Advances in GEOPHYSICS

VOLUME 49

Advances in

GEOPHYSICS

VOLUME 49

This page intentionally left blank

Advances in
GEOPHYSICS

VOLUME 49

Edited by

RENATA DMOWSKA
Harvard University
Cambridge, MA
USA

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
84 Theobald's Road, London WC1X 8RR, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2008

ISBN: 978-0-12-374231-5

ISSN: 0065-2687

For information on all Academic Press publications
visit our website at books.elsevier.com

Printed and bound in USA

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

CONTENTS

Chapter 3

Seismic Ray Tracing and Wavefront Tracking in Laterally Heterogeneous Media

N. RAWLINSON, J. HAUSER AND M. SAMBRIDGE

This page intentionally left blank

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin

HAUSER, J. (203) Research School of Earth Sciences, Australian National University, Canberra ACT 0200, Australia

MIGNAN, A. (67) Science & Technology Research, Risk Management Solutions, Peninsular House, 30 Monument Street, London EC3R 8NB, UK and Laboratoire Tectonique, Institut de Physique du Globe de Paris, 4, place Jussieu, 75252 Paris Cedex 05, France

OKAL, E.A. (1) Department of Geological Sciences, Northwestern University, Evanston, IL 60201, USA

RAWLINSON, N. (203) Research School of Earth Sciences, Australian National University, Canberra ACT 0200, Australia

SAMBRIDGE, M. (203) Research School of Earth Sciences, Australian National University, Canberra ACT 0200, Australia

This page intentionally left blank

ADVANCES IN GEOPHYSICS, VOL. 49, CHAPTER 1

THE GENERATION OF *T* WAVES BY EARTHQUAKES

EMILE A. OKAL

Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL 60201, USA

ABSTRACT

T waves propagate in the so-called SOFAR channel of minimum sound velocity acting as a waveguide for acoustic energy in the world's oceans. They can be excited by sources in the solid Earth such as earthquakes through conversion of seismic energy into acoustic waves at the solid–liquid interfaces. We present a historical perspective of the investigations of such conversions. In the context of geometrical optics, a sloping interface provides a mechanism for the penetration of the SOFAR channel after a series of reflections in the liquid wedge. This process, known as "downslope conversion", successfully explained many characteristics of earthquake-generated *T* waves, but has severe limitations, notably regarding "abyssal" *T* phases, generated under flat oceanic basins. We review theoretical developments based on mode theory which describe coupling between elastic and acoustic modes under scattering by structural heterogeneities located at the ocean bottom, and which are becoming increasingly successful at modeling the waveshapes of abyssal *T* phases.

As a particular form of seismic wave emanating from an earthquake, *T* waves can provide insight into seismic sources in the oceanic environment. We review the application of *T* waves to the detection of small earthquakes in marine basins, discuss the retrieval of seismic source properties from *T*-phase waveforms, and show that several algorithms combining measurements of their amplitude and duration can yield information on source rupture, and more specifically detect the presence of source slowness. In particular, anomalously slow earthquakes such as the so-called "tsunami earthquakes" are poor *T*-wave generators, and more generally, *T*-phase amplitudes and tsunami generation are not found to correlate.

In the context of the Comprehensive Nuclear-Test Ban Treaty, hydroacoustics has been recognized as a monitoring technology, and the deployment of state-of-the-art receivers at eleven sites will significantly improve long-range detection capabilities and open up new opportunities for the investigation of acoustic sources, including earthquakes, in the oceanic environment.

ISSN: 0065-2687
DOI: 10.1016/S0065-2687(07)49001-X

1. INTRODUCTION

This paper examines the generation, by earthquake sources, of the so-called T waves guided in the water body of the world's oceans through a channel of minimum sound velocity generally centered around 1200 meters depth. It discusses the mechanisms of conversion of elastic energy in the solid Earth into acoustic energy in the water, describes the potential for improving our knowledge of the seismicity of remote ocean provinces using T waves, and shows that recent developments in the quantification of teleseismic T phases can help shed new light on the high-frequency properties of earthquake source spectra.

The variation of sound velocity with depth in the ocean was originally investigated in order to convert shipboard acoustic measurements into accurate depth soundings (e.g., Langevin, 1924; Kuwahara, 1939). In simple terms, the velocity of acoustic waves in water is controlled by the thermodynamic variables describing its state, namely composition, pressure and temperature. Under ambient oceanic conditions, the sound velocity increases with pressure, at a rate of about $1.8 \times 10^{-6}\ \mathrm{m\,s^{-1}\,Pa^{-1}}$, with temperature at a rate of $2.1\ \mathrm{m\,s^{-1}\,K^{-1}}$, and with salinity at a rate of $1.3\ \mathrm{m\,s^{-1}}$ per part per 1000 (Sverdrup *et al.*, 1942). In the oceanic column, the pressure is hydrostatic, increasing linearly with depth, which translates into a positive velocity gradient of $0.018\ \mathrm{s^{-1}}$, or 1.8 m/s per 100 m of oceanic column. The effect of temperature is far more complex, since the thermal gradient in the world's oceans is highly variable, both spatially and seasonally (Teague *et al.*, 1990; Levitus *et al.*, 1994). In most areas, the temperature drops significantly through the thermocline layer, and in particular in the first 150 m, and then much more slowly at greater depths. Salinity can have a highly variable behavior in the thermocline, and stabilizes around 35 parts per thousand in the deeper ocean. As a whole, the effect of temperature on sound velocity prevails in the thermocline and that of pressure in the deep ocean, creating a minimum in sound velocity at around 1200 m, where the resulting low-velocity channel can trap acoustic energy with wavelengths shorter than its width, in practice with frequencies greater than 2.5 Hz. The existence of this waveguide was recognized and its structure investigated early on (Swainson, 1936), but its full potential for efficient long-range propagation was realized and thoroughly studied only during World War II, as summarized upon post-war declassification by Ewing *et al.* (1946), who coined the acronym SOFAR (for **SO**und **F**ixing **A**nd **R**anging) for the technique and, by extension, for the low-velocity channel itself.

Irrespective of the presence of the SOFAR channel, it had long been recognized, apparently ever since Leonardo da Vinci (Urick, 1975), that the propagation of sound in water is particularly efficient, especially as compared to propagation in the atmosphere. Modern experimental research (Urick, 1963; Thorp, 1965; Urick, 1966) has indeed documented the virtual absence of anelastic attenuation in seawater in the 1–100 Hz frequency range, which serves to further enhance the

efficiency of the SOFAR channel, and sets the finite size of the ocean basins as the only physical limit to the range of SOFAR propagation.

Despite the variability of the detailed characteristics of the sound velocity profile (depth of SOFAR axis, minimum velocity, reciprocal depth, etc.), the existence of the channel is a quasi-universal feature, the only possible exceptions involving the cold waters of the extreme Southern Ocean from which the thermocline is absent. Its ability to propagate energy from exceptionally small sources over remarkably large distances has found applications world-wide in fields as varied as the tracking of submarines, the discovery and monitoring of distant volcanoes (Dietz and Sheehy, 1954; Norris and Johnson, 1969; Talandier and Okal, 1982, 1996; Fox *et al.*, 1995), the detection of explosions in the oceanic environment, notably in the context of the Comprehensive Nuclear-Test Ban Treaty (CTBT) (Milne, 1959; Adams, 1979; Okal, 2001a; Wallace and Koper, 2002; Reymond *et al.*, 2003; Talandier and Okal, 2004a), the detection of iceberg collisions in the Southern Ocean (Talandier *et al.* 2002; 2006), the study of the vocalization patterns of large cetaceans (Reysenbach de Haan, 1966; Stafford *et al.*, 1998), and the monitoring of global warming (Munk *et al.*, 1994).

While SOFAR propagation is limited to the oceanic column, acoustic energy in the water can be transformed to and from elastic energy at the solid–liquid interfaces marking the boundaries of an oceanic basin, and this mechanism, whose degree of complexity can vary widely, allows both the excitation of acoustic energy by sources in the solid Earth, such as earthquakes and underground explosions, and its recording by land-based seismic stations located near the shore or even, under exceptional circumstances, hundreds of kilometers away from the coastlines (Båth and Shahidi, 1971; Cansi and Béthoux, 1985; Pasyanos and Romanowicz, 1997). Indeed, when the seismic wave resulting from the receiver-side conversion is of sufficient amplitude, it can be felt by the population close to the shore, even though the source of the acoustic energy may be many thousands of kilometers away. Examples include the underwater nuclear explosion WIGWAM on 14 May 1955, felt in Hawaii (J.P. Eaton, pers. comm., 1979) and even in Torishima, 8750 km from the source (Wadati, 1960); the Fairweather, Alaska earthquake on 10 July 1958 felt in Hawaii (J.P. Eaton, pers. comm., Eaton, 1979); the Tonga earthquake on 22 June 1977 felt in Tahiti (Talandier and Okal, 1979); the South Island, New Zealand earthquake of 21 August 2003 felt in Sydney, Australia (Leonard, 2004); and as discussed below, the great 2004 Sumatra earthquake felt in the Maldives (A.C. Yalçıner, pers. comm., 2005). In addition, *T* waves from certain Venezuelan earthquakes are routinely felt in Puerto Rico at a somewhat shorter distance (not exceeding 900 km) (C.G. von Hillebrandt-Andrade, pers. comm., 1998). Note that coupling is also possible, at least in principle, with the atmospheric column; however, the mechanics of the conversion of air waves into acoustic phases are poorly understood, due to the scant number of adequate sources, consisting exclusively of large atmospheric nuclear tests in the early 1960s (Talandier and Okal, 2001, 2004b).

As discussed more in detail below, earthquake-generated hydroacoustic wavetrains were first identified as a far-field phase on seismograms (if not correctly interpreted) in 1935 by scientists at Harvard Observatory upon recording of Caribbean earthquakes (Collins, 1936), and given the name T group (for "*Tertiary*" arrival, following P, primary and S, secondary) during informal discussions of these phases by the Harvard staff (Leet *et al.*, 1951). To our knowledge, this name first appeared in print in Linehan (1940).

When T waves were later correctly identified as water-borne waves guided by the SOFAR channel, a major problem arose, as the excitation of T waves trapped in the channel from earthquake sources located in the solid Earth, and thus by necessity outside the waveguide, is theoretically impossible in the framework of geometrical optics applied to a simple flat-layered Earth model. Indeed, all acceptable sound velocity profiles predict that only rays inclined less than $\sim 12^\circ$ on the horizontal can be trapped in the channel, while the sharp contrast in seismic and acoustic velocities in the source region precludes the penetration of the oceanic column at all but the steepest incidences.

In this framework, the present paper offers a largely chronological review of five decades of observational developments and theoretical efforts, seeking to resolve the apparent paradox of the ubiquitous observation of efficient excitation and propagation of far-field T phases from earthquake sources, through a series of modifications to the simple flat-layered model based on geometrical optics. It further discusses various efforts at quantifying the energy present in T phases, including recent developments that use their characteristics to explore the dynamics of the seismic source.

The discussion of the excitation of T waves by acoustic sources in the water column, such as man-made underwater explosions, magmatic episodes of volcanism delivering magma to the ocean floor, or the generation of cryosignals during collision between large icebergs, is intentionally left out of the present review.

2. Geometrical Optics: Understanding T Waves in a Simple Context

Early Observations

To our knowledge, the first published report of a teleseismic T wavetrain goes back to Jaggar (1930), who describes high-frequency oscillations recorded at Hawaii Volcano Observatory (HVO) in the coda of a major earthquake which took place on 24 October 1927 on the Fairweather Fault of the Alaska panhandle (Fig. 1). However, Jaggar interprets the record as local volcanic tremor "possibly touched off by the big earthquake waves". A modern examination of the spindle-shaped waveform, frequency content and timing of the phase (quoted as 06:20 a.m. HST or 16:50 GMT) definitely identifies it as the T wave of the

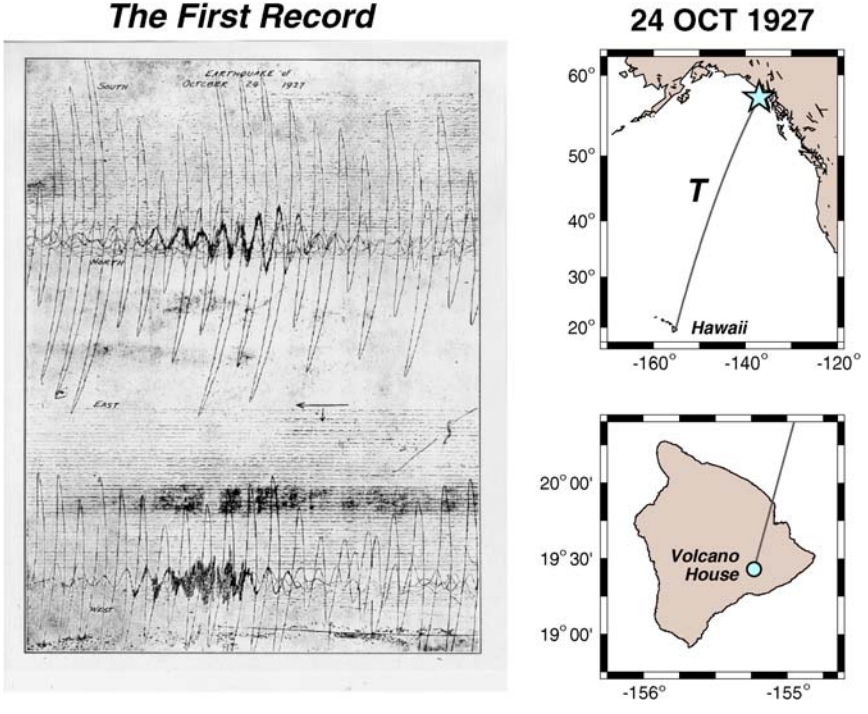


FIG. 1. *Left*: Seismic recording of the Alaskan earthquake of 24 October 1927 at Hawaii Volcano Observatory, after Jaggar (1930). This is believed to be the first published record of a teleseismic *T* phase. The top trace is the S–N component, the lower one, the E–W component; time increases from right to left and from top to bottom. The window is approximately 200 seconds long. *Right*: Maps sketching the transpacific path of the *T* phase from source (star) to receiver, and the location of the receiver station inside the "Big" Island of Hawaii. Adapted from Jaggar (1930).

Alaskan earthquake. Note that Jaggar mentions that the phase was "feebly felt" in Hawaii, as would later be the larger 1958 earthquake, in a very similar geometry.

Several years later, Collins (1936) noted in the Seismological Bulletin of Harvard Observatory that the record of the Caribbean earthquake of 15 September 1935 featured a third arrival, following P and S, detected exclusively on short-period channels. This constitutes the first correct interpretation of a T wave as the independent phase of a distant earthquake. Our relocation of that earthquake based on travel-times published in the International Seismological Summary (ISS) and the algorithm of Wysession *et al.* (1991) converges on 19.03°N, 64.85°W and a depth of 64 km, although the Monte Carlo ellipsoid (run with a standard deviation $\sigma_G = 4$ s for the Gaussian noise injected into the dataset) does intersect the Earth's surface. It is nevertheless probable that the earthquake occurred at depth in the Puerto Rico trench as suggested by reports of weak surface waves, and in general agreement with the local patterns of seismicity (Fischer and McCann, 1984). Based on this hypocenter, we have verified that the time given by Collins (1936) as the emergence of the "third group" (04:28:20 GMT) supports its interpretation as a T phase converted to a seismic wave approximately 3°South of the receiver. Subsequent detections of T phases are also reported in the Harvard Bulletin for several Caribbean earthquakes, notably on 18 September and 10 November 1935 (the latter relocating about 40 km West of Montserrat), and 12 December 1936.

Such observations were analyzed in greater detail by Linehan (1940) who only offered speculation as to their nature; in particular, his published illustration of the waveshapes (Fig. 2) appears to be no more than a hand-drawn rendition of their timing. More insightful is Ravet's (1940) contemporaneous but obviously independent report on very short period signals recorded in Tahiti in the wake of distant major earthquakes: the author correctly establishes their association with the epicenter, their generation by the seismic source and their propagation along the surface of the Earth as opposed to through its body. He does estimate a group velocity of 1.5 km/s, but stops short of identifying the latter as the velocity of sound in water, which had been known since the work of Beudant in the Mediterranean Sea in the 1810s and the landmark experiments of Colladon and Sturm (1827) in Lake Geneva.

Indeed, the correct interpretation of T waves as water-borne phases of the seismic source would come only in the wake of the considerable progress in hydroacoustics achieved during World War II, as summarized in a special volume of the Geological Society of America, in which Ewing and Worzel (1948) published the basics of long range propagation in the SOFAR channel, and Pekeris (1948) developed elementary models of the structure of T waves, based on modal theory. These works were complemented by the experimental results of Worzel and Ewing (1948). Simultaneously and independently, Brekhovskikh (1949) published a model for the reverberation of sound in the channel, based on experimental results by Rozenberg (1949).

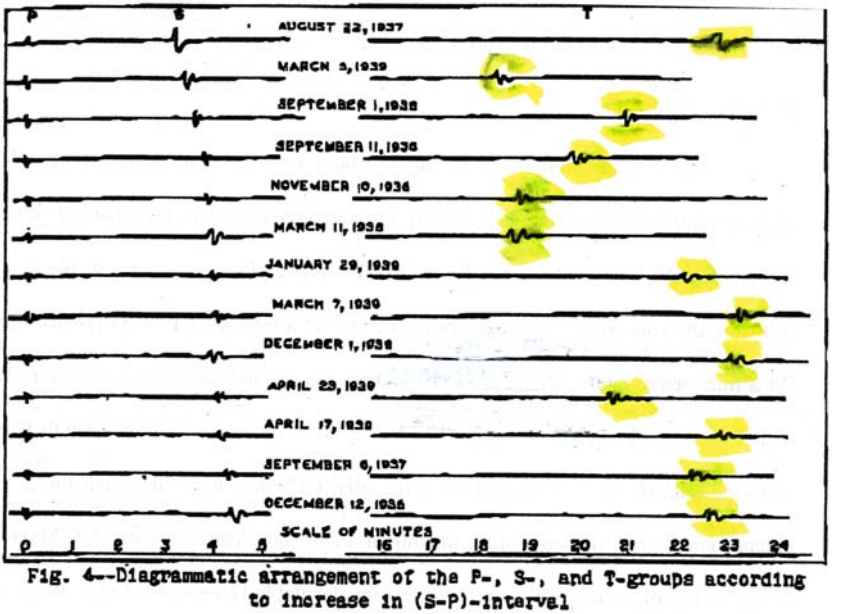


FIG. 2. Reproduction of Linehan's (1940) Fig. 4 (complete with original caption), showing a sketch of T phases from Caribbean earthquakes recorded at Harvard Observatory in the late 1930s. Note that even though the records have been arranged by distance (based on $S - P$ times), the T phases (emphasis added) arrive at irregular times, illustrating the variability of the source-side conversion process; this certainly hindered any simple interpretation of the phase by the author. The diagram was probably hand-drawn, and does not reproduce the waveshape of the T phases; note 10-minute gap in the time scale between the left and right parts of the figure. Adapted from Linehan (1940).

T Phases as Seismo-Acoustic Conversions

In a number of seminal papers published in the early 1950s, W. Maurice Ewing and his collaborators established the bases of the generation of T phases by earthquakes, as resulting from the conversion of seismic energy to acoustic waves at a source-side solid Earth–liquid ocean interface, followed by the reverse acoustic-to-seismic conversion at the receiver side (Tolstoy and Ewing, 1950). A fundamental aspect of this model is that the existence of sloping interfaces allows the trapping of acoustic energy inside the low-velocity waveguide and hence its efficient propagation to teleseismic distances. In particular, using Caribbean, North and South American earthquakes recorded at Bermuda, Shurbet and Ewing (1957) conclusively modeled their T waves as resulting from conversion of P or L_g phases at an isobath which they selected as 1800 m.

Such models were initially not without detractors, and in particular, Leet (1951) and Leet *et al.* (1951) argued, using occasionally vehemently forceful rhetoric,

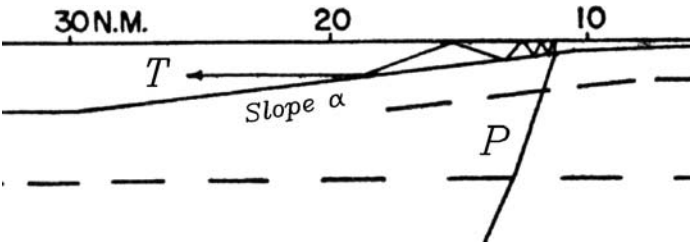


FIG. 3. Principle of the generation of a T phase from an underground source by downslope conversion. The source-side seismic ray (labeled P) is converted at the ocean bottom into a steep acoustic ray which is reflected back and forth between the sea surface and the sea floor. In the presence of an interface sloping at an angle α, the incidence of the ray is decreased by 2α at each bottom reflection, and penetration becomes possible after a handful of reverberations. Adapted (cropped and new labels) from Johnson *et al.* (1963).

that T phases rather represent energy channeled through ocean-bottom sedimentary layers, an idea already expressed by Coulomb and Molard (1949). Later work (e.g., Okal and Talandier, 1981) has indeed documented the possibility of long-range propagation of such waves, but at generally lower frequencies, and with considerable group dispersion (from 0.5 to 2 km/s), not observed in T phases. In their response to Leet's criticism, Ewing *et al.* (1952) correctly pointed out the need for a careful decomposition of the path into its seismic and acoustic portions. When the former become significant, they invalidate any estimate of the velocity of the T phase simply averaged over the entire distance from the source to the receiver.

Despite these early controversies, the work of Ewing and collaborators has withstood the trial of time, in particular concerning the timing of earthquake-generated T waves, which they successfully modeled by summing the contributions of the various seismic and acoustic segments (Shurbet, 1955). In their second paper on T waves in the Lesser Antilles, Coulomb and Molard (1952) reached similar conclusions, but emphasized the importance of $S \rightarrow T$ conversions in specific geometries. At the same time, Wadati and Inouye (1953) underscored the role of steep slopes for efficient conversions, and again of source-side $S \rightarrow T$ conversions notably for moderately deep (70 km) Japanese earthquakes.

In retrospect, it is remarkable that Ewing's group successfully developed their model in the Caribbean-to-New England geometry, which requires complex and long seismic conversions at both ends of the paths. In particular, Fig. 2 (from Linehan (1940)) clearly illustrates the lack of correlation for such paths between the precise timing of T and distance expressed by $S - P$ intervals. By contrast, Ravet (1940) was helped in his interpretation by the generally greater epicentral distances and minimal converted paths at the receiver side in Tahiti.

The Downslope Conversion Model

Considerable insight was gained throughout the 1960s from the operation, principally by the University of Hawaii, of a wide-aperture network of hydrophone stations throughout the Pacific Basin (Wake, Enewetak, Midway, Oahu, Point Sur). In this framework, Johnson *et al.* (1963) soon developed the concept of "*downslope conversion*" (Fig. 3), originally proposed by Officer (1958). In the presence of a beach sloping at an angle α, reverberations of acoustic rays between the seafloor and the ocean surface are deflected by 2α at each successive cycle, which eventually allows trapping of the energy inside the SOFAR channel at an angle inclined less than 12° on the horizontal, even though all acoustic rays at the original conversion point would remain much steeper.

The downslope conversion model, based entirely on geometrical optics, proved highly successful at explaining many attributes of the observed T wavetrains. In particular, Johnson *et al.* (1963) modeled lateral changes in the strength and waveshapes of T phases observed at Kaneohe Bay, Oahu from a series of hypocenters in the Andreanof Islands, based on variations in the number of reflections necessary for trapping through downslope conversion. Milne (1959) had used a similar concept to constrain the source of two nuclear tests to the interior of Enewetak lagoon, based on the spectral properties of their teleseismic T waves, which he interpreted as resulting from downslope conversion.

More recently, Talandier and Okal (1998) contrasted T waves received at the French Polynesian Seismic Network (Reseau Sismique Polynésien or RSP), from various types of earthquakes occurring on the Southern part of the Island of Hawaii (Fig. 4). For shallow earthquakes originating near the shoreline, characterized in that area by steep (up to 50°) subaerial and underwater cliffs (the "palis"), the T wavetrains are impulsive, of high amplitude, and feature a fast group time, as a result of a simple conversion process, requiring no more than one reflection to penetrate the SOFAR channel (Fig. 4A). In contrast, for deeper events occurring under the more gentle slopes (typically 15°) of the nearby Loihi volcanic edifice, the T waves are emergent, spindle-shaped, much lower in amplitude, and feature positive delays, as the conversion process requires as many as five back-and-forth reverberations, and is delocalized along the liquid–solid interface (Fig. 4B). In the former geometry, Talandier and Okal (1998) showed that the waveshape of the teleseismic T phase is a simple transposition of the ground motion at the conversion point, as recorded for example by a seismic station in the near field. It is dominated by the $P \rightarrow T$ and $S \rightarrow T$ conversions, the latter featuring a characteristic time delay and a lower frequency content.

Such differences in the efficiency of the conversion process can be transposed to the case of the acoustic-to-seismic conversion at the receiver side, where they can govern the siting of the so-called T-phase seismic stations designed to provide high-quality seismic recording of hydroacoustic phases, as mandated under the CTBT (Okal, 2001a). In particular, Okal and Talandier (1998) provided guidelines to define station corrections, in order to account for the receiver-side seismic

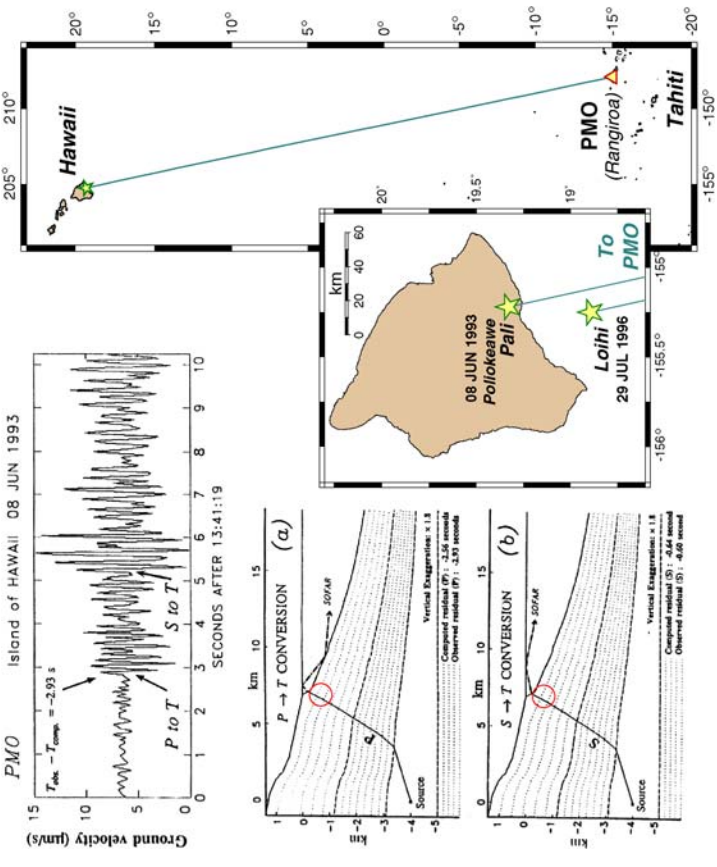

PMO
Island of HAWAII 08 JUN 1993
Ground velocity (μm/s)
SECONDS AFTER 13:41:19
P to T
S to T
Hawaii
PMO
(Rangiroa)
Tahiti
08 JUN 1993
Pollokeawe
Pali
29 JUL 1996
Loihi
To PMO
km
P → T CONVERSION (a)
S → T CONVERSION (b)
SOFAR
Source
Vertical Exaggeration: x 1.8

FIG. 4A. Comparison of the T waves recorded at Pomariorio (PMO) from two earthquakes on and near the Big Island of Hawaii. *Right*: The maps sketch the general geometry of the focal areas and the path to Polynesia. The 1993 earthquake is located under the steep Poliokeawe "pali" [cliff], while the 1996 event took place deeper and under gentler slopes at Loihi seamount. *Left*: Seismogram of the T phase of the 1993 Pali event at PMO (*Top*) and modeling of the conversion *(Bottom)*. Note the impulsive character of the phase, and the clear separation of the $P \rightarrow T$ and $S \rightarrow T$ wave packets. The modeling shows an efficient conversion, requiring only one reverberation, at the steep slope immediately off the shoreline (circles). In this geometry, the T phase is predicted to arrive early relative to acoustic propagation along the whole path, as observed on the record. Adapted (combined; map added) from Talandier and Okal (1998).

path, and showed that $S \rightarrow T$ conversions could become prominent under certain combinations of structure and distance between shore and receiver leading to shadowing for the converted P wave.

The 1994 Bolivian T Phases

The generation of T phases by intermediate-depth ($70 \leqslant h \leqslant 300$ km) or deep ($h \geqslant 300$ km) earthquakes was mentioned in the very earliest studies by Linehan (1940), Ewing *et al.* (1952) and Shurbet (1955), the latter noticing that a deep focus actually favors T-phase generation for large South American earthquakes recorded at Bermuda. We now attribute such observations to the absence of a source-side asthenospheric path otherwise generally responsible for the strong anelastic attenuation, in teleseismic or even regional S waves, of the high-frequency components exclusively capable of penetrating the SOFAR channel.

This remark led to an unexpected development upon the recording (Fig. 5) of spectacular T waves across the entire Pacific Basin following the great 1994 deep Bolivian earthquake (Kirby *et al.*, 1995), an observation itself intriguing given the considerable distance of its 631-km deep source from the oceanic water mass.[†] Okal and Talandier (1997, 1998) further documented that the timing of the Bolivian T waves at receivers across the Pacific required source-side conversion of regional S waves, principally at the Arica Bight. This observation was important from a structural point of view, since the delivery by an S wave, 920 km away from the source, of the high frequencies needed to penetrate the SOFAR (by necessity $f \geqslant 2.5$ Hz; $f \approx 5$ Hz as observed) requires propagation through material with exceptionally low anelastic attenuation. In turn, this implies the continuity of the cold slab throughout the upper mantle, despite the presence of a gap in seismic activity between 300 and 600 km depth in the local Benioff zone, which had led early investigators to propose a broken or detached slab (Isacks and

[†] Indeed, there exists some scant anecdotal evidence that the Bolivian T phase may have been felt on the Southern shore of the "Big" Island of Hawaii (G.J. Fryer, pers. comm., 2000).

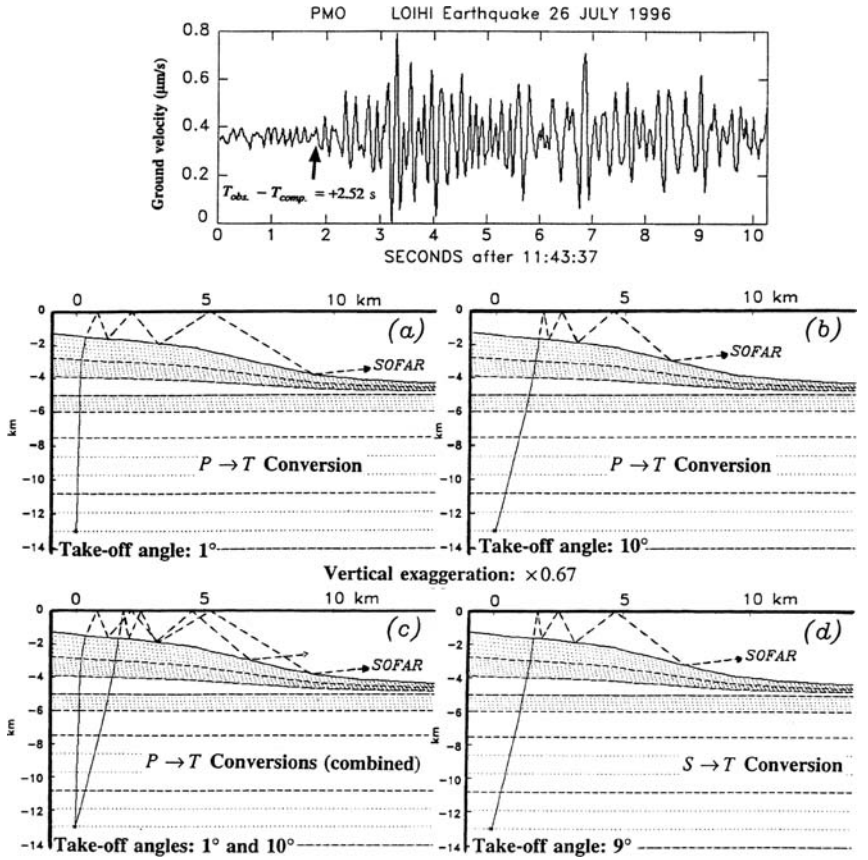


FIG. 4B. Same as Fig. 4A for the 1996 Loihi earthquake. *Top*: T phase record at PMO. Note the emergent phase, with no clear separation of its various components. Also, the phase is significantly late relative to a pure acoustic path from epicenter to receiver. *Bottom*: Modeling of the $P \rightarrow T$ and $S \rightarrow T$ conversions in the presence of a gently sloping beach. In (a) and (b), we show rays departing the focus at 1 and 10° incidence angles. Each requires several reverberations to penetrate the SOFAR channel, and when combined in (c), this results in a wavetrain of longer duration. Frame (d) shows similar characteristics for the $S \rightarrow T$ conversion. Adapted (combined) from Talandier and Okal (1998).

Molnar, 1971). An $S \rightarrow T$ conversion from a deep earthquake then constitutes a proxy for a thermally, and hence mechanically, continuous, low-attenuating slab in the source region, and Okal (2001b) later applied the concept to the so-called "detached" or "outboard" earthquakes, proving in most instances (e.g., Sakhalin, Bonin Islands) that these occurred in warped, rather than detached, segments of the slab. In this framework, T phases can provide an unexpected insight into the deep structural properties of the Earth's mantle.

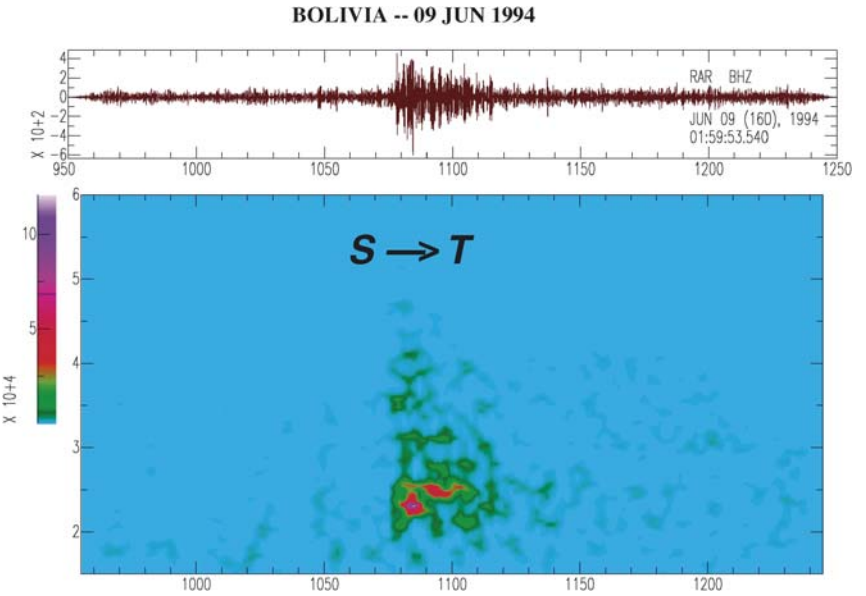


FIG. 5. Examples of T phases generated by deep earthquakes at the bottom of subduction zones. *Left*: Case of the 1994 Bolivian earthquake. The top figure presents a high-pass filtered waveform at Rarotonga and the corresponding spectrogram. Note the presence of a single wave packet, identified by Okal and Talandier (1997, 1998) as an $S \rightarrow T$ conversion at the Arica Bight. The bottom figure shows the geometry of the source side conversion. *Right*: Case of the so-called "detached" 1990 deep earthquake under Sakhalin, recorded at Pomariorio (French Polynesia). Note the presence of two separate wave packets, corresponding to $P \rightarrow T$ and $S \rightarrow T$ conversions, respectively. The presence of the latter requires the continuity of a finger of cold slab, despite the location of the earthquake outboard of the conventional location of the slab. Also, in both instances, note the location of the conversion points at preferential sites along the coast line. Adapted (combined; new material added) from Okal and Talandier (1997) and Okal (2001b).

Preferential Conversion Sites; On the Road to Scattering

A by-product of Okal's (2001b) investigation was the documentation of preferential sites of source-side conversion, such as the Arica Bight for the deep Bolivian earthquake, and Cape Erimo in Southern Hokkaido, for the Sakhalin one (Fig. 5). This observation was in line with Johnson and Norris' (1968) study of the T waves generated by the 1965 Rat Island aftershock series, which identified loci of preferential conversion for an otherwise largely uniform field of seismic epicenters. Similarly, Walker *et al.* (1992) documented occasional precursors to T phases, attributed to conversion at seamounts neighboring Alaskan and Aleutian epicenters. While observations such as Okal and Talandier's (1997) are still reconcilable with geometrical optics by invoking focusing due to strong

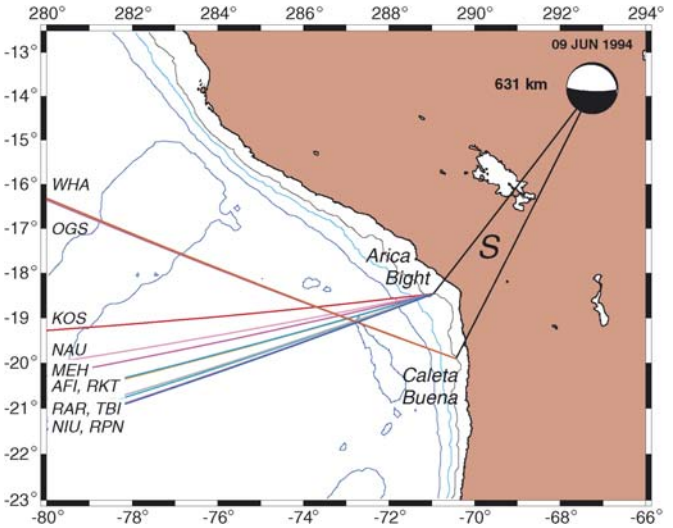


FIG. 5. Continued.

curvature of the mid-SOFAR isobath, Walker *et al.*'s (1992) results suggest the influence of scattering in the generation process.

The Paradox of the Abyssal T Phase

Furthermore, non-geometrical processes are clearly required to explain the paradox of the so-called "abyssal T phase" sketched on Fig. 6. In this instance, a small intraplate earthquake occurring in a flat abyssal plain, far away from any documented shallow slope, generates a strong T wave throughout the Pacific, whose group times are compatible with generation at the time and epicenter of the seismic source. Note in particular that this property holds not only in the far field (RAR), but also at regional distances (PTCN; 400 km). This indicates that the acoustic wave can be generated in a variety of azimuths in the immediate vicinity of the epicenter, in the absence of a relief appropriate for downslope conversion, and with a minimal, or absent, source-side seismic path. This scenario is reminiscent of Johnson *et al.*'s (1968) observation (Fig. 7) of two components to the T wave of an outer rise Aleutian earthquake ($m_b = 6.3$; 29 July 1965), the relatively impulsive, low-frequency and high-amplitude "downslope" arrival being expectedly late (by $\sim$2 minutes), but preceded by an "abyssal" arrival featuring high frequencies, a low amplitude, and an emergent wavetrain, which appears to originate at the time and location of the seismic source.

