC.

Vidale, J. E., and H. M. Benz (1992) Upper-mantle seismic discontinuities and the thermal structure of subduction zones, *Nature*, *356*, 678–83.

Von Huene, R., L. D. Kulm, and J. Miller (1985) Structure of the frontal part of the Andean continental margin, *J. Geophys. Res.*, *90*, 5429–42.

Walck, M. C. (1984) The P-wave upper mantle structure beneath an active spreading center: The Gulf of California, *Geophys. J. R. Astron. Soc.*, *76*, 697–723.

Wald, D. J., H. Kanamori, D. V. Helmberger, and T. H. Heaton (1993) Source study of the 1906 San Francisco earthquake, *Bull. Seism. Soc. Am.*, *83*, 981–1019.

Wald, D. J., T. H. Heaton, and K. Hudnut (1996) The slip history of the 1994 Northridge, California, earthquake determined from strong-motion, teleseismic, GPS, and leveling data, *Bull. Seism. Soc. Am.*, *86*, S49–70.

Wallace, T. C. (1985) A reexamination of the moment tensor solutions of the 1980 Mammoth Lakes earthquakes, *J. Geophys. Res.*, *90*, 11,171–6.

Wallace, T. C., A. Velasco, J. Zhang, and T. Lay (1991) Broadband seismological investigation of the 1989 Loma Prieta earthquake, *Bull. Seism. Soc. Am.*, *81*, 1622–46.

Ward, P. D., and D. Brownlee (2000) *Rare Earth*, Copernicus Press, New York.

Ward, S. (1989) Tsunamis, in D. E. James (ed.), *The Encyclopedia of Solid Earth Geophysics*, Van Nostrand-Reinhold, New York, pp. 1279–92.

Waters, K. H. (1981) *Reflection Seismology: a tool for energy resource exploration*, John Wiley, New York.

Weber, J., S. Stein, and J. Engeln (1998) Estimation of strain accumulation in the New Madrid seismic zone from repeat Global Positioning System surveys, *Tectonics*, *17*, 250–66.

Weertman, J., and J. R. Weertman (1975) High temperature creep of rock and mantle viscosity, *Ann. Rev. Earth Planet. Sci.*, *3*, 293–315.

Weidner, D. J. (1986) Mantle model based on measured physical properties of minerals, in S. K. Saxena (ed.), *Chemistry and Physics of Terrestrial Planets*, Advances in Physical Geochemistry, 6, Springer-Verlag, New York, pp. 251–74.

Wells, D. L., and K. J. Coppersmith (1994) New empirical relations among magnitude, rupture length, rupture width, rupture area and surface displacement, *Bull. Seism. Soc. Am.*, *84*, 974–1002.

Widmer, R., G. Masters, and F. Gilbert (1992) Observably-split multiplets – data analysis and interpretation in terms of large-scale aspherical structure, *Geophys. J. R. Astron. Soc.*, *111*, 559–76.

Wiegel, R. L. (ed.) (1970) *Earthquake Engineering*, Prentice-Hall, Englewood Cliffs, NJ.

Wiemer, S., and M. Wyss (1997) Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times?, *J. Geophys. Res.*, *102*, 15,115–28.

Wiens, D. A., and S. Stein (1983) Age dependence of oceanic intraplate seismicity and implications for lithospheric evolution, *J. Geophys. Res.*, *88*, 6455–68.

Wiens, D. A., and S. Stein (1984) Intraplate seismicity and stresses in young oceanic lithosphere, *J. Geophys. Res.*, *89*, 11,442–64.

Wiens, D. A., and S. Stein (1985) Implications of oceanic intraplate seismicity for plate stresses, driving forces and rheology, *Tectonophysics*, *116*, 143–62.

Wiens, D. A., *et al.* (1985) A diffuse plate boundary model for Indian Ocean tectonics, *Geophys. Res. Lett.*, *12*, 429–32.

Wiggins, R. A. (1972) The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure, *Rev. Geophys. Space Phys.*, *10*, 251–85.