In attempting to explain the abyssal T phase, we note that Biot (1952) had initially proposed that a strong coupling could develop between "SOFAR waves" and

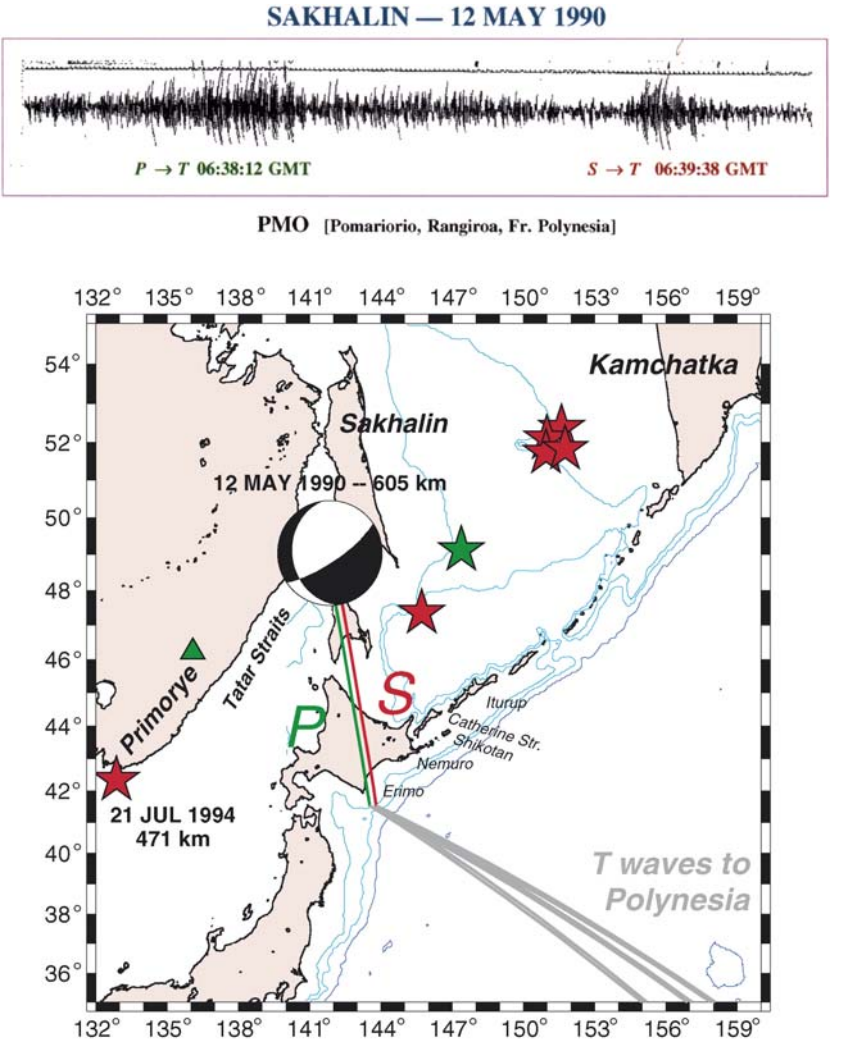


FIG. 5. Continued.

Stoneley waves generated at the solid–liquid interface, which could provide the mechanism for the excitation of an abyssal *T* phase. However, he did not elaborate beyond the observation that the two systems may have comparable dispersion curves in the high-frequency limit, which may be a necessary but not sufficient condition for the actual development of coupling. This concept was revived recently by the observation by Butler and Lomnitz (2002) of a so-called *Ti* wave, which they define as featuring the propagation characteristics of *T* phases, while

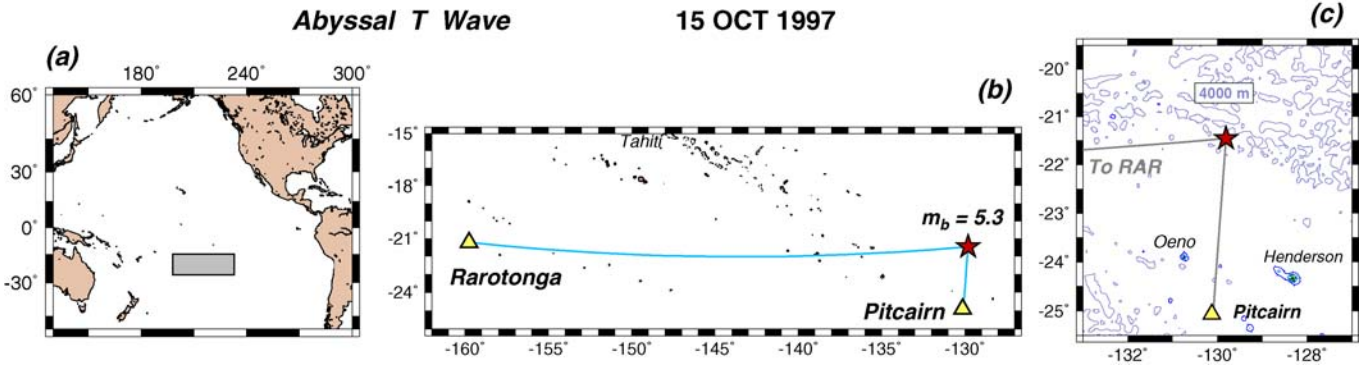


FIG. 6. Example of abyssal T phases generated by a "true" intraplate earthquake, occurring outside any large scale bathymetric feature, (a), (b) and (c): Situation maps of the relevant source and paths. The shaded box in (a) outlines Frame (b). Note the smooth bathymetry of the epicentral area, (d) and (e): Seismograms observed at Rarotonga (RAR) and Pitcairn (PTCN), respectively, high-pass filtered at 1.5 Hz, with relevant spectrograms. Note the long duration (80 s), and emergent character of the T phase, which arrives at a group time in agreement with acoustic propagation along the whole path. The vertical arrows indicate group times calculated from the epicenter for $v = 1.483$ km/s, and in the case of PTCN, the origin time, and *Pn* and *Sn* arrivals. Note that the latter is the dominant phase.

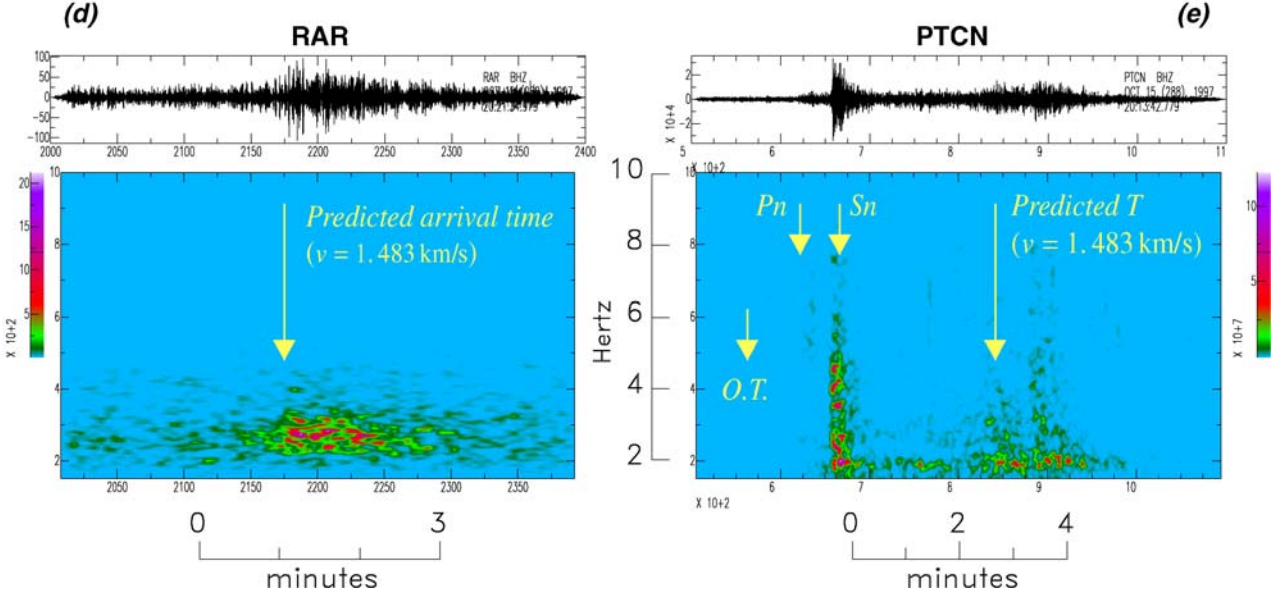


FIG. 6. Continued.

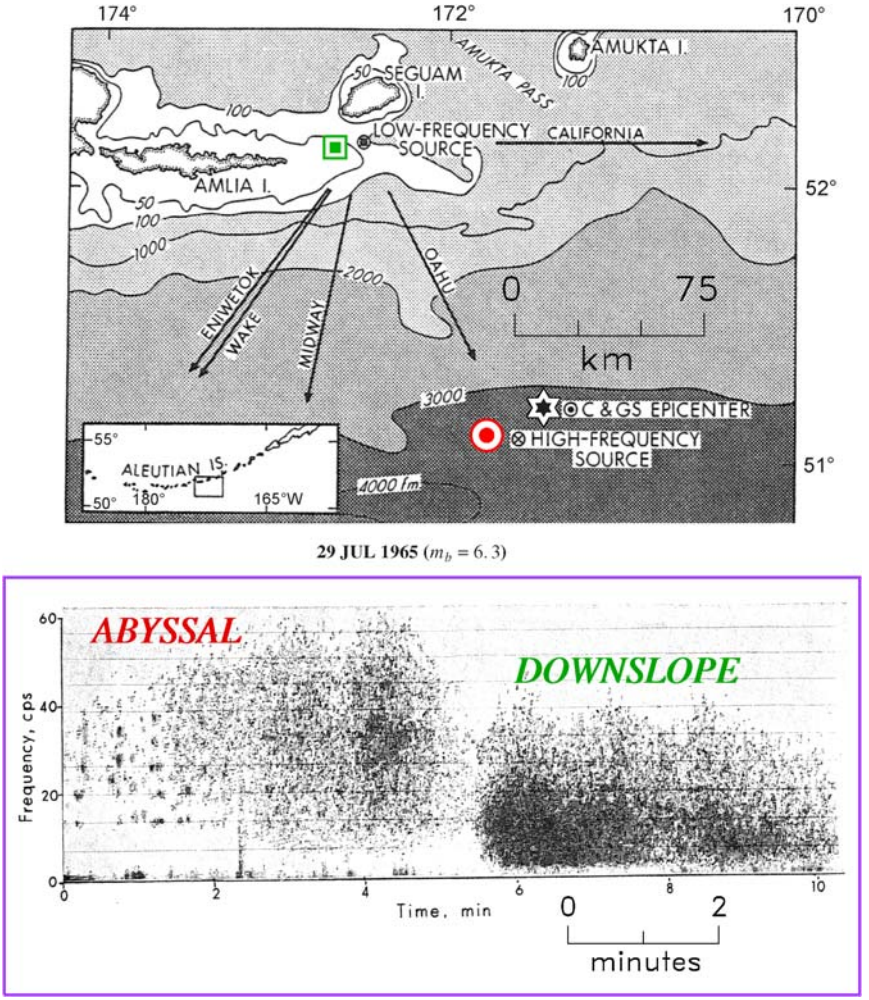


FIG. 7. Combination of abyssal and downslope generation of the T phase for the Aleutian earthquake of 29 July 1965 recorded at Kaneohe, Oahu (after Johnson *et al.*, 1968). *Top*: Map of the source area, showing the epicentral location (star), and the proposed sites of downslope (square) and abyssal (circle) conversions. *Center*: Spectrogram of the T phase contrasting the late, impulsive, high-amplitude and low-frequency downslope arrival with the earlier, emergent, higher-frequency but lower-amplitude abyssal arrival. *Bottom*: Model of abyssal generation as proposed by Johnson *et al.* (1968). The T phase is generated from scattering by an irregular sea surface, in the immediate vicinity of the epicenter, and thus suffers practically no delay, as compared to the downslope-converted phase which first backtracks as a P wave an estimated 150 km. Adapted (combined; cropped; re-labeled) from Johnson *et al.* (1968).

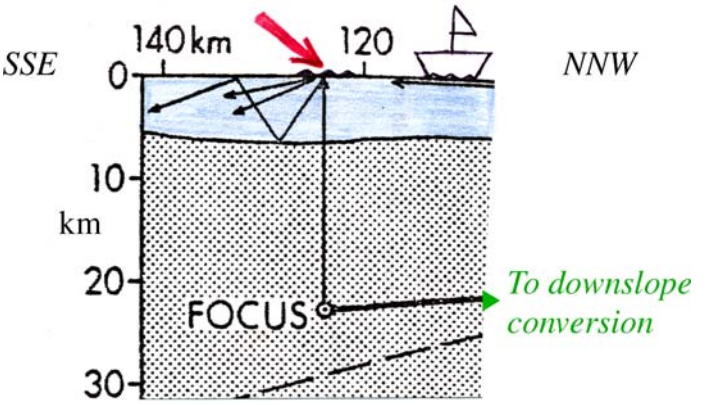


FIG. 7. Continued.

being detected on ocean-bottom or borehole instruments, i.e., in a geometrical shadow for waves guided in the SOFAR channel. However, their model of the propagation of *Ti* in a layer of relatively loose sediments on the ocean floor raises the question of the effect of anelastic attenuation (expected to be high in such structures) on the high frequencies characteristic of the phase.

This leaves a non-geometrical process, such as scattering, as the probable mechanism of generation of the abyssal T phase. Johnson *et al.* (1968) initially attributed its origin to sea surface roughness (Fig. 7), while Keenan and Merriam (1991) later suggested underside scattering by sea ice in the Arctic Ocean, an interpretation obviously limited to polar regions. Later, and most decisively, Fox *et al.* (1994) invoked broad-band scattering by a rough seafloor in the general framework of modal propagation, which forms the basis of the presently consensual interpretation of the abyssal T phase.

3. T WAVES IN THE MODE FORMALISM

Modal theory envisions T waves as the superposition of a discrete, albeit in principle infinite, number of *modes* of surface waves guided by the oceanic column, and in particular by the SOFAR channel. The fundamental framework of this approach was developed by Pekeris (1948). Given a flat-layered structure featuring translational symmetry along the horizontal x and y directions, the potential ϕ of the elastic wave in the water is sought at each angular frequency ω as a cylindrical wave radiating out of the polar axis $r = 0$:

$$\phi(r, z, t) = \Phi(z) H_0^{(1)}(kr)\, \mathrm{e}^{-\mathrm{i}\omega t}, \tag{1}$$

where $H_0^{(1)}$ is the Hankel function of first kind and order 0, and Φ satisfies

$$\frac{\mathrm{d}^2\Phi(z)}{\mathrm{d}z^2} + \left[\frac{\omega^2}{v^2(z)} - k^2\right]\Phi(z) = 0, \tag{2}$$

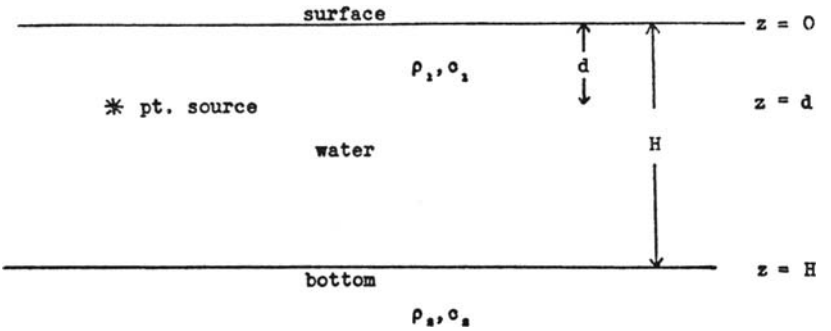
v being the local velocity of sound, which can be a priori a function of depth z. The boundary conditions, imposing zero pressure at the ocean surface, and zero displacement either at a rigid bottom or as $z \to \infty$, control the existence, at each frequency, of a finite number of *eigenmodes* expressing the relationship between ω and the wave number k, and analogous to the fundamental and overtones of a classical Rayleigh wave.

In his initial study, predating the era of high computational power, Pekeris (1948) first considered the oversimplified model of a single oceanic layer over a *liquid* half-space (Fig. 8a), and explored analytically the properties of the fundamental and first few overtones, closely following Love's (1911) classical analysis of the transverse shear modes of a solid layer over a solid half space. He then extended his investigation to a model featuring two layers over a half space (Fig. 8b), more representative of the SOFAR channel, but still involving exclusively liquids. Pekeris was able to describe the essential dispersive properties of the various branches, in particular the existence of one or several group velocity minima, for various contrasts in sound velocities and layer thicknesses. Most remarkably, he also laid the groundwork for the computation of the excitation of the various modes by an explosive source located in the water, by following Lamb's (1904) classical decomposition of the solution into an integral in the complex wavenumber plane. Such results were remarkably insightful given the simplified representation of the oceanic column as a single layer over a liquid half space, and for that reason, this model, known as a "*Pekeris waveguide*", has remained a benchmark reference for comparisons with modern, much more sophisticated and accurate models of the oceanic column.

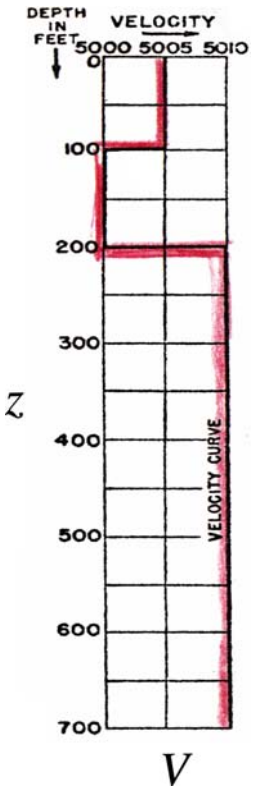
Over the next decades, and as detailed for example in Jensen *et al.*'s (1994) comprehensive monograph, improvements in computational techniques, as well as the systematic three-dimensional surveying of the properties of the ocean, have led to more sophisticated models of the structure of guided propagation in the SOFAR channel and of the excitation of acoustic modes by oceanic or land-based sources. Obviously, these developments owed much to the parallel investigation of the theory of seismic waves in layered structures, notably for surface waves by Haskell (1953) (and Harkrider (1964) regarding their excitation by various systems of forces), and for body waves as summarized by Kennett (1983). The following are the milestones most important to our understanding of the excitation and propagation of T waves by earthquake sources.

• *The use of realistic velocity profiles*

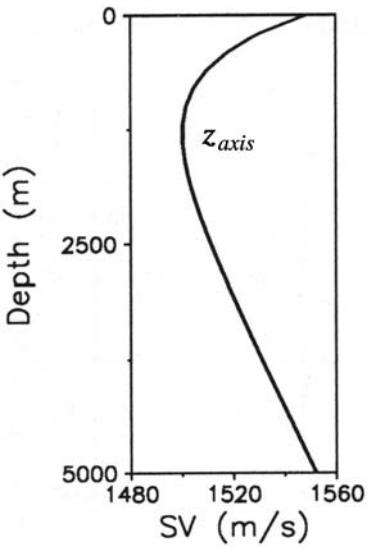
Based on a combination of experimental and theoretical work (Tolstoy and Clay, 1966; Pedersen, 1969), Munk (1979) provided an analytical representation



(a) Assumed model for a two-layered liquid half-space



(b)



(c)

FIG. 8. Reference models used in theoretical studies of the structure and generation of acoustic modes of the oceanic column. (a): The Pekeris waveguide, consisting of a liquid layer over a liquid half-space; all parameters can be widely varied. (b): Pekeris' later model of a low velocity waveguide, consisting of two layers over a substratum, all media remaining liquid; velocities in ft/s; (c): The Munk sound speed profile described by Eq. (3). Adapted (rescaled; combined; re-labeled) from Pekeris (1948) and Jensen *et al.* (1994).

of a reference velocity profile in the oceanic column,

$$v(z) = v_0\left[1 + \varepsilon\left(\bar{z} - 1 + \mathrm{e}^{-\bar{z}}\right)\right]; \quad \bar{z} = 2\left(\frac{z}{z_{\mathrm{axis}}} - 1\right) \tag{3}$$

where $v_0 = 1500$ m/s, $\varepsilon = 0.00737$, and $z_{\mathrm{axis}} = 1300$ m. This model, known as the Munk profile and shown on Fig. 8c, features a well-developed SOFAR channel centered around the depth z_{axis}. The Sturm–Liouville nature of the problem then allows the use of efficient algorithms (e.g., Wilkinson, 1965) for the systematic computation of literally hundreds of overtones over a wide range of frequencies. The duality between the modal solutions and the optical ray representation can then be investigated in the framework of the WKB approximation (e.g., Ewing *et al.*, 1957; Tindle, 1979). In general, as the mode (overtone) number n is increased at any given frequency, the phase velocity increases, which amounts to steepening the incidence of the equivalent ray.

• *The use of realistic boundary conditions*

The effect of the rigidity of the ocean bottom was addressed early on by Press and Ewing (1950), who considered the case of a single liquid layer over a solid half-space, in particular for an explosive source in the water column. They obtained fundamental results, especially regarding the influence of the finite value of the rigidity of the bottom on the apparent dispersion of the acoustic wave. More systematic investigations of this model (Brekhovskikh and Lysanov, 1982; Jensen *et al.*, 1994) have suggested that the reflection on the solid bottom is well approximated by a Pekeris model in which the sound velocity in the liquid substratum would equal the shear velocity of the solid half-space. This important remark justifies a posteriori the use of the oversimplified Pekeris model.

Figure 9, adapted from Park *et al.* (2001), contrasts, at a single frequency (in this case 5 Hz), the eigenfunction of the first overtone mode (not counting the interface Stoneley mode), with that of a higher ("hybrid") overtone. The former has its energy concentrated around 1000 m depth, in the axis of the SOFAR channel; its properties are equivalent to those described by Pekeris (1948), and in particular, its phase velocity, $C = 1.483$ km/s, expresses propagation of the acoustic energy in the channel. But because its eigenfunction has become essentially negligible by the time it reaches the solid substratum, such a mode cannot be excited by any source in the solid Earth. By contrast, Mode 32 has a well-developed eigenfunction both in the first 12 km of the solid Earth and in the water column; however its phase velocity, $C = 3.438$ km/s, indicates that its energy mostly reverberates at a steep incidence (25°) between the surface and bottom of the ocean, and does not propagate laterally in the SOFAR channel. In this framework, modal theory cannot explain the excitation of an abyssal T phase in a flat-layered structure any better than its geometrical optics dual, namely ray theory.

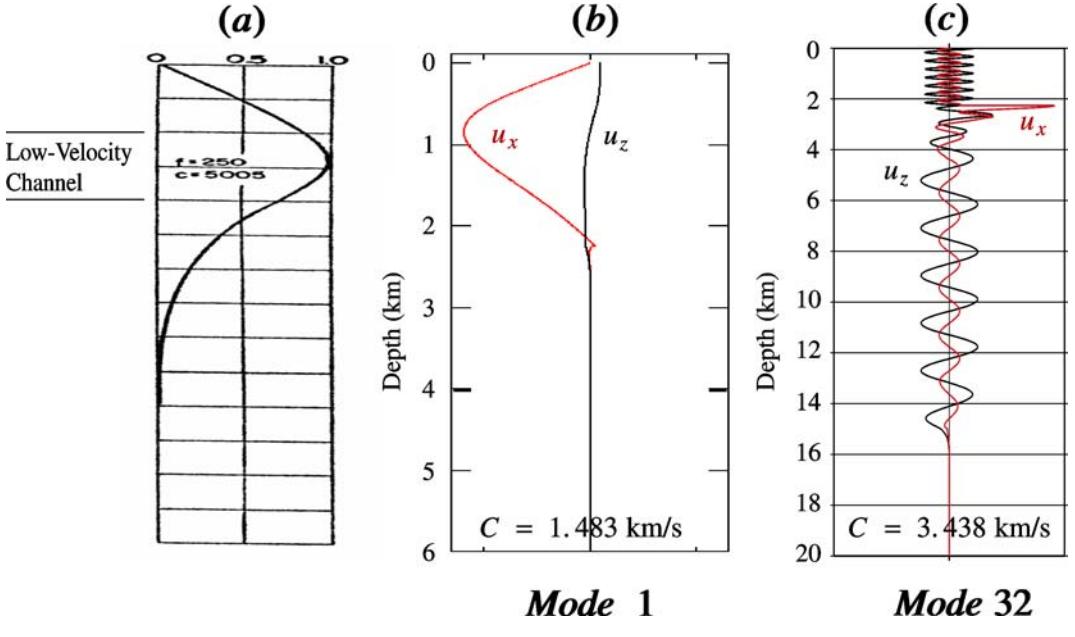


FIG. 9. Theoretical modal solutions reproduced from Pekeris (1948) and Park *et al.* (2001). (a): Pekeris' solution, obtained for his waveguide model shown on Fig. 8b; the plot shows the depth variation of the pressure eigenfunction. (b): First overtone solution at a frequency of 5 Hz, computed by Park et al. for a 2.25-km deep ocean featuring a SOFAR channel; the black line is the eigenfunction of the vertical displacement, the red one that of the horizontal displacement. Note that the mode does not penetrate appreciably the solid Earth. (c): Same as (b) for the 32nd overtone mode, plotted using a different vertical scale. Note that energy is present both in the liquid and solid, but the larger phase velocity expresses the inefficient lateral propagation of the energy in the liquid layer. Adapted (combined; rescaled; re-labeled) from Pekeris (1948) and Park *et al.* (2001).

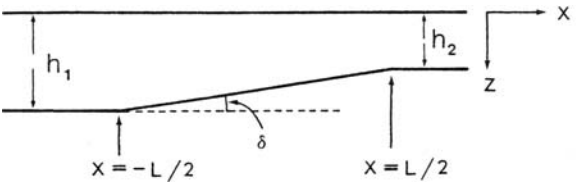
• *The computation of the excitation of the modes*

At any given frequency, the excitation of the various modes by any source in the ocean column (taken as a point source explosion) or in the solid Earth (taken as an elastic dislocation) is readily computed in the formalisms of Saito (1967) or Gilbert (1970). The summation of the various modes, weighted by their appropriate excitation, allows the synthesis of the wavetrains propagating away from the source, just as the summation of the free oscillations of the Earth excited by an earthquake yields a synthetic seismogram featuring the familiar seismic phases observed at teleseismic distances (Brune, 1964).

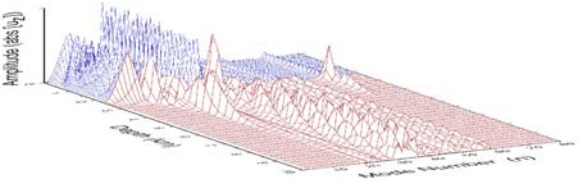
• *The application of perturbation theory in the modal framework*

This allows the study of *mode coupling* induced by [weak] lateral heterogeneity of the layered structure under consideration. Most such efforts have considered two types of heterogeneity: one involves two flat basins with different water depths, with a smooth transition extending over several wavelengths, the other a single basin with localized irregularities of the ocean floor, of an amplitude comparable to the acoustic wavelengths (Fig. 10a, b). Based on earlier theoretical work by Shevchenko (1962), Pierce (1965) and Milder (1969), Odom (1986) investigated a number of such scenarios, and in particular, for the second case, derived coupling coefficients between the various modes of the water column. In lay terms, this means that a corrugated structure of the type shown on Fig. 10b can provide a mechanism to leak energy from modes with little if any amplitude in the SOFAR channel (but strongly excited by underwater earthquake sources) into modes representing propagation of energy trapped into the channel (but not excitable by underground sources). This approach provides the key to a satisfactory explanation of the generation of the abyssal T phase by scattering at the ocean bottom.

More recently, Park and Odom (1999) extended the concept to the case of a stochastic field of heterogeneities on the seafloor, and Park *et al.* (2001) applied their results to a number of scenarios involving different geometries and depths of earthquakes under both a homogeneous ocean–bottom interface, and a heterogeneous one, which could be deterministic (as in the case of the sloping interface) or stochastic (i.e., featuring random roughness). Figure 10c (adapted from Park *et al.* (2001) quantifies, for a flat-layered structure and at $f = 5$ Hz, the seismic amplitude excited into the various overtone branches, and illustrates that only those modes with overtone numbers $n = 18$–55 are significantly excited. Figure 10d similarly shows the kernels for conversion of the elastic field in (c) by various scatterers located at the ocean bottom. Once multiplied by a scattering function expressing the density of heterogeneity in models such as (a) or (b), the kernels yield the cross-over amplitudes converted into the various modes. Note that the horizontally propagating water modes ($n = 1$–10) are now excited with a finite amplitude; this expresses the key result of the model, namely that it can indeed predict the excitation of an abyssal T phase.



(a)



(b)

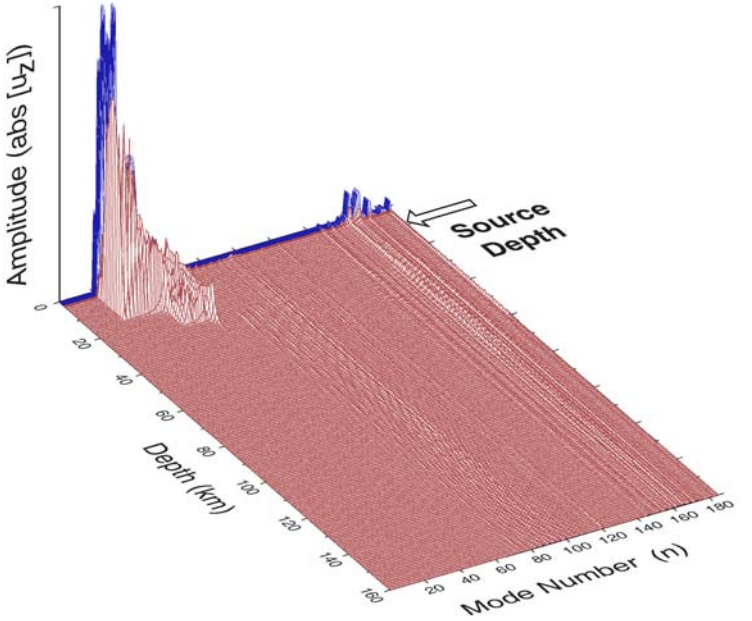
FIG. 10. (a) Example of lateral perturbation of a flat-layered ocean model: Two basins of depths h_1 and h_2 are connected by a slope. (b): Model of a corrugated segment of ocean floor in an otherwise laterally homogeneous flat-layered structure. (c): Initial excitation, at $f = 5$ Hz, of the various overtone modes of a flat-layered structure by a double-couple source located at 9 km depth (the ocean layer, 2.25 km deep, is shown in darker tone). Note that only overtones of order 18–55 are substantially excited. In particular the lower overtones corresponding to propagation in the water column are not. (d): Mode excitation by sea-bottom scattering. Note that the energy of the higher modes in (c) has been converted into the lower overtones (of order less than 10), propagating in the SOFAR channel. See text for details. Adapted (combined; re-labeled) from Odom (1986) and Park *et al.* (2001).

Park *et al.* (2001) further documented that relatively deep earthquakes ($h = 80$ km) can excite T waves through scattering by a rough ocean bottom, and also proposed that strike–slip earthquakes seem especially efficient at generating T phases, as observed by Dziak (2001), and further discussed below.

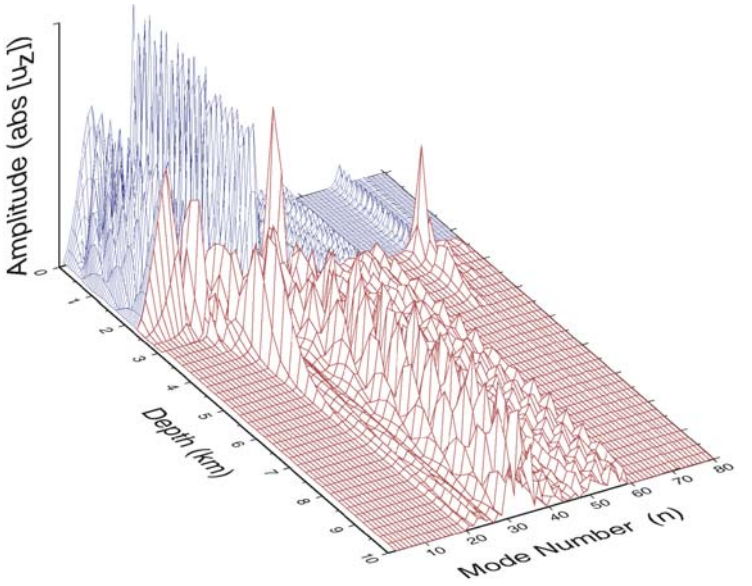
Finally, Odom and Soukup (2004) examined the amplification of the scattering process in the presence of an intermediate layer of low-rigidity sediments, which can be regarded as the mode dual of the well-known amplification of strong motion by sedimentary structures routinely observed in classical seismology (e.g., Gutenberg, 1957; Bard *et al.*, 1988).

Following a different strategy, Schmidt *et al.* (2004) model the scattering of acoustic energy into the water column through the Virtual Source Approach, which uses the Rayleigh–Kirchhoff approximation. They show that abyssal T phases can be interpreted as resulting from scattering of seismic energy trapped in sedimentary layers in the form of Scholte waves (Scholte, 1947).

Using a simplified methodology, deGroot-Hedlin and Orcutt (1999) modeled scattering into a given acoustic mode by sea-floor heterogeneities as directly proportional to the product of the amplitude of the mode at the ocean–sediment interface and of the ground motion produced by the dislocation source at the relevant location. They emphasize that the bottom modal amplitude is strongly



(c)



(d)

FIG. 10. Continued.

dependent upon the depth of the water column (mostly decreasing with increasing depth below a few hundred meters), and that the scattering process generating the *T* phase is thus strongly controlled by bathymetry. When combined with differences in propagation times for *T* waves generated at various locations on the ocean floor, this model successfully explains differences in the shape of envelopes observed at Wake Island for *T* phases originating at various subduction zones of the Northern Pacific (Fig. 11).

Yang and Forsyth (2003) later expanded on deGroot-Hedlin and Orcutt's (1999) model by including the contribution of *S* waves to the ground motion at the conversion point, and by assigning only a small fraction (typically 1%) of the incident amplitude to scattering in a horizontal direction; however, they also consider the effect of scattering when reverberating rays traveling quasi-vertically in the water column hit the ocean bottom at later times, thus contributing to the extended duration of the *T* phases. As shown on Fig. 12, this approach allowed Yang and Forsyth (2003) to produce very realistic renditions of the envelopes of abyssal *T* phases recorded at regional distances by an array of ocean-bottom seismometers, and in turn to improve on the definition of group times for such signals, and eventually on the precision of location algorithms using *T* phases.

4. Using *T* Phases to Detect and Locate Seismic Sources

Since *T* phases can transmit energy from small sources over large distances, they provide an exceptional opportunity to detect and locate small seismic events which would otherwise have gone unsuspected. In this context, *T* phases have been used primarily to refine our knowledge of the low-level seismicity of mid-oceanic ridge systems and of intraplate abyssal basins.

As a result of the operation by the University of Hawaii of a wide aperture hydrophone array in the Pacific, Duennebier and Johnson (1967) located more than 20,000 *T*-phase sources in the Pacific Basin in 1964–1967, and compared them with the dataset of USCGS epicenters for the same period (numbering roughly half as many events). The authors identified regional trends in spaciotemporal differences between solutions derived from *T* phases and conventional seismic waves, which could often be ascribed to the influence of source-side processes such as downslope conversion. They presented maps of the Pacific Basin divided in quadrangles where they compared seismicity defined by seismic and *T* phases, which suggest that the latter reveal many more events at mid-oceanic ridges and occasionally in abyssal plains; Walker (1989) later included 206 such epicenters in a catalog of intraplate seismic events.

Using the same database, Northrop *et al.* (1968) focused on the Juan de Fuca and Gorda Ridge systems where they described earthquake epicenters in the framework of small scale ridge-and-transform segments. Later, Hammond and

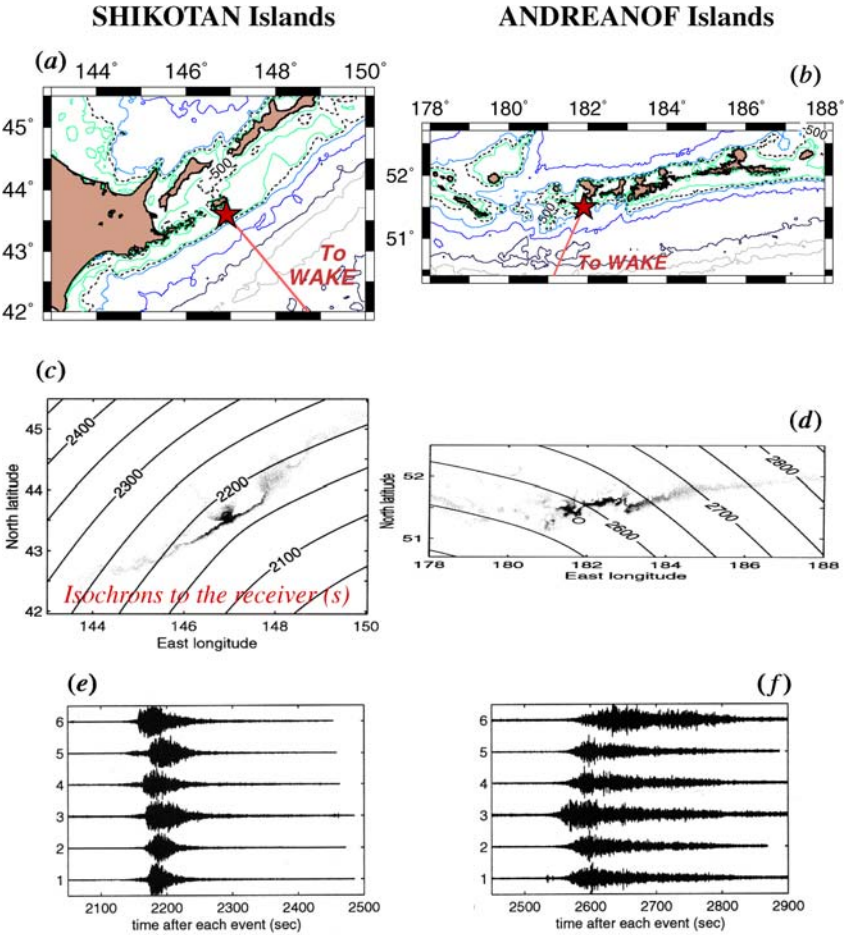


FIG. 11. Modeling of T-wave generation by scattering at continental or island arc slope, after deGroot-Hedlin and Orcutt (1999). The authors contrast the case of T phases received at the Wake Island hydrophone from earthquakes in the Shikotan (a) and Andreanof (b) Islands. In (c) and (d), they map the intensity of the scattered T phase on the ocean floor as a function of the seismic amplitude reaching the epicenter and of the excitation of acoustic modes (essentially mode 1) at the relevant depth. The contours represent isochrons to the receiving station at Wake. The aspect ratio of frame (d) has been corrected to make the map conformal and give it the same scale as the Mercator projection in (b). Synthetic waveforms are presented in frames (e) and (f) by summing up the contributions of the various scatterers. Note the greater geometrical scatter of the secondary sources in the Andreanof case, which combines with the different orientation of the path to Wake to give the signal a greater dispersion in time, and hence a more spindled shape. Adapted (combined; rescaled; maps added) from deGroot-Hedlin and Orcutt (1999).

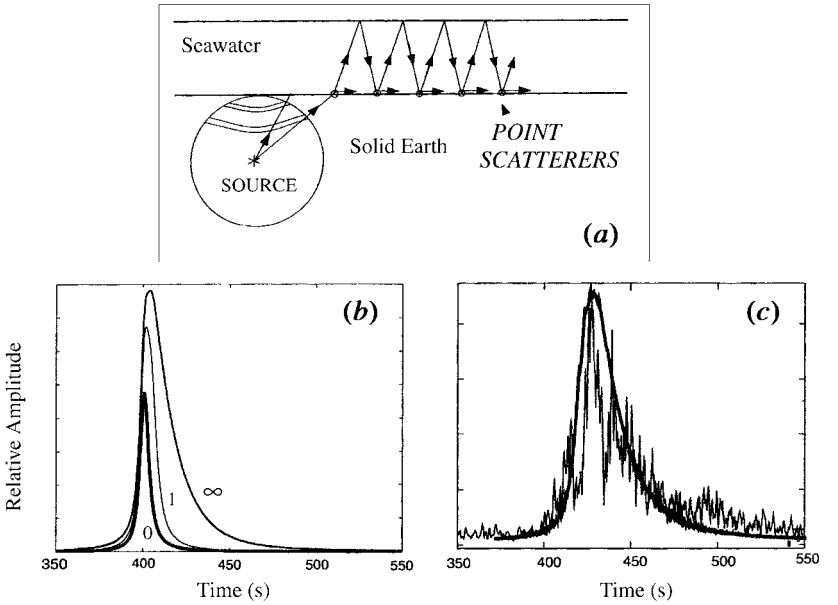


FIG. 12. Generation of the abyssal T phase by a field of scatterers on the ocean floor, after Yang and Forsyth (2003). (a): General layout of the concept: the ocean floor is illuminated both directly by the seismic source, and through reverberation of steep acoustic rays in the water column. At each reflection or transmission point, a small fraction (1%) of the elastic energy is scattered in a horizontal direction, and thus made available for penetration of the SOFAR channel. (b) T-phase envelopes synthetized from scattering by the P-wave near field of a symmetric (explosive) source. The thick line (labeled "0") is the direct field including no reverberation, the intermediate one (labeled "1") includes the effect of a single reverberation and the outer curve (labeled "∞") includes a very large number of reverberations. (c): Comparison of a T-wave envelope synthetized using the geometry of the full field of a strike–slip dislocation (including S waves; thick line), and of an observed waveshape; note the generally good agreement of the waveshapes. Adapted (combined; rescaled; re-labeled) from Yang and Forsyth (2003).