Wilson, J. T. (1976) *Continents Adrift and Continental Aground*, W. H. Freeman, San Francisco.

Wong, J., N. Bregman, G. West, and P. Hurley (1987) Cross-hole seismic scanning and tomography, *Leading Edge*, *6*, 36–41.

Wood, B. J., and D. G. Fraser (1977) *Elementary Thermodynamics for Geologists*, Oxford University Press, Oxford.

Woodhouse, J. H., and A. M. Dziewonski (1984) Mapping the upper mantle: Three dimensional modeling of Earth structure by inversion of seismic waveforms, *J. Geophys. Res.*, *89*, 5953–86.

Woods, M. T., and E. A. Okal (1987) Effect of variable bathymetry on the amplitude of teleseismic tsunamis: A ray-tracing experiment, *Geophys. Res. Lett.*, *14*, 765–8.

Wyllie, P. J. (1971) *The Dynamic Earth: textbook in geosciences*, John Wiley, New York.

Wysession, M. E. (1996a) Imaging cold rock at the base of slabs: The sometimes fate of slabs?, in G. E. Bebout, D. W. Scholl, S. H. Kirby, and J. P. Platt (eds), *Subduction: top to bottom*, Am. Geophys. Un., Washington, DC, pp. 369–84.

Wysession, M. E. (1996b) Large-scale structure at the core–mantle boundary from core-diffracted waves, *Nature*, *382*, 244–8.

Wysession, M. E., and P. J. Shore (1994) Visualization of whole mantle propagation of seismic shear energy using normal mode summation, *Pure Appl. Geophys.*, *142*, 295–310.

Wysession, M. E., J. Wilson, L. Bartko, and R. Sakata (1995) Intraplate seismicity in the Atlantic Ocean basin: A teleseismic catalog, *Bull. Seism. Soc. Am.*, *85*, 755–74.

Wysession, M. E., B. C. Hicks, D. A. Wiens, and P. J. Shore (1996) Determining the frequency–magnitude–depth relations of seismicity in the Tonga subduction zone. *Seism. Res. Lett.*, *67*, 62.

Wysession, M. E., T. Lay, J. Revenaugh, Q. Williams, E. J. Garnero, R. Jeanloz, and L. H. Kellogg (1998) Implications of the D″ discontinuity, in M. Gurnis, M. E. Wysession, E. Knittle, and B. Buffett (eds), *The Core–Mantle Boundary Region*, Am. Geophys. Un., Washington, DC, pp. 273–97.

Wyss, M., and R. Koyanagi (1992) Seismic gaps in Hawaii, *Bull. Seism. Soc. Am.*, *82*, 1373–87.

Wyss, M., R. K. Aceves, and S. K. Park (1997) Cannot earthquakes be predicted?, *Science*, *278*, 487–90.

Yeats, R. S., K. Sieh, and C. R. Allen (1997) *The Geology of Earthquakes*, Oxford University Press, New York.

Yilmaz, O. (1987) *Seismic Data Processing*, Society of Exploration Geophysicists, Tulsa, OK.

Young, C. J., and T. Lay (1990) Multiple phase analysis of the shear velocity structure in the D″ region beneath Alaska, *J. Geophys. Res.*, *95*, 17,385–402.

Young, G. B., and L. W. Braile (1976) A computer program for the application of Zoeppritz's amplitude equations and Knott's energy equations, *Bull. Seism. Soc. Am.*, *66*, 1881–6.

Youngs, R. R., and K. J. Coppersmith (1985) Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates, *Bull. Seism. Soc. Am.*, *75*, 939–64.

Yu, G.-K., and B. Mitchell (1979) Regionalized shear velocity models of the Pacific upper mantle from observed Rayleigh and Love wave dispersion, *Geophys. J. R. Astron. Soc.*, *57*, 311–42.

Zhao, L., T. Jordan, and C. Chapman (2000) Three-dimensional Frechet differential kernels for seismic delay times, *Geophys. J. Int.*, *141*, 558–76.

Zoback, M. L. (1992) First and second order patterns of stress in the lithosphere: The world stress map project, *J. Geophys. Res.*, *97*, 11,703–2104.

Solutions to selected odd-numbered problems

Note that in many problems (as in reality), the solution varies depending on the interpretation of the data or the assumptions used.

Chapter 2

(1) $R_{12}=0, T_{12}=1$.

(3a) $(3,-2,5)$.

(3b) $(2,1,3)$.

(3c) $(5,3,0)/\sqrt{14}$.

(5a) $\sigma_1=2, \sigma_2=0, \sigma_3=-2$; $\mathbf{n}^{(1)}=(1,1,0)/\sqrt{2}$, $\mathbf{n}^{(2)}=(0,0,1)$, $\mathbf{n}^{(3)}=(1,-1,0)/\sqrt{2}$.

(5b) $\tau=2$. Planes have normals $(1,0,0)$ and $(0,1,0)$.

(7b) -150 kbar.

(7c)
$$D=\begin{bmatrix} 0 & -2 & 1 \\ -2 & -5 & 3 \\ 1 & 3 & 5 \end{bmatrix}.$$

(7d) 450 km.

(9) 2%.

(11a)
$$\begin{bmatrix} 6\lambda+6\mu & 0 & 0 \\ 0 & 6\lambda+2\mu & 2\mu \\ 0 & 2\mu & 6\lambda+4\mu \end{bmatrix}.$$

(11b) $18\lambda+16\mu$.

(13a) $[(\lambda+2\mu)/\rho]^{1/2}$.

(13b) $(\mu/\rho)^{1/2}$.

(15) $\sqrt{3}$.

(17a) $\bar{\alpha}=11.25$ km/s; $\bar{\beta}=6.18$ km/s.

(17b) $\bar{\alpha}/\bar{\beta}=1.82$.

(19a) 0.8 km, 8 km, 800 km.

(19b) 0.000125 s, 0.125 s, 12.5 s; 8000 Hz, 8 Hz, 0.08 Hz.

(21) $i_2=13°$; $i_3=17°$; $i_c=37°$.

(23a) For the $i_1=0°$ wave: $i_2=0°$, $l_1=2$ km, $l_2=2$ km, $T=3.3$ s. For the $i_1=30°$ wave: $i_2=49°$, $l_1=2.3$ km, $l_2=3.0$ km, $T=4.3$ s.

(23b) For the $i_1=0°$ wave: $\mathbf{s}_1=(0,1)$ s/km, $|\mathbf{s}_1|=1/v_1$, $\mathbf{s}_2=(0,2/3)$ s/km, $|\mathbf{s}_2|=1/v_2$. For the $i_1=30°$ wave: $\mathbf{s}_1=(0.5,\sqrt{3}/2)$ s/km, $|\mathbf{s}_1|=1/v_1$, $\mathbf{s}_2=(0.5,0.44)$ s/km, $|\mathbf{s}_2|=1/v_2$.

(25a)
$$\Psi_I=B_1\exp[i(\omega t-k_x x+k_x r_\beta z)],$$
$$\Psi_R=B_2\exp[i(\omega t-k_x x-k_x r_\beta z)],$$
$$\Phi_R=A_2\exp[i(\omega t-k_x x-k_x r_\alpha z)].$$

(25b)
$$\sigma_{xz}=\mu\left(2\frac{\partial^2\Phi}{\partial z\partial x}+\frac{\partial^2\Psi}{\partial x^2}-\frac{\partial^2\Psi}{\partial^2 z}\right)=0,$$
$$\sigma_{zz}=\lambda\left(\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial z^2}\right)+2\mu\left(\frac{\partial^2\Phi}{\partial z^2}+\frac{\partial^2\Psi}{\partial x\partial z}\right)=0.$$
$$2r_\alpha A_2+(1-r_\beta)B_1+(1-r_\beta)B_2=0,$$
$$A_2(\lambda+\lambda r_\alpha^2+2\mu r_\alpha^2)-2\mu r_\beta B_1+2\mu r_\beta B_2=0.$$

(25c) $B_2/B_1=-1$, $A_2/B_1=0$.