Walker (1991) located 54 otherwise undetected earthquakes (as well as 4 reported ones) on the Juan de Fuca and Endeavor Ridges, which they attributed to volcanic activity during episodes of sea-floor spreading.

The partial declassification of the Sound Surveillance System (SOSUS) network of U.S. Navy hydrophones in 1991 resulted in much enhanced location capabilities in the Northeast Pacific, which allowed Fox *et al.* (1995) to closely monitor, in real time, the spatial evolution of a swarm of more than 600 earthquakes on the Juan de Fuca Ridge, starting on 26 June 1993. This was interpreted as part of a lateral dike injection by Dziak *et al.* (1995), and resulted in the "rapid response" dispatch of a multidisciplinary team of investigators for what turned out

to be the first in situ observation of volcanic activity on a segment of mid-oceanic ridge (Embley *et al.*, 1995).

Following these developments, systematic monitoring of mid-oceanic ridges was instigated through the deployment, illustrated on Fig. 13, of two long-term, wide aperture so-called "autonomous" hydrophone arrays, around the fast-spreading Equatorial Pacific Rise (Fox *et al.*, 2001) and the slow-spreading North Atlantic Ridge (Smith *et al.*, 2002). By providing extremely detailed catalogues of the spaciotemporal distribution of seismicity on and around the ridge systems, these still ongoing projects have shed considerable new light on the characteristics of ridge earthquakes, both of tectonic and volcanic origin. Among the most important results is the confirmation of the generally aseismic character, even at low source levels, of fast-spreading segments, with all tectonic activity on the EPR located on the transform faults (Fox *et al.*, 2001), in contrast to the case of slow-spreading ridges (Smith *et al.*, 2003). In addition, Bohnenstiehl *et al.* (2002, 2003) provided very detailed imaging of aftershock sequences in several geometries, and documented high values, in the median valleys, for the parameter p used to fit the Modified Omori Law to temporal distributions of aftershocks (Ogata, 1983), which these authors attributed to a local field of high temperatures.

Using an independent approach, Forsyth *et al.* (2003) combined near-field seismic and T phases recorded by an array of ocean bottom seismometers to precisely locate a swarm of activity along the Equatorial East Pacific Rise system, and documented coupling between nearby short transform fault segments; they tentatively interpreted the swarm as the only manifestation of an otherwise silent episode of "aseismic" slip comparable to creep events detected on the San Andreas system (Linde *et al.*, 1996).

On a teleseismic, basin-wide scale, it is expected that deployments such as the autonomous Equatorial Pacific Hydrophone Array (EPHA) will lead to a better definition of the level of seismicity of abyssal basins. However, systematic tests performed by Fox *et al.* (2001) showed that precision and accuracy in epicentral range (as opposed to azimuth) quickly deteriorate when the source moves far outside the recording array. Figure 14 (adapted from Fox *et al.*, 2001) clearly illustrates this smearing of epicentral location on great circles radiating from the array, but would still suggest, over and beyond this effect and the possible influence of acoustic blockage to certain provinces, recognizable patterns of variability in the intraplate activity of the Pacific Basin. For example, the relatively large level of activity inside the Nazca plate (at distances not exceeding 1.5 times the maximum dimension of the array) would agree with the conclusions of Wysession *et al.* (1991).

On the other hand, it is clear that the location capabilities of the EPHA could be drastically improved by the addition of a single hydroacoustic receiver such as the CTBT hydrophones at Juan Fernandez, or T-phase stations such as those of the Polynesian Seismic Network (Fig. 14). Also, the synergetic use of acoustic and seismic phases should significantly improve location capabilities of small

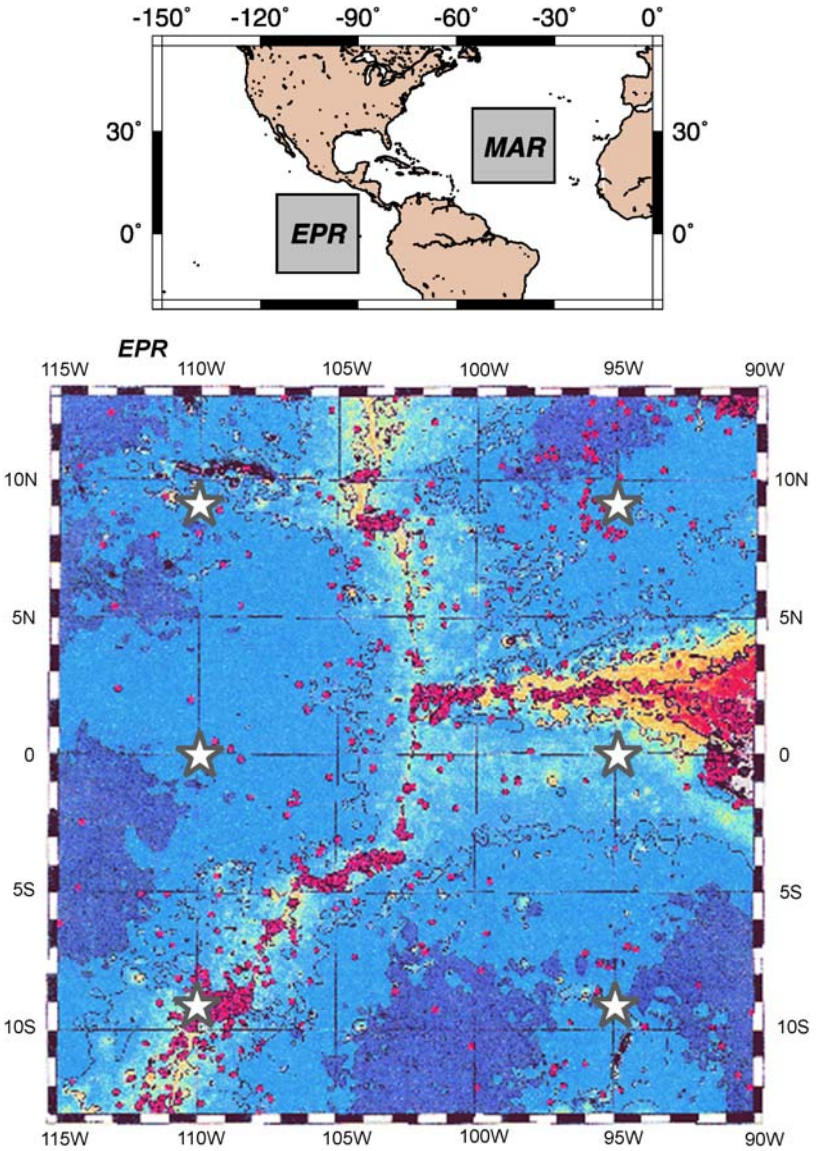


FIG. 13. Regional epicenters located by the autonomous arrays of hydrophones (large stars) recently deployed around the Equatorial East Pacific Rise (EPR) and the Northern Mid-Atlantic Ridge (MAR). The bottom maps contrast the seismicity patterns along the slow-spreading MAR, where earthquakes are ubiquitous along both transform and spreading segments, and along the fast-spreading EPR, where the latter are essentially silent, except for occasional events presumably associated with magmatic swarms at axial volcanoes. Adapted (combined; enhanced; map added) from Fox *et al.* (2001) and Bohnenstiehl *et al.* (2003).

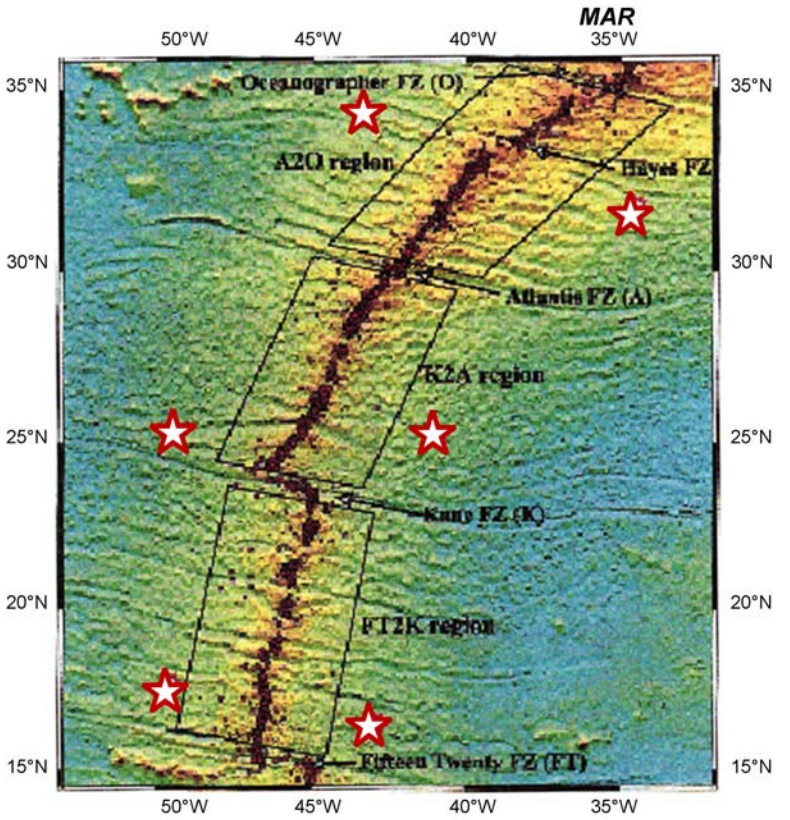


FIG. 13. Continued.

events even in the far field. In this respect, Talandier and Okal (2004b) showed that T waves could be used to compensate for the systematic bias introduced in locations of underwater explosions achieved from ground-based seismic stations, and due to the lateral heterogeneity of crustal phases at continental margins.

5. USING T WAVES TO EXPLORE THE SEISMIC SOURCE

We discuss in this section several approaches to extracting from a T-phase waveform quantitative information relative to the earthquake source. Despite the complexity of their generation process, T waves are but one of the numerous seismic phases generated by an elastic dislocation. As such, they should carry information about earthquake source spectra in the high-frequency range

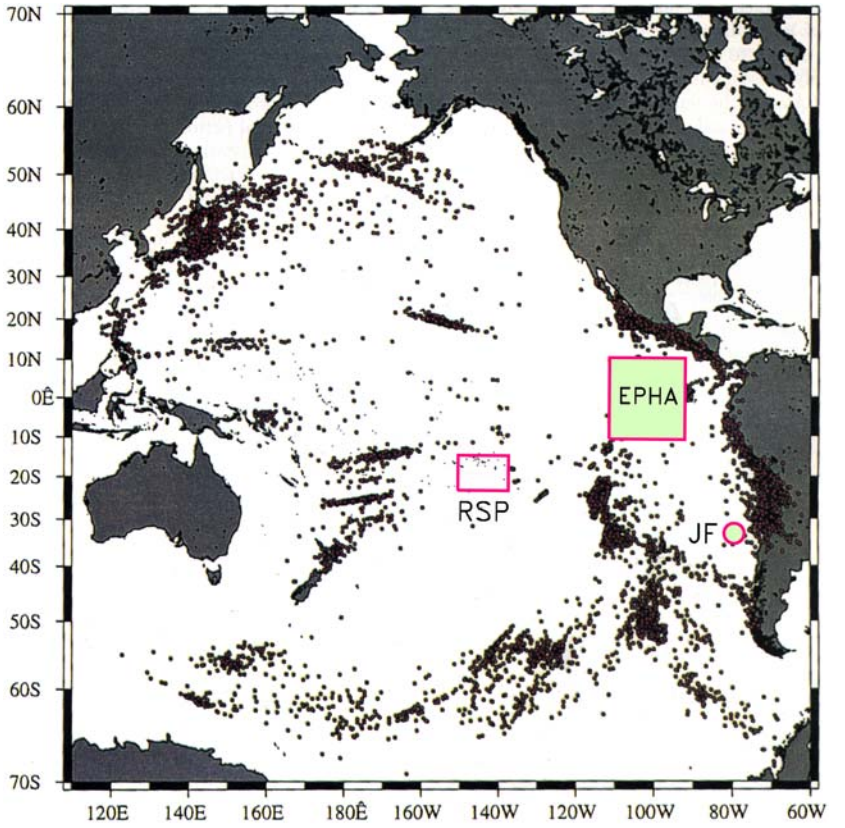


FIG. 14. Far-field epicenters located by Fox *et al.* (2001) from detections at the Equatorial Pacific Hydrophone Array (EPHA). The extent of the array is shown by the shaded box. Note the streaks of epicenters along great circles radiating from the array, which express the loss of range resolution for very distant events. The location of the new hydrophone station at Juan Fernandez Island (JF) is shown by the large circle and that of the Polynesian Seismic Network (RSP) by the open box. Adapted (re-labeled) from Fox *et al.* (2001).

($f \geqslant 2.5$ Hz), otherwise poorly sampled due to anelastic attenuation of conventional teleseismic waves. However and as detailed below, for larger earthquakes, T waves are strongly affected by interference resulting from source finiteness, and their amplitude alone cannot be a good proxy for source size. For this reason, the independent measurement of T-wave *duration* is necessary to retrieve an adequate estimate of an earthquake's size. We discuss below the joint use of amplitude and duration, which can take two forms: by combining their values, in the form of the T-Phase Energy Flux (TPEF), Okal *et al.* (2003) extended to T waves the concept of radiated seismic energy applied to body waves by Boatwright and Choy (1986); by contrasting them in the form of an amplitude-duration criterion,

Talandier and Okal (2001) allowed the discrimination of explosive and dislocative sources, and more generally the identification of unusual seismic sources, such as slow earthquakes and possibly landslides.

Background: Source Finiteness for T Waves

While early investigators sought an interpretation of T-wave amplitudes in terms of earthquake magnitudes, the high-frequency nature of the T phase makes it particularly vulnerable to destructive interference due to source finiteness which leads, for large sources, to the eventual saturation of spectral amplitudes measured at any fixed frequency. This concept was introduced in the case of conventional seismic waves by Ben-Menahem (1961) and studied in detail by Geller (1976) in his seminal paper on the saturation of magnitude scales. Essentially, the growth of a seismic source in a material with given ("invariant") elastic properties requires extending the source both in time and space. Any measurement taken on a seismic wave at a given frequency f will suffer from destructive interference as soon as its period $1/f$ (or wavelength Λ) becomes comparable to, or shorter than, the duration (or spatial extent) of the source. The combined effect of the finite strain release on a 2-dimensional fault and of the limited velocity of the actual slip between the fault walls leads to a total saturation of spectral amplitudes once the source corner frequencies recess below the frequency f of interest in the wave group under study. On this basis, Geller (1976) justifies the well-known saturation of the 20 s surface-wave magnitude M_s around 8.2 for moments $M_0 \geqslant 10^{28}$ dyn cm, and similarly of the body-wave magnitude m_b around 6.3 ($M_0 \geqslant 1.8 \times 10^{26}$ dyn cm) when properly measured on P waves at 1 second (larger values of m_b reported in bulletins usually stem from measurements taken at longer periods, or on S waves, or relate to deep events allowing greater strain release). The destructive nature of the interference will be exacerbated for T waves, which are limited to $f \geqslant 2.5$ Hz by the geometry of the SOFAR channel, and saturation would be expected to take place as early as $M_0 > 2 \times 10^{25}$ dyn cm (equivalent to $m_b = 6.0$ or $M_s \geqslant 6.4$) (Geller, 1976; Talandier and Okal, 1979). Note also that the use of source scaling laws assumes the invariance of the seismic-to-acoustic processes, which can be a gross oversimplification, given their complexity and their variability for any population of sources (e.g., Johnson *et al.*, 1963, 1968).

In this general framework, we review early results on the quantification of T waves, as well as more recent and promising developments.

Amplitude Measurements

An early attempt at relating T-phase amplitudes to earthquake magnitudes is found in Johnson and Northrop (1966), who studied 49 records from Aleutian earthquakes. They quantified T phases through their so-called *strength* (*TS*), expressed in dB over a reference pressure of 0.1 μbar, and proposed a relation of the

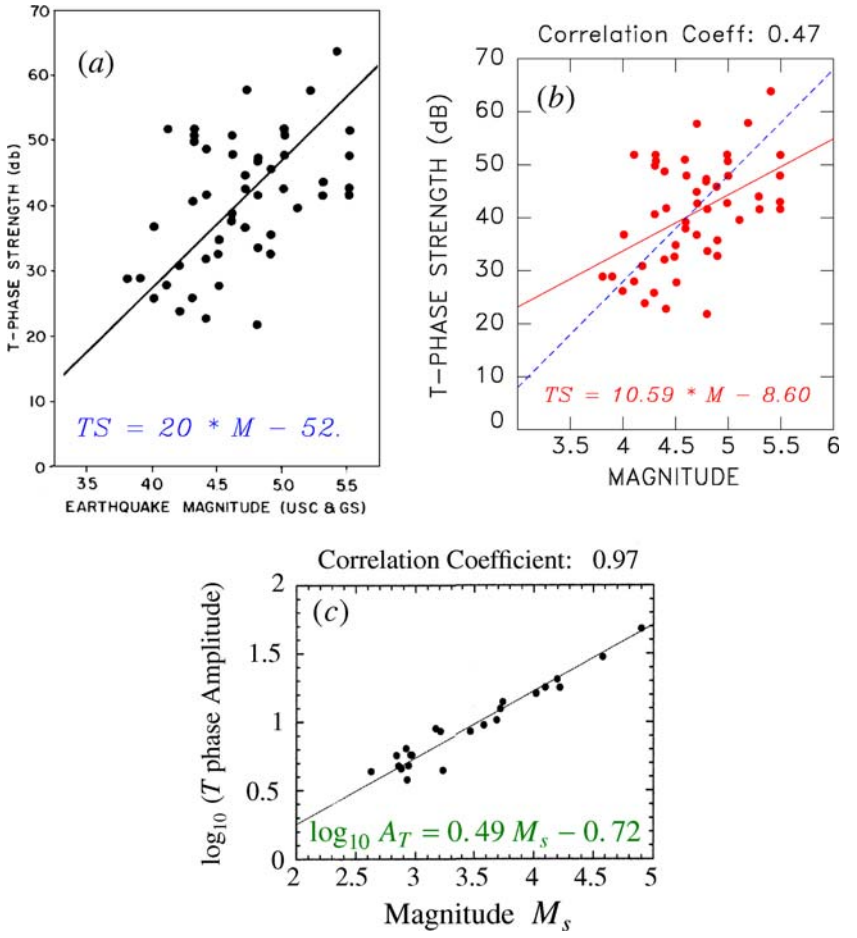


FIG. 15. Examples of correlation between earthquake magnitude and T-phase amplitudes. (a): Correlation with T-phase strength after Johnson and Northrop (1966). In (b), the same dataset is fit by least-squares. Note the mediocre correlation coefficient, and the much reduced slope, when the latter is left to float; the dashed line reproduces the regression proposed in (a). (c): Correlation with surface-wave magnitude for 24 East Pacific Rise earthquakes (Yang and Forsyth, 2003). This diagram has been replotted using the exact the same orientation and scales as in (a) and (b). Note the much improved correlation. Combination (re-labeled, rotated, combined) of new material (b) and material (a) from Johnson and Northrop (1966), and (c) from Yang and Forsyth (2003).

form

$$TS = 20M - 52 \quad \text{(Fig. 15a)}. \tag{4}$$

However, this approach has several shortcomings. First, the constant 20 was fixed, under the assumption that the T-phase amplitude should parallel seismic

amplitude; second, it is not clear which magnitude (m_b; M_s) is used in (4). We conducted an independent regression of Johnson and Northrop's (1966) dataset allowing the slope to float (Fig. 15b), which yields a much gentler slope: $TS = 10.59M - 8.60$, but a mediocre correlation coefficient of only 0.47. This lower value of the best-fitting slope indicates that T-phase amplitudes are indeed affected by destructive interference in a range of magnitudes where M_s (and probably m_b) should not be. Fox *et al.* (2001) also reported a lower value (7.84 ± 2.04) of the slope of acoustic source level (in dB) vs. m_b for a dataset of 87 events recorded at the Autonomous Equatorial Array. Such results are also in agreement with those of Yang and Forsyth (2003), reproduced on Fig. 15c, who regress the amplitude A_T of T phases as $\log_{10} A_T = 0.49M_s - 0.72$ (equivalent to a slope of 9.8 between TS and M_s in (4)) for a very homogeneous dataset of 24 East Pacific Rise earthquakes at regional distances. The factor 0.49 similarly expresses that T-wave amplitudes do suffer from interference effects in a range of magnitudes ($M_s \leqslant 5$) where 20-second waves are immune from them.

Walker *et al.* (1992) later investigated the correlation of T-phase strength (which they restricted to the frequency range 10–35 Hz) with seismic moment M_0 for a dataset of 25 Pacific-rim earthquakes recorded at the Wake hydrophone array. The use of M_0 should ensure a more robust estimation of the true size of large events, and Walker *et al.* (1992) indeed observe an excellent correlation for certain sub-datasets (their Fig. 17), but their results are difficult to interpret since most of their large moments involve strike–slip sources in the Gulf of Alaska; as discussed below, this geometry may be a preferential T-phase generator (Dziak, 2001), and those strike–slip earthquakes may thus not be directly comparable to the remainder of the dataset, consisting almost exclusively of thrust and normal events. Hiyoshi *et al.* (1992) similarly studied 17 Japanese earthquakes recorded at Wake. They observe a strong correlation between TS and $\log_{10} M_0$, but the resulting slopes are only half of those suggested by Walker *et al.* (1992), and the earthquakes in their dataset are generally smaller.

Duration: Another Measure of Source Size

As source size is increased, so is the time it takes for the rupture to take place at any given point on the fault, and more significantly to propagate along the full extent of the fault zone. As a result, scaling laws predict that the duration of any seismic wavetrain should also grow, in principle linearly with the dimension of the source, or like $M_0^{1/3}$. This idea forms the basis of the use of duration magnitudes (e.g., Lee *et al.*, 1972; Real and Teng, 1973), and it would be expected that the duration of T waves should also grow with source size.

Indeed, a number of early studies sought to measure the duration of the T phase of large shocks in relation to their size. Eaton *et al.* (1961) noted that the T phase of the 1960 Chilean earthquake lasted as much as 6 minutes, while Ben-Menahem

and Toksöz (1963) proposed 4 minutes for the 1958 Fairweather, Alaska earthquake. However, neither of these authors gave a formal definition of duration. Following an idea suggested by Johnson (1970), Okal and Talandier (1986) analyzed a large dataset of T waves from Pacific shocks recorded in Tahiti and Hawaii, and obtained reasonable correlations between their durations and various magnitude scales. They defined the duration of the wavetrain either as that of sustained maximum amplitude, or that of saturation on paper records at Polynesian stations. In particular, they showed that T-wave duration keeps growing with seismic moment, even for the very largest seismic events (Alaska, 1964; Chile, 1960). Consequently, Okal and Talandier (1986) suggested the use of T-phase duration as an indicator of the tsunamigenic potential of large earthquakes, with a threshold of 100 s for the excitation of a destructive transoceanic tsunami ($M_0 \geqslant 5 \times 10^{28}$ dyn cm). They noted, however, the presence of outliers in their dataset, notably strike–slip earthquakes on the Alaskan Fairweather Fault.

T-Phase Energy Flux: The Parameter γ

In an attempt to characterize the total energy generated by an earthquake source into a T wave, Okal *et al.* (2003) introduced the concept of the *T-Phase Energy Flux* (TPEF), which mimics the estimated energy E^E developed by Newman and Okal (1998) for body waves. The latter was itself inspired by Boatwright and Choy's (1986) algorithm for the computation of radiated seismic energy, but with the philosophy of a magnitude measurement, i.e., a real-time "quick-and-dirty" single-station estimate ignoring such source details as focal geometry and exact depth.

Specifically, given a seismic record of the vertical ground motion $u(t)$ of a T phase, Okal *et al.* (2003) define

$$TPEF = \rho\alpha \int_W \left[\dot{u}(t)\right]^2 \mathrm{d}t, \tag{5}$$

where ρ and α are the density and P-wave velocity of the receiver medium, and W is an appropriate time window containing the T phase. Using Parseval's theorem, *TPEF* can be more readily computed in the Fourier domain as

$$TPEF = \frac{\rho\alpha}{\pi} \int_{\omega_{\min}}^{\omega_{\max}} \omega^2 \left|U(\omega)\right|^2 \mathrm{d}\omega, \tag{6}$$

where the spectral amplitude $U(\omega)$ is the Fourier transform of $u(t)$, and $\omega_{\min}$ and $\omega_{\max}$ are adequate bounds expressing the natural filtering resulting from propagation in the SOFAR channel and seismic recording.

Because of the complexity of the conversion mechanisms near the source and receiver, it is not possible to derive a universal correction and to interpret *TPEF* in terms of the absolute energy radiated by the source into the T phase. Also, such a correction would require a full understanding of the response of the receiving

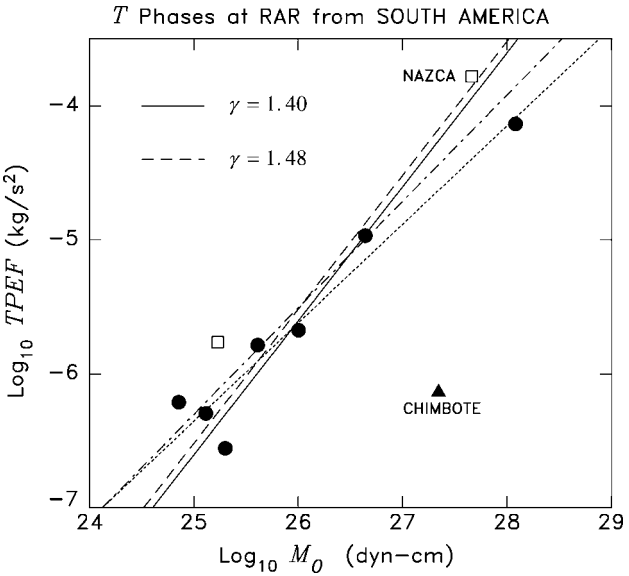


FIG. 16. Correlation between seismic moment M_0 and *TPEF*, as recorded at Rarotonga for earthquakes along a 250-km stretch of the coast of Central Chile (solid dots). The dotted line, with a slope of 0.74, is the best linear regression of the dataset of 7 points. The solid line, with a slope constrained to 1, defines the best-fitting value of γ (1.40). The two open squares show the large 1996 Nazca event, and one of its strong aftershocks; note that the best regression (dash-dot line, with a slope of 0.79) and the best-fitting γ (1.48; long dashes) are essentially unchanged, even though the new events are located in Peru, more than 1000 km away from the original seven. By contrast, the 1996 Chimbote "tsunami earthquake" (triangle) features an extreme deficiency in TPEF of two logarithmic units in γ, or about six times the scatter in the Chilean data. Adapted (re-labeled) from (Okal *et al.*, 2003).

shore to an acoustic wavetrain, and could therefore depend on such parameters as the back-azimuth of the wavetrain at the station. For this reason, Okal *et al.* (2003) limit the scope of the parameter *TPEF* to the comparison of records obtained at the same station under equivalent receiving conditions, i.e., from similar epicentral areas. In particular, no distance correction is applied in (6).

As in the case of body waves, the quantity *TPEF* would then be expected to scale linearly with seismic moment M_0 and thus the parameter $\Gamma = TPEF/M_0$ to remain invariant for sets of comparable seismic sources. (To avoid the use of exceedingly small numbers, we further introduce $\gamma = \log_{10}\Gamma + 30$, where Γ is in cm^{-2}.)

This is verified on Fig.16 which examines the relationship between *TPEF* and M_0 for ten South American earthquakes recorded at Rarotonga. Note the generally good agreement between the two parameters, suggesting a constant γ, the only earthquake exhibiting a pronounced deficiency in γ (by 1.5 logarithmic units) being the 1996 Chimbote, Peru, "tsunami earthquake". We recall that

Kanamori (1972) defined "tsunami earthquakes" as those events whose tsunamis are much larger than expected from their conventional magnitudes, typical examples being the 1896 Sanriku, 1946 Aleutian, 1975 Kuriles and more recently, 1992 Nicaraguan earthquakes. These have been shown to feature an extremely slow propagation of the rupture along the fault plane, at velocities as low as 1 km/s, itself interpreted as involving rupture in sedimentary wedges (as proposed for Sanriku (Tanioka and Satake, 1996) or Kuriles (Fukao, 1979)), or conversely in a jerky mode along a corrugated fault plane, in the case of a sediment-starved environment such as Nicaragua (Tanioka *et al.*, 1997; Polet and Kanamori, 2000).

The clear violation of source scaling laws by "tsunami earthquakes" is illustrated by their strong $m_b : M_s$ anomalies, recast in more quantitative terms by Newman and Okal (1998) as a deficiency of more than 1 logarithmic unit in the slowness parameter $\Theta = \log_{10}[E^E/M_0]$. Such a source slowness is expected to affect the parameter γ as well, and Okal *et al.* (2003) documented that all tsunami earthquakes are strongly deficient generators of T waves, by comparing them with regular events located in the same general area (Fig. 17). The case of the tsunami earthquakes of the 1990s (Nicaragua, 1992; Chimbote, Peru, 1996) for which high-quality digital records are available, is most compelling (γ deficient by an average of 1.8 and 1.7 logarithmic units, respectively), but the technique can be extended to analog records in the case of the 1963 Kuriles main aftershock, 1975 Kuriles earthquake and 1982 Tonga event, all of which were described as "tsunami earthquakes" (Fukao, 1979; Talandier and Okal, 1989). Also, Okal *et al.* (2003) and Okal (2004) showed that a faint T wave detected at Hawaii Volcano Observatory in the aftermath of the 1946 Aleutian earthquake arrived too late to be associated with the main shock, and thus confirmed that the latter had not generated a detectable T phase, which further supports its character as an exceptionally slow tsunami earthquake (Kanamori, 1972; López and Okal, 2006).

These results settled a controversy regarding the correlation between T-phase excitation and the tsunamigenic character of an earthquake source, which had been longstanding ever since Ewing *et al.* (1950) suggested the use of T phases in tsunami warning, based on the assumption that both kinds of waves would be favored by a very shallow seismic source involving strong coupling with the oceanic column. While their model was disputed by Leet (1951) and Wadati and Inouye (1953), Talandier (1966) and later Walker *et al.* (1992) and Walker and Bernard (1993) have supported the proposed correlation. However, such studies were mostly based on restricted datasets, and predated the large tsunamis of the 1990s, and in particular the "tsunami earthquakes" of 1992 (Nicaragua), 1994 (Java) and 1996 (Chimbote, Peru). As mentioned above, modern developments have rendered Ewing *et al.*'s (1950) model precarious since we now understand that T waves can be generated by the very deepest events (Northrop, 1974; Okal, 2001b), while the generation of far-field tsunami energy is only mildly controlled by source depth for reasonably shallow earthquakes ($h \leqslant 70$ km) (Ward, 1980; Okal, 1988). In addition, Ewing *et al.* (1950) were simply overlooking the extreme

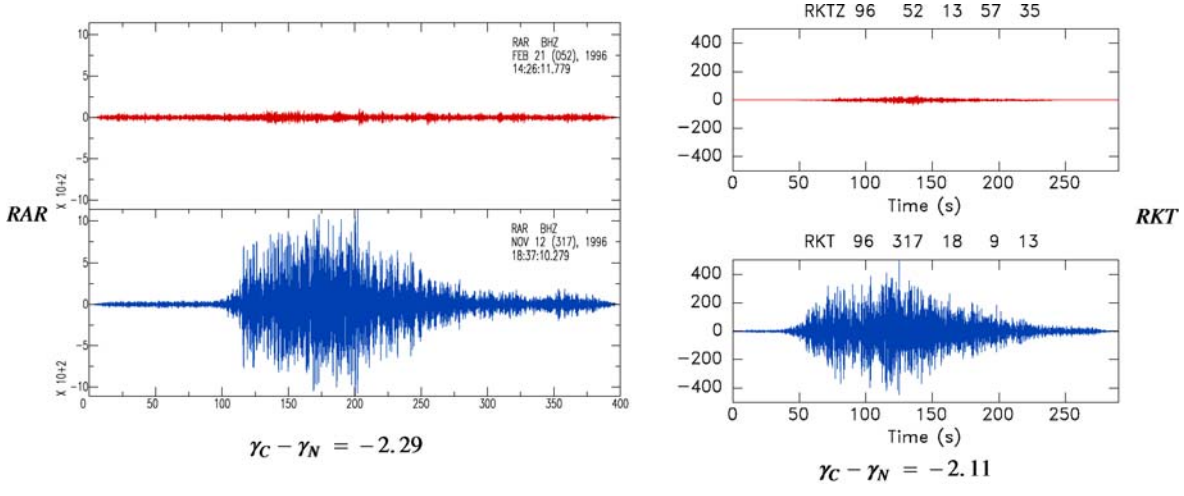


FIG. 17. Comparison of T phases from the Chimbote, Peru "tsunami earthquake" of 21 February 1996 ($M_0 = 2.2 \times 10^{27}$ dyn cm; top frames) and the regular Nazca, Peru event of 12 November 1996 ($M_0 = 4.6 \times 10^{27}$ dyn cm; bottom frames). Vertical scales (in digital units) are arbitrary, but for each station (RAR: Rarotonga, Cook Is.; RPN: Rapa Nui, Easter Is.; RKT: Rikitea, Gambier; PATS: Pohnpei, Caroline Is.), the two T wavetrains are plotted on the same scale. The strong deficiency of the Chimbote T waves is quantified by the difference in the parameters γ of the two events at each station. Adapted (re-labeled) from (Okal *et al.*, 2003).

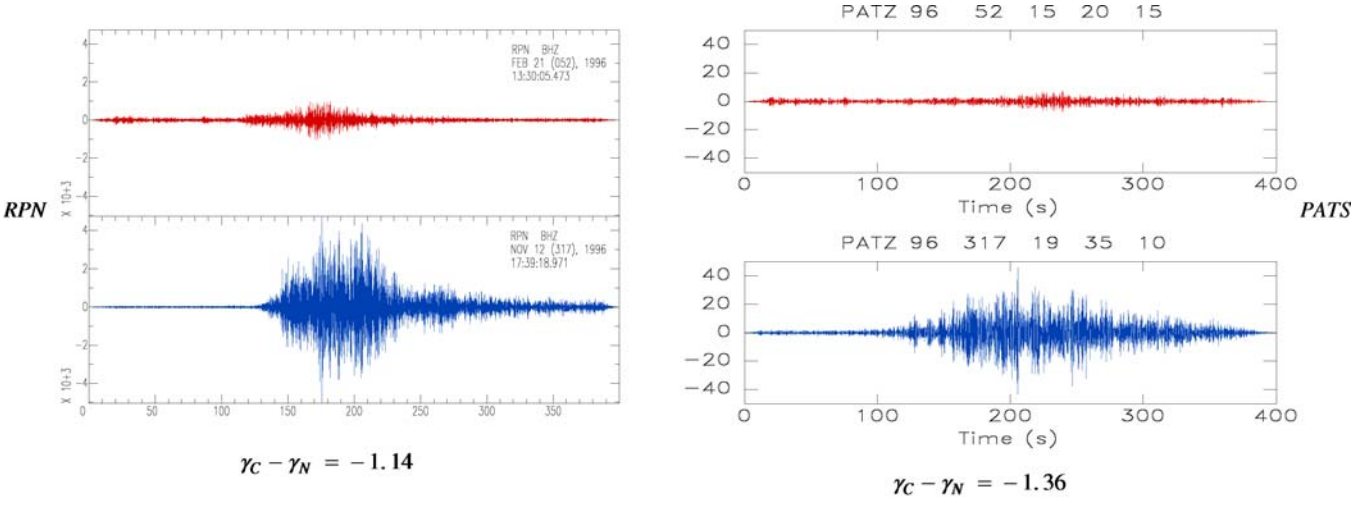


FIG. 17. Continued.

and fundamental difference in the frequencies characteristic of the two types of waves. Indeed, the most efficient tsunami generators, the so-called "tsunami earthquakes", are essentially silent in the acoustic band, just as they often have minimum amplitudes at 1 Hz. As a result, they can remain unfelt by the local residents, who then lack any harbinger of the upcoming disaster, and do not realize the need to urgently evacuate beaches. Such was the tragic case in Nicaragua (1992) where the tsunami killed more than 160 people (Satake *et al.*, 1993), and in Java (1994 and 2006), where the human toll reached 223 and about 700, respectively (Tsuji *et al.*, 1995, Fritz *et al.*, 2007).

The failure of the proposed correlation between the excitation of tsunamis and T waves is further illustrated by the fact that those few earthquakes whose T waves were actually felt at teleseismic distances in the Pacific Basin (the 1958 Fairweather, Alaska, 1977 Tonga, 1994 deep Bolivian and 2003 New Zealand earthquakes; see above) all had at most weak teleseismic tsunamis, while the major catastrophic transpacific tsunamis (1946 Aleutian; 1960 Chile) were not, to our knowledge, accompanied by T waves of sufficient amplitude to be felt in the far field. The 2004 Sumatra T waves were felt in the Maldives, at comparatively short distances.

In conclusion, the concept of the *TPEF* and the associated parameter γ can be a powerful tool for the investigation of the dynamic properties of a seismic source. In particular, it proves the deficiency of tsunami earthquakes as T-phase generators and establishes firmly this property in the context of seismic source theory, by invoking rupture slowness, itself resulting in a deviation from generally accepted source scaling laws. *TPEF* and hence γ involve straightforward, one-station measurements which can be carried out in real time in an observatory environment, and offer immediate insight into the dynamic properties of the source of a distant earthquake.

The Amplitude-Duration Criterion D

In a recent contribution, Talandier and Okal (2001) introduced a source discriminant D comparing the duration and amplitude of teleseismic T wavetrains recorded at island stations. In the context of the monitoring of the CTBT, their motivation was primarily to devise a robust algorithm capable of discriminating between explosive sources in the water column and elastic dislocations in the solid Earth. Specifically, and given the time series $x(t)$ of the ground velocity recorded by a seismic station, they characterize its amplitude through the maximum, e_{Max}, expressed in μm/s, of the envelope $e(t)$ of $x(t)$, after high-pass filtering x above 2 Hz. The duration $\tau_{1/3}$ of the wavetrain is taken as the total time (in s) during which the envelope $e(t)$ maintains a value of at least $\frac{1}{3}e_{\mathrm{Max}}$ (Fig. 18). The discriminant D is then computed as

$$D = \log_{10} e_{\mathrm{Max}} - 4.9 \log_{10} \tau_{1/3} + 4.1. \tag{7}$$

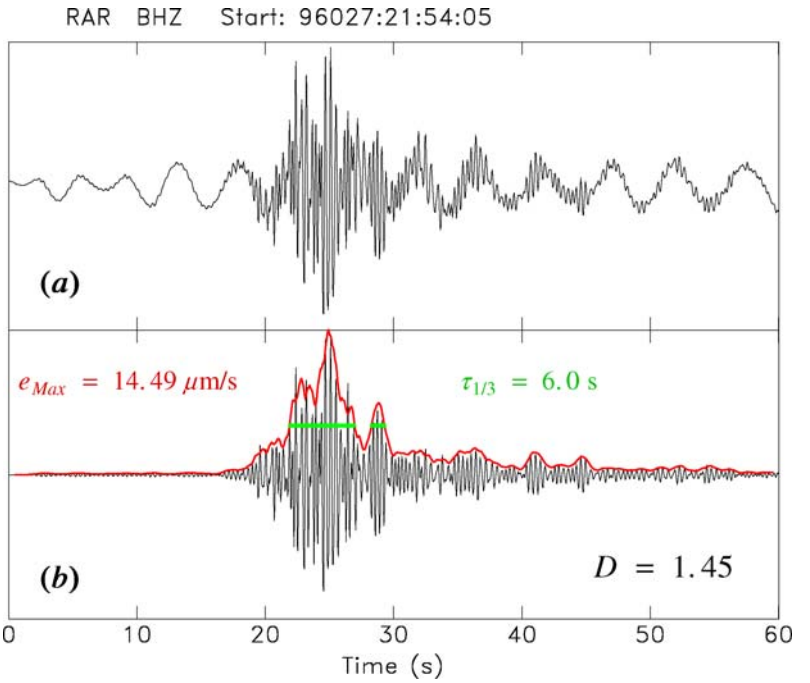


FIG. 18. Computation of the source discriminant D illustrated in the case of the Polynesian nuclear test of 27 January 1996, recorded at Rarotonga. (a): Original broad-band seismogram. (b): Seismogram high-pass filtered at 2 Hz. The thick trace represents the envelope of the record, whose maximum amplitude e_{Max} is retained. The thick horizontal segments identify those time intervals during which the envelope is sustained at 1/3 of its maximum value (estimated above ambient noise); their total duration is summed to obtain $\tau_{1/3}$. Finally, the discriminant D is computed using (7).