(25d)
$$\frac{|\mathbf{u}|_{SR}}{|\mathbf{u}|_{SI}}=\frac{|B_2|}{|B_1|};\quad \frac{|\mathbf{u}|_{PR}}{|\mathbf{u}|_{SI}}=\frac{\beta|A_2|}{\alpha|B_1|}.\quad \frac{\dot{\mathrm{E}}_{SR}}{\dot{\mathrm{E}}_{SI}}=\frac{B_2^2}{B_1^2},\quad \frac{\dot{\mathrm{E}}_{PR}}{\dot{\mathrm{E}}_{SI}}=\frac{\eta_\alpha A_2^2}{\eta_\beta B_1^2}.$$

(27a) For *ScS*: $j_{\mathrm{slab}}=26°$, $j_{\mathrm{surf}}=4°$. For *ScSp*: $i_{\mathrm{slab}}=50°$, $i_{\mathrm{surf}}=20°$.

(27b) 50 km.

(29) $\omega_1=0.58$, $\omega_2=1.16$.

(33) $(3/2)(5\cos^2\theta\sin\theta-\sin\theta)$.

(35a) 2591 s.

(35b) 4.4 km/s.

(35c) 8.3 km/s, 13.8 km/s, 74.5 km/s.

(35d) $c=5.36$ km/s, 5.01 km/s, 4.05 km/s; $\lambda=11{,}440$ km, 1312 km, 307 km.

(37a) $\Delta\omega/\omega$ observed: 0.095.

(37b) $\Delta\omega/\omega$ predicted: 0.038.

(C-3) For *P* waves at the CMB, $T_{mc}=0.975$, $R_{mc}=-0.025$, $T_{cm}=1.025$, $R_{cm}=0.025$, $\dot{\mathrm{E}}_{\mathrm{R}}/\dot{\mathrm{E}}_{\mathrm{I}}=0.0006$, $\dot{\mathrm{E}}_{\mathrm{T}}/\dot{\mathrm{E}}_{\mathrm{I}}=0.9994$. For *S* waves at the CMB, $T_{mc}=2$, $R_{mc}=1$, $\dot{\mathrm{E}}_{\mathrm{R}}/\dot{\mathrm{E}}_{\mathrm{I}}=1$, $\dot{\mathrm{E}}_{\mathrm{T}}/\dot{\mathrm{E}}_{\mathrm{I}}=0$.

Chapter 3

(1) $\alpha_0=5.7$ km/s, $\alpha_1=7.8$ km/s, $h_0=23$ km.

(3a) $\alpha_c=6.7$ km/s, $\alpha_m=7.8$ km/s.

(3b) 3.1 km.

(3c) 6.1 km.

(5) $\alpha_c=6.5$ km/s, $\alpha_m=8$ km/s, dip $=4°$, $h_u=50$ km, $h_d=30$ km.

(11) 24,000,000.

(13a) 9.34 km/s.

(13b) 11.24 km/s.

(15) For $D = 0$ km: $p_{40} = 8.3$ s/degree, $p_{60} = 6.9$ s/degree, $i_{40} = 26°$, $i_{60} = 21°$. For $D = 600$ km: $p_{40} = 7.9$ s/degree, $p_{60} = 6.6$ s/degree, $i_{40} = 52°$, $i_{60} = 41°$.

(17a) 4.5 s/degree.

(17b) 13.4 km/s.

(17c) *SKKS*.

(19a) $t_{1/e}({}_0T_2) = 58.3$ hr, $t_{1/e}({}_0T_{30}) = 3.0$ hr, $t_{1/e}({}_0S_{30}) = 4.2$ hr.

(19b) 54,300 km for ${}_0T_{30}$, 76,450 km for ${}_0S_{30}$.

(21a) 3%.

(21b) 0.3%.

(23) $M_c = 1.94 \times 10^{24}$ kg, $\bar{\rho}_c = 11$ g/cm³.

Chapter 4

(3) Earthquake a: $(\phi, \delta, \lambda)_1 = (310°, 65°, 90°)$ (thrust); $(\phi, \delta, \lambda)_2 = (130°, 25°, 90°)$ (thrust); P axis (azimuth, plunge) = (40°, 20°); T axis = (220°, 70°); B axis = (130°, 0°).
Earthquake b: $(\phi, \delta, \lambda)_1 = (176°, 80°, 195°)$ (right-lateral strike-slip); $(\phi, \delta, \lambda)_2 = (83°, 75°, 350°)$ (left-lateral strike-slip); P axis (azimuth, plunge) = (40°, 18°); T axis = (309°, 3°); B axis = (209°, 72°).
Earthquake c: $(\phi, \delta, \lambda)_1 = (9°, 90°, 180°)$ (right-lateral strike-slip); $(\phi, \delta, \lambda)_2 = (99°, 90°, 0°)$ (left-lateral strike-slip); P axis (azimuth, plunge) = (234°, 0°); T axis = (144°, 0°); B axis = (undefined, 90°).
Earthquake d: First solution: $(\phi, \delta, \lambda)_1 = (16°, 85°, 90°)$ (dip slip); $(\phi, \delta, \lambda)_2 = (196°, 5°, 90°)$ (thrust); P axis (azimuth, plunge) = (106°, 40°); T axis = (286°, 50°); B axis = (196°, 0°). Second solution: $(\phi, \delta, \lambda)_1 = (78°, 66°, 25°)$ (left-lateral strike-slip); $(\phi, \delta, \lambda)_2 = (337°, 67°, 154°)$ (right-lateral strike-slip); P axis (azimuth, plunge) = (28°, 1°); T axis = (297°, 34°); B axis = (119°, 56°).

(7) 0.

(9a)
$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M_{zz} \end{pmatrix} = \begin{pmatrix} M_{zz}/3 & 0 & 0 \\ 0 & M_{zz}/3 & 0 \\ 0 & 0 & M_{zz}/3 \end{pmatrix} + \begin{pmatrix} -M_{zz}/3 & 0 & 0 \\ 0 & -M_{zz}/3 & 0 \\ 0 & 0 & 2M_{zz}/3 \end{pmatrix}.$$

(9b)
$$\begin{pmatrix} -2.14 & 0 & 0 \\ 0 & 2.01 & 0 \\ 0 & 0 & 0.13 \end{pmatrix} = \begin{pmatrix} -2.075 & 0 & 0 \\ 0 & 2.075 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -0.065 & 0 & 0 \\ 0 & -0.065 & 0 \\ 0 & 0 & 0.13 \end{pmatrix}.$$

(Double-couple scalar moment)/(original scalar moment) = 0.999.
(CLVD scalar moment)/(original scalar moment) = 0.054.

(9c) There are two solutions:

Solution 1:

$$\begin{pmatrix} -2.14 & 0 & 0 \\ 0 & 2.01 & 0 \\ 0 & 0 & 0.13 \end{pmatrix} = \begin{pmatrix} -1.135 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1.135 \end{pmatrix} + \begin{pmatrix} -1.005 & 0 & 0 \\ 0 & 2.01 & 0 \\ 0 & 0 & -1.005 \end{pmatrix}.$$

(Double-couple scalar moment)/(original scalar moment) = 0.546.
(CLVD scalar moment)/(original scalar moment) = 0.838.

Solution 2:

$$\begin{pmatrix} -2.14 & 0 & 0 \\ 0 & 2.01 & 0 \\ 0 & 0 & 0.13 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.94 & 0 \\ 0 & 0 & -0.94 \end{pmatrix} + \begin{pmatrix} -2.14 & 0 & 0 \\ 0 & 1.07 & 0 \\ 0 & 0 & 1.07 \end{pmatrix}.$$

(Double-couple scalar moment)/(original scalar moment) = 0.452.
(CLVD scalar moment)/(original scalar moment) = 0.892.

(11) 4.24 mm/yr, 0.85 mm/yr, 0.42 mm/yr.