Figure 19 shows that D effectively separates most earthquakes ($D < 0$) from documented explosions ($D > 0$).

In developing this discriminant, Talandier and Okal (2001) were motivated by the failure of several other techniques, applied either in the time or frequency domains, to adequately separate populations of earthquakes and underwater explosions. By comparing an instantaneous measurement (the envelope maximum e_{Max}) and an integrated one (the duration $\tau_{1/3}$), they contrast the high- and low-frequency properties of the source, in a way reminiscent of the time-honored $m_b : M_s$ discriminant used to identify underground explosions based on their classical seismic waves (Marshall and Basham, 1972).

In order to justify the performance of the discriminant D, Talandier and Okal (2001) have discussed in very general terms the scaling laws governing the growth with source size of underground dislocations and underwater explosions, and the resulting differences in the power laws controlling the duration and amplitude

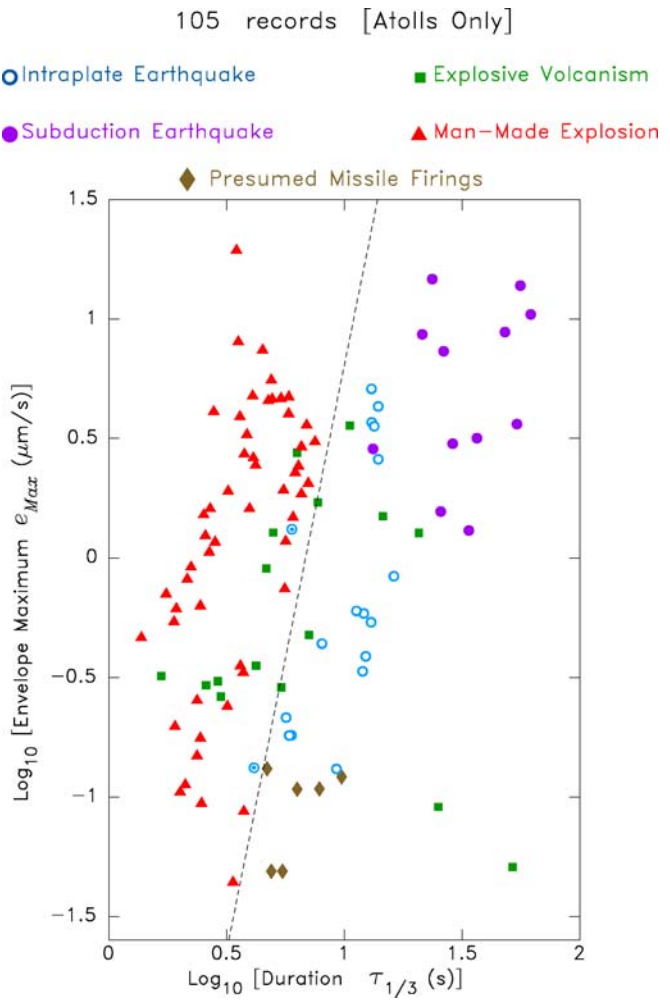


FIG. 19. Performance of the discriminant D (Eq. (7)) for the initial dataset of 105 records at atoll stations studied by Talandier and Okal (2001). The dashed line separates positive and negative values of D. Note that all chemical explosions in the oceanic column (triangles) have positive D, while most earthquakes (circles) feature negative discriminants. The exceptions (shown as bull's eye symbols) involve two small earthquakes originating in the Hawaiian volcanic structure; see text for details. Adapted (re-labeled) from Talandier and Okal (2001).

of the acoustic wave generated in the SOFAR channel, which indeed lead to a predictable separation of the two populations in the $e_{\rm Max}$: $\tau_{1/3}$ plane for large enough sources (magnitude $m_b \geqslant 2$ or yield $Y \geqslant 10$ kg).

We refer to Talandier and Okal (2001) for a full discussion of the operational details of the algorithm, notably regarding the definition of a necessary distance correction, the handling of background noise, the optimization (by trial and error) of the factor $1/3$ for the definition of duration, and the preferential use of receiving stations located either on atolls or in the vicinity of a steep, "pali"-like, conversion zone, in order to prevent a long receiver-side upslope conversion from lengthening the wavetrain and diminishing its amplitude.

An important feature of Fig. 19 is the observation that, while documented explosions are well identified by the parameter D, the performance of the discriminant is degraded for a small number of earthquakes occurring in the vicinity of intraplate volcanic edifices, such as Hawaii (and Mehetia in the Society Islands), and dubbed "hotspot earthquakes". These events feature an essentially null, or even slightly positive, value of D, which Talandier and Okal (2001) explain by noting that, for small earthquakes occurring in the immediate vicinity of the source-side conversion point, the ground motion giving rise to the acoustic wave may belong to the seismological near field, rather than to the far field, and thus violate the scaling laws providing the basis for the discriminant D. While this mis-identification of a few intraplate earthquakes may appear as a failure of the proposed discriminant, it actually shows that D is a powerful proxy for the characteristics of T-wave sources, and suggests its use to explore systematically the source properties of certain classes of earthquakes.

Note also on Fig. 19 that D fails to identify T-phase sources documented as belonging to episodes of underwater volcanic activity. This merely reflects the broad variability of the nature of these sources during the course of a volcanic eruption (e.g., Talandier and Okal, 1987). These may include explosive events, corresponding to the initial unplugging of magmatic conduits, which not surprisingly can share the characteristics of underwater explosions and feature positive values of D, as well as slower phenomena occurring later in the sequence and associated with the progression of magma within the conduit and its delivery to the ocean floor; the latter can then feature some of the most negative values of D in Talandier and Okal's (2001) dataset.

Talandier and Okal (2004b) recently complemented their initial study by substantially enlarging their dataset through inclusion of specially targeted families of sources. Among their most significant results, they confirmed the special behavior of many, but not all, "hotspot earthquakes" which show enhanced values of the discriminant D (Fig. 20a). They were able to associate those records featuring positive D with sources in the immediate vicinity of steep conversion zones, which provide an efficient conversion resulting in high amplitudes and short durations.

By contrast, Talandier and Okal (2004b) studied 10 records from true intraplate earthquakes, located in abyssal areas where no significant bathymetry is documented. They all feature negative values of D, and are thus unequivocally identified as earthquakes (Fig. 20b). These include some of the most negative D

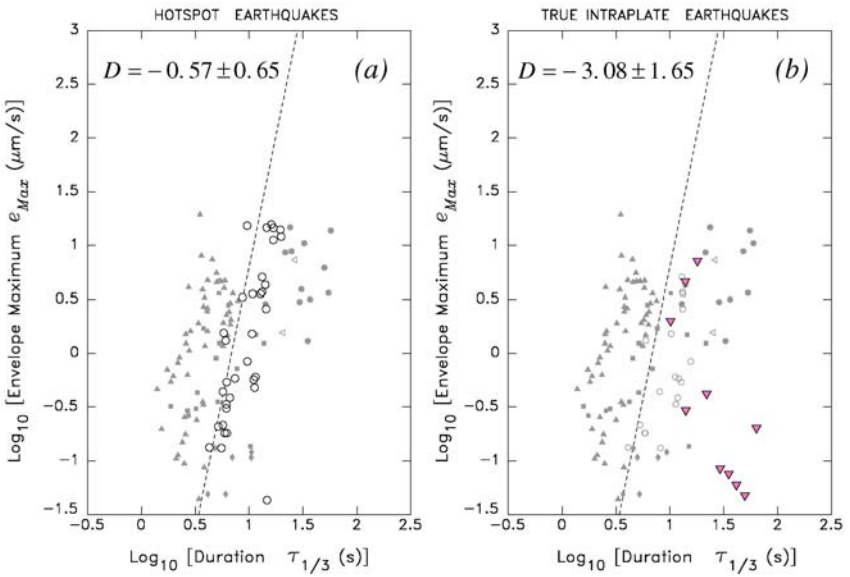


FIG. 20. Discriminant D applied to intraplate earthquakes located inside (or close to) volcanic edifices ("Hotspot earthquakes"; (a): open circles) or in abyssal plains with no known bathymetric features ("True intraplate"; (b): downward-pointing triangles). The background symbols in half-tone are the original dataset from Talandier and Okal (2001). $D = 0$ is shown as the dashed line. Note the occasional positive values of D for hotspot earthquakes, corresponding to sources located close to steep slopes in the volcanic edifice. By contrast, the true intraplate earthquakes all have negative values of D, some of them among the lowest measured. After Talandier and Okal (2004b).

values in the authors' catalogue, which expresses the general inefficiency of the scattering mechanisms required for abyssal conversion, especially at the relatively low acoustic frequencies typical of seismic recording. In addition, the delocalization of the scatterers on the ocean floor can contribute to an increase in the duration of the signal, as modeled by deGroot-Hedlin and Orcutt (1999) and Yang and Forsyth (2003). The occasionally larger but still negative values of D shown on Fig. 20b could result from conversions involving an unchartered bathymetric feature.

Figure 21 further compares the properties of T waves generated from earthquakes at the subduction zones of the Pacific "Ring of Fire" and along transform faults of the Pacific Mid-Oceanic Ridge systems. While the average value of D is slightly greater for the latter, there is no clear separation of the two populations. Anomalous events, featuring normal faulting on the Eltanin Transform Faults (Okal and Langenhorst, 2000), are not identifiable within the group. On the other hand, this dataset does support the observation that strike–slip earthquakes (from the Transform Fault dataset) may be nominally more efficient T-wave generators than thrust or normal faulting ones, especially at lower magnitudes. Figure 22

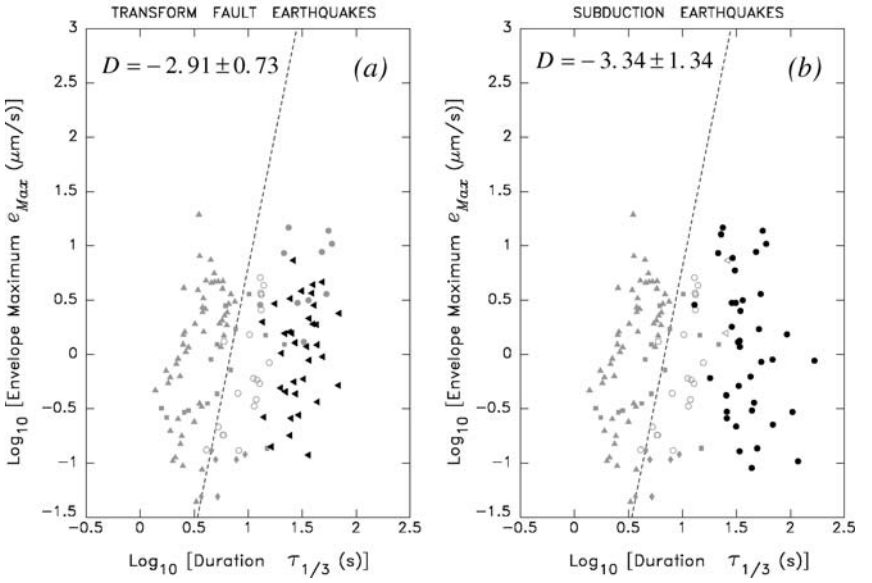


FIG. 21. Comparison of the performance of the discriminant D for 35 records from earthquakes on transform faults ((a): left-pointing triangles) and 53 records of subduction zone events ((b): solid circles). The background symbols in half-tone are the original dataset from Talandier and Okal (2001). $D = 0$ is shown as the dashed line. After Talandier and Okal (2004b).

(Talandier and Okal, 2004b) documents and quantifies this trend on a population of 23 T phases recorded at Rarotonga from the Eltanin and Kurile epicentral areas. These results are generally compatible with those of Dziak (2001), who had first reported this property using hydrophone records of regional events in the area of the Juan de Fuca and Gorda Ridges (Fig. 22a). While the trends of the two datasets are largely similar, the slopes of the best-fitting regressions differ significantly. However, in Talandier and Okal's (2004b) study, the regression slopes had to be computed using body-wave magnitude m_b in the absence of a sufficient number of published moment values, whereas Dziak (2001) used moment magnitudes obtained by inversion of regional waveforms (Nábělek and Xia, 1995). In the relevant range of magnitudes, the slope of m_b vs. $\log_{10} M_0$ is expected to vary between $2/3$ ($m_b \leqslant 5$) and $1/3$ ($m_b > 5$) (Geller, 1976), and consequently, Talandier and Okal's (2004b) slopes should be adjusted by a factor which could reach 0.5, in order to be compared with those of Dziak (2001). Additional differences between the two studies may reflect the variation in frequency band resulting from the difference in instrumentation.

Park *et al.* (2001) have suggested that the preferential excitation of T waves by strike–slip faults may express the dependence of Rayleigh wave excitation on focal mechanism and depth (e.g., Kanamori and Stewart, 1976). A simple physi-

cal analysis of this conjecture is not forthcoming, as the extremely short seismic wavelengths involved (i.e., the very large equivalent angular orders in the spherical Earth, on the order of $l = 10^4$ to 10^5 for the modes on Fig. 10) preclude the use of asymptotic expansions valid for shallow sources or in the limit of the seismic near field; note in particular that the scatterers are bound to range several, if not many, wavelengths away from the epicenter. In addition, the apparent decrease of the effect with earthquake size, recognizable in both Dziak's (2001) and Talandier and Okal's (2004b) datasets (Fig. 22), would suggest that source finiteness plays a crucial role; however, its effect on the excitation of seismic surface waves does not lend itself to a simple analysis at very high frequencies, both in the directivity formalism (Ben-Menahem, 1961) or the higher-moment one (Okal, 1982). By contrast, the formalism of Yang and Forsyth (2003) may, at least conceptually, be more directly applicable to further exploring the effect of focal mechanism on the generation of abyssal T waves; at any rate, it is clear that more systematic theoretical investigations will be necessary before the origin of this remarkable observation is fully understood.

Talandier and Okal (2004b) also explored the T-wave characteristics of underwater landslides, by considering the case of the Papua New Guinea (PNG) earthquake series of 1998 (Fig. 23). We recall that a major local tsunami, which killed more than 2200 people, was generated following a relatively weak earthquake ($M_0 = 3.7 \times 10^{26}$ dyn cm), which did not feature the slow character of "tsunami earthquakes" (Synolakis *et al.*, 2002). The origin of the tsunami was later attributed to an underwater landslide, triggered 13 minutes after the mainshock and identified from its T wave recorded throughout the Pacific Basin, and in particular on the Wake Island hydrophone (Okal, 2003). The T waveform was exceptionally long, given the small classical magnitude ($m_b = 4.4$) assigned to its source. This can be interpreted as a disparity between amplitude and duration of the source, which we formally examine on Fig. 23, where we plot D values computed from waveforms recorded at the seismic station on Wake Island. In addition to the tsunamigenic slump (with origin time 09:02 GMT on 17 July 1998), the dataset on that figure consists of the 1998 PNG mainshock (08:49), of a number of small aftershocks (09:06, 09:40), and of the mainshock of the 2002 PNG event, which occurred 120 km to the Southeast, was larger than the 1998 earthquake, but generated only a benign tsunami (Borrero *et al.*, 2003). Note that the 09:02 event ($D = -3.55$) effectively features a deficiency in D of more than one logarithmic unit with respect to most of the neighboring events, suggesting a source process of a different physical nature. By contrast, the 1998 and 2002 mainshocks feature comparable values of D (-1.58 and -1.26 respectively), supporting a commonality of seismic source, which again requires that the catastrophic 1998 tsunami must have been generated by an auxiliary source.

In addition, we point out a small event on 12 December 2002, located 6 km off the southwestern shore of the Island of Hawaii, and featuring $D = -3.18$. While this value is not significantly different from those of truly intraplate earthquakes,

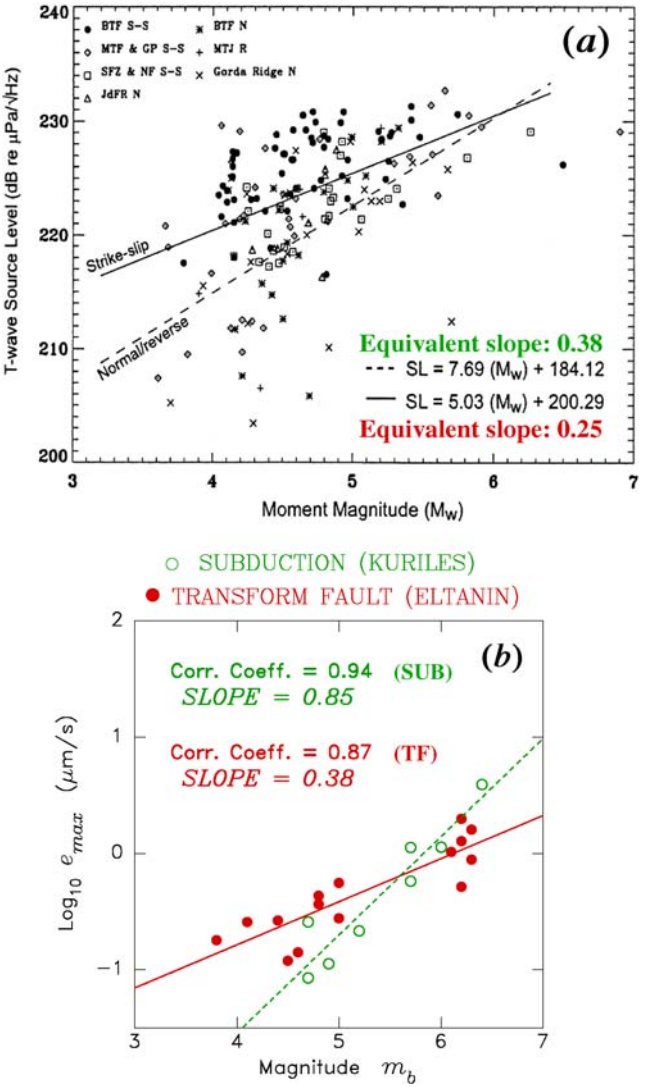


FIG. 22. Preferential excitation of T waves by strike–slip earthquakes. (a): Dziak's (2001) dataset of 179 regional earthquakes recorded on hydrophones in the Northeast Pacific. The regression slopes have been reinterpreted in terms of linear physical units, rather than decibels, as 0.25 (strike–slip) and 0.38 (other), (b): Talandier and Okal's (2004b) dataset of 8 T phases from subduction events in the Kuriles (open symbols; regressed as dashed line), and 15 transform fault earthquakes from the Eltanin Fracture Zone (solid symbols; regressed as solid line), recorded at Rarotonga. Note the generally similar pattern of preferential excitation of T phases by the latter at low magnitudes. See text for discussion. Combination of new material with figure adapted from Dziak (2001).

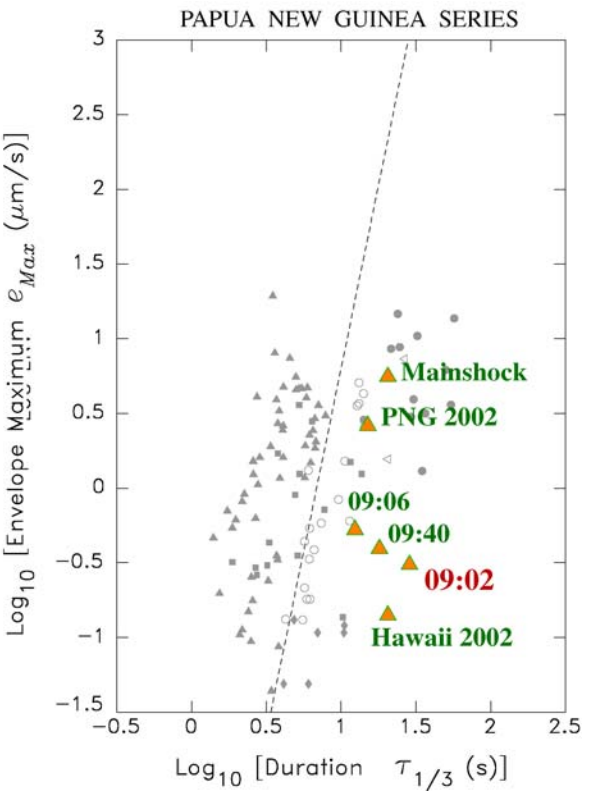


FIG. 23. Discriminant D computed at the Wake Island seismic station for a series of events in Papua New Guinea on 17 July 1998. Note that the 09:02 aftershock, interpreted as a major underwater slump (Synolakis *et al.*, 2002), features a deficit of more than one unit ($D = -3.55$) with respect to neighboring earthquakes. Also shown are the regular event nearby on 08 September 2002, and a small 2002 event near the coast of Hawaii interpreted by Reymond *et al.* (2004) as an underwater landslide. The background symbols in half-tone are the original dataset from Talandier and Okal (2001). $D = 0$ is shown as the dashed line. After Talandier and Okal (2004b).

it is substantially deficient among Hawaiian earthquakes, and reminiscent of that of the 1998 PNG underwater landslide. This observation, and the comparison of the regional seismic waveforms recorded on the island with those of documented subaerial landslides, suggests that the event in question is indeed an underwater landslide, taking place on a steeply sloping section of the island structure identifiable in the local bathymetry (Reymond *et al.*, 2004). This tentative interpretation would support the contention that landslides are probably rather inefficient, if not occasionally silent, sources of hydroacoustic energy, even though their duration may be substantial. Incidentally, such signals differ significantly from those described by Caplan-Auerbach *et al.* (2001) as originating from bench collapses

at the Kilauea eruption ocean entry point and recorded on hydrophones of the autonomous equatorial array (Fox *et al.*, 2001). The latter have much shorter durations, which probably reflects the difference in the nature of the slide (the simple collapse of a cohesive block) and the shorter length of its total underwater path.

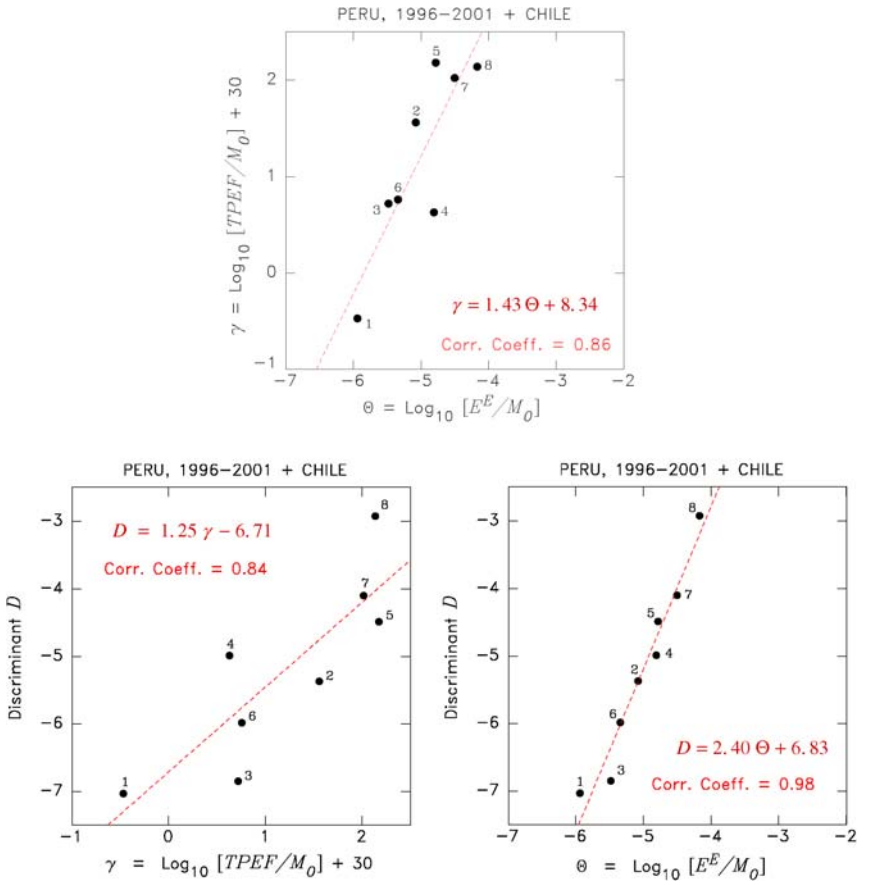
Finally, in Fig. 24, we compare on a dataset of eight South American earthquakes the performance of the three parameters D, γ, and Θ, introduced by Talandier and Okal (2001), Okal *et al.* (2003) and Newman and Okal (1998) to detect any possible slowness in the rupture characteristics of seismic sources. The first two are based on the properties of T waves, and the third one on the energy of body waves. The eight events span two logarithmic units of Θ, including both slow earthquakes such as the 1996 Chimbote, Peru "tsunami earthquake", and fast or "snappy" ones such as the Chilean event of 15 October 1997 (Okal and Kirby, 2002). We find remarkable correlations between γ and Θ (86%), and between D and γ (84%), and a spectacular one (98%) between D and Θ. In particular, the 1996 Chimbote, Peru earthquake features the lowest value of D (-7.03) measured by Talandier and Okal (2004b). The 2001 Peruvian mainshock, identified as trending toward slowness (Okal *et al.*, 2002) also features a low value of D, and to a lesser extent, of γ. By contrast, all three "snappy" earthquakes with large values of Θ also have high parameters γ and D. These include the two deeper than normal intraplate events in Chile and the normal faulting Peruvian shock triggered on 05 July 2001 by stress transfer from the Peruvian mainshock.

These observations serve to confirm, if need be, that D can be a reliable descriptor of the source properties of an earthquake, and more generally that the T-phase waveform carries valuable information on the source properties of earthquakes, which can be extracted through the use of analytical algorithms paralleling the methodologies traditionally utilized on conventional seismic waves.

The Case of the 2004 Sumatra Earthquake

With a seismic moment of at least 10^{30} dyn cm (Stein and Okal, 2005; Tsai *et al.*, 2005), the disastrous Sumatra–Andaman earthquake of 26 December 2004 was the largest one in 40, and perhaps 44, years. A fascinating aspect of this event is that its recordings transcended the conventional boundaries of instrumentation. For example, its tsunami was recorded by technologies as diverse as satellite altimetry (Scharroo *et al.*, 2005), infrasound (Le Pichon *et al.*, 2005), hydrophones of the International Monitoring System (IMS) of the CTBT Organization (Hanson and Bowman, 2005), detection of ionospheric perturbation using arrays of GPS recorders (Occhipinti *et al.*, 2005), and even on the horizontal components of near-shore broadband seismometers at teleseismic distances (Yuan *et al.*, 2005; Okal, 2007). In this context, and given the unique character of the earthquake, a description of its T waves is in order.

The source characteristics of the Sumatra earthquake were investigated in the framework of the parameters γ and D introduced above. In this respect, we recall



Number	Date	Event	Remarks
1	1966:052	Chimbote, Peru	Tsunami Earthquake
2	1996:317	Nazca, Peru	
3	2001:174	Peru, Main shock	
4	2001:177	Peru, First Large Aftershock	
5	2001:186	Peru, Triggered Normal Faulting	Snappy
6	2001:188	Peru, Largest Aftershock	
7	1997:288	Ovalle, Chile	Snappy (Okal and Kirby, 2002)
8	1998:210	Chile (Outboard, Intraplate)	Snappy

FIG. 24. Correlation between various scale-invariant measures of the slowness of an earthquake source, investigated on T wave records at Rarotonga of eight South American earthquakes, listed in the box. *Top*: Parameter γ *vs.* slowness parameter Θ introduced by Newman and Okal (1998). *Center left*: Discriminant D vs parameter γ. *Center right*: Discriminant D *vs.* Θ. Note the excellent correlations, especially in the latter case. See text for detailed interpretation.

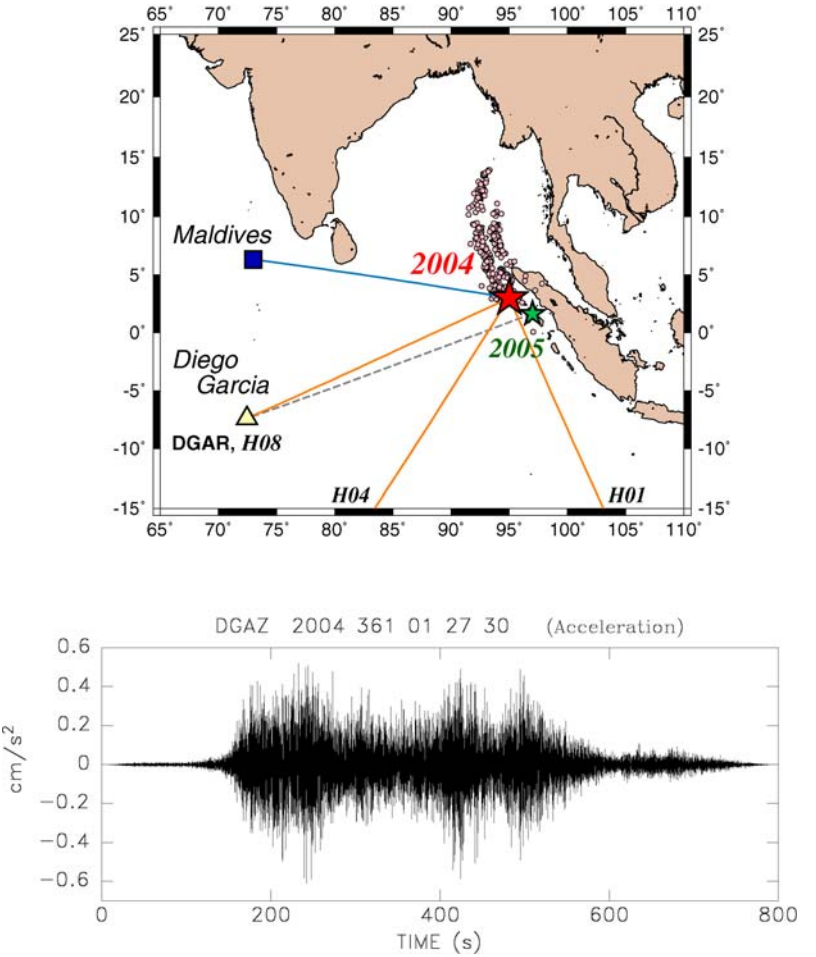


FIG. 25. *Top*: Map of the Northeastern Indian Ocean Basin showing the 2004 and 2005 Sumatra events (stars) and paths of T phases to the Maldives and Diego Garcia, with partial rays to Crozet (H04) and Cape Leeuwin (H01). The small symbols represent the same-day aftershocks of the 2004 Sumatra earthquake, and illustrate the extent of the fault rupture. *Bottom*: Accelerogram of the T phase deconvolved from the broadband seismometer record at the seismic station DGAR on Diego Garcia. Note that the duration of the record does not extend beyond 450 seconds.

that only relative evaluations at the same receiver are legitimate, and proceed to compare the T wavetrains recorded at the seismic station DGAR on Diego Garcia atoll from the 2004 Sumatra–Andaman earthquake and the Simeulue–Nias event of 28 March 2005 to the Southeast (Fig. 25). At $M_0 = 1.1 \times 10^{29}$ dyn cm, the latter, apparently triggered by Coulomb stress transfer from the 2004 event

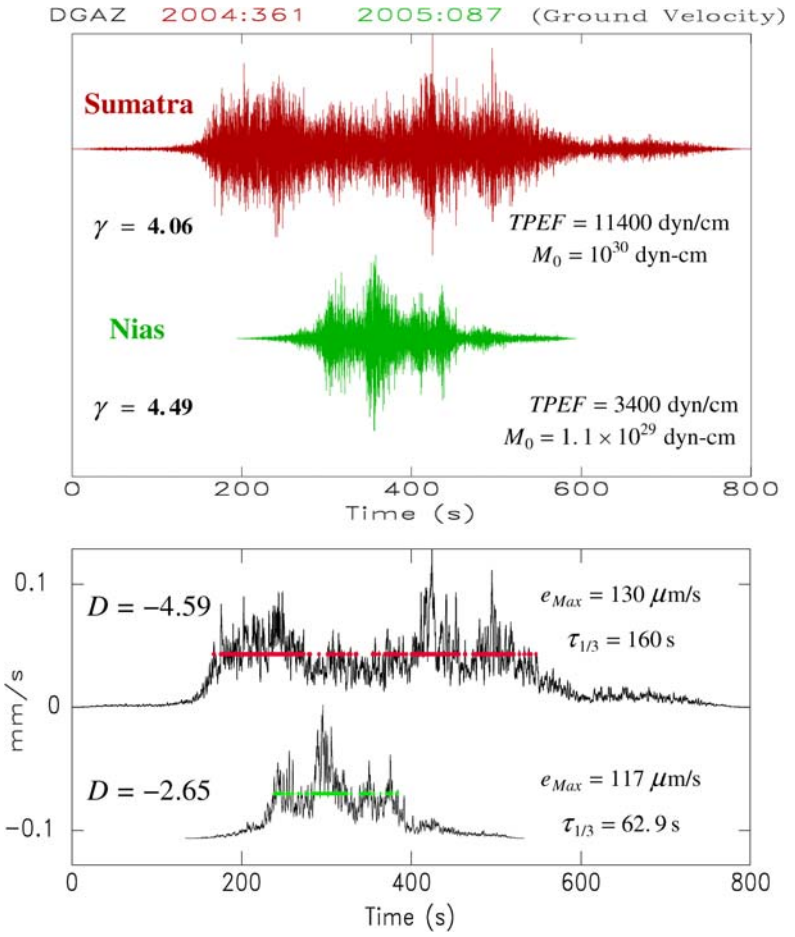


FIG. 26. Analysis of T phases recorded at the seismic station DGAR (Diego Garcia) from the two large 2004 and 2005 Sumatra earthquakes. *Top*: Broadband seismograms high-pass filtered at 2 Hz (yielding essentially ground velocity) plotted on common time and amplitude scales, with parameters relevant to the calculation of *TPEF* and γ. *Bottom*: Calculation of the discriminant D. Note the relatively comparable values of e_{Max}, but the very much longer $\tau_{1/3}$ in the case of the 2004 event, resulting in a much lower value of D.

(McCloskey *et al.*, 2005) is itself truly gigantic, and would have been the largest recorded event since 1965, if not for the 2004 Sumatra earthquake. Figure 26 compares the relevant waveforms and demonstrates a significant difference (of more than two units) in the discriminants D. Note that the Northern extremity of the 2004 rupture is generally masked in the direction of Diego Garcia, resulting in a somewhat shortened T wave, not exceeding 450 s in duration (Fig. 25),

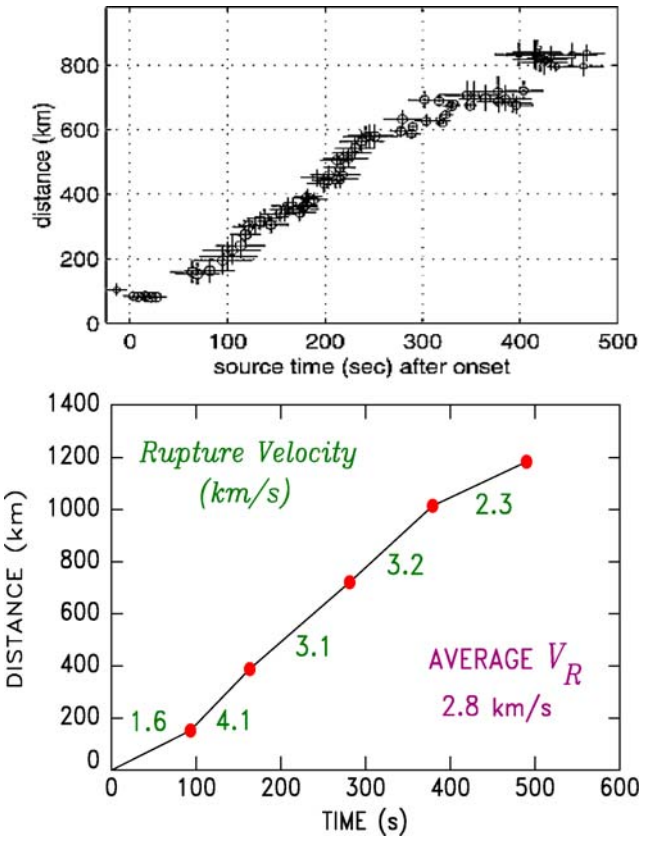


FIG. 27. Source tomography of the 2004 Sumatra earthquake. *Top*: Time history of the progression of the rupture along the fault line as mapped by backtracking beamed signals from Indian Ocean hydrophone arrays (deGroot-Hedlin, 2005). *Bottom*: Spaciotemporal evolution of the components of the multi-source CMT inversion of mantle waves by Tsai *et al.* (2005). The two diagrams use the exact same scales, allowing direct comparison. Note the remarkable agreement, especially regarding the slowing down of the rupture to the North, and also the tentatively slow initiation of the rupture in the South. Adapted (combined; re-scaled; re-labeled) from deGroot-Hedlin (2005) and Tsai *et al.* (2005).

rather than the generally accepted value of at least 500 s obtained from seismic source tomography (Ishii *et al.*, 2005). This also results in an underestimation of $\tau_{1/3}$, which would suggest an even more deficient value of D. By contrast, the Simeulue–Nias 2005 earthquake featured bilateral rupture (Walker *et al.*, 2005), resulting in an increased value of D. As for the parameters γ of the two events, they are found largely comparable.

T waves from the 2004 Sumatra earthquake were also recorded by hydrophones of the IMS; by a stroke of luck, the deployment of the hydrophone part of the sys-

tem started with the Indian Ocean sites, with stations operational in 2004 at Diego Garcia (H08), Crozet (H04) and Cape Leeuwin (H01). The availability of triads of hydrophones then allowed directional beaming of the signals using sophisticated techniques, like the PMCC method (Cansi, 1995), which tracks parameters such as incident azimuth, frequency, signal strength and coherence among array sites, as a function of time in the wave packet. By back-tracking this information to the source, T waves were successfully used in a series of studies (deGroot-Hedlin, 2005; Guilbert *et al.*, 2005; Tolstoy and Bohnenstiehl, 2005) to map the progression of the acoustic source, and hence of the earthquake rupture along the fault plane, in what amounted to the first applications of T phases to detailed earthquake source tomography. These studies confirmed the large spatial extent and long duration of the source, and documented a generally slow, but subtly variable, rupture velocity (Fig. 27). Note that because of the slower propagation velocity of T waves in the ocean, hydroacoustic signals offer in this respect a better resolution than comparable algorithms based on conventional body waves, such as used by Ishii *et al.* (2005) and Ohrnberger and Krüger (2005). The rupture velocity (on the average 2.7 km/s) was found to be significantly deficient with respect to that predicted by scaling laws (typically 3.5 km/s), but without reaching the very slow velocities, generally less than 2 km/s, typical of "tsunami earthquakes" (Polet and Kanamori, 2000; López and Okal, 2006). The emerging picture for the 2004 Sumatra event is that of a definitive trend towards slowness, but falling short of the truly anomalous behavior of the classic tsunami earthquakes, a property also confirmed by Weinstein and Okal's (2005) analysis of the parameter Θ of the 2004 earthquake.