(13) $M_S = 5.2$.

(15) assuming $\mu = 3 \times 10^{11}$; 200,000 km; 43,333 km.

(17) ~0.04 Hz.

(19b) 0.003–0.03.

(21) Japan: 8 mo. ($M \geq 6$); 7 yr ($M \geq 7$); 65 yr ($M \geq 8$). S. California: 1 yr ($M \geq 6$); 8 yr ($M \geq 7$); ≈ 100 yr ($M \geq 8$). New Madrid: 92 yr ($M \geq 6$); 920 yr ($M \geq 7$); 9200 yr ($M \geq 8$).

(C-1) Earthquake a: $\hat{\mathbf{n}} = (0.453, -0.785, 0.423)$;
$\hat{\mathbf{d}} = (0.098, 0.515, 0.852)$.
Earthquake b: $\hat{\mathbf{n}} = (0.853, -0.150, 0.500)$;
$\hat{\mathbf{d}} = (0.492, -0.087, -0.866)$.
Earthquake c: $\hat{\mathbf{n}} = (0.853, -0.150, 0.500)$;
$\hat{\mathbf{d}} = (-0.492, 0.087, 0.866)$.
Earthquake d: $\hat{\mathbf{n}} = (-0.633, -0.754, 0.174)$;
$\hat{\mathbf{d}} = (0.758, -0.559, 0.337)$.
Earthquake e: $\hat{\mathbf{n}} = (-0.633, -0.754, 0.173)$;
$\hat{\mathbf{d}} = (-0.758, 0.559, -0.337)$.

(C-3) Earthquake a: $\begin{pmatrix} 0.088 & 0.157 & 0.427 \\ 0.157 & -0.808 & -0.451 \\ -0.427 & -0.451 & 0.720 \end{pmatrix}$.

Earthquake b: $\begin{pmatrix} 0.840 & -0.148 & -0.492 \\ -0.148 & 0.026 & 0.087 \\ -0.492 & 0.087 & -0.866 \end{pmatrix}$.

Earthquake c: $\begin{pmatrix} -0.840 & 0.148 & 0.492 \\ 0.148 & -0.026 & -0.087 \\ 0.492 & -0.087 & 0.866 \end{pmatrix}$.

Earthquake d: $\begin{pmatrix} -0.960 & -0.218 & -0.082 \\ -0.218 & 0.843 & -0.351 \\ -0.082 & -0.351 & 0.117 \end{pmatrix}$.

Earthquake e: $\begin{pmatrix} 0.960 & 0.218 & 0.082 \\ 0.218 & -0.843 & 0.351 \\ 0.082 & 0.351 & -0.117 \end{pmatrix}$.

(C-7b) 0.95.

(C-9a) Gaussian: 0.1%; Poisson: 4%.

(C-9b) Gaussian: 0.3%; Poisson: 3%.

(C-9c) Gaussian: 0.5%; Poisson: 2%.

(C-11a) $\tau = 21.8$ yr; $\sigma = 7.2$ yr; Poisson: $p = 37\%$; Gaussian: C(1993, 1985) = 64%.

(C-11b) $\tau = 21.8$ yr; $\sigma = 1.5$ yr; Poisson: $p = 37\%$; Gaussian: C(1993, 1985) = 99%.

(C-11c) $\tau = 25.5$ yr; $\sigma = 11.1$ yr; Poisson: $p = 31\%$; Gaussian: C(2018, 2010) = 82%.

(C-11d) $\tau = 27.2$ yr; $\sigma = 14.7$ yr; Poisson: $p = 29\%$; Gaussian: C(2028, 2020) = 74%.

Chapter 5

(1a) 0.77 m; $M_w = 6.8$, length = 31 km.

(1b) 4.62 m; $M_w = 7.8$, length = 240 km.

(3a) 40 mm/yr.

(3b) 125, 250, 500 yr.

(3c) For 25%: 500, 1000, 2000 yr; for 50%: 250, 500, 1000 yr.