In this context, it is interesting to note that tremors were reported widely felt in the Maldives around 06:30 local time (GMT + 5), which coincides with the predicted arrival time of T waves. There was no seismic station operational on the Maldives, but the T phases recorded at DGAR on Diego Garcia, 1530 km to the South (Fig. 25), correspond to a maximum ground acceleration of 0.5 cm/s^2. This is slightly below the generally accepted threshold for feeling by humans (1 cm/s^2), but we note the unfavorable recording conditions at DGAR, where the seismic station faces the Southeastern shore, thus requiring a significant land path for the converted seismic wave. It is thus almost certain that the Sumatra T waves were felt in the Maldives, four hours before the arrival of the tsunami, and in hindsight that their correct interpretation could have led to a warning of the population of this atoll nation, which suffered about 100 casualties. In this respect, the 2004 Sumatra earthquake may be the only reported case of an earthquake having generated a catastrophic oceanwide tsunami, and which had been felt in the far field through its T waves. The fact that the T waves were of high enough amplitude to be felt is also consistent with the relatively high value of γ, relative to the 2005 event: the 2004 earthquake did release significant amounts of high-frequency energy which found its way into T waves, and it was anomalous only in that its source lasted longer than expected, even for an event of its size. These remarks

further illustrate the different source characteristics controlling the parameters D and γ.

6. Conclusion

In conclusion of this review, it is interesting to reflect on the role of earthquake-generated T phases in the development of hydroacoustics. Underwater explosions obviously constitute simpler sources of acoustic energy in the ocean, and indeed have been used in a broad range of calibration experiments, including Worzel and Ewing's (1948), Weigel's (1990), and more recently Blackman *et al.*'s (2003).

However, the ubiquitous character of earthquake sources, both in space and time, has made them the primary tool of the steady progress of our command of hydroacoustics in the past 70 years. In this respect, and from the strict standpoint of unclassified geophysical research, several eras emerge as milestones, which can be conveniently identified by the individual laboratories where the crucial developments took place.

- The Lamont years (1950s), under the leadership of W. Maurice Ewing, saw the emergence of the fundamental picture, establishing firmly the nature of the T wave and the relevant interactions between seismic and acoustic waves at source and receiver slopes. The resulting model would pave the way for a decade of glorified applications of standard geometrical optics.
- The Hawaii years (1960s) involved the operation of a Pacific-wide array of hydrophones by the Hawaii Institute of Geophysics. Some of the most far-reaching results obtained by R.H. Johnson included the discovery of Macdonald Seamount, and the identification of the limits of validity of strict geometrical optics. The dataset collected was truly massive by those days' standards, and would fuel many years of seminal studies, notably by D.A. Walker, on various topical subjects. However, the 1970s and 1980s did not see major developments in instrumentation.
- The PMEL years (1990s) resulted from partial declassification of SOSUS, which allowed C.G. Fox to detect and monitor in real time an episode of volcanic eruption on a segment of the Mid-Oceanic Ridge system. In turn, this led to the development and deployment of the autonomous arrays in the vicinity of the East Pacific in Mid-Atlantic Ridges.
- The 2000s should become the years of the International Monitoring System (IMS), as hydroacoustics were defined in 1996 as one of four technologies to be used in the monitoring of the CTBT, which mandated the deployment of six hydrophones and five land-based seismic T-phase stations. While hydroacoustic monitoring is clearly aimed at the detection of man-made events in the water column, the growing database flowing from the IMS stations

(and from other networks such as the autonomous arrays and the RSP) emanates primarily from natural sources, some of which would not have been envisioned a decade ago, e.g., cryosignals generated by calving or drifting icebergs (Talandier *et al.*, 2002, 2006). However, earthquakes remain a premium source of hydroacoustic energy, and the new detection capabilities provided by the implementation of the IMS will allow the furtherance of our understanding of earthquake sources in the oceanic environment, notably through the use of the analytical methods being developed to interpret the hydroacoustic spectrum in terms of source properties.

In this context, T waves are truly coming of age as an equal partner to conventional seismic waves.

ACKNOWLEDGEMENTS

I am grateful to Ralph Stephen and Robert Odom, conveners of the workshop on "Seismo-acoustic applications in Marine Geology and Geophysics", held at Woods Hole in March 2004, for inviting me to present a keynote lecture, and encouraging me to write up a review paper on its topic. My work on T phases has benefited from many years of regular cooperation with the Commissariat à l'Energie Atomique (France), and particularly from joint research with Jacques Talandier, Dominique Reymond and Olivier Hyvernaud. Several aspects of T-phase quantification were supported by the Department of Defense, most recently under Contract DTRA01-00-C-0065, and regarding tsunami hazard by the National Science Foundation, under Grant CMS-03-01054. I thank Library Staff at Northwestern University, the University of Chicago and Harvard University for help with the retrieval of historical references. Most maps were drafted using the GMT software (Wessel and Smith, 1991).

REFERENCES

Adams, R.D. (1979). T-phase recordings at Rarotonga from underground nuclear explosions. *Geophys. J. Roy. Astr. Soc.* **58**, 361–369.

Bard, P.-Y., Campillo, M., Chavez-García, F.J., Sanchez-Sesma, F. (1988). The Mexico earthquake of September 19, 1985; A theoretical investigation of large- and small-scale amplification effects in the Mexico City valley. *Earthq. Spectra* **4**, 609–633.

Båth, M., Shahidi, M. (1971). T phases from Atlantic earthquakes. *Pure Appl. Geophys.* **92**, 74–114.

Ben-Menahem, A. (1961). Radiation of seismic surface waves from finite moving sources. *Bull. Seismol. Soc. Amer.* **51**, 401–435.

Ben-Menahem, A., Toksöz, M.N. (1963). Source mechanism from spectra of long-period seismic surface waves. 3. The Alaska earthquake of July 10, 1958. *Bull. Seismol. Soc. Amer.* **53**, 905–919.

Biot, M.A. (1952). The interaction of Rayleigh and Stoneley waves in the ocean bottom. *Bull. Seismol. Soc. Amer.* **42**, 81–93.

Blackman, D.K., Mercer, J.A., Andrew, R., deGroot-Hedlin, C.D., Harben, P.E. (2003). Indian Ocean Calibration tests: Cape Town–Cocos Keeling, 2003. In: Proc. 25th Seismic Res. Review – Nuclear Monitoring: Building the Knowledge Base, Tucson, AZ, September 23–25, 2003, pp. 517–523. Nat. Nuclear Security Admin. (abstract).

Boatwright, J., Choy, G.L. (1986). Teleseismic estimates of the energy radiated by shallow earthquakes. *J. Geophys. Res.* **91**, 2095–2112.

Bohnenstiehl, D.R., Tolstoy, M., Dziak, R.P., Fox, C.G., Smith, D.K. (2002). Aftershock sequences in the mid-ocean ridge environment: an analysis using hydroacoustic data. *Tectonophysics* **354**, 49–70.

Bohnenstiehl, D.R., Tolstoy, M., Smith, D.K., Fox, C.G., Dziak, R.P. (2003). Time-clustering behavior of spreading-center seismicity between 15 and 35° N on the Mid-Atlantic Ridge: observations from hydroacoustic monitoring. *Phys. Earth Planet. Inter.* **138**, 147–161.

Borrero, J.C., Bu, J., Saiang, C., Uslu, B., Freckman, J., Gomer, B., Okal, E.A., Synolakis, C.E. (2003). Field survey and preliminary modeling of the Wewak, Papua New Guinea earthquake and tsunami of September 9, 2002. *Seismol. Res. Lett.* **74**, 393–405.

Brekhovskikh, L.M. (1949). O rasprostranenii zvuka v podvodnom zvukovom kanale. *Dokl. Akad. Nauk SSSR* **69**, 157–160 (in Russian).

Brekhovskikh, L.M., Lysanov, Yu. (1982). Fundamentals of Ocean Acoustics. Springer-Verlag, Berlin, 250 pp.

Brune, J.N. (1964). Travel times, body waves, and normal modes of the Earth. *Bull. Seismol. Soc. Amer.* **54**, 2099–2128.

Butler, R., Lomnitz, C. (2002). Coupled seismoacoustic modes on the seafloor. *Geophys. Res. Lett.* **29** (10), 571–574.

Cansi, Y. (1995). An automatic seismic event processing for detection and location: The P.M.C.C. method. *Geophys. Res. Lett.* **22**, 1021–1024.

Cansi, Y., Béthoux, N. (1985). T waves with long inland paths: Synthetic seismograms. *J. Geophys. Res.* **90**, 5459–5465.

Caplan-Auerbach, J., Fox, C.G., Duennebier, F.K. (2001). Hydroacoustic detection of submarine landslides on Kilauea Volcano. *Geophys. Res. Lett.* **28**, 1811–1813.

Colladon, J.D., Sturm, J.K.F. (1827). La vitesse du son dans les liquides. *Annales de Chimie et de Physique, Série 2* **36**, 236–257.

Collins, M.P. (1936). Bulletin Number 5, Harvard University Seismograph Station, 5, 23 pp.

Coulomb, J., Molard, P. (1949). Ondes séismiques au fond de la mer des Antilles. *Ann. Geophys.* **5**, 212–214.

Coulomb, J., Molard, P. (1952). Propagation des ondes séismiques T dans la mer des Antilles. *Ann. Géophys.* **8**, 264–266.

deGroot-Hedlin, C.D. (2005). Estimation of the rupture length and velocity of the great Sumatra earthquake of Dec 26, 2004 using hydroacoustic signals. *Geophys. Res. Lett.* **32** (11), L11303, 4 pp.

deGroot-Hedlin, C.D., Orcutt, J.A. (1999). Synthesis of earthquake-generated T waves. *Geophys. Res. Lett.* **26**, 1227–1230.

Dietz, R.S., Sheehy, M.J. (1954). Transpacific detection of Myojin Volcanic eruptions by underwater sound. *Geol. Soc. Amer. Bull.* **65**, 942–956.

Duennebier, F.K., Johnson, R.H. (1967). T-phase sources and earthquake epicenters in the Pacific Basin, *Hawaii Inst. Geophys. Rept*. 67–24, Univ. Hawaii, Honolulu, 104 pp.

Dziak, R.P. (2001). Empirical relationship of T-wave energy and fault parameters of Northeast Pacific ocean earthquakes. *Geophys. Res. Lett.* **28**, 2537–2540.

Dziak, R.P., Fox, C.G., Schreiner, A.E. (1995). The June–July 1993 seismo-acoustic event at CoAxial segment, Juan de Fuca Ridge: Evidence for a lateral dike injection. *Geophys. Res. Lett.* **22**, 135–138.

Eaton, J.P., Richter, D.H., Ault, W.U. (1961). The tsunami of May 23, 1960 on the island of Hawaii. *Bull. Seismol. Soc. Amer.* **51**, 135–157.

Embley, R.W., Chadwick Jr., W.W., Jonasson, I.R., Butterfield, D.A., Baker, E.T. (1995). Initial results of the rapid response to the 1993 CoAxial event: Relationships between hydrothermal and volcanic processes. *Geophys. Res. Lett.* **22**, 143–146.

Ewing, W.M., Worzel, J.L. (1948). Long-range sound transmission. *Geol. Soc. Amer. Mem.* **27**. Part 3, 39 pp.

Ewing, W.M., Woollard, G.P., Vine, A.C., Worzel, J.L. (1946). Recent results in submarine geophysics. *Geol. Soc. Amer. Bull.* **57**, 909–934.

Ewing, W.M., Tolstoy, I., Press, F. (1950). Proposed use of the T phase in tsunami warning systems. *Bull. Seismol. Soc. Amer.* **40**, 53–58.

Ewing, W.M., Press, F., Worzel, J.L. (1952). Further study of the T phase. *Bull. Seismol. Soc. Amer.* **42**, 37–51.

Ewing, W.M., Jardetzky, W.S., Press, F. (1957). Elastic Waves in Layered Media. McGraw-Hill, New York, 380 pp.

Fischer, K.M., McCann, W.R. (1984). Velocity modeling and earthquake relocation in the Northeast Caribbean. *Bull. Seismol. Soc. Amer.* **74**, 1249–1262.

Forsyth, D.W., Yang, Y., Mangriotis, M.-D., Shen, Y. (2003). Coupled seismic slip on adjacent oceanic transform faults. *Geophys. Res. Lett.* **30** (12), 201–204.

Fox, C.G., Dziak, R.P., Matsumoto, H., Schreiner, A.E. (1994). Potential for monitoring low-level seismicity on the Juan de Fuca Ridge using military hydrophone arrays. *Mar. Tech. Soc. J.* **27** (4), 22–30.

Fox, C.G., Radford, W.E., Dziak, R.P., Lau, T.-K., Matsumoto, H., Schreiner, A.E. (1995). Acoustic detection of a seafloor spreading episode on the Juan de Fuca Ridge using military hydrophone arrays. *Geophys. Res. Lett.* **22**, 131–134.

Fox, C.G., Matsumoto, H., Lau, T.-K.A. (2001). Monitoring Pacific Ocean seismicity from an autonomous hydrophone array. *J. Geophys. Res.* **106**, 4183–4206.

Fritz, H.M. and 17 co-authors (2007). Extreme runup from the 17 July 2006 Java tsunami. *Geophys. Res. Lett.* **34** (2), L12602, 5 p.

Fukao, Y. (1979). Tsunami earthquake and subduction processes near deep sea trenches. *J. Geophys. Res.* **84**, 2303–2314.

Geller, R.J. (1976). Scaling relations for earthquake source parameters and magnitudes. *Bull. Seismol. Soc. Amer.* **66**, 1501–1523.

Gilbert, F. (1970). Excitation of the normal modes of the Earth by earthquake sources. *Geophys. J. Roy. Astr. Soc.* **22**, 223–226.

Guilbert, J., Vergoz, J., Schisselé, E., Roueff, A., Cansi, Y. (2005). Use of hydroacoustic and seismic arrays to observe rupture propagation and source extent of the $M_w = 9.0$ Sumatra earthquake. *Geophys. Res. Lett.* **32** (15), L15310, 5 pp.

Gutenberg, B. (1957). Effects of ground on earthquake motion. *Bull. Seismol. Soc. Amer.* **47**, 221–250.

Hammond, S.R., Walker, D.A. (1991). Ridge event detection: T-phase signals from the Juan de Fuca spreading center. *Mar. Geophys. Res.* **13**, 331–348.

Hanson, J.A., Bowman, J.R. (2005). Dispersive and reflected tsunami signals from the 2004 Indian Ocean tsunami observed on hydrophones and seismic stations. *Geophys. Res. Lett.* **32** (17), L17608, 5 pp.

Harkrider, D.G. (1964). Surface waves in multilayered elastic media. Part I: Rayleigh and Love waves from buried sources in a multilayered elastic half-space. *Bull. Seismol. Soc. Amer.* **54**, 627–679.

Haskell, N.A. (1953). The dispersion of surface waves in multilayered media. *Bull. Seismol. Soc. Amer.* **43**, 17–34.

Hiyoshi, Y., Walker, D.A., McCreery, C.S. (1992). T-phase data and regional tsunamigenesis in Japan. *Bull. Seismol. Soc. Amer.* **82**, 2213–2223.

Isacks, B.L., Molnar, P. (1971). Distribution of stresses in the descending lithosphere from a global survey of focal-mechanism solutions of mantle earthquakes. *Rev. Geophys. Space Phys.* **9**, 103–174.

Ishii, M., Shearer, P., Houston, H., Vidale, J. (2005). Rupture extent, duration and speed of the 2004 Sumatra–Andaman earthquake imaged by the Hi-Net array. *Nature* **435**, 933–936.

Jaggar, T.A. (1930). How the seismograph works. *The Volcano Letter* **268**, 1–4.

Jensen, F.B., Kuperman, W.A., Porter, M.B., Schmidt, H. (1994). Computational Ocean Acoustics. AIP Press, Woodbury, NY, 612 pp.

Johnson, R.H. (1970). Estimating earthquake rupture length from T waves. In: Adams, W.A. (Ed.), *Tsunamis in the Pacific Ocean*. Univ. Hawaii, Honolulu, pp. 253–259.

Johnson, R.H., Norris, R.A. (1968). T-phase radiators in the Western Aleutians. *Bull. Seismol. Soc. Amer.* **58**, 1–10.

Johnson, R.H., Northrop, J. (1966). A comparison of earthquake magnitude with T-phase strength. *Bull. Seismol. Soc. Amer.* **56**, 119–124.

Johnson, R.H., Northrop, J., Eppley, R. (1963). Sources of Pacific T phases. *J. Geophys. Res.* **68**, 4251–4260.

Johnson, R.H., Norris, R.A., Duennebier, F.K. (1968). Abyssally generated T phases. *Amer. Geophys. Un. Geophys. Monog.* **12**, 70–78.

Kanamori, H. (1972). Mechanisms of tsunami earthquakes. *Phys. Earth Planet. Inter.* **6**, 346–359.

Kanamori, H., Stewart, G.S. (1976). Mode of the strain release along the Gibbs fracture zone, Mid-Atlantic Ridge. *Phys. Earth Planet. Inter.* **11**, 312–332.

Keenan, R.E., Merriam, L.R.L. (1991). Artic abyssal T phases: Coupling seismic energy to the ocean sound channel via under-ice scattering. *J. Acoust. Soc. Amer.* **89**, 1128–1133.

Kennett, B.L.N. (1983). Seismic Wave Propagation in Stratified Media. Cambridge Univ. Press, 342 pp.

Kirby, S.H., Okal, E.A., Engdahl, E.R. (1995). The 09 June 1994 great Bolivian deep earthquake: An exceptional deep earthquake in an extraordinary subduction zone. *Geophys. Res. Lett.* **22**, 2233–2236.

Kuwahara, S. (1939). Velocity of sound in sea water and calculation of the velocity for use in sonic sounding. *Hydrog. Rev.* **16**, 123–140.

Lamb, H. (1904). On the propagation of tremors over the surface of an elastic solid. *Phil. Trans. Roy. Soc. London, Ser. A* **203**, 1–42.

Langevin, P. (1924). The employment of ultra-sonic waves for echo sounding. *Hydrog. Rev.* **2**, 57–91.

Lee, W.H.K., Bennett, R.E., Meagher, K.L. (1972). A method of estimating magnitude of local earthquakes from signal duration, U.S. Geol. Surv. Open File Rept., 28 pp.

Leet, L.D. (1951). Discussion of 'Proposed use of the T phase in tsunami warning systems'. *Bull. Seismol. Soc. Amer.* **41**, 165–167.

Leet, L.D., Linehan, S.J., Berger, P.R. (1951). Investigations of the T phase. *Bull. Seismol. Soc. Amer.* **41**, 123–141.

Leonard, M. (2004). T-phase perception: The August 2003, $M_w = 7.1$ New Zealand earthquake felt in Sydney, 1800 km away. *Seismol. Res. Lett.* **75**, 475–480.

Le Pichon, A., Kerry, P., Mialle, P., Vergoz, J., Brachet, N., Garcés, M., Drob, D., Ceranna, L. (2005). Infrasound associated with 2004–2005 large Sumatra earthquakes and tsunami. *Geophys. Res. Lett.* **32** (19), L19802. 5 pp.

Levitus, S., Boyer, T.P., Antonov, J., Burgett, R., Conkright, M.E. (1994). World Ocean Atlas 1994, NOAA/NESDIS, Silver Springs, MA.

Linde, A.T., Gladwin, M.T., Johnston, M.S., Gwyther, R.L., Bilham, R.G. (1996). A slow earthquake sequence on the San Andreas Fault. *Nature* **383**, 65–68.

Linehan, D.S.J. (1940). Earthquakes in the West Indian region. *Trans. Amer. Geophys. Un.* **21**, 229–232.

López, A.M., Okal, E.A. (2006). A seismological reassessment of the source of the 1946 Aleutian "tsunami" earthquake. *Geophys. J. Int.* **165**, 835–849.

Love, A.E.H. (1911). Some Problems of Geodynamics. Cambridge Univ. Press, 180 pp.

Marshall, P.D., Basham, P.W. (1972). Discrimination between earthquakes and underground explosions using an improved M_S scale. *Geophys. J. Roy. Astr. Soc.* **28**, 431–458.

McCloskey, J., Nalbant, S.S., Steacy, S. (2005). Earthquake risk from co-seismic stress. *Nature* **434**, 291.

Milder, D.M. (1969). Ray and wave invariants for SOFAR channel propagation. *J. Acoust. Soc. Amer.* **46**, 1259–1263.

Milne, A.R. (1959). Comparison of spectra of an earthquake T phase with similar signals from nuclear explosions. *Bull. Seismol. Soc. Amer.* **49**, 317–329.

Munk, W.H. (1979). Sound channel in an exponentially stratified ocean with applications to SOFAR. *J. Acoust. Soc. Amer.* **55**, 220–255.

Munk, W.H., Spindel, R.C., Baggeroer, A., Birdsall, T.G. (1994). The Heard Island feasibility test. *J. Acoust. Soc. Amer.* **96**, 2330–2342.

Nábělek, J., Xia, G. (1995). Moment-tensor analysis using regional data; application to the 25 March 1993, Scotts Mills, Oregon earthquake. *Geophys. Res. Lett.* **22**, 13–16.

Newman, A.V., Okal, E.A. (1998). Teleseismic estimates of radiated seismic energy: The E/M_0 discriminant for tsunami earthquakes. *J. Geophys. Res.* **103**, 26885–26898.

Norris, R.A., Johnson, R.H. (1969). Submarine volcanic eruptions recently located in the Pacific by SOFAR hydrophones. *J. Geophys. Res.* **74**, 650–664.

Northrop, J. (1974). T phases from the Hawaiian earthquake of April 26, 1973. *J. Geophys. Res.* **79**, 5478–5481.

Northrop, J., Menard, H.W., Duennebier, F.K. (1968). Seismic and bathymetric evidence of a fracture zone on Gorda Ridge. *Science* **161**, 688–690.

Occhipinti, G., Lognonné, P., Kherani, A., Hébert, H. (2005). Modeling and detection of ionospheric perturbation associated with the Sumatra tsunami of December 26th, 2004. *Eos, Trans. Amer. Geophys. Un.* **86** (52), U11A-0829 (abstract).

Odom, R.I. (1986). A coupled mode examination of irregular waveguides including the continuum spectrum. *Geophys. J. Roy. Astr. Soc.* **86**, 425–453.

Odom, R.I., Soukup, D.J. (2004). Model scattering of T waves: Sediment amplification and source effects. *J. Acoust. Soc. Amer.* **115**, 2445 (abstract).

Officer, C.B. (1958). Introduction to the Theory of Sound Transmission, with Application to the Ocean. McGraw-Hill, New York, 284 pp.

Ogata, Y. (1983). Estimation of the parameters of the modified Omori formula for aftershocks frequencies by the maximum likelihood procedure. *J. Phys. Earth* **31**, 115–124.

Ohrnberger, M., Krüger, F. (2005). Imaging of large earthquake rupture processes using multiple tele-seismic arrays: Application to the Sumatra–Andaman Islands earthquake. *Eos, Trans. Amer. Geophys. Un.* **86** (52), U11A-819 (abstract).

Okal, E.A. (1982). Higher moment excitation of normal modes and surface waves. *J. Phys. Earth* **30**, 1–31.

Okal, E.A. (1988). Seismic parameters controlling far-field tsunami amplitudes: A review. *Natural Hazards* **1**, 67–96.

Okal, E.A. (2001a). T-phase stations for the international monitoring system of the comprehensive nuclear-test ban treaty: A global perspective. *Seismol. Res. Lett.* **72**, 186–196.

Okal, E.A. (2001b). "Detached" deep earthquakes: Are they really?. *Phys. Earth Planet. Inter.* **127**, 109–143.

Okal, E.A. (2003). T waves from the 1998 Papua New Guinea earthquake and its aftershocks: Timing the tsunamigenic slump. *Pure Appl. Geophys.* **160**, 1843–1863.

Okal, E.A. (2004). Comment on "Source of the great tsunami of 1 April 1946: a landslide in the upper Aleutian forearc", by G.J. Fryer *et al.*. *Marine Geology* **209**, 363–369.

Okal, E.A. (2007). Seismic records of the 2004 Sumatra and other tsunamis: A quantitative study. *Pure Appl. Geophys.* **164**, 325–353.

Okal, E.A., Kirby, S.H. (2002). Energy-to-moment ratios for damaging intraslab earthquakes: Preliminary results on a few case studies. USGS Open File Rept., 02–328, pp. 127–131.

Okal, E.A., Langenhorst, A.R. (2000). Seismic properties of the Eltanin Transform System, South Pacific. *Phys. Earth Planet. Inter.* **119**, 185–208.

Okal, E.A., Talandier, J. (1981). Dispersion of one-second Rayleigh modes through oceanic sediments following shallow earthquakes in the southcentral Pacific Ocean basin. In: Kuperman, W.A., Jensen, K.B. (Eds.), *Bottom-Interacting Ocean Acoustics*. In: *NATO Confer. Ser.*, vol. IV:5. Plenum, New York, pp. 345–358.

Okal, E.A., Talandier, J. (1986). *T*-wave duration, magnitudes and seismic moment of an earthquake; application to tsunami warning. *J. Phys. Earth* **34**, 19–42.

Okal, E.A., Talandier, J. (1997). *T* waves from the great 1994 Bolivian deep earthquake in relation to channeling of *S* wave energy up the slab. *J. Geophys. Res.* **102**, 27421–27437.

Okal, E.A., Talandier, J. (1998). Correction to "*T* waves from the great 1994 Bolivian deep earthquake in relation to channeling of *S* wave energy up the slab". *J. Geophys. Res.* **103**, 2793–2794.

Okal, E.A., Dengler, L., Araya, S., Borrero, J.C., Gomer, B., Koshimura, S., Laos, G., Olcese, D., Ortiz, M., Swensson, M., Titov, V.V., Vegas, F. (2002). A field survey of the Camana, Peru tsunami of June 23, 2001. *Seismol. Res. Lett.* **73**, 904–917.

Okal, E.A., Alasset, P.-J., Hyvernaud, O., Schindelé, F. (2003). The deficient *T* waves of tsunami earthquakes. *Geophys. J. Int.* **152**, 416–432.

Park, M., Odom, R.I. (1999). The effect of stochastic rough interfaces on coupled-mode elastic waves. *Geophys. J. Roy. Astr. Soc.* **136**, 123–143.

Park, M., Odom, R.I., Soukup, D.J. (2001). Modal scattering: A key to understanding oceanic *T* waves. *Geophys. Res. Lett.* **28**, 3401–3404.

Pasyanos, M.E., Romanowicz, B.A. (1997). Observations of *T* phases across Northern California using the Berkeley digital seismic network. *Eos, Trans. Amer. Geophys. Un.* **78** (46), 461–462 (abstract).

Pedersen, M.A. (1969). Theory of the axial ray. *J. Acoust. Soc. Amer.* **45**, 157–176.

Pekeris, C.L. (1948). Theory of propagation of explosive sound in shallow water. *Geol. Soc. Amer. Mem.* **27**. Part 2, 117 pp.

Pierce, A.D. (1965). Extension of the method of normal modes to sound propagation in an almost stratified medium. *J. Acoust. Soc. Amer.* **37**, 19–27.

Polet, Y., Kanamori, H. (2000). Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys. J. Int.* **142**, 684–702.

Press, F., Ewing, W.M. (1950). Propagation of explosive sound in a liquid layer overlying a semi-infinite elastic solid. *Geophysics* **15**, 426–446.

Ravet, J. (1940). Remarques sur quelques enregistrements d'ondes à très courte période au cours de tremblements de terre lointains à l'observatoire du Faiere, Papeete, Tahiti. In: Sixth Pacific Congr., vol. 1, pp. 127–130.

Real, C.R., Teng, T.–L. (1973). Local Richter magnitude and total signal duration in Southern California. *Bull. Seismol. Soc. Amer.* **63**, 1809–1827.

Reymond, D., Hyvernaud, O., Talandier, J., Okal, E.A. (2003). *T*-wave detection of two underwater explosions off Hawaii on April 13, 2000. *Bull. Seismol. Soc. Amer.* **93**, 804–816.

Reymond, D., Hyvernaud, O., Talandier, J., Okal, E.A. (2004). Detection and identification of two anomalous events on the Big Island of Hawaii from teleseismic *T* phases. *Eos, Trans. Amer. Geophys. Un.* **85** (28), WP112 (abstract).

Reysenbach de Haan, F.W. (1966). Listening under water: thoughts on sound and cetacean hearing. In: Schevill, W.E. (Ed.), *The Whale Problem: A Status Report*. Univ. California Press, Berkeley, pp. 583–596.

Rozenberg, L.D. (1949). Ob odnom novom yavlenii v gidroakustike. *Dokl. Akad. Nauk SSSR* **69**, 175–176 (in Russian).

Saito, M. (1967). Excitation of free oscillations and surface waves by a point source in a vertically heterogeneous Earth. *J. Geophys. Res.* **72**, 3689–3699.

Satake, K., Bourgeois, J., Abe, Ka., Abe, Ku., Tsuji, Y., Imamura, F., Iio, Y., Katao, H., Noguera , E., Estrada, F. (1993). Tsunami field survey of the 1992 Nicaragua earthquake. *Eos, Trans. Amer. Geophys. Un.* **74**, 145 and 156–157.

Scharroo, R., Smith, W.H.F., Titov, V.V., Arcas, D. (2005). Observing the Indian Ocean tsunami with satellite altimetry. *Geophys. Res. Abstr.* **7**, 230 (abstract).

Schmidt, H., Baggeroer, A.B., Sperry, B.J. (2004). Wave-theory modeling of oceanic *T*-phase coupling at continental margin and seamounts. *J. Acoust. Soc. Amer.* **115**, 2444 (abstract).

Scholte, J.G.J. (1947). The range of existence of Rayleigh and Stoneley waves. *Mon. Not. Roy. Astr. Soc., Geophys. Supp.* **5**, 120–126.

Shevchenko, V.V. (1962). Irregular acoustic waveguides. *Sov. Phys. Acoust.* **7**, 389–397.

Shurbet, D.H. (1955). Bermuda *T* phases with large continental paths. *Bull. Seismol. Soc. Amer.* **45**, 23–35.

Shurbet, D.H., Ewing, W.M. (1957). *T* phases at Bermuda and transformation of elastic waves. *Bull. Seismol. Soc. Amer.* **47**, 251–262.

Smith, D.K., Tolstoy, M., Fox, C.G., Bohnenstiehl, D.R., Matsumoto, H., Fowler, M.J. (2002). Hydroacoustic monitoring of seismicity at the slow-spreading Mid-Atlantic Ridge. *Geophys. Res. Lett.* **29** (11), 131–134.

Smith, D.K., Escartin, J., Cannat, M., Tolstoy, M., Fox, C.G., Bohnenstiehl, D.R., Bazin, S. (2003). Spatial and temporal distribution of seismicity along the northern Mid-Atlantic Ridge (15°–35°N). *J. Geophys. Res.* **108** (B3), EPM8-1–EPM8-22.

Stafford, K.M., Fox, C.G., Clark, D.S. (1998). Long-range acoustic detection and localization of blue whale calls in the Northeast Pacific Ocean. *J. Acoust. Soc. Amer.* **104**, 3616–3625.

Stein, S., Okal, E.A. (2005). Size and speed of the Sumatra earthquake. *Nature* **434**, 581–582.

Sverdrup, H.U., Johnson, M.W., Fleming, R.H. (1942). The Oceans, Their Physics, Chemistry and General Biology. Prentice-Hall, New York, 1087 pp.

Swainson, O.W. (1936). Velocity and ray paths of sound waves in sea water. *U.S. Coast Geod. Surv. Field Eng. Bull.* **10**, 64 pp.

Synolakis, C.E., Bardet, J.-P., Borrero, J.C., Davies, H.L., Okal, E.A., Silver, E.A., Sweet, S., Tappin , D.R. (2002). The slump origin of the 1998 Papua New Guinea tsunami. *Proc. Roy. Soc. (London) Ser. A* **458**, 763–789.

Talandier, J. (1966). Contribution à la prévision des tsunamis. *C. R. Acad Sci. Paris B* **263**, 940–942.

Talandier, J., Okal, E.A. (1979). Human perception of *T* waves: the June 22, 1977 Tonga earthquake felt on Tahiti. *Bull. Seismol. Soc. Amer.* **69**, 1475–1486.

Talandier, J., Okal, E.A. (1982). Crises sismiques au volcan Macdonald (Océan Pacifique Sud). *C. R. Acad. Sci. Paris, Sér. II.* **295**, 195–200.

Talandier, J., Okal, E.A. (1987). Seismic detection of underwater volcanism: the example of French Polynesia. *Pure Appl. Geophys.* **125**, 919–950.

Talandier, J., Okal, E.A. (1989). An algorithm for automated tsunami warning in French Polynesia, based on mantle magnitudes. *Bull. Seismol. Soc. Amer.* **79**, 1177–1193.

Talandier, J., Okal, E.A. (1996). Monochromatic *T* waves from underwater volcanoes in the Pacific Ocean: Ringing witnesses to geyser processes?. *Bull. Seismol. Soc. Amer.* **86**, 1529–1544.

Talandier, J., Okal, E.A. (1998). On the mechanism of conversion of seismic waves to and from *T* waves in the vicinity of island shores. *Bull. Seismol. Soc. Amer.* **88**, 621–632.

Talandier, J., Okal, E.A. (2001). Identification criteria for sources of *T* waves recorded in French Polynesia. *Pure Appl. Geophys.* **158**, 567–603.

Talandier, J., Okal, E.A. (2004a). Hydroacoustic signals from presumed CHASE explosions off Vancouver Island in 1969–70: A modern perspective. *Seismol. Res. Lett.* **75**, 188–198.

Talandier, J., Okal, E.A. (2004b). Amplitude-duration and other discriminants for seismically recorded hydroacoustic phases. *Eos, Trans. Amer. Geophys. Un.* **85** (47), F1297 (abstract).

Talandier, J., Hyvernaud, O., Okal, E.A., Piserchia, P.-F. (2002). Long-range detection of hydroacoustic signals from large icebergs in the Ross Sea, Antarctica. *Earth Planet Sci. Lett.* **203**, 519–534.

Talandier, J., Hyvernaud, O., Reymond, D., Okal, E.A. (2006). Hydroacoustic signals generated by parked and drifting icebergs in the Southern Indian and Pacific Oceans. *Geophys. J. Int.* **165**, 817–834.

Tanioka, Y., Ruff, L.J., Satake, K. (1997). What controls the lateral variation of large earthquake occurrence along the Japan Trench?. *The Island Arc* **6**, 261–266.

Tanioka, Y., Satake, K. (1996). Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys. Res. Lett.* **23**, 1549–1552.

Teague, W.J., Carron, M.J., Hogan, P.J. (1990). A comparison between the generalized digital environmental model and Levitus climatologies. *J. Geophys. Res.* **95**, 7167–7183.

Thorp, W.H. (1965). Deep ocean sound attenuation in the sub- and low-kilocycle per second region. *J. Acoust. Soc. Amer.* **38**, 648–654.

Tindle, C.T. (1979). The equivalence of bottom loss and mode attenuation per cycle in underwater acoustics. *J. Acoust. Soc. Amer.* **66**, 250–255.

Tolstoy, L., Clay, C.S. (1966). Ocean Acoustics: Theory and Experiment in Underwater Acoustics. McGraw-Hill, New York, 293 pp.

Tolstoy, L., Ewing, W.M. (1950). The T phase of shallow-focus earthquakes. *Bull. Seismol. Soc. Amer.* **40**, 25–52.

Tolstoy, M., Bohnenstiehl, D.R. (2005). Hydroacoustic constraints on the rupture duration, length, and speed of the great Sumatra–Andaman earthquake. *Seismol. Res. Lett.* **76**, 419–425.

Tsai, V.C., Nettles, M., Ekström, G., Dziewoński, A.M. (2005). Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys. Res. Lett.* **32** (17), L17304, 4 pp.

Tsuji, Y., Imamura, F., Matsutomi, H., Synolakis, C.E., Nang, P.T., Jumadi, Harada, S., Han, S.S., Arai, K., Cook, B. (1995). Field survey of the East Java tsunami of June 3, 1994. *Pure Appl. Geophys.* **144**, 839–854.

Urick, R.J. (1963). Low-frequency sound attenuation in the deep ocean. *J. Acoust. Soc. Amer.* **35**, 1413–1422.

Urick, R.J. (1966). Long-range deep-sea attenuation measurement. *J. Acoust. Soc. Amer.* **39**, 904–906.

Urick, R.J. (1975). Principles of Underwater Sound, 2nd ed. McGraw-Hill, New York. 384 pp.

Wadati, K. (1960). On the T phases observed at Torishima. *Geophys. Mag.* **30**, 1–18.

Wadati, K., Inouye, W. (1953). On the T phase of seismic waves observed in Japan. *Proc. Japan Acad.* **29**, 47–54.

Walker, D.A. (1989). Seismicity of the interiors of plates in the Pacific Basin. *Eos, Trans. Amer. Geophys. Un.* **70**, 1543–1544.

Walker, D.A., Bernard, E.N. (1993). Comparison of T-phase spectra and tsunami amplitudes for tsunamigenic and other earthquakes. *J. Geophys. Res.* **98**, 12557–12565.

Walker, D.A., McCreery, C.S., Hiyoshi, Y. (1992). T-phase spectra, seismic moments, and tsunamigenesis. *Bull. Seismol. Soc. Amer.* **82**, 1275–1305.

Walker, K.T., Ishii, M., Shearer, P.M. (2005). Rupture details of the 28 March 2005 Sumatra $M_w = 8.6$ earthquake imaged with teleseismic P waves. *Geophys. Res. Lett.* **32** (24), L24303, 4 pp.

Wallace, T.C., Koper, K.D. (2002). Forensic analysis of seismic events in the water: submarines, explosions and impacts. *Eos, Trans. Amer. Geophys. Un.* **83** (47), F1049 (abstract).

Ward, S.N. (1980). Relationship of tsunami generation and an earthquake source. *J. Phys. Earth* **28**, 441–474.

Weigel, W. (1990). Bericht über die SONNE-Expedition SO65-2, Papeete–Papeete, 7.-28. Dez. 1989. Universität Hamburg, Institut für Geophysik.

Weinstein, S.A., Okal, E.A. (2005). The mantle wave magnitude M_m and the slowness parameter Θ: Five years of real-time use in the context of tsunami warning. *Bull. Seismol. Soc. Amer.* **95**, 779–799.

Wessel, P., Smith, W.H.F. (1991). Free software helps map and display data. *Eos, Trans. Amer. Un.* **72**, 441 and 445–446.

Wilkinson, J.H. (1965). The Algebraic Eigenvalue Problem. Clarendon Press, Oxford, 662 pp.

Worzel, J.L., Ewing, W.M. (1948). Explosion sounds in shallow water. *Geol. Soc. Amer. Mem.* **27**, Part 1, 63 pp.

Wysession, M.E., Okal, E.A., Miller, K.L. (1991). Intraplate seismicity of the Pacific Basin, 1913–1988. *Pure Appl. Geophys.* **135**, 261–359.

Yang, Y., Forsyth, D.W. (2003). Improving epicentral and magnitude estimation of earthquakes from T phases by considering the excitation function. *Bull. Seismol. Soc. Amer.* **93**, 2106–2122.

Yuan, X., Kind, R., Pedersen, H.A. (2005). Seismic monitoring of the Indian Ocean tsunami. *Geophys. Res. Lett.* **32** (15), L15308, 4 pp.

This page intentionally left blank

ADVANCES IN GEOPHYSICS, VOL. 49, CHAPTER 2

THE STRESS ACCUMULATION MODEL: ACCELERATING MOMENT RELEASE AND SEISMIC HAZARD

A. MIGNAN

Science & Technology Research, Risk Management Solutions, Peninsular House, 30 Monument Street, London EC3R 8NB, UK
Laboratoire Tectonique, Institut de Physique du Globe de Paris, 4, place Jussieu, 75252 Paris Cedex 05, France

PREFACE

A progressive increase of seismic activity over a wide region around a future earthquake epicenter is termed accelerating moment release (AMR). This has been observed in several studies, although there is no consensus about the physical origin of the effect. In the recent Stress Accumulation model (SAM), AMR is thought to result from loading during the earthquake cycle. In this view, AMR is due to minor stress release as the whole region becomes sufficiently stressed for the mainshock to occur. The SAM makes predictions about the spatial distribution of AMR. Because AMR results from loading on the main fault, the precursory activity should be concentrated in the positive lobes of the pre-stress field of the mainshock. This is clearly observed in California seismicity.

The power-law behaviour of AMR is often thought to be a consequence of critical processes. However, a new mathematical formulation is obtained based on stress transfer without criticality. In this view (following the SAM), AMR is due to the decrease of the size of a stress shadow. A power-law time-to-failure equation can be expressed as a function of the loading rate on the fault that is going to rupture and the m-value (power exponent) defined as $m = D/3$ with D a parameter that takes into account the geometrical shape of the stress lobes and the distribution of active faults.

The SAM suggests that AMR can help to identify when a stretch of fault is approaching failure without any knowledge of the recurrence time. This has been applied to the Sumatra–Java subduction zone as a new forecasting method and new statistics have been added to improve the quantification of AMR.