(3e) $M_w \approx 8.4$; $M_0 = 5 \times 10^{28}$ dyn-cm.

(5) 6×10^{31} dyn-cm; M_w 10.5.

(7a) 47 mW/m^2.

(7b) 33 mW/m^2.

(7c) 84 mW/m^2.

(9) ~1 Ga.

(11) −21.5 bar/°C.

(13) $vL^3/(24\kappa^2 t)$; 28 (for $v = 10$ cm/yr and $t = 150$ Ma).

(17) 58°; 251 MPa.

(C-1b) San Andreas: 46 mm/yr at 324°; Aleutian: 53 mm/yr at 346°.

(C-3b) $(\theta, \phi, |\omega|) = (-63.0°, 107.4°, 0.641$ °/my).

(C-3c) Hawaii: 66 mm/yr at 299°.

Chapter 6

(1a) $a_0 = 0, a_k = 0, b_k = (2/k\pi)(1 - \cos(k\pi))$.

(1b) $a_0 = 0, a_k = 0, b_k = -\cos(k\pi)/k\pi = (-1)^{k+1}/k\pi$.

(3a) −1.

(3b) $4i$.

(3c) $-i$.

(3d) $1.5 + 2.6i$.

(7a) $\pi e^{-i\pi/2}[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)]$.

(9a) $a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2$.

(9b) $a^2v^2\sigma_u^2 + a^2u^2\sigma_v^2 + 2a^2uv\sigma_{uv}^2$.

(9c) $(a^2/v^2)\sigma_u^2 + (a^2u^2/v^4)\sigma_v^2 - 2(a^2u/v^3)\sigma_{uv}^2$.

(9d) $a^2b^2u^{(2b-2)}\sigma_u^2$.

(11a) $v\Delta t/(2\cos i)$.

(11b) 10 km.

(11c) $(\sigma_v^2\Delta t^2 + \sigma_{\Delta t}^2 v^2 + \sigma_i^2 v^2 \Delta t^2 \tan^2 i)/4\cos^2 i$.

(11d) 4 km.

Appendix

(1) 21°.

(5) $a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$.

(7) $A^{-1} = \begin{pmatrix} -2/3 & 1/3 \\ 5/6 & -1/6 \end{pmatrix}$.

(13a) $6xy^2 + 2x^3 + 2$.

(13b) (0, 0, 0).

(13c) $(6y^2 + 6x^2, 12xy, 0)$.

(13d) $xz + x^3y^2 + y^2 +$ constant.

(15) Hint: use Eqn A.7.4.

(C-5) $|(1, 4, 2)| = \sqrt{21}$; $|(2, 3, 1)| = \sqrt{14}$; sum = (3, 7, 3); $\mathbf{a} \cdot \mathbf{b} = 16$; $\mathbf{a} \times \mathbf{b} = (-2, 3, -5)$; $\theta = 21.1°$.

(C-7b) (0, −1, 1).

(C-9b) $\begin{pmatrix} 0.7 & -0.1 & 0.3 \\ 0.5 & -0.5 & 0.5 \\ -0.8 & 0.4 & -0.2 \end{pmatrix}$.

(C-11b) 1, 3, 4.