ISSN: 0065-2687
DOI: 10.1016/S0065-2687(07)49002-1

General Introduction

One of the most important goals of modern geophysics is to predict large earthquakes. Since the 1990s, very important advances have been made in the study of earthquake populations. Interactions between large earthquakes and between mainshocks and their aftershocks can at present be explained by the stress transfer theory. However, precursory seismicity patterns that may permit a new mainshock to be forecast, such as accelerating moment release (i.e. accelerating seismicity through time), have been poorly understood.

Strangely, many studies (but not all) of large earthquake precursors based on seismicity rate changes do not involve simple stress transfer models, but a variety of criticality models based on stress triggering at different scales (i.e. cascading seismicity). A common idea is that earthquake populations are complex systems with a chaotic behaviour. This idea is accepted by a large part of the scientific community because of the Gutenberg–Richter frequency-magnitude relation and by other power-laws that emerge from seismicity populations (e.g. accelerating moment release).

However, a new model has been recently proposed to explain the evolution of precursory seismicity, based on Reid's Elastic Rebound, with no need of critical processes. In that view, accelerating moment release can be easily related to the incoming mainshock by stress loading during the seismic cycle of the fault that is going to fail. This is the Stress Accumulation model, which is the main topic of this thesis.

The purpose of this work is to test the validity of the Stress Accumulation model, to improve the underlying physics and to determine its possible role in future earthquake forecasting methods. The structure of this thesis can be viewed at two different levels. On one side, a parallel is made between accelerating moment release and aftershocks. These two seismicity patterns are considered as pre- and post-mainshock respectively, meaning that accelerating moment release can be studied in the same way that aftershocks are. Therefore, the spatial distribution and the changes of the rate of events that compose precursory seismicity are explained in terms of stress transfer. On the other hand, the Stress Accumulation model is different from models based on criticality, meaning that stress transfer imposed by loading is not the same as stress transfer by triggering and their relative role is discussed.

Section 1 presents the stress transfer theory and its role in the behaviour of earthquake populations. Interactions between large earthquakes and between mainshocks and their aftershocks are first illustrated using famous examples. Then accelerating moment release and the Stress Accumulation model are described in detail.

Section 2 presents a test of validity of the Stress Accumulation model, by studying the spatial distribution of accelerating moment release. It is linked to the article "An Observational Test of the Origin of Accelerating Moment Release

Before Large Earthquakes" by A. Mignan, D. Bowman and G.C.P. King (J. Geophys. Res. 111, doi:10.1029/2006JB004374, 2006).

Section 3 presents a review of different explanations of the origin of accelerating moment release based on critical processes and a new mathematical formulation of the time-to-failure power-law of accelerating moment release based on the theory of the Stress Accumulation model. It is linked to the article "A Mathematical Formulation of Accelerating Moment Release based on the Stress Accumulation Model" by A. Mignan, G.C.P. King and D. Bowman (J. Geophys. Res., 112, doi:10.1029/2006JB004671, 2007).

Section 4 presents a review of existing forecasting methods, including accelerating moment release analysis and new possibilities offered by the Stress Accumulation model. The model is applied in the Sumatra–Java region to create a forecast map based on accelerating moment release. It is linked to the article "Seismic activity in the Sumatra–Java region prior to the December 26, 2004 (Mw = 9.0–9.3) and March 28, 2005 (Mw = 8.7) earthquakes" by A. Mignan, G.C.P. King, D.D. Bowman, R. Lacassin and R. Dmowska (Earth. Planet. Sci. Lett., 244, 639–654, 2006). Then a statistical study is carried out to test the different parameters that are important in the calculation of the probability that an observed accelerating moment release is real. Finally the possible extensions of the Stress Accumulation model are discussed for a systematic method of space-, time-, and magnitude-prediction.

1. Stress Transfer Theory

> "The 1990s witnessed tremendous advances in our understanding of earthquake sequences and earthquake interactions. Refinement of the most simple stress-triggering theory enhanced our understanding of aftershocks and helped scientists grasp the intricacies of large earthquake interactions. Further development of these theories continues apace."
>
> S.E. Hough, Earthshaking Science (2002)

There is now an internationally distributed set of published work about static stress changes. At the time of writing, there are at least 350 papers that can be considered relevant to stress interactions, too many to refer to in any detail. For more complete reference lists, the reader is referred to reviews by Harris (1998), Stein (1999), King and Cocco (2000), Freed (2005) and Steacy et al. (2005).

In this section, the principles of stress transfer are explained and illustrated using examples that made the stress-triggering theory become a successful technique adopted by the scientific community (Section 1.1). These examples show interactions between large earthquakes but stress interactions can also occur at the scale of one seismic cycle to explain changes in the rate of seismicity, in particular

"preshocks" (accelerating moment release) and aftershocks observed respectively before and after large earthquakes (Section 1.2). If interactions between a mainshock and its aftershocks are at present quite well understood (Section 1.3), this is not the case concerning interactions between a mainshock and its "preshocks". Many explanations have been proposed but none of them have been proved. One of the most promising approaches to explain accelerating moment release is the Stress Accumulation model, recently proposed by King and Bowman (2003), which differs greatly from the other models and is based on simple stress transfer considerations (Section 1.4). Other explanations are described later, in Section 3.

1.1. Principles of Static Stress Transfer

The static stress transfer theory is based on Coulomb stress calculations applied to the study of earthquake interactions. Coulomb stress changes are often referred to as the Coulomb Failure Function (CFF):

$$\Delta CFF = \Delta\tau + \mu(\Delta\sigma_n + \Delta P) \tag{1.1}$$

where $\Delta\tau$ are the shear stress changes on the failure plane, $\Delta\sigma_n$ are the normal stress changes (positive for extension), ΔP indicates the pore pressure changes and μ the coefficient of friction. Note that CFF can also be termed Coulomb Failure Stress (CFS).

When an earthquake occurs, it releases the stress accumulated during its seismic cycle on the fault, by changing the stress field in the surrounding area (static stress changes attenuate roughly inversely to the cube of the distance from the causative fault). Future events can be advanced or retarded depending if they are in regions of enhanced stress or decreased stress, respectively (Fig. 1.1). Note that the role of triggered events is to relax the excess stresses generated by a previous event.

It is important to note that static stress changes generally refer to changes that occur instantaneously and permanently. Most commonly this refers to time-independent deformation. Some transient processes can be added to the stress transfer theory using rate-and-state functions (Section 1.3.2). Fluid migrations and viscoelastic effects are also likely to be important, but are not discussed in the present work. Dynamic stresses associated with propagating waves clearly play a role in triggering earthquakes, but are also not reviewed in this chapter.

1.1.1. Clock Advance in the Stress Field of a Large Earthquake

A triggered earthquake is defined as one whose failure time has been advanced by Δt (clock advance) due to a stress perturbation (Gomberg et al., 1998). These stress perturbations correspond to an increase of stress. Stress perturbations can be due to one or a few earthquakes, like for example the triggering of the 1992

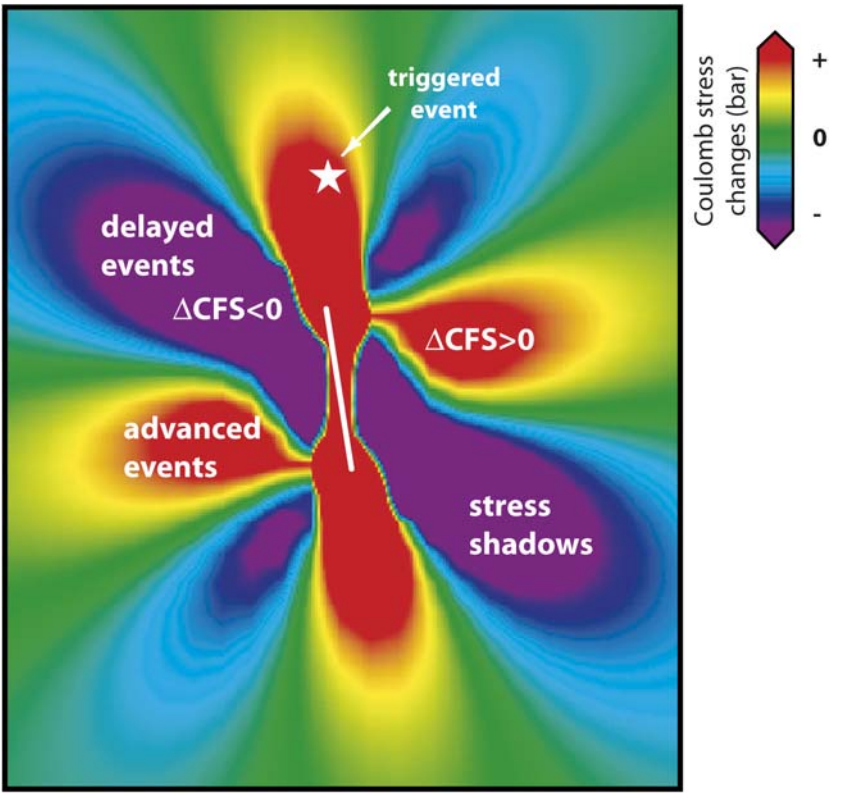


FIG. 1.1. Schematic illustration of the principles of stress transfer after a large earthquake (strike-slip rupture represented in white). Events localised in regions of increased stress ($\Delta CFS > 0$) are advanced, meaning that they are triggered by the previous earthquake. Events localised in regions of decreased stress ($\Delta CFS < 0$), known as stress shadows, are delayed.

$M = 7.3$ Landers earthquake (Section 1.1.4). However, the regional stress field can be more heterogeneous, due to a long history of active faulting and it is best, where possible, to take into account earthquakes that occurred over a long time period (Section 1.1.4).

Although stress triggering has been most clearly revealed for strike slip faults, evidence for large event interactions has also been inferred for normal faults (Jacques et al., 1996; Cocco et al., 2000; Nostro et al., 2005), between subduction and normal faulting earthquakes (Robinson, 2003; Lin and Stein, 2004), between strike-slip and thrust faults (Lin and Stein, 2004) and between strike-slip and normal faults (Nalbant et al., 1998). If some large earthquakes can be triggered by previous ones, Coulomb stress changes are also applicable to smaller events such as aftershock sequences (Section 1.3).

1.1.2. Clock Delay in the Stress Field of a Large Earthquake

When a large earthquake occurs, it also releases stress in some regions, now commonly called stress shadows (Harris and Simpson, 1993) (Fig. 1.1). This can substantially modify the earthquake history of a given region by delaying events and Coulomb stress change models can predict where subsequent large earthquakes will not occur (Harris and Simpson, 1998).

The first studies described the apparent deficit of moderate earthquakes following large events in the San Fransisco Bay area in 1868, 1906, and 1989, and the existence of a stress shadow following the 1906 earthquake has become widely accepted (Kenner and Segall, 1999; Reasenberg et al., 2003; Pollitz et al., 2004). A history of southern California stress shadows carried out by Deng and Sykes (1997) shows that only 15% of 36 events of different magnitudes occur in regions of decreased stress and that these events fall very close to boundaries between positive and negative stress changes. Stress shadows also have been inferred in other studies (Nalbant et al., 1998; Parsons et al., 1999; Stein, 1999; Ogata et al., 2003; Toda and Stein, 2003).

When a fault is located in a zone of decreased stress, one can perform simple determinations of the approximate time that it should take for long-term tectonic loading to recover the static stress change (Harris, 1998). Time delay calculations have been performed in particular for the San Andreas system. It has been studied in the case of the 1983 $M = 6.7$ Coalinga earthquake (Simpson et al., 1988) and of the 1989 $M = 7.0$ Loma Prieta earthquake (Simpson and Reasenberg, 1994; Lienkaemper et al., 1997).

It is important to note that there are exceptions with events occurring in stress shadows whereas they should have been delayed. Part of the reason might be that stress calculations are done for idealised, vertical, single-fault systems, while real faults have more complex structures.

Although widely accepted, the existence of stress shadows has been questioned by a recent study of changes of the rate of seismicity. The authors conclude that stress shadows either do not exist or cannot be distinguished from other processes (Felzer and Brodsky, 2005).

1.1.3. "Pre-Stress Field" of a Large Earthquake

The main source of stress on any fault results from the loading processes before a large earthquake occurs. Figure 1.2a shows the evolution of Coulomb stresses prior to an earthquake due to loading at depth and to displacement on adjacent segments. However, neither the regional tectonic loading nor the history and slip distributions of previous events can be determined in detail except under exceptional circumstances such as West Turkey (Pondard et al., submitted for publication).

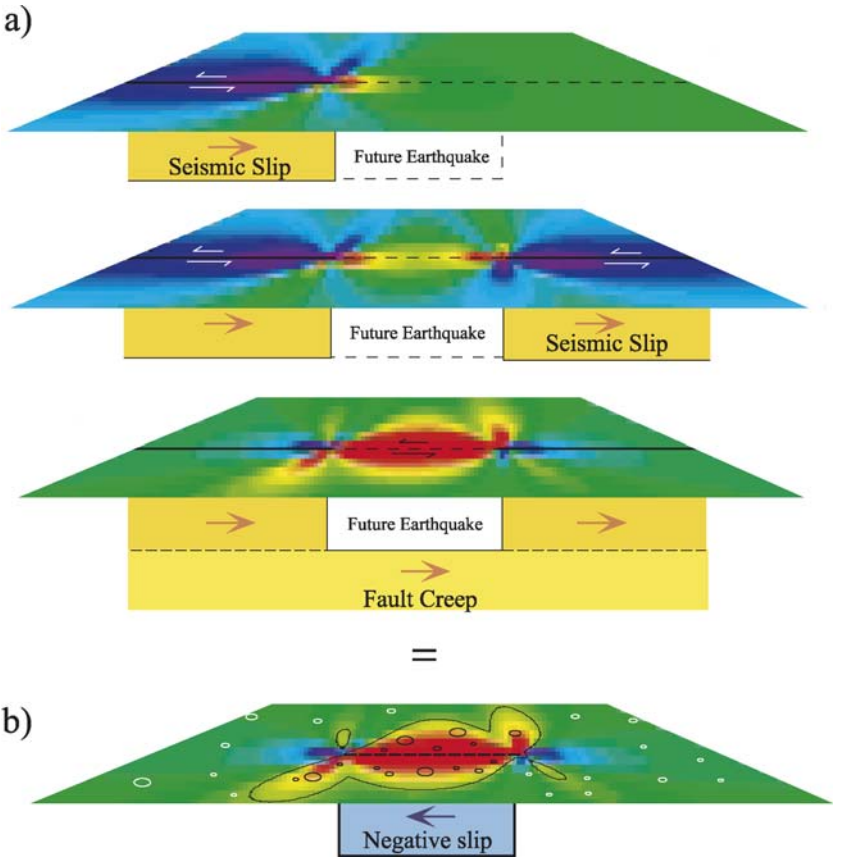


FIG. 1.2. Evolution of Coulomb stresses prior to an earthquake. Light grey colours indicate positive static stress change, while dark grey colours indicate negative static stress change; (a) Perspective view of the static stress changes calculated at the Earth's surface. The fault slip at depth is indicated below the calculated stress change field. The three figures depict the cumulative stress changes due to slip on adjacent segments of the fault (top two panels) and loading by aseismic slip on the fault below seismogenic depths (bottom panel); (b) For a strike-slip fault, this stress can be modelled by reversing the sense of slip observed to occur in the earthquake. From Bowman and King (2001a).

An alternative approach is to calculate the stress field required to move a fault with the orientation, displacement, and rake observed for the mainshock (Bowman and King, 2001a; King and Bowman, 2003). This approach rests on the concept of elastic rebound, which is that the stress released in an earthquake must have existed as a pre-stress prior to the event (Fig. 1.2b). This can be determined by calculating the stress that results from slipping the fault backwards by the amount that it moved in the earthquake (Savage, 1983; Matsuura et al., 1986). The Coulomb stress field from this "negative earthquake"

can then be calculated using established techniques (e.g. King and Cocco, 2000). This produces the same field that would result from an accurate knowledge of the tectonic history and loading mechanism in the region (Fig. 1.2a). It is important to note that dip-slip earthquakes are associated with permanent deformation of the crust and not simple elastic rebound. However we will only be concerned with stresses from tectonic loading, which generally lie in a horizontal plane and vertical motion on the fault can be neglected (see Bowman and King, 2001a, for further detail).

The technique proposed by Bowman and King (2001a) permits stress interactions between loading before a mainshock and pre-mainshock seismicity to be studied (see Section 1.4).

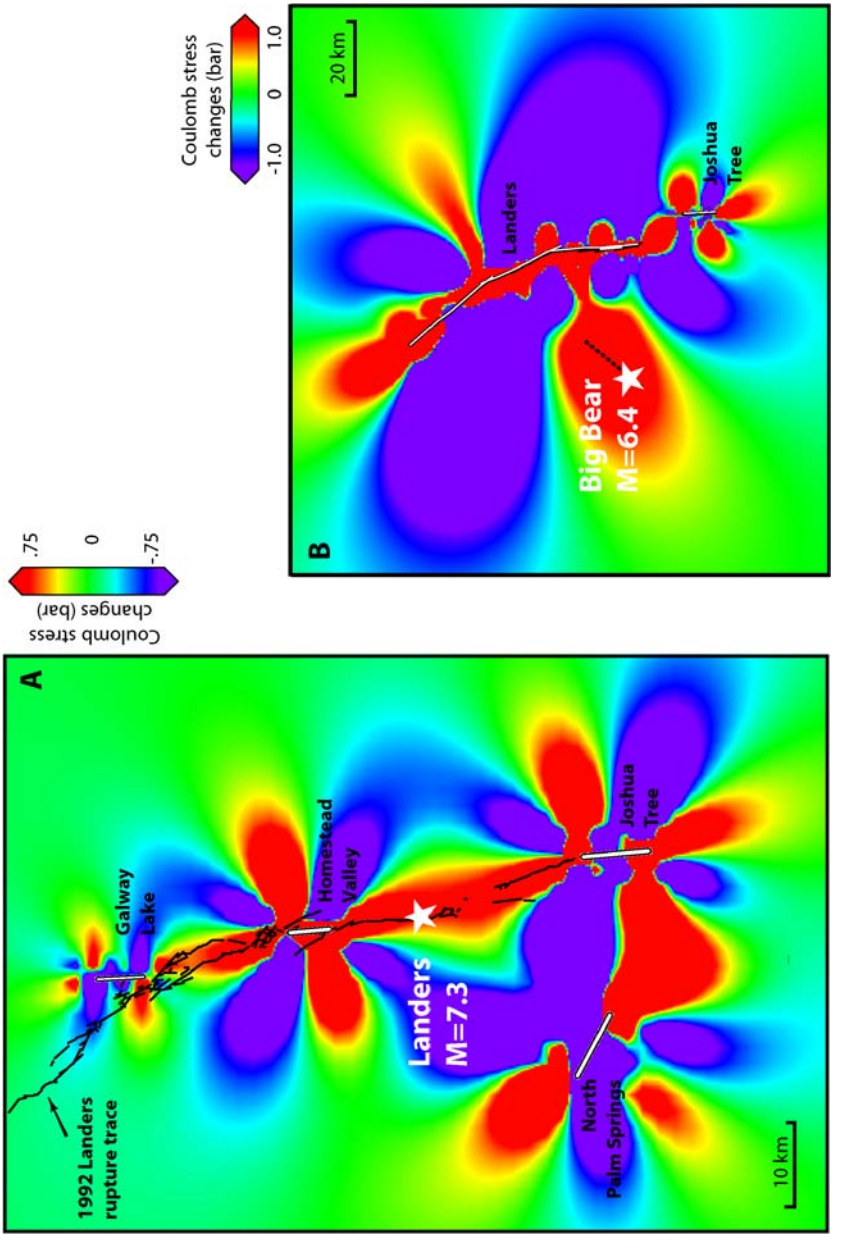
1.1.4. Spatial Distribution of Large Earthquakes

1992, $M_w = 7.3$, Landers earthquake (California): Although stress interactions had been previously observed for other events, the clear correlations associated with the 1992 $M_w = 7.3$ Landers earthquake suggested to the scientific community that Coulomb calculations might indeed prove an effective method of relating large events with each other. The Landers earthquake remains the best example for illustrating the technique and the modelling processes proposed by King et al. (1994) are reproduced in Fig. 1.3a.

Figure 1.3a shows the Coulomb stress changes caused by the four $M > 5$ earthquakes within 50 km of the epicentre that preceded the Landers earthquake. These increased Coulomb stresses by about 1 bar at the future Landers epicentre. Together they also produced a narrow zone of Coulomb stress increase of 0.7–1.0 bars, which the future 70-km-long Landers rupture followed for 70% of its length (King et al., 1994).

The stress changes caused by the Landers event are shown in Fig. 1.3b. The largest lobe of increased Coulomb stress is centred on the epicentre of the future $M_L = 6.5$ Big Bear event, where stresses were raised 2–3 bars. The Big Bear earthquake apparently initiated by this stress rise occurred 3 hr 26 min after the Landers main shock.

1999, $M_w = 7.4$, Izmit earthquake (Turkey): Another classic example of mainshock triggering is the 1999 $M_w = 7.4$ Izmit, Turkey, earthquake that occurred along a section of the North Anatolian fault. This section has been identified by Stein et al. (1997) as being highly loaded due to both secular loading and coseismic stress transfer from previous large earthquakes ($M \geqslant 6.7$) along the North Anatolian fault and by Nalbant et al. (1998), who computed stress changes due to 29 regional events ($M \geqslant 6$) in West Turkey and in the Aegean region, from 1912 to 1999 and identified Izmit as one of three places likely to host a future catastrophic event (Fig. 1.3c).


A
1992 Landers rupture trace
Galway Lake
Homestead Valley
Landers M=7.3
Joshua Tree
North Palm Springs
10 km
Coulomb stress changes (bar)
.75
0
-.75
B
Big Bear M=6.4
Landers
Joshua Tree
20 km
Coulomb stress changes (bar)
-1.0
0
1.0

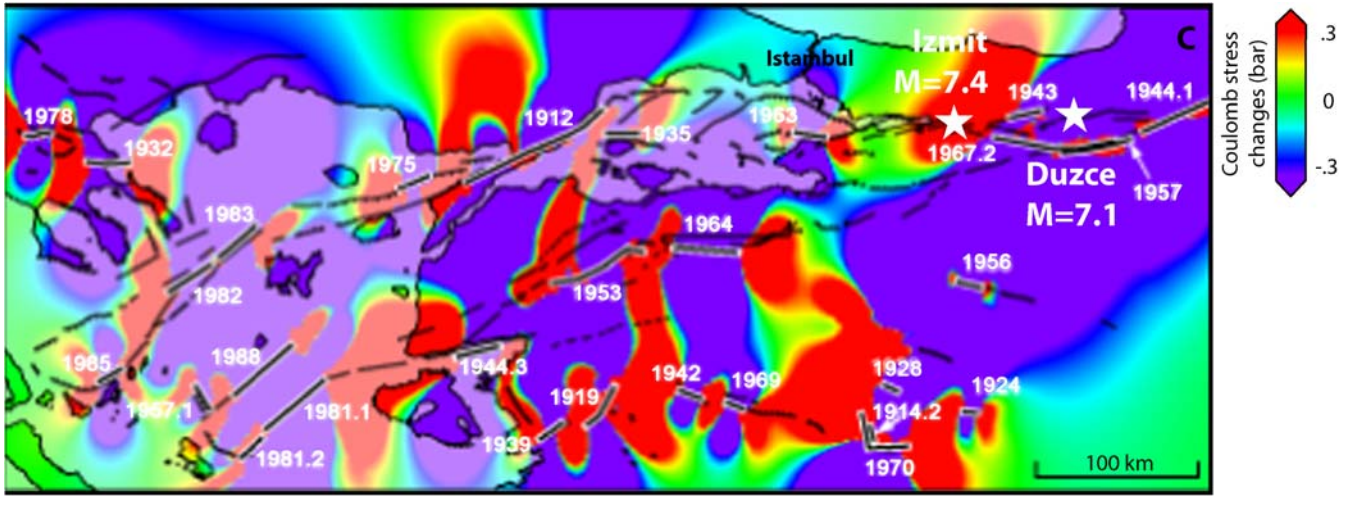


FIG. 1.3. Stress transfer between large earthquakes. Increased Coulomb stress are represented in light grey and decreased Coulomb stress in dark grey. Triggered events are localised by a white star; (a) Coulomb stress changes caused by the four earthquakes that preceded the 1992 Landers main shock within 50 km of the epicentre. The source parameters used in this calculation are listed in King et al. (1994). These events raised the stress at the future Landers epicentre and on most (70%) of the Landers rupture. Redrawn after King et al. (1994); (b) Coulomb stress changes caused by the 1992 Joshua Tree and Landers earthquakes. The location of the Big Bear aftershock and its faulting mechanism are consistent with the induced coseismic stress field. Redrawn after King et al. (1994); (c) Coulomb stress changes caused by 29 $M \geqslant 6.0$ events from 1912 to 1998. The 1999 Izmit earthquake occurred where the North Anatolian fault and increased Coulomb stress coincided. When the Izmit earthquake struck, Coulomb stress increased about 1 bar to the east (Hubert-Ferrari et al., 2000) where the $M_w = 7.1$ Duzce earthquake occurred a few months later. Redrawn after Nalbant et al. (1998).

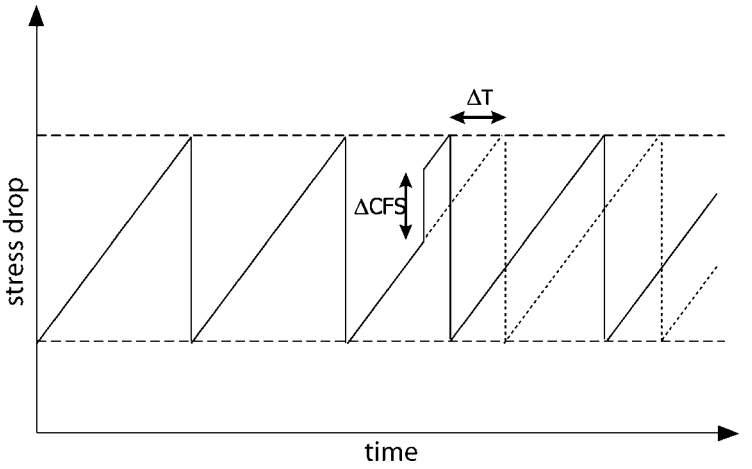


FIG. 1.4. Example of stress triggering. Schematic model of Coulomb failure for a fault under constant loading rate. If an external factor, such as a static stress change ΔCFS generated by a nearby earthquake, occurs, the next earthquake on the fault is advanced of ΔT. Modified from Gomberg et al. (1998).

Modelling studies show that stress redistribution from the Izmit earthquake may have led to positive Coulomb stress changes on the fault both to the west, near Istanbul (Hubert-Ferrari et al., 2000) and to the east (Barka, 1999). The $M_w = 7.1$ Duzce earthquake that occurred 3 months later, was located in a region of enhanced stress. It is important to note that this region was in a stress shadow prior to the Izmit earthquake which changed its status. Thus while stress shadows are in general less likely to host future earthquakes, their status can change rapidly if an event occurs in an adjacent region of high Coulomb stress. Other examples are discussed by Nalbant et al. (1998).

1.1.5. Temporal Distribution of Large Earthquakes

It is widely held that the time of occurrence of an earthquake on a fault undergoing tectonic loading is controlled both by the stress and frictional properties on that fault and by earthquakes on other faults nearby (Stein, 1999). Figure 1.4 shows implications of stress transfer on Coulomb failure models for a fault loaded at a constant background rate (modified from Gomberg et al. (1998)). The failure time is changed (time shift ΔT), independently of the seismic cycle of the fault.

A time shift can be determined from the stress transfer theory in the simple way indicated in Fig. 1.4. However this assumes that failure depends only on a failure stress level being reached. If stress changes abruptly, transient effects substantially modify failure conditions. More complex failure criteria described as rate-and-state friction have been used to quantify this effect (Dieterich, 1994).

Initially proposed to explain the origin of aftershocks (Section 1.3), rate-and-state equations summarise in empirical expressions laboratory observations that fault strength is dependent on slip rate and slip and time history of a fault. Rate-and-state equations have appeal because they allow for a time delay before the onset of failure and can be applied for the study of stress triggering (e.g. Gomberg et al., 1997, 1998) as well as stress shadows (e.g. Harris and Simpson, 1998). They also describe how some faults may slide stably, whereas others may experience "stick-slip" motion.

Whereas Coulomb friction invokes the simple equation (1.1) to describe the failure limit, rate-and-state formulations use a more complex relation which also describes the evolution of quantities as they approach failure. Many formulations of rate and state dependent friction law have been proposed (e.g. Scholz, 1998 and references therein). A commonly used form is

$$\tau = \sigma\left[\mu_0 + A\ln(\dot{\delta}/\dot{\delta}^*) + B\ln(\theta/\theta^*)\right] \tag{1.2}$$

where $\dot{\delta}$ is the sliding speed, θ is a state variable that can be interpreted to represent the effects of contact time between the two surfaces of a fault, and μ_0 and A and B are empirically determined coefficients. $A - B > 0$ leads to stable sliding; $B - A > 0$ leads to instability (stick–slip behaviour). The asterisked terms are normalising constants.

In the rate-and-state formalism an earthquake nucleates on a fault when the sliding speed δ "runs away" and increases dramatically above interseismic values (to speeds of the order of centimetres per second). Dieterich (1994) gives the time of failure in terms of the sliding speed and other variables:

$$t = \frac{A\sigma}{\dot{\tau}}\ln\left(1 + \frac{\dot{\tau}}{H\sigma\dot{\delta}}\right) \tag{1.3}$$

where

$$H = \frac{-k}{\sigma} + \frac{B}{D_c} \tag{1.4}$$

t is the time to failure, $\dot{\tau}$ is the long-term stressing rate, k is the effective stiffness for source nucleation, B is a fault constitutive parameter, and D_c is the characteristic sliding distance.

If an external factor, such as a static stress change generated by a nearby earthquake, perturbs the sliding speed, it also changes the eventual nucleation time, or time to failure. Following Dieterich (1994), during the long interval of self-driven accelerating slip and before an externally applied stress change, a fault's sliding speed depends on the initial stress and state. After a stress step occurs, the slip speed depends on the new stress and state. Dieterich (1994) explicitly describes how a stress change perturbs the sliding speed (and, thereby, the time to failure):

$$\frac{\dot{\delta}}{\dot{\delta}_0} = \left(\frac{\sigma}{\sigma_0}\right)^{\alpha/A}\exp\left(\frac{\tau}{A\sigma} - \frac{\tau_0}{A\sigma_0}\right) \tag{1.5}$$

TABLE 1.1. Comparison between aftershocks and "preshocks" (i.e. accelerating moment release) characteristics

Events	Rate change	Location	Duration
Aftershocks	Deceleration	Near-field (<fault length)	Days–years
"Preshocks" (AMR)	Acceleration	Far-field (<10s fault lengths)	Years–10s years

where $\dot{\delta_0}$, τ_0, and σ_0 are the sliding speed, shear stress, and effective normal stress (which includes pore pressure effects) on the nearby fault before the stress change and $\dot{\delta}$, τ, and σ are the sliding speed, shear stress, and effective normal stress (positive for compression) after the stress change; α is a sum of parameters governing normal stress dependence of the state variables.

1.2. Seismicity Rate Changes

For the purpose of this work, seismicity is separated into different classes of earthquakes, depending of their occurrence in the seismic cycle of a fault and of their spatio-temporal behaviour. These are "preshocks", foreshocks, mainshocks, aftershocks, and background seismicity. Mainshocks correspond to the largest events, located on recognisable faults. "Preshocks" and aftershocks are events that precede and follow a mainshock, respectively. "Preshocks" (Accelerating Moment Release – AMR) correspond to an increase of the seismicity rate whereas aftershocks correspond to a decay of the seismicity rate (Table 1.1). Foreshocks correspond to events that occur just before a mainshock, with epicentres close to the future epicentre of the main event, and they are parts of the nucleation process. Background seismicity is composed of random events (constant seismicity rate) with possible interactions independent of the seismic cycle on a given fault, contrary to "preshocks", foreshocks and aftershocks which are linked to the mainshock. Note that other seismicity patterns possibly exist and other definitions can be proposed but these are not used in the present work.

1.2.1. Aftershocks

Aftershocks are the numerous smaller events that follow a mainshock and are a transient response to the abrupt change of stress associated with the main event. Rate-and-state friction provides a popular explanation, but transient effects, due to viscous flow, fault creep, and fluid flow have yet to be fully explored. Other seismicity is not a transient effect which explains why "preshocks" and foreshocks are less dramatic than aftershocks. A classic preconception is that aftershocks occur only on the fault that produced the mainshock. Indeed, the earliest aftershocks sometimes appear to delineate the perimeter of a mainshock rupture, and they are often used to infer the spatial extent of the rupture. As a technical definition, an event counts as an aftershock if it is within one fault length of

the mainshock rupture in any direction, on or off the mainshock fault (Hough, 2002). An aftershock sequence continues until the rate of earthquakes in an area returns to its pre-mainshock levels (background seismicity). The statistical properties of aftershock sequences are associated with three empirical scaling relations: (1) the Gutenberg–Richter frequency-magnitude relation, (2) Bath's law, and (3) the modified Omori law (see, e.g. Scherbakov et al., 2005, for a review of statistical properties of aftershocks and Kisslinger (1996) for a general review of aftershocks).

The size-frequency distribution of earthquakes is commonly described in terms of the b-value, the slope of the Gutenberg–Richter relation (Ishimoto and Ida, 1939; Gutenberg and Richter, 1956). The empirical relation is given by the equation

$$\log N = a - bM \tag{1.6}$$

where M is the magnitude, N the cumulative number of earthquakes with magnitude $>$ M and a a constant depending on the seismic activity. The b-value varies in space and time (e.g. Wiemer and Wyss, 2002) and can range between 0.5 and 1.5 (Turcotte, 1992). The most commonly observed b-value for crustal earthquakes is close to unity (Aki and Richards, 1980; King, 1983).

One of the oldest empirical laws of aftershock sequences, known as Bath's law, holds that the largest aftershock in a sequence will generally be one magnitude unit smaller than the mainshock. Modified versions of Bath's law have been proposed based on an extrapolation of the Gutenberg–Richter statistics (e.g. Scherbakov et al., 2005).

The modified Omori relation is a law of aftershock distributions which describes the decay in aftershock rate with time after a mainshock (Utsu, 1961) (extension of the Omori law (Omori, 1894)). An aftershock sequence will decay, on average, with the reciprocal of time $(1/t)$ as

$$\lambda_0(t) = K(t + c)^{-p} \tag{1.7}$$

with K, c and p constants. The p-value is usually near to unity (Fig. 1.5).

Aftershocks are best observed after large earthquakes but can also be observed in background seismicity, which is clustered in both space and time (Zhuang et al., 2005). The Epidemic-Type Aftershock Sequence (ETAS) model suggests that the normal occurrence rate of earthquakes in a region results from the triggering effect of previous events (e.g. Ogata, 2005; Ogata and Zhuang, 2006). In this view, all earthquakes are simultaneously mainshocks, aftershocks, and possibly foreshocks. The ETAS model extends the modified Omori function (Utsu, 1961) to the case where any shock j that occurs at time t_j triggers offspring events, so that the seismicity rate at time t is given by the linear superposition of the aftershock effect in the past in such a way that

$$\lambda(t) = \lambda_0 + \sum_{j;t_j \leqslant t} K_j(t - t_j)^{-p} \tag{1.8}$$

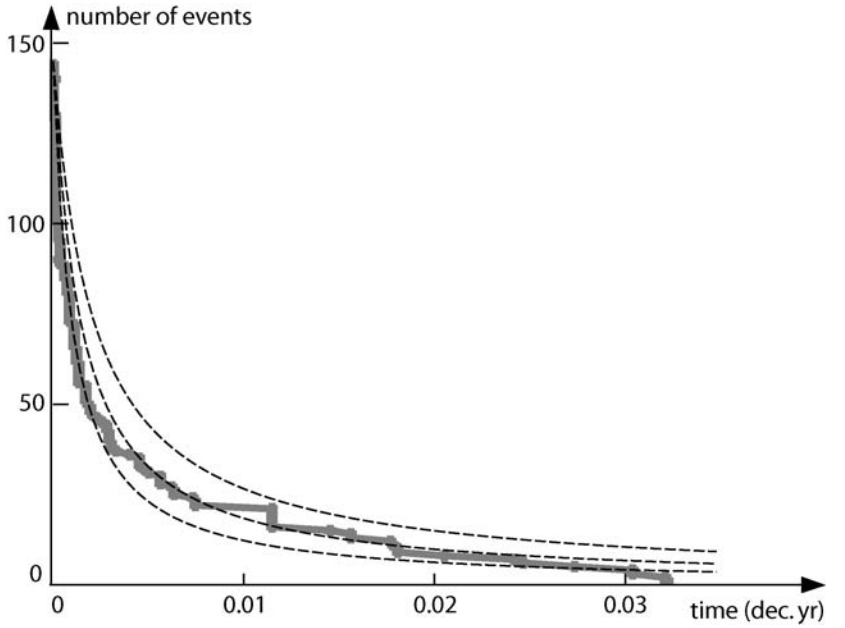


FIG. 1.5. Aftershocks decay of the 1994 $M = 6.7$ Northridge earthquake (California) following the modified Omori law (dotted curves with $p = 0.95$, $p = 1$ and $p = 1.05$). The best p-value corresponds to the standard Omori parameter $p = 1$. Note that the law is not followed after a certain amount of days due to the limit of the declustering method (see details in Section 2).

where the sum is taken for all shocks j occurring before time t and the coefficient K_j for each shock j contributes to the size of the corresponding offspring, or aftershocks in a wide sense. The constant λ_0 (so-called the background seismic activity) represents the occurrence rate of events that cannot be explained as the aftershock effect of past events whose record is available in the data (Ogata, 1989). In other words, the ETAS model is fundamentally a "branching" model where each event is triggered by a unique mainshock but each mainshock triggers several events depending of its magnitude.

1.2.2. "Preshocks": Accelerating Moment Release (AMR)

Accelerating moment release or AMR (Fig. 1.6) has been identified for a substantial number of earthquakes and observed for years to tens of years before a mainshock over tens to hundreds of kilometres from the future epicentre (e.g. Sykes and Jaumé, 1990; Bufe and Varnes, 1993; Knopoff et al., 1996; Bowman et al., 1998; Brehm and Braile, 1998; Jaumé and Sykes, 1999; Robinson, 2000; Bowman and King, 2001a; Zoller et al., 2001; Papazachos et al., 2002a; Bowman and Sammis, 2004).

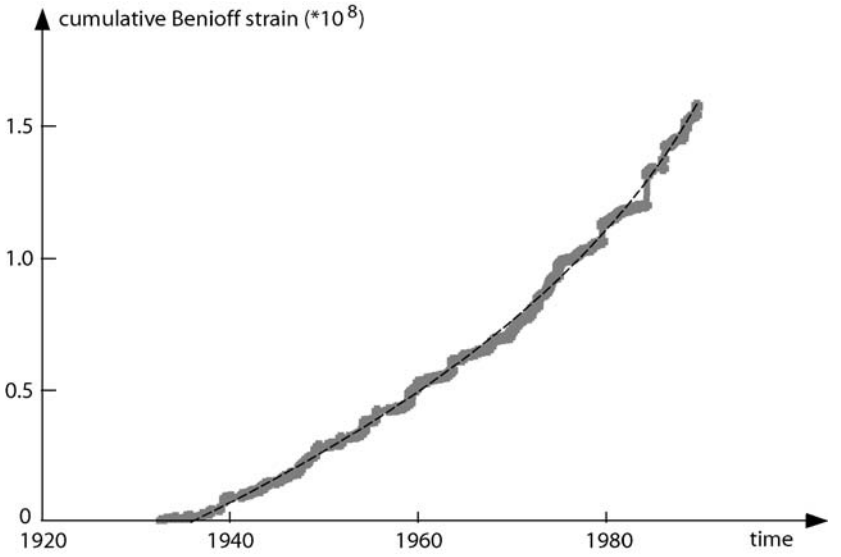


FIG. 1.6. Accelerating moment release (AMR) observed before the 1989, $M_w = 7.0$, Loma Prieta, California, earthquake (figure from Section 2). AMR is best fitted by a power-law curve (Bufe and Varnes, 1993) (dotted curve).