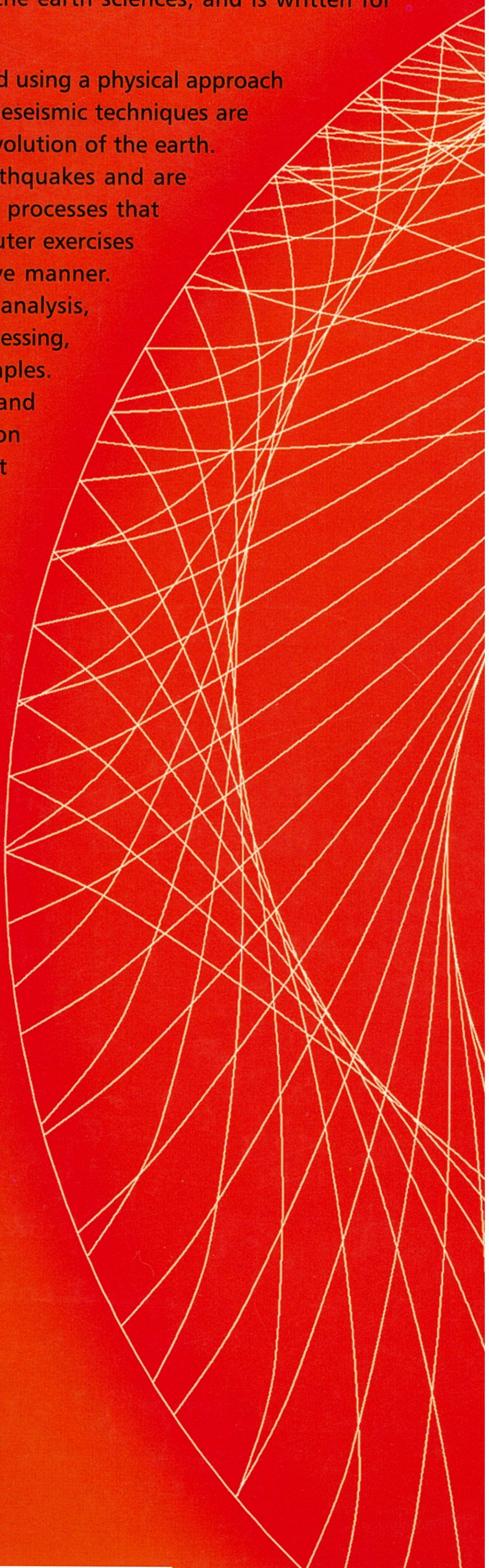
(C-13b) i: $\Delta = 93°$, $\zeta = 48°$.
ii: $\Delta = 54°$, $\zeta = 49°$.
iii: $\Delta = 87°$, $\zeta = 148°$.
iv: $\Delta = 35°$, $\zeta = 232°$.

Index

This book is an introduction to seismology and its role in the earth sciences, and is written for advanced undergraduate and beginning graduate students.

The fundamentals of seismic wave propagation are developed using a physical approach and then applied to show how refraction, reflection, and teleseismic techniques are used to study the structure and thus the composition and evolution of the earth. The book shows how seismic waves are used to study earthquakes and are integrated with other data to investigate the plate tectonic processes that cause earthquakes. Figures, examples, problems, and computer exercises teach students about seismology in a creative and intuitive manner. Necessary mathematical tools including vector and tensor analysis, matrix algebra, Fourier analysis, statistics of errors, signal processing, and data inversion are introduced with many relevant examples. The text also addresses the fundamentals of seismometry and applications of seismology to societal issues. Special attention is paid to help students visualize connections between different topics and view seismology as an integrated science.

An Introduction to Seismology, Earthquakes, and Earth Structure gives an excellent overview for students of geophysics and tectonics, and provides a strong foundation for further studies in seismology.



Seth Stein is Professor of Geological Sciences at Northwestern University. He has received the James B Macelwane Medal of the American Geophysical Union, been elected a Fellow of the American Geophysical Union and Geological Society of America, and named to the Institute for Scientific Information Highly Cited Researchers list. He served as Scientific Director of the University Navstar Consortium and on the Incorporated Research Institutions for Seismology's Executive Committee, and started Northwestern's Environmental Science program.

Michael Wysession is an Associate Professor in the Department of Earth and Planetary Sciences at Washington University. He is the recipient of the Packard Foundation and NSF Presidential Faculty Fellowships for his research into the structure of the Earth's deep interior.

Titles of related interest

An Introduction to Geophysical Exploration, 3rd edn
Philip Kearey, Michael Brooks & Ian Hill
2002, 272 pages, 295 illustrations
0 632 04929 4

Principles of Geophysics
NH Sleep & K Fujita
1997, 608 pages, 259 illustrations
0 865 42076 9

Cover design by Code 5 Design Associates Ltd

ISDN 0-86542-078-5
90000>

http://www.blackwellpublishing.com