Accelerating moment release is typically modelled by a simple power-law time-to-failure equation (Bufe and Varnes, 1993). This is a relation of the form

$$\sum \epsilon(t) = A + B(t_f - t)^m \tag{1.9}$$

where t_f is the time of the large event, B is negative and m is usually about 0.3. A is the value of $\epsilon(t)$ when $t = t_f$. Equation (1.9) is equivalent to the rate-dependent failure equation of Voight (1989) as noted by Sykes and Jaumé (1990) and Bufe and Varnes (1993). The physical relation of this expression to accelerating moment release is not verified and other power-laws can fit the pre-event seismicity rate changes (see Section 3). $\epsilon(t)$ is the cumulative Benioff strain at time t and defined as

$$\epsilon(t) = \sum_{i=1}^{N(t)} \sqrt{E_i(t)} \tag{1.10}$$

where E_i is the energy of the ith event and $N(t)$ is the number of events at time t. The cumulative Benioff strain is preferred over the cumulative number of earthquakes because the multitude of smaller earthquakes would dominate the data, whereas the moderate to large events would dominate if the cumulative seismic energy alone was used. Therefore the choice of Eq. (1.10) has no physical meaning and is only an observational tool.

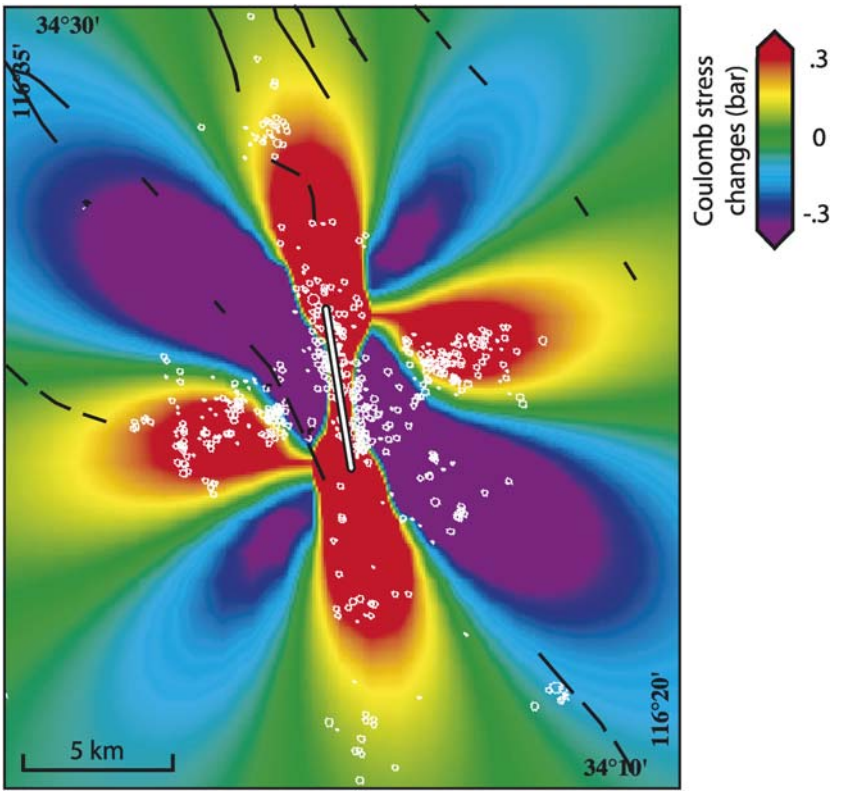


FIG. 1.7. Coulomb stress changes caused by the 15 March 1979, right-lateral Homestead Valley earthquake sequence and aftershock distribution. The calculations are carried out in a half-space with the values of Coulomb stress plotted in the figure being calculated at half the depth to which the faults extend. The fault, simplified as a single vertical plane, is 5.5 km long and 6 km wide and slip is 0.5 m and it is tapered at the fault ends. Redrawn after King et al. (1994).

1.3. Triggering of Aftershocks by Mainshocks

1.3.1. Spatial Distribution

The stress triggering methods outlined previously (Section 1.1) can be applied to aftershock distributions. Figure 1.7 shows the spatial distribution of the aftershocks of the 1979 Homestead Valley earthquake in the stress field caused by the mainshock. The stress field is composed of four lobes of Coulomb stress rise and four lobes of Coulomb stress drop. The distributions of aftershocks are consistent with these patterns (King et al., 1994). Many events are associated with increases of Coulomb stress of less than 1 bar, while reductions of the same amount apparently suppress them. Relatively few events fall in the regions of lowered Coulomb

stress and the clusters of off-fault aftershocks are separated from the fault itself by a region of diminished activity.

A spatial correspondence between aftershocks and positive Coulomb stress has been documented by a number of researchers (e.g. Karakostas et al., 2003, 2004; Wang and Chen, 2001; Toda and Stein, 2000). Hardebeck et al. (1998) quantitatively evaluated the role of static stress changes for the apparent triggering of aftershocks by the 1992 $M = 7.3$ Landers and 1994 $M = 6.7$ Northridge earthquakes (Southern California). For the Landers earthquake, $\sim 85\%$ of events between 5 and 75 km distance from the mainshock fault plane are consistent with static stress change triggering. However it cannot explain the first month of the Northridge aftershock sequence. Although the aftershocks broadly follow the distribution predicted by stress calculations, some events occur where they are not expected. Another example is given by Parsons (2002) showing that only about 60% of aftershocks are located where the stress increased after a mainshock, meaning that other processes have to be taken into account. Part of such discrepancies could be also blamed on the simplification of assuming simple planar geometries of the main faults, from which stress fields are calculated, as well as simplifications of slip distributions. Other studies have been done by Beroza and Zoback (1993), Simpson and Reasenberg (1994), Harris et al. (1995), Deng and Sykes (1997), Kilb et al. (1997), Toda et al. (1998), Stein (1999). Some studies also note that aftershocks sometimes expand in space, probably related to diffusion of stresses with time (e.g. Tajima and Kanamori, 1985).

1.3.2. Temporal Distribution

Aftershocks are caused by the change of stress that occurs at the time of the mainshock. From rate-and-state equations (Section 1.1.5), Dieterich (1994) obtains expressions to quantify the seismicity rate as a function of time after a stress step (i.e. mainshock)

$$R = \frac{r\dot{\tau}/\dot{\tau}_r}{[\dot{\tau}/\dot{\tau}_r \exp(-\Delta\tau/(A\sigma)) - 1]\exp[-t/t_a] + 1}, \quad \dot{\tau} \neq 0 \tag{1.11}$$

and

$$R = \frac{r}{\exp(-\Delta\tau/(A\sigma)) + t\dot{\tau}_r/A\sigma}, \quad \dot{\tau} = 0 \tag{1.12}$$

where $\dot{\tau}_r$ and $\dot{\tau}$ are the stressing rate prior to and following the step, respectively, t is set to zero at the time of the stress step, and

$$t_a = \frac{A\sigma}{\dot{\tau}} \tag{1.13}$$

is the characteristic relaxation time for the perturbation of earthquake rate.

It can be noted that Eq. (1.12) has the form of Omori's law for aftershock decay:

$$R = \frac{a}{b + t}. \tag{1.14}$$

Similarly, Eq. (1.11) gives Omori's law for $t/t_a < 1$, but seismicity rate merges to the steady state background rate for $t/t_a > 1$. This characteristic of Eq. (1.11) makes it a somewhat preferable representation of aftershock decay for many applications. For further details, see Dieterich (1994).

Rate-and-state friction laws are well accepted and used in many studies concerned with the behaviour of aftershocks. Gross and Kisslinger (1997) solved the rate-and-state equations for some variables (e.g. for A and σ from Eq. (1.12)) by examining the aftershocks of the 1992 Landers earthquake. Gross and Rundle (1998) have followed a similar tactic and investigated the aftershocks of the 1989 Loma Prieta earthquake. Toda et al. (1998) used formulations proposed by Dieterich (1994) and investigated the aftershock rates of the 1995 Kobe earthquake.

1.4. Triggering of "Preshocks" by Loading at Depth: The Stress Accumulation Model

To explain the origin of accelerating moment release observed before large earthquakes, some studies point out the role of stress loading processes (Bowman and King, 2001a; King and Bowman, 2003). The concept of "pre-stress field" (Section 1.1.3) permits events which are part of the regional seismicity to be related to a future mainshock. In this section, loading processes are described at the scale of the seismic cycle of one fault or fault segment. Possible stress triggering from a nearby earthquake (Section 1.1) are not taken into account and correspond to independent processes that superimpose their effects on stress loading.

1.4.1. The Seismic Cycle

Fedotov (1968) was perhaps the first to describe a "seismic cycle" or "earthquake cycle" in the occurrence of smaller earthquakes between great plates rupturing events in the Kurile–Kamchatka region. It is widely accepted that the concept of a "seismic cycle" is a convenient fiction; reality is more complicated. It does however summarise in a useful way behaviour observed to occur before, during and after an earthquake. More recent studies describe the evolution of stresses during the earthquake cycle, in particular for large subduction events, and look for the relation to smaller earthquakes (e.g. Dmowska et al., 1988, 1996; Taylor et al., 1996, 1998).

King and Bowman (2003) describe an "earthquake cycle" showing the evolution of the stress field which includes the pre-event, inter-event and post-event stress fields to simulate seismicity patterns, in particular accelerating moment release. An idealised loading cycle for an infinitely long vertical strike-slip fault is shown in Fig. 1.8. This idealised fault has constant slip along strike. It is localised and extends through both the crust and upper mantle, which are both assumed to

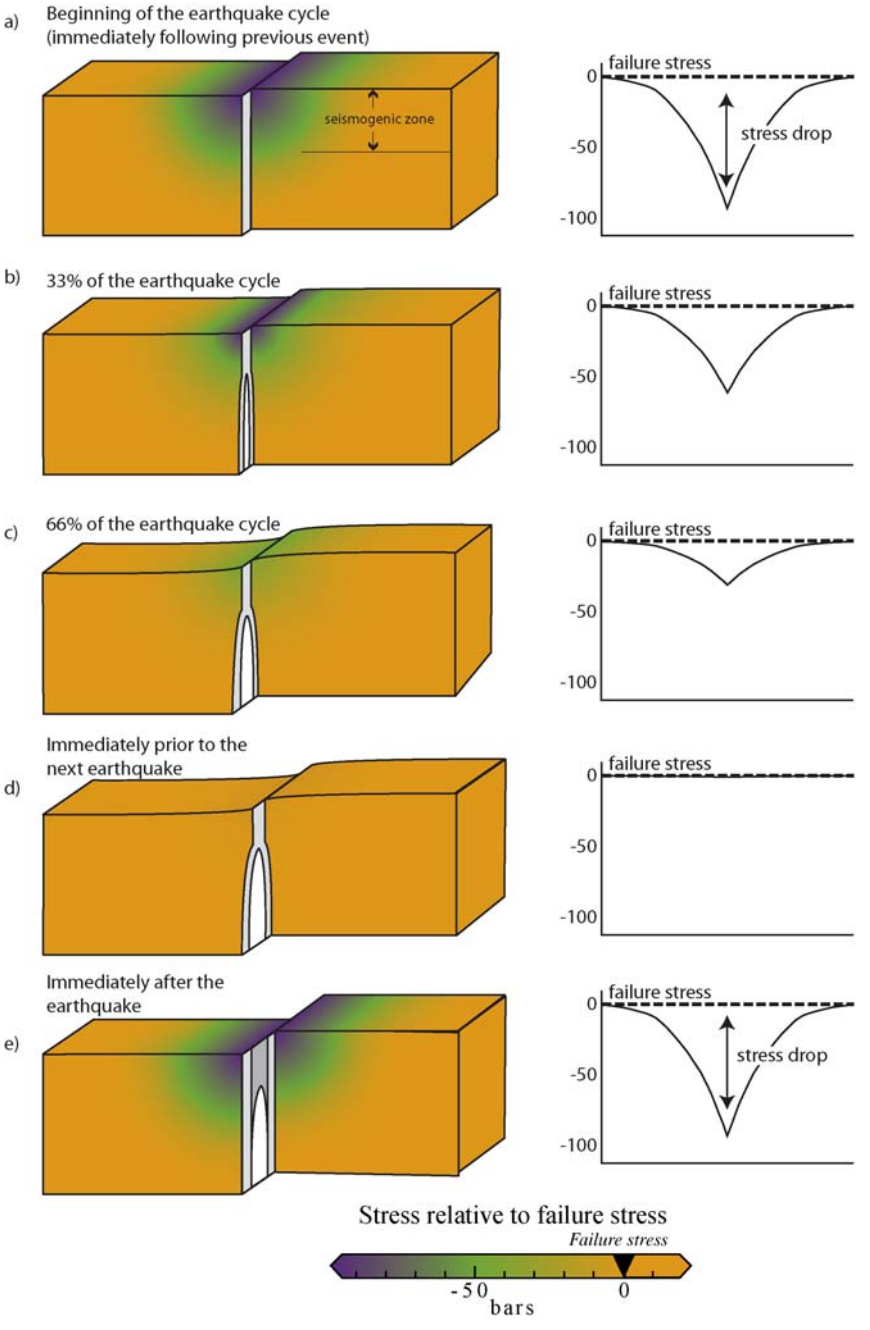

a)
Beginning of the earthquake cycle
(immediately following previous event)
seismogenic zone
failure stress
0
-50
-100
stress drop
b)
33% of the earthquake cycle
failure stress
0
-50
-100
c)
66% of the earthquake cycle
failure stress
0
-50
-100
d)
Immediately prior to the
next earthquake
failure stress
0
-50
-100
e)
Immediately after the
earthquake
failure stress
0
-50
-100
stress drop
Stress relative to failure stress
Failure stress
-50
0
bars

FIG. 1.8. Schematic illustration of the stress evolution around an infinitely long fault during the earthquake cycle. The system is driven by distant displacement (plate) boundary conditions. Light grey colour indicates stresses (shear stress driving fault motion) close to failure. Grey to dark grey indicates lower stresses. Stress values are indicated assuming an earthquake stress drop of about 100 bars; (a) Stress condition immediately following the earthquake; (b) At 33% of the earthquake cycle; (c) At 66% of the earthquake cycle; (d) Immediately prior to the next event; (e) Stress conditions restored in Fig. 1.8a, but with a further increment of slip. During loading, the stress "hole" (shadow) formed by the last earthquake fills progressively and the stress shadow reduces. From King and Bowman (2003).

retain long-term strength. King and Bowman (2003) use a purely elastic model to simulate the evolution of the stress in the earthquake cycle and argue that distributed viscous relaxation in the lower crust or upper mantle is not required at this time scale by either geological (e.g. Hubert-Ferrari et al., 2003) or geodetic results (e.g. Meyer et al., 1998).

In the seismogenic zone, the evolution of the stress field is largely controlled by earthquakes. Near the surface, static friction (τ_s) is greater than dynamic friction (τ_d), resulting in the stick–slip cycling of stress associated with earthquakes. Beneath the seismogenic zone, static velocity-strengthening occurs (e.g. Tse and Rice, 1986), resulting in stable sliding. The boundary between stick–slip and stable sliding is not abrupt. In the "Schizosphere", earthquake rupture that has initiated in a stick–slip zone can propagate into a stable sliding zone (Scholz, 1990b). The system is driven by the displacement boundary conditions associated with plate motions, thus subjecting the whole system to the stress necessary to drive the fault. At greater depth where plastic flow occurs, the stress required to move the fault becomes independent of normal stress. In Fig. 1.8, a light grey colour indicates stress close to that necessary to cause static failure or localised plastic flow. Beneath the seismogenic zone, this remains constant. However, in the seismogenic zone, the stress field drops to the level required to sustain dynamic rupture following an earthquake. Figure 1.8a shows the stress conditions immediately following an earthquake. The amount that the stress level drops along the fault is equal to the stress drop in the earthquake, and is shown in dark grey in Fig. 1.8a. As the earthquake cycle progresses, continued stable sliding at depth, as a result of plate loading, steadily restores the region of reduced stress (Fig. 1.8b–d). At the end of the seismic cycle, the seismogenic zone again approaches static failure stress, and a second event occurs (Fig. 1.8e). Figure 1.8 also shows the difference between the absolute stress and the failure stress (for a section perpendicular to the fault) at each step in the loading cycle. This is a simplified vision of the earthquake process. However it is adequate for the discussions which follow that are only concerned with off-fault processes in the far-field.

1.4.2. Absolute Stress Changes

Unlike conventional Coulomb stress techniques which calculate stress changes, the evolution of the stress field relative to the failure stress is modelled by King and Bowman (2003). To examine seismicity through the earthquake cycle, it is necessary to calculate the stress distribution promoting activity around the main fault (Fig. 1.9). The stresses are calculated for the cycle of an $M = 7.5$ event that repeats with a stress drop of 100 bars (mean slip of 3 m, fault length of 80 km, downdip extent of 15 km).

In Coulomb stress models, regions of increased stress are shown as having positive stresses. The modified stress field presented here accommodates the fact that, as a result of the previous tectonic history, no region ever exceeds the failure stress without an earthquake occurring. Closeness to failure is indicated by a less negative stress. This can be considered to be an absolute stress since it is related to a specific stress level (failure stress) and is not just the stress change of classical Coulomb stress calculations.

1.4.3. Tectonic Memory

For real faults, the fault system has been subjected to a tectonic history (leaving a stress memory) that results in an inhomogeneous pre-stress, which must have the following characteristics: (1) When rupture ceases on the main fault segment, stress at the fault ends and aftershock regions can be brought close to, but not beyond failure stress. If this were not true, the main fault rupture would continue or main events would occur immediately in the side lobes. (2) When the main earthquake occurs, the stress on the main fault plane can have reached, but not exceeded the failure stress. If not, the event would have occurred earlier.

To account of these effects, a "tectonic memory stress" must be added. It must have a form that annuls the positive stresses in the first column of Fig. 1.9. The "tectonic memory" is shown in the centre of Fig. 1.9 and the result of adding it is shown in the second column of Fig. 1.9. Unlike the first column of Fig. 1.9 which describes only stress changes, the second one describes what can be defined as "stress relative to failure stress" (or absolute stress).

1.4.4. Simulation of the Background Seismicity

The stress field shown in Fig. 1.8 and in Fig. 1.9 (the 2 first columns) vary smoothly, but this is not an accurate description of the stress field in the seismogenic crust. Since faults are finite in length and active regions are subject to background seismicity, some locations must have stresses approaching the failure stress (or background seismicity would not occur), while other areas must have stresses reduced from failure stress by approximately the order of an earthquake stress drop (as a result of previous seismicity).

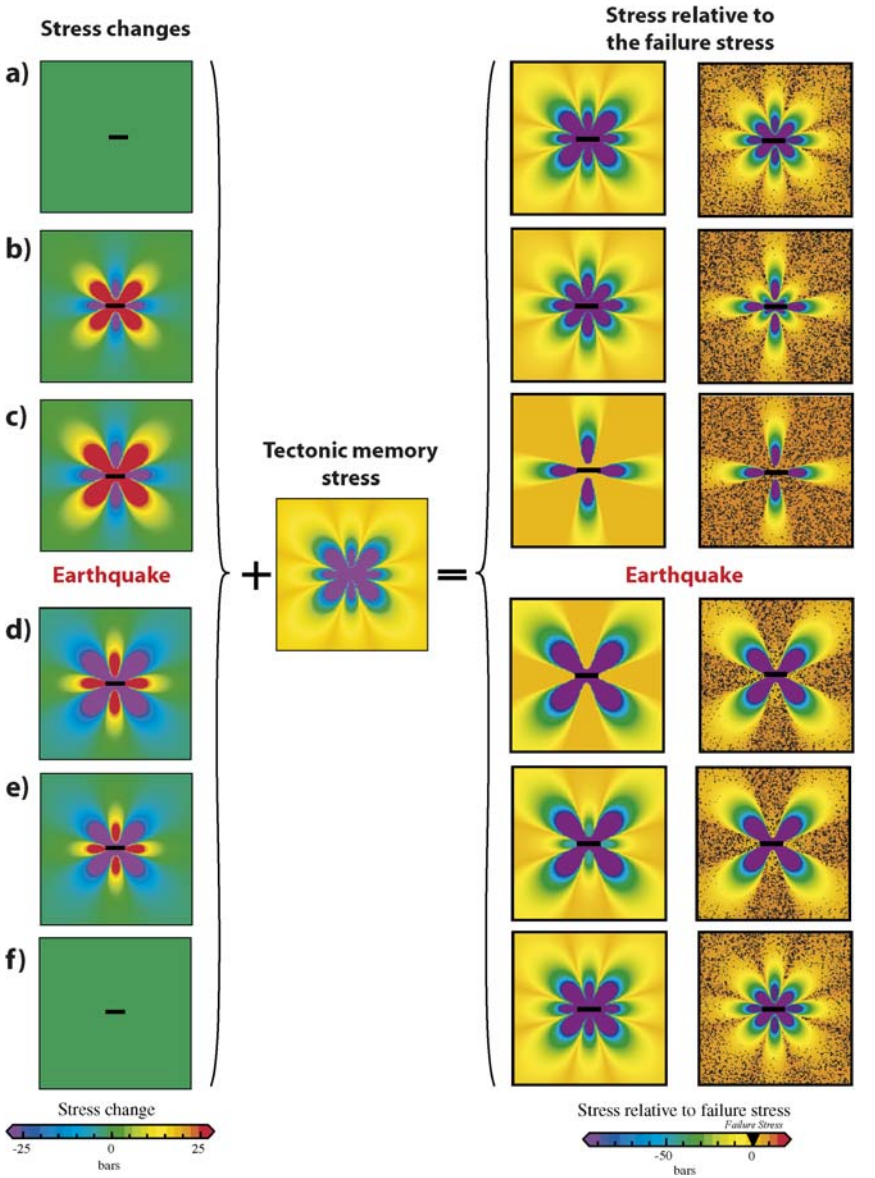


FIG. 1.9. Stress Accumulation model for a complete earthquake cycle. Stress changes through the earthquake cycle, but not the actual stress levels in the system, since at no place can the model exceed the failure stress without an earthquake occurring. The fault region must have a "memory" of its previous deformation history, represented by the "Tectonic memory stress". Stress changes are represented in column 1, stress relative to the failure stress are represented in column 2 and in column 3 where background seismicity is added. See text for details. Modified from King and Bowman (2003).

Figure 1.10 combines the stress evolution through the earthquake cycle of Fig. 1.8 with an irregular stress distribution, that respects the Gutenberg–Richter law (see King and Bowman, 2003, for details). The stress drop for the main event is 100 bars as before, and the amplitude of the stress variations is similar. It becomes clear from this illustration why pre-event seismicity should appear at a distance from the future epicentre. In Fig. 1.10a, immediately following the big event, no peaks in the stress field approach failure in the epicentral region, which is in agreement with the concept of stress shadow. As stress levels around the past and future epicentre are reestablished, the peaks first reach failure at a distance from the future earthquake (Fig. 1.10b–c). Stress levels near the future epicentre remain too low for any failure. Only immediately prior to the future event are stress levels near to the main fault restored (Fig. 1.10e).

Small events through the cycle can consequently be attributed to the inhomogeneous stress distribution that must exist in the seismogenic crust as a consequence of tectonic history. The nature of these inhomogeneities must be such as to create and perpetuate a stress regime that results in the Gutenberg–Richter scaling relations when stress concentrations are relieved by earthquakes (see Fig. 1.10). By combining suitable inhomogeneous stress fields with the overall stress evolution due to the earthquake cycle (Fig. 1.9, column 3), stress peaks can be identified as earthquakes, permitting the creation of artificial catalogues of the location and magnitude of background seismicity (King and Bowman, 2003).

1.4.5. Simulation of Accelerating Moment Release

The Stress Accumulation model provides a physical explanation for changes in the rate of seismicity observed before the mainshock during a given earthquake cycle. Figure 1.9 (last column) shows a simulated distribution of seismicity through the seismic cycle.

Immediately following a large event, the model produces regions of increased seismicity (aftershocks) where the overall stress field has been elevated and regions of reduced seismicity where the stress field has been reduced (stress shadows).

The high stress levels in the aftershock regions decrease due to loading following the main event. Combined with the stress shadow from the main event, the result is a broad seismically quiet region of lowered stress around the epicentre. Pre-event seismicity appears as the original stress shadow finally fills as a result of loading, The increase in seismicity initially occurs several fault lengths away from the main fault and moves inward as the event approaches, forming a “Mogi Doughnut” of activity (Mogi, 1969, 1981).

As a result of this effect, the seismic moment release in the region around the future epicentre increases as the event approaches. Synthetic catalogues can be generated that are virtually indistinguishable from real earthquake sequences giving simulated accelerating moment release similar to real patterns (Fig. 1.11).

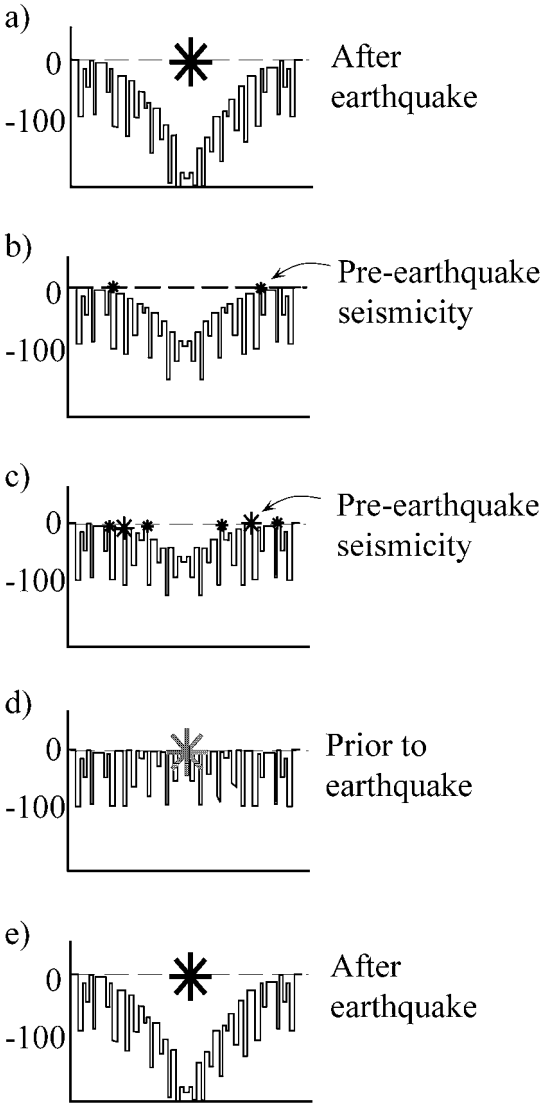


FIG. 1.10. Changing stress through the earthquake cycle of an infinite length fault for inhomogeneous stress fields; (a) Immediately after the earthquake, all stress peaks near to the main event are reduced to below the failure threshold; (b) As the stress hole starts to fill, peaks first reach the threshold at large distances from the fault; (c) Progressively this preearthquake activity becomes more intense and approaches the future epicentre; (d) Immediately prior to the next major event, peaks on or near the main fault plane reach failure. The later modelling shows that this phase occurs very rapidly, explaining why stages shown in *b* and *c* are visible for real earthquakes, but the stage in *d* (classical foreshocks) is not unequivocally observed; (e) The stress conditions shown in *a* are restored. Modified from King and Bowman (2003).

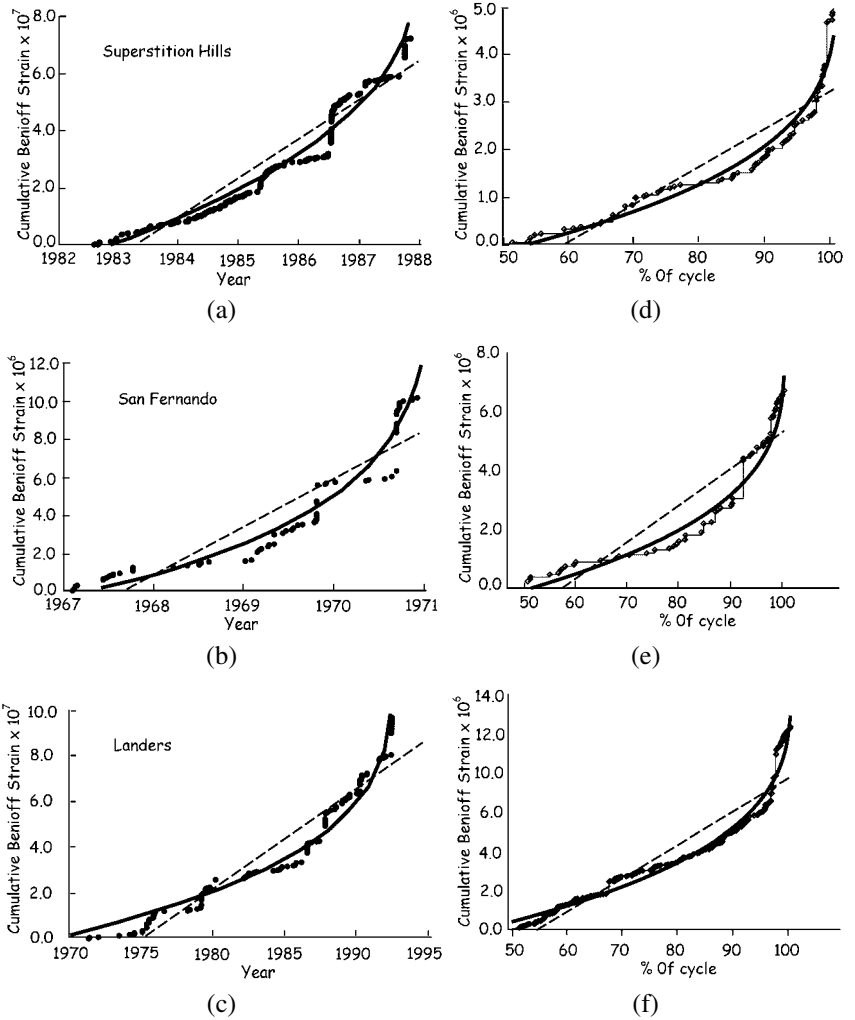


FIG. 1.11. Examples of accelerating moment release observed in real data (a)–(c) and in synthetic catalogues created from the Stress Accumulation model (d)–(f). Simulated accelerating moment release is similar to real patterns. Redrawn after King and Bowman (2003).

However some aspects of the Stress Accumulation model still need to be confirmed. Levin et al. (2006) test the "Mogi Doughnut" prediction by examining seismicity rates as a function of distance before the Landers earthquake but results cannot conclusively be distinguished from those obtained for a random earthquake catalogue. The expected spatial distribution of accelerating moment release in regions of enhanced stress for real seismicity catalogues is considered in Section 2.

A mathematical formulation for accelerating moment release in the Stress Accumulation model to replace the power-law time-to-failure is described in Section 3.

2. Spatial Distribution of Accelerating Moment Release

There are now many studies of accelerating moment release (e.g. Bowman et al., 1998; Brehm and Braile, 1998; Karakaisis et al., 2002; Papazachos and Papazachos, 2000; Robinson, 2000; Zoller et al., 2001). These all adopt different geometrical windows in which the effect is identified. These are all centered around the epicentre of the event that has already occurred. This approach has been criticised on the grounds that an arbitrary choice of shape can arbitrarily be used to select a region where accelerating seismicity occurs. That it must be centred on the epicentre of the future event has been regarded as a weak constraint. An advance was made by systematically searching for optimum circle size and time window. A further approach has been to examine random catalogues to determine the probability that a given acceleration could occur at random (discussed in Section 4).

A prediction of the Stress Accumulation model is that the events should be found in the regions where stress must accumulate prior to the forthcoming earthquake and represent a slight unloading of the future fault. Models that invoke accelerating moment release as a process that adds load to the future earthquake fault predict that the events will not be in these lobes.

In this section the location of the events is examined for earthquakes in California and are found to fall in the stress regions predicted by the Stress Accumulation model. This result has two important implications. First that the shape of the region is not chosen in an arbitrary way and second, that the Stress Accumulation Model is consistent with observations while other hypotheses are not.

2.1. Introduction

Here an observational test is reported on the origin of accelerating moment release, following Mignan *et al*. (2006b): A progressive increase of seismic activity distributed over a wide region around a future earthquake epicenter is termed Accelerating Moment Release (AMR). This phenomenon has been observed in several studies over the last 15 years, although there is no consensus about the physical origin of the effect. In a recent hypothesis known as the Stress Accumulation (SA) model, the AMR is thought to result from the last stage of loading in the earthquake cycle. In this view, the increasing seismicity is due to minor stress release as the whole region becomes sufficiently stressed for the major event to occur. The stress accumulation model makes specific predictions about the distribution of events in an AMR sequence. Because the AMR is predicted to be a

result of loading on the main fault, the precursory activity should be concentrated in the positive lobes of the far-field stresses calculated by a backslip dislocation model of the mainshock. To test this model, AMR is first found in optimal circular regions around the epicenters of each of the $M_w \geqslant 6.5$ earthquakes in central and southern California since 1950. A backslip dislocation model is then used to determine which of the precursory events occur in the regions predicted by Stress Accumulation. AMR is shown to occur preferentially in the lobes of the backslip stress field predicted by the Stress Accumulation model.

It is generally accepted that large earthquakes strongly perturb the background seismicity rate in large regions. The most obvious of these perturbations are aftershocks, which are generally thought to be triggered by stress transferred from the mainshock (e.g. Karakostas et al., 2003, 2004; King et al., 1994; Tod and Stein, 2002a; Wang and Chen, 2001) and may persist for years to decades after a great event (e.g. Utsu, 1961). There is also mounting evidence that "stress shadows" from large events can also depress the seismicity rate over broad regions (e.g. Harris and Simpson, 1996; Harris, 1998; Harris and Simpson, 2002; Ma et al., 2005; Nalbant et al., 1998; Pollitz et al., 2004; Toda and Stein, 2002b). In contrast, both observational and theoretical studies suggest that seismicity rate changes before large earthquakes are not as clearly defined. Various workers have documented evidence of both quiescence (e.g. Wiemer and Wyss, 1994; Katsumata and Kasahara, 1999; Ogata, 2005) and activation (e.g. Bufe and Varnes, 1993; Jaumé and Estabrook, 1992; Knopoff et al., 1996) of regional seismicity before large events.

Although the importance of precursory quiescence was recognised by workers as early as Mogi (1969), relatively few studies in recent years have explored this phenomenon (for example, see Holliday et al., 2006; Ogata, 2005; Tiampo and Anghel, 2006). This is primarily due to the difficulty of obtaining statistically significant results over timescales of months or years. Instead, recent work on precursory seismicity rate changes before large earthquakes has focused on observations of increased seismic activity over a broad region, known as "accelerating moment release" (AMR). The earliest studies of accelerating moment release documented seismicity rate changes in the San Francisco Bay region (Bufe and Varnes, 1993; Sykes and Jaumé, 1990). Since then, studies have documented AMR in a variety of regions, including California (Bowman et al., 1998; Bowman and King, 2001a; Knopoff et al., 1996; Sammis et al., 2004), Turkey and the Aegean (Karakaisis et al., 2002, 2004, Papazachos and Papazachos, 2000; Papazachos et al., 2002a, 2002b, 2002c), New Zealand (Robinson, 2000), the central United States (Brehm and Braile, 1998), Alaska (Bufe et al., 1994; Jaumé and Estabrook, 1992), and Sumatra (Jiang and Wu, 2005; Mignan et al., 2006a).

Two broad classes of models have been proposed to explain the origin of accelerating moment release before large earthquakes. Several authors have suggested that AMR is the result of a cascade of small earthquakes progres-

sively building up stress before a large event (e.g. Bowman and Sammis, 2004; Helmstetter et al., 2003; Helmstetter, 2003; Jaumé and Sykes, 1999; Saleur et al., 1996; Sammis and Smith, 1999; Smalley et al., 1985; Sornette et al., 1990; Sornette and Sammis, 1995). In this view, the primary interaction is static stress transfer, and the accelerating activity is due to the self-organisation of the stress field in a process similar to a phase transition (e.g. Smalley et al., 1985; Sornette et al., 1990). In contrast, King and Bowman (2003) suggested that AMR is a secondary result of the loading of a large fault primarily by aseismic creep on the fault in the lower crust. In their so-called "Stress Accumulation" model, the accelerating activity is associated with a small amount of stress release in a volume of crust surrounding the future earthquake during the progressive loading of the main fault prior to a major earthquake.

In this paper, we test the model of King and Bowman (2003) using observed seismicity in central and southern California. Because the loading process described by King and Bowman (2003) produces a spatially heterogeneous stress field (Fig. 2.1), the AMR predicted by the model should be concentrated in regions determined by the geometry of the eventual mainshock. Therefore, a technique that uses the mechanism of the mainshock to constrain the region where accelerating seismicity is observed (e.g. Bowman and King, 2001a) should provide better constraints on the observation than a technique that ignores the mechanics of fault loading (e.g. Bowman et al., 1998). To test this hypothesis, we use the seismicity catalog for central and southern California to search for AMR before events $M_w \geqslant 6.5$ since 1950, similar to the studies of Bowman et al. (1998) and Bowman and King (2001a). Unlike the earlier studies, the seismicity catalog is first declustered to minimize spurious results in the pattern recognition algorithm. The procedure of Bowman et al. (1998) is then used to find a circular region that optimises the observed AMR. Finally, the backslip model of Bowman and King (2001a) is adopted to discriminate the regions within the circular area where seismicity should be increasing due to loading from below. If the Stress Accumulation model of King and Bowman (2003) is correct, then accelerating seismicity should be found preferentially in the regions identified by this procedure. However, if the events causing the AMR are distributed either preferentially in the negative lobes or randomly throughout the circular region, then another mechanism, such as stress triggering, must play an important role in the origin of AMR. Although earlier work has addressed this topic qualitatively for selected events (e.g. Sammis et al., 2004), this is the first study that systematically addresses both the importance of correctly identifying the loading regions when studying AMR and the relative role of stress triggering versus stress loading.

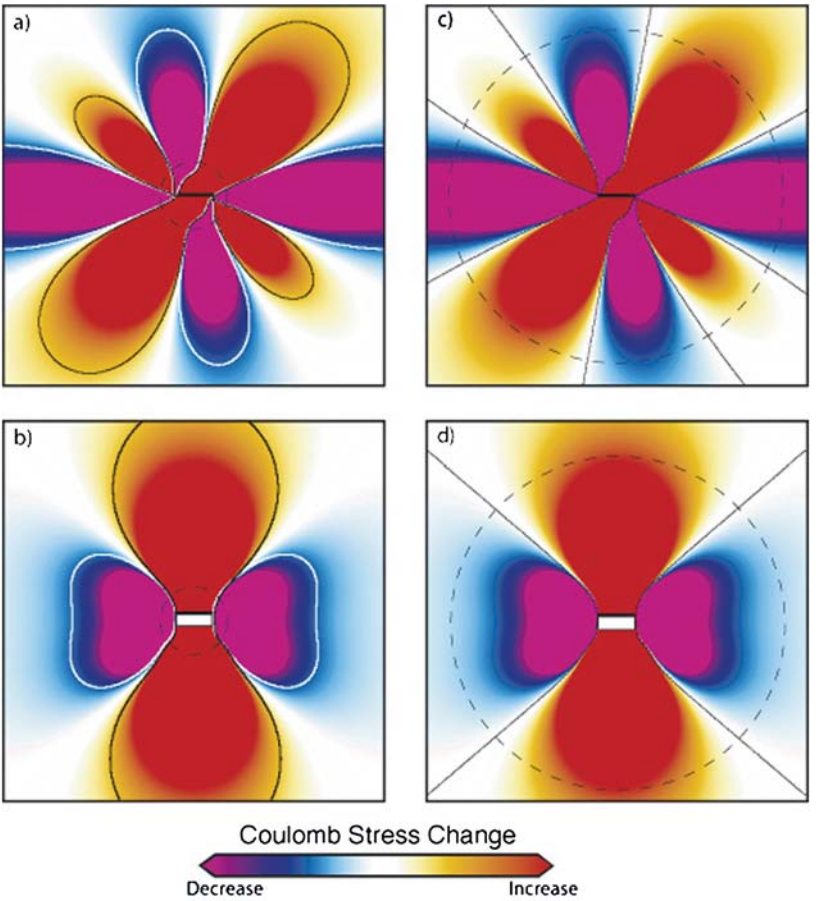


FIG. 2.1. Stress distribution prior to strike-slip (a)–(c) and dip-slip earthquakes (b)–(d) calculated by the backslip method. This fully describes the far-field stress distribution used in the Stress Accumulation model. Although the near-field stresses may not be completely represented in regions where dip-slip faulting modifies the topography (e.g. King et al., 1988; Stein et al., 1988), the affected region is never larger than the area indicated by small dashed circles in (a)–(b) and has no significant influence on the AMR studies discussed here. In earlier papers (e.g. Bowman and King, 2001a; Mignan et al., 2006a; Sammis et al., 2004) AMR was sought by finding an optimum stress contour (black line) in the lobes that must be loaded prior to the earthquake. For each event, a regional stress is chosen to be consistent with the mechanism of the mainshock (Table 2.1). The actual value of the optimum stress is not important in this procedure. The loading model is only important to provide the azimuthal distribution defining where AMR is expected. In this paper, AMR is first optimised for circles. The dashed circles (c)–(d) indicate a typical radius within which AMR is found. The black lines correspond to the zero stress contour, and are used to distinguish the parts of the circular region enclosing events that reduce or increase stress on the future earthquake fault. In the upper panels, the motion is left-lateral strike-slip on a vertical fault (grey line) and in the lower panels reverse motion occurs on a 45° dipping fault (outlined).

2.2. Methods

2.2.1. Quantifying AMR

Seismicity rate changes are typically modelled by a simple power-law time-to-failure equation (e.g. Bufe and Varnes, 1993). This is a relation of the form

$$\sum \epsilon(t) = A + B(t_f - t)^m \tag{2.1}$$

where t_f is the time of the large event, B is negative and $m = 0.3$. A is the value of $\epsilon(t)$ when $t = t_f$. The cumulative Benioff strain at time t is defined as

$$\epsilon(t) = \sum_{i=1}^{N(t)} \sqrt{E_i(t)} \tag{2.2}$$

where E_i is the energy of the ith event and $N(t)$ is the number of events at time t. This phenomenon, known as Accelerating Moment Release, has been identified for a substantial number of earthquakes (e.g. Sykes and Jaumé, 1990; Bowman et al., 1998; Bowman and King, 2001a; Bowman and Sammis, 2004; Jaumé and Sykes, 1999; Papazachos et al., 2002a).

We use the algorithms originally developed by Bowman et al. (1998) to identify domains of accelerating seismicity. In order to quantify the degree of acceleration, Bowman et al. (1998) define a curvature parameter C (c-value), where

$$C = \frac{\text{power-law fit RMS error}}{\text{linear fit RMS error}}. \tag{2.3}$$

Therefore, when the data are best characterised by a power law curve, the parameter C will be small. Following the approach of Bowman et al. (1998), the exponent m of Eq. (2.1) is held at 0.3. As a result, when the seismicity is either decelerating or linearly increasing, C will be at or near unity, whereas accelerating sequences will have $C < 1$.

In this study we use a Levenberg–Marquardt least-squares routine to fit the seismicity data to Eq. (2.1). Vere-Jones et al. (2001) pointed out that the least-squares technique leads to instabilities in the determination of the parameters of Eq. (2.1). However, they also note that the least squares method will produce a "reasonable visual fit to the data". In this work, the purpose of Eqs. (2.1) and (2.2) are merely to provide a logical metric that can be used to identify accelerating activity in the seismicity data. Because the Stress Accumulation model of King and Bowman (2003) does not specify the precise functional form of accelerating moment release, the power-law time-to-failure is used in this paper both for convenience and to be consistent with previous works.

2.2.2. Region Optimisation

To detect accelerating moment release in catalog data, it is necessary to isolate the precursory events in a space/time/magnitude window defined by the model of

interest. Many studies use simple geometrical shapes such as circles (Bowman et al., 1998; Brehm and Braile, 1998; Robinson, 2000; Zoller et al., 2001) or ellipses (Karakaisis et al., 2002; Papazachos and Papazachos, 2000) to define the region of interest. The *c*-value provides a simple, computationally inexpensive parameter that can be used to optimise this region. Typically, a search algorithm finds the size and/or orientation of the region that produces the lowest *c*-value, which is then defined as the precursory region. Bowman and King (2001a) extended this technique to search for precursory regions defined by a backslip model for the loading region. In their technique, the region is optimised by searching for the contour of the backslip stress field within which the seismicity produces the lowest *c*-value. Sammis et al. (2004) further refined the technique to search for both the best backslip loading region and the duration of the acceleration.

Since the far field stress distribution before a large earthquake is fully described by a backslip dislocation loading model, Bowman and King (2001a) suggested that it provides a better identification of the precursory region than simple circular regions. They documented that this technique yields lower *c*-values for accelerating sequences in the resulting regions. However, because King and Bowman (2003) demonstrated that the loading process should also produce regions where the seismicity rate is unchanged or even decreasing, a better test would be to assess the spatial distribution of the events contributing to the observed AMR. Figure 2.1 illustrates this concept for a simple vertical strike-slip fault and a 45° dipping reverse fault. A precursory region of interest defined by the backslip model of Bowman and King (2001a) (black line in Fig. 2.1a–c) will select events in the area where the stress accumulation model of King and Bowman (2003) shows the greatest increase in background seismicity. In contrast, the low-stress regions in the backslip model (outlined in white on Fig. 2.1a–c) include the regions where King and Bowman (2003) predict AMR should not be observed. Therefore, a simple test of the model would be to compare the *c*-value of the AMR in these two regions.

Such an approach is incorrect for two main reasons. First, the test implicitly assumes that the precursory seismicity must be within either the black or white contours in Fig. 2.1a–c. This will not be the case if the signal is distributed throughout an arbitrary volume (e.g. in a circular region). Second, this approach creates an inherent procedural bias. In the most straightforward application, the regions to be compared would be defined by the equivalent "positive" and "negative" contours of the backslip stress field. However, this value is found by optimising the contour in the positive lobes of the backslip stress field. This could be solved by also optimising the negative lobes. However, this would result in the comparison of fundamentally different datasets. To avoid this procedural bias, any test of the distribution of seismicity should not include optimisation of the stress field.

In this study, AMR is identified in circular areas using the region optimisation algorithm of Bowman et al. (1998). The region of interest is found by calculating the *c*-value in concentric circular regions up to 500 km in radius centered

TABLE 2.1. Characteristics of all earthquakes of $m \geqslant 6.5$ in Central and Southern California between 1950 and 2000 and their AMR-search parameters

Earthquake	Mainshock source parameters						Reg. stress		AMR region	
	Date	Mag	Strike	Dip	Length (km)	Width (km)	σ_1	σ_3	Δt (years)	Radius (km)
Kern County	07/21/1952	7.5	60	60	70	15	−30	60	9.75	102
Borrego Mountain	04/08/1968	6.5	135	90	30	20	0	90	22.44	23
San Fernando	02/09/1971	6.6	125	25	20	15	35	125	14.76	150
Coalinga	05/02/1983	6.7	135	15	38	7	45	135	47.84	90
Superstition Hills	11/24/1987	6.6	130	90	28	15	−5	85	5.72	136
Loma Prieta	10/18/1989	7.0	135	90	17	20	0	90	56.80	200
Landers	06/28/1992	7.3	150	90	80	15	15	105	35.05	212
Northridge	01/17/1994	6.7	120	35	20	15	30	120	37.45	231
Hector Mine	10/14/1999	7.1	145	90	40	15	10	100	47.93	228

on the mainshock epicenter. The duration of the AMR signal is also optimised as described by Sammis et al. (2004). For each region radius, the beginning of the AMR signal is found by calculating the c-value at an array of starting times ranging from the beginning of the catalog (see catalog description below) until the year prior to the event. The precursory region is defined as the combination of region radius and starting time that produces the optimal (lowest) c-value. After the precursory region is defined, we apply the backslip model of Bowman and King (2001a) to identify the subset of the circular region that was loaded by slip at depth (Fig. 2.1b–d). Because the region size under consideration is much greater than the thickness of the seismogenic crust, the region optimisation is effectively a 2-D problem (Bowman and King, 2001a); the Coulomb stress fields used in this study are calculated on optimally oriented planes at a depth of 5 km with a coefficient of friction of 0.5. The source mechanism used for each event is listed in Table 2.1. With this approach, the identified precursory region is not dependent on any physical model, and event discrimination using the calculated stress field only occurs after optimisation of the AMR. If slip at depth is the primary mechanism driving the observed AMR, then the c-value for events in the positive pre-stress sub-region should be significantly lower than the c-value in the negative pre-stress sub-region.

2.2.3. Statistical Significance

Given sufficiently flexible search parameters, most optimisation techniques designed in this manner will find some degree of acceleration in catalog data. We have quantified this bias by using random synthetic seismicity catalogs to determine the likelihood that a seismicity sequence with any given c-value could have arisen stochastically (Bowman et al., 1998). We use a Monte Carlo approach, where 1000 catalogues each composed of 500 events placed randomly in space

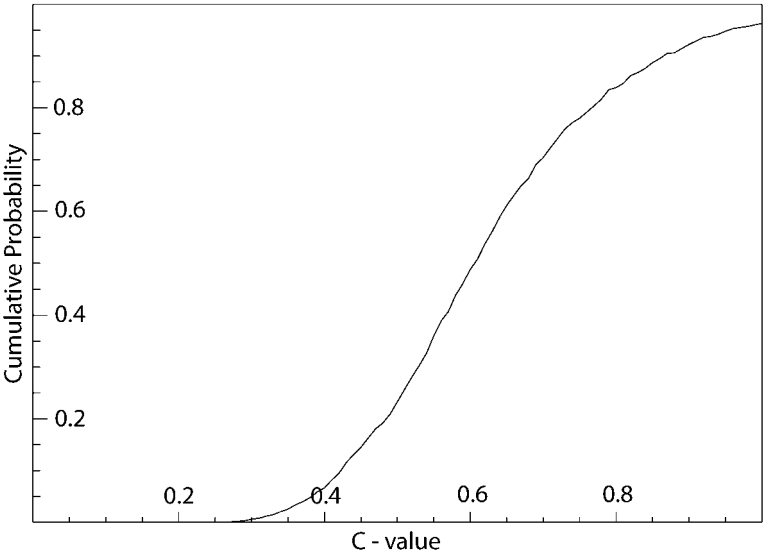


FIG. 2.2. Cumulative probability of generating a given C-value by stochastic processes. The figure suggests that any given synthetic catalog had a 50% chance of generating an optimal $C \geqslant 0.6$, and a 25% chance of generating $C \leqslant 0.5$.

and time are created. The events are assigned a uniformly distributed random "longitude" and "latitude" in the range [0 : 1] and a uniformly distributed random time in the range [0 : 1]. Each point is indexed to a magnitude that respects the Gutenberg–Richter law in the magnitude range [3.5 : 6.0]. Energy for each point is calculated using the magnitude-energy scaling law ($E = 10^{4.8+1.5M}$) developed by Kanamori and Anderson (1975). This is then used to generate the cumulative Benioff strain curve. We test the false-alarm rate in circular regions determined by the same search algorithm used for the Californian seismicity catalog. For each synthetic catalog, the AMR search is done within concentric circles centered on the point (0.5, 0.5). The c-value is calculated for regions with radii ranging from [0.05 : 0.5] and starting times ranging from [0.0 : 0.90]. AMR cannot be reliably observed in regions containing fewer than fifteen events. The optimisations scheme automatically rejects these regions by assigning them a c-value of 1. The best (i.e. minimum) c-value from the search is kept and the process is repeated for a new synthetic catalog.

Figure 2.2 shows the result of this analysis expressed as the cumulative probability of observing a given c-value or higher in the random synthetic catalog. The figure suggests that any given synthetic catalog had a 50% chance of generating an optimal c-value of 0.6 or higher. In contrast, the optimisation procedure found $C \leqslant 0.5$ in less than 25% of the cases, and $C \leqslant 0.4$ in less than 7% of the synthetic catalogs. Because there is a high likelihood of generating $C \geqslant 0.6$

through stochastic processes, any seismicity sequence with $C \geqslant 0.6$ is classified in this study as having "unclear" AMR. Lower c-values are more likely to represent AMR driven by nonrandom processes. We also note that these synthetic seismicity tests only produce stable c-values for subregions with more than 25 events.

2.3. Data

To facilitate comparison with earlier studies, we will analyse AMR before the 9 $M \geqslant 6.5$ earthquakes in southern and central California studied by (Bowman and King, 2001a) (Table 2.1). The seismicity catalog is provided by the Advanced National Seismic System (ANSS) (available at http://quake.geo.berkeley.edu/anss/catalog-search.html). We used events larger than a magnitude threshold $m_0 = 3.5$ within the time window January 1, 1930 to October 16, 1999 (the date of the last $M > 6.5$ event, Hector Mine). The AMR search was limited to a rectangular area $30^\circ <$ latitude $< 40^\circ$ and $-124^\circ <$ longitude $< -112^\circ$.

Unlike previous studies of accelerating moment release, the seismicity catalog in this study was declustered prior to searching for AMR. Although the precise physical origin of aftershocks is still poorly understood, most explanations invoke some form of stress transfer (static and/or dynamic) from a larger event (King and Cocco, 2000; King et al., 1994; Stein, 1999; Toda and Stein, 2000, 2002a). Because we are testing a model of seismicity rate changes driven by slip at depth, the unrelated transient effect of aftershocks will contaminate the results. Furthermore, the model of accelerating moment release described by King and Bowman (2003) results in a broad, steady increase of activity within the precursory region. Aftershocks to intermediate magnitude events in the precursory region represent a burst of activity that may produce anomalous c-values in the search algorithm (Fig. 2.3). A detailed description of the declustering technique is provided in Appendix.

Changes in the lower magnitude of completeness of the catalog can cause artificial rate changes unrelated to any geologic process. In AMR studies, this generally is manifested as apparent acceleration through the entire catalog as the sensitivity of seismic networks improves. To evaluate this, the C-value is calculated for the

FIG. 2.3. Influence of the declustering procedure on the calculation of C. (a) A major jump in 1971 (San Fernando earthquake and its aftershocks) obscures the background acceleration ($C = 0.82$). Once major events ($m \geqslant 6.4$, see Appendix 1) and their associated aftershocks are removed, the C-value ($C = 0.36$) provides a better reflection of changes in the background seismicity rate. (b) Aftershocks cause a jump in the cumulative Benioff strain curve. The C-value for this sequence is low ($C = 0.64$), even though the background seismicity is clearly not accelerating. After declustering the sequence gives $C = 0.91$, more consistent with the linear trend of the background seismicity.

a.
c-value = 0.82
Cumul. Benioff Strain (x10^8 Joules^1/2)
jump
aftershocks
large event
1960
1970
1980
1990
Date
DECLUSTERING
c-value = 0.36
Cumul. Benioff Strain (x10^8 Joules^1/2)
1960
1970
1980
1990
Date
b.
c-value = 0.64
Cumul. Benioff Strain (x10^8 Joules^1/2)
jump
1960
1970
1980
1990
Date
DECLUSTERING
c-value = 0.91
Cumul. Benioff Strain (x10^8 Joules^1/2)
1960
1970
1980
1990
Date

entire declustered catalog during the time window in which AMR is observed for the nine earthquakes being studied. In all 9 cases, the evolution of the cumulative Benioff strain is clearly linear for the entire region ($C_{\text{mean}} = 0.96$) (see Fig. 2.4 and Table 2.2).

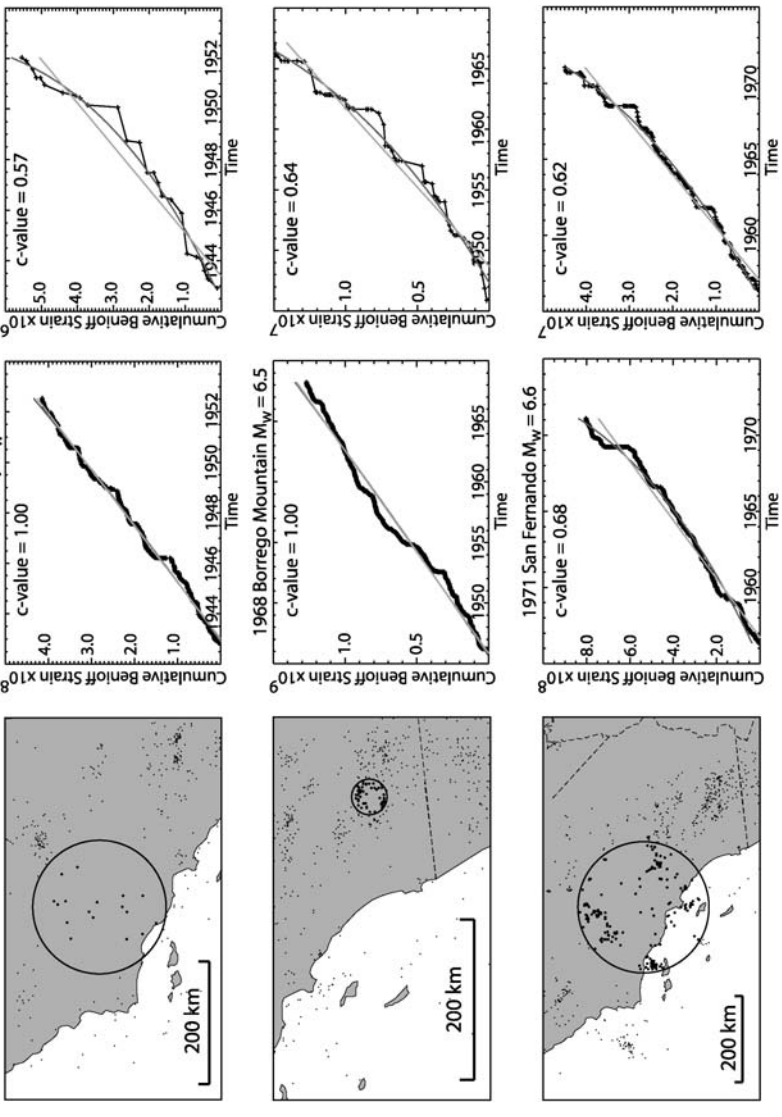
2.4. Results and Discussion

For each of the 9 earthquakes studied, it was possible to find clear AMR in optimised circles centered on their epicenter ($C_{\text{mean}} = 0.49$, details in Table 2.2). The observed C-values for the circular regions ranged from 0.36 for the 1994 Northridge earthquake to 0.64 for the 1968 Borrego Mountain event (Fig. 2.4). Note that the size of the regions found here are slightly different from those found by previous studies (e.g. Bowman et al., 1998; Zoller et al., 2001). This is because the current study used declustered seismicity and the search algorithm jointly optimised the region size and the duration of the AMR.

The backslip stress models of Bowman and King (2001a) were next used to discriminate between the "positive" and "negative" pre-stress regions within each optimal circular region. Figure 2.5 shows the resulting regions and seismicity curves for each of the studied events. The dashed lines in Fig. 2.5a represent the 0-bar stress change contour. The areas shaded dark grey are within the optimal circular radius and in the positive lobes of the pre-stress field; this is where the stress accumulation model of King and Bowman (2003) predicts seismicity leading to AMR should be concentrated. The light grey regions are within the optimal circular radius but in the negative lobe of the pre-stress field; King and Bowman (2003) predict that AMR in this region should be weak or nonexistent. The seismicity curves for the positive and negative pre-stress zones are shown in Figs. 2.5a and 2.5b respectively, and summaries in Table 2.2.

In the positive pre-stress regions, c-values range from $C = 0.36$ for the 1992 Landers earthquake to $C = 0.63$ for the 1968 Borrego Mountain event, with $C_{\text{mean}} = 0.49$. The seismicity in the positive stress regions produced AMR much more reliably than the random synthetic catalogs. Only two events had $C > 0.6$, Borrego Mountain ($C = 0.63$) and Kern County ($C = 0.61$); random catalogs have an approximately 50% chance of producing similar acceleration. Six of the events (San Fernando, Coalinga, Superstition Hills, Loma Prieta, Landers, and

FIG. 2.4. Results for the AMR-search by circles. The panels on the left show the optimal circular regions for the Kern County, Borrego Mountain, San Fernando, Coalinga, Superstition Hills, Loma Prieta, Landers, Northridge, and Hector Mine earthquakes. The center panels show the cumulative Benioff strain for the entire region from the beginning of the observed AMR until the time of the respective mainshock. This curve should be linear, indicating that there is no systematic change in the completeness of the catalog for the selected magnitude and time windows. The panels on the right show the cumulative Benioff strain in the optimal circular region for each event.


1952 Kern County $M_w = 7.5$
c-value = 1.00
Cumulative Benioff Strain x10^8
Time
1944 1946 1948 1950 1952
1.0 2.0 3.0 4.0
1968 Borrego Mountain $M_w = 6.5$
c-value = 1.00
Cumulative Benioff Strain x10^9
Time
1950 1955 1960 1965
0.5 1.0
1971 San Fernando $M_w = 6.6$
c-value = 0.68
Cumulative Benioff Strain x10^8
Time
1960 1965 1970
2.0 4.0 6.0 8.0
c-value = 0.57
Cumulative Benioff Strain x10^6
Time
1944 1946 1948 1950 1952
1.0 2.0 3.0 4.0 5.0
c-value = 0.64
Cumulative Benioff Strain x10^7
Time
1950 1955 1960 1965
0.5 1.0
c-value = 0.62
Cumulative Benioff Strain x10^7
Time
1960 1965 1970
1.0 2.0 3.0 4.0
200 km
200 km
200 km

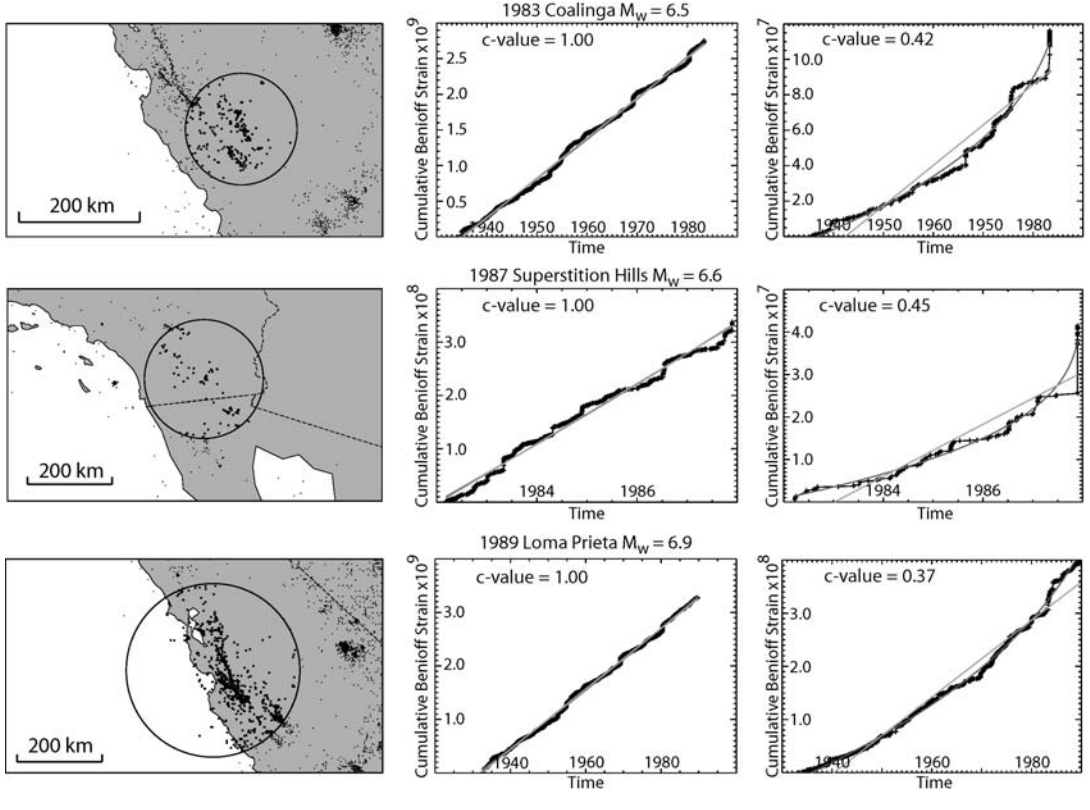


FIG. 2.4. Continued.

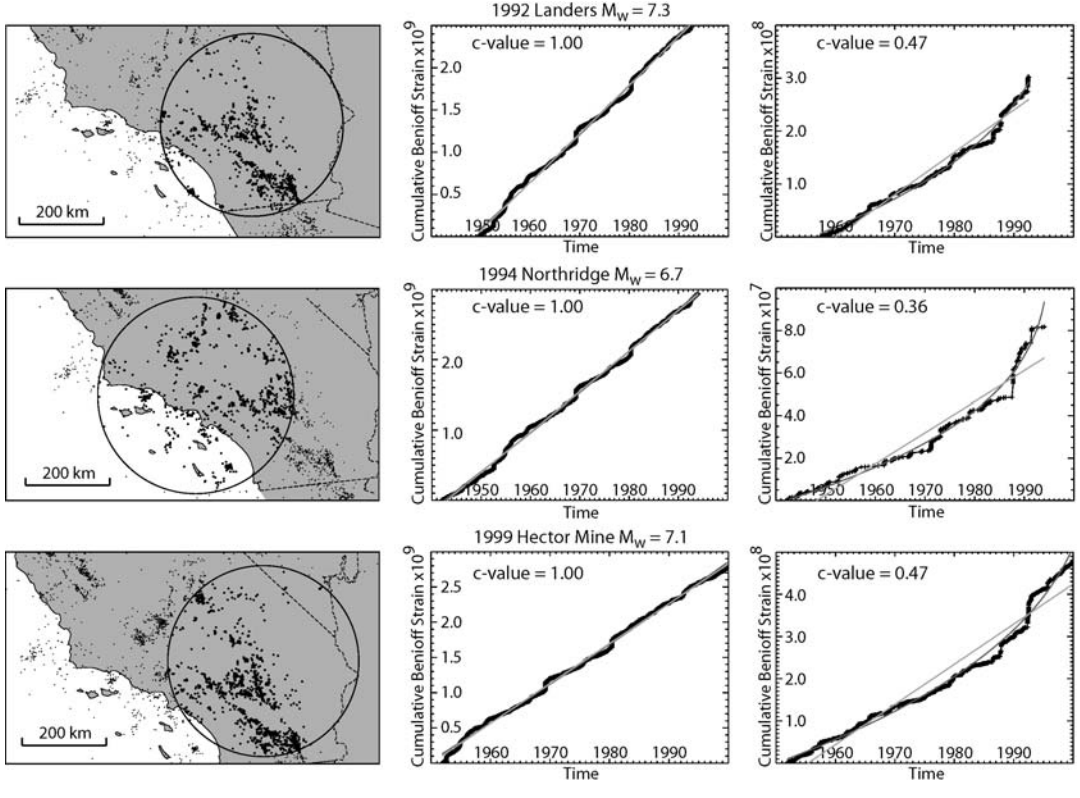


FIG. 2.4. Continued.

TABLE 2.2. C-values for all earthquakes of $m \geqslant 6.5$ in Central and Southern California between 1950 and 2000. Values in bold characters correspond to reliable observations of Accelerating Moment Release (AMR)

Earthquake	C (region)	C (circle)	C (pos. stress)	C (neg. stress)
Kern County	1.00	**0.57**	0.61	1.00
Borrego Mountain	0.94	0.64	0.63	1.00
San Fernando	0.93	0.62	**0.49**	0.93
Coalinga	1.00	**0.42**	**0.47**	0.63
Superstition Hills	1.00	**0.45**	**0.44**	1.00
Loma Prieta	1.00	**0.37**	**0.40**	0.62
Landers	1.00	**0.47**	**0.36**	0.65
Northridge	1.00	**0.36**	**0.47**	**0.54**
Hector Mine	1.00	**0.47**	**0.57**	0.80
MEAN	0.99	**0.49**	**0.49**	0.80

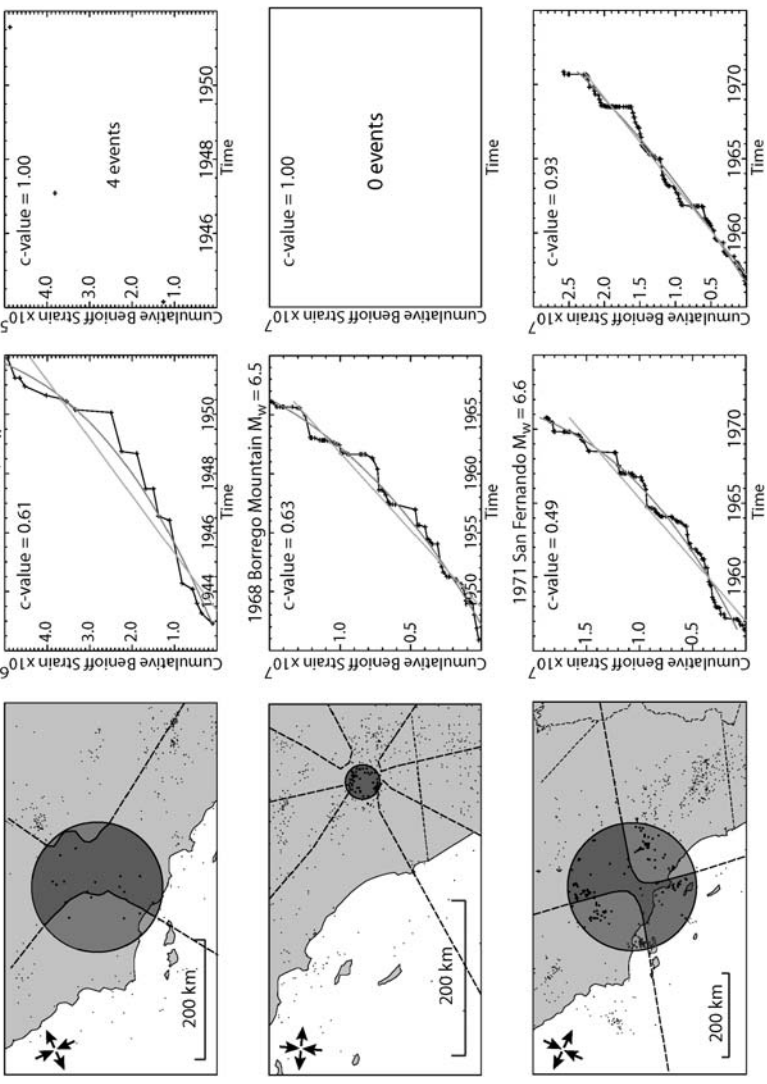
Northridge) had $C < 0.5$; random catalogs have less than 23% chance of producing any of these individual sequences. The probability of random catalogs producing all of these sequences is small.

The negative pre-stress region had c-values ranging from 0.54 for the 1994 Northridge earthquake to $C = 1.0$ for the 1952 Kern County, 1967 Borrego Mountain, and 1987 Superstition Hills events, with $C_{\text{mean}} = 0.80$. Only one event had $C < 0.6$ (Northridge).

We note that the circular search by itself appears to outperform the model of Bowman and King (2001a). Furthermore, the regions found here are significantly different from those observed in earlier works (e.g. Bowman et al., 1998; Bowman and King, 2001a). However, a direct comparison between earlier works and the results of this paper is not possible for two main reasons. First, Bowman and King (2001a) performed an optimisation of the stress contour used to define the region of interest, whereas no optimisation of the stress field was permitted here. Second, Bowman and King (2001a) did not decluster the seismicity, unlike the data used in this study. Therefore, the results of this study should not be interpreted to indicate that circular regions outperform the backslip loading model.

What do these results suggest about the physical mechanism driving observed accelerating moment release before large earthquakes? If cascading interactions

FIG. 2.5. Results for the AMR-search by regions of positive and negative pre-stress. The panels on the left show the search regions. The optimal circular region (of Fig. 2.4) is shown. The dashed line indicated the 0-bar stress change contour; regions of positive stress are shaded dark grey and regions of negative stress are shaded light grey. The orientation of the regional stress field used to calculate the pre-stress is indicated by the arrows on the map. The central panels show the cumulative Benioff strain in the positive stress regions. The panels on the right show the cumulative Benioff strain in the negative stress regions. See text for discussion.


1952 Kern County M_w = 7.5
c-value = 0.61
4 events
1968 Borrego Mountain M_w = 6.5
c-value = 0.63
0 events
c-value = 1.00
1971 San Fernando M_w = 6.6
c-value = 0.49
c-value = 0.93
Cumulative Benioff Strain x10^5
Cumulative Benioff Strain x10^6
Cumulative Benioff Strain x10^7
Time
200 km

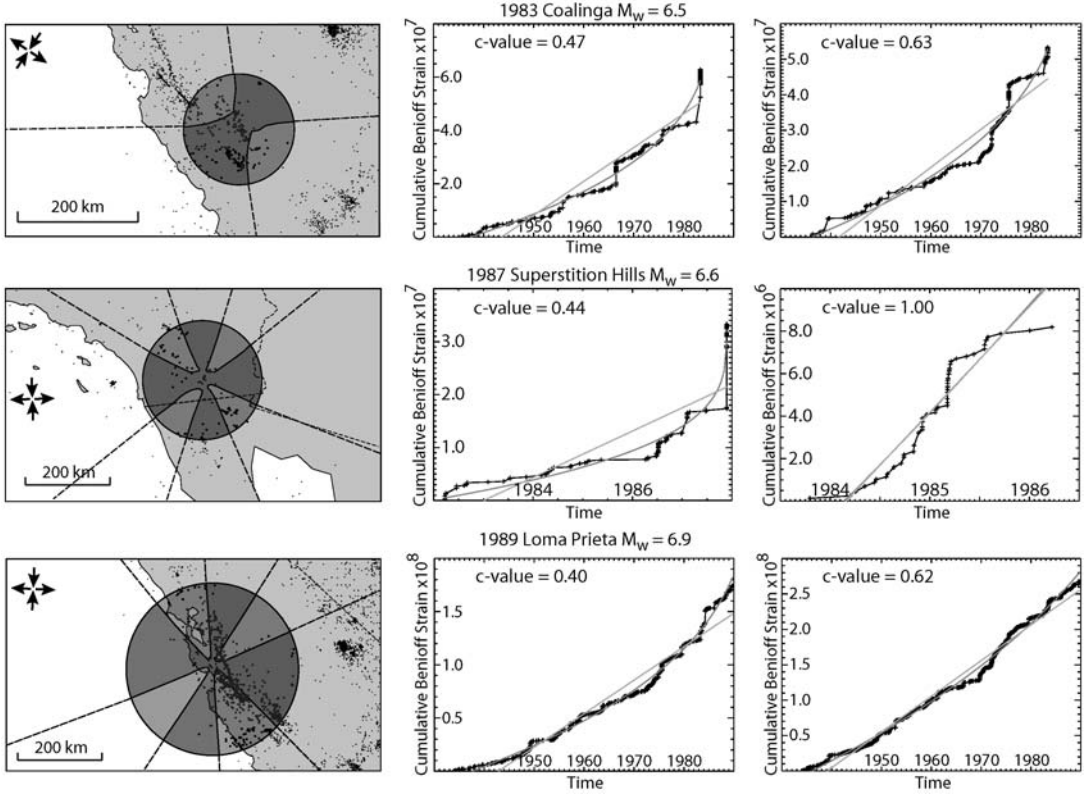


FIG. 2.5. Continued.

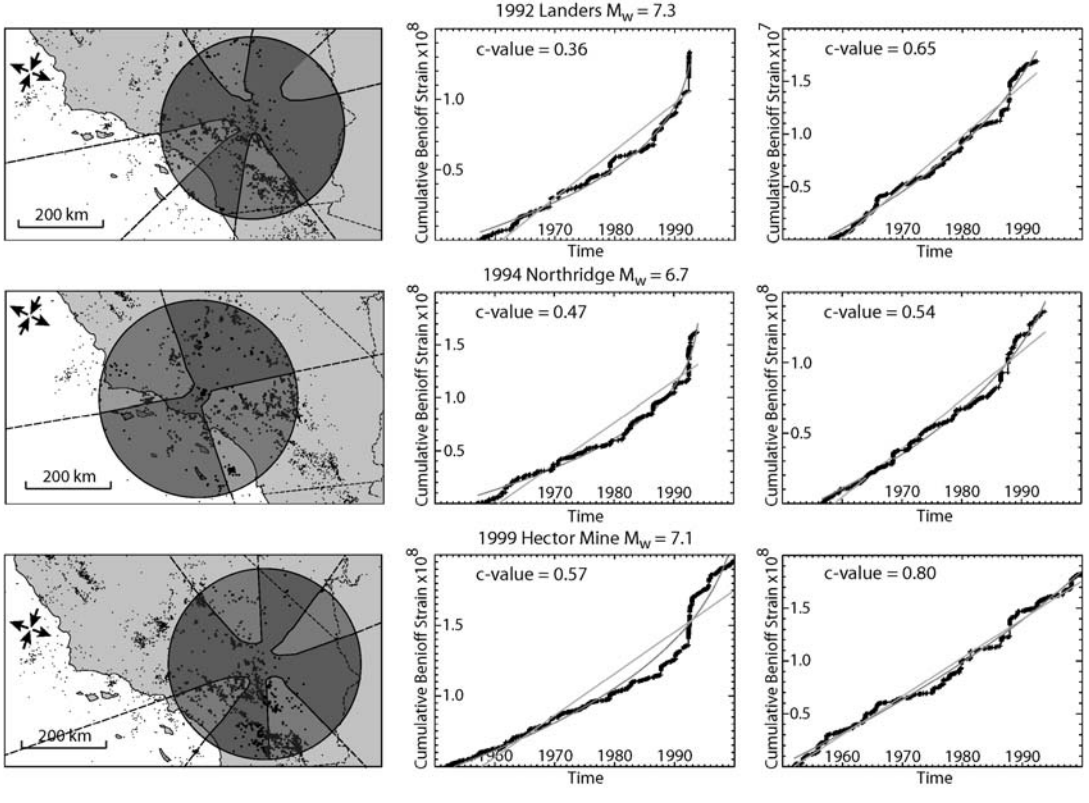


FIG. 2.5. Continued.

driven by stress transfer from smaller events is the driving mechanism of AMR, then the precursory rate changes should be expected to be strongest in the negative lobes of the Coulomb pre-stress field. In contrast, if the stress accumulation model of King and Bowman (2003) is correct, then AMR should be more apparent (smaller c-values) in the positive lobe of the pre-stress field. The latter prediction is true for all 9 earthquakes in the study region, consistent with the predictions of King and Bowman (2003). This suggests that slip at depth has a strong, if not dominant, influence on the development of accelerating moment release. While this alone does not rule out cascading seismicity as an important factor in the evolution of regional seismicity, it does suggest that static stress triggering does not play an important role in the generation of accelerating moment release.

The relative importance of stress triggering and (stress) loading in generating precursory AMR has implications not only for the spatial distribution of accelerating seismicity in the region of an impending earthquake, but also for the size of events that might be expected to exhibit precursory AMR. Simple static stress interaction models (Fig. 2.6) demonstrate the relative role of stress triggering and stress loading for events with different ratios of fault length, L, to width, W. Figure 2.6 shows the fault plane viewed face-on, with a dotted rectangle indicating the site of an upcoming event. The grey region at the base of each figure represents the fault in the lower crust, which moves aseismically to load the overlying seismogenic portion of the fault. The white rectangles represent the rupture area of earlier events along strike with sizes comparable to the impending earthquake; the colouring represents the static stress change in the seismogenic zone from both the earlier adjacent events and creep on the fault in the lower crust.

For events with $L/W \sim 1$ (Fig. 2.6a), stress transfer from adjacent events contributes more to the overall stress on the fault than the loading caused by aseismic creep in the lower crust. This can be seen by comparing the 10-bar stress contour for adjacent event (solid line) and the loading from below (dashed line). In contrast, the stress field on the fault plane of events with $L/W > 1$ (Fig. 2.6b) is dominated by aseismic creep on the fault in the lower crust. Although static stress triggering certainly increases the stress on the plane of the impending earthquake, the effect is restricted to only a small section of the fault. Comparing the 10-bar stress contour for both the adjacent event and loading from below suggests that large earthquakes are dominantly loaded by aseismic slip on the fault at depth rather than stress transfer from nearby earthquakes, consistent with the results presented here.

Because stress triggering is the dominant loading mechanism for small and intermediate ($L/W \sim 1$) events, slip at depth will not necessarily be the dominant loading mechanism on the rupture plane of the small earthquakes. Unlike large events, it is likely that any precursory seismicity before small earthquakes is dominated by traditional stress transfer. However, any systematic test of this concept for small events would require much higher precision catalog locations than are routinely available at present. Furthermore, the stress-based approach used in this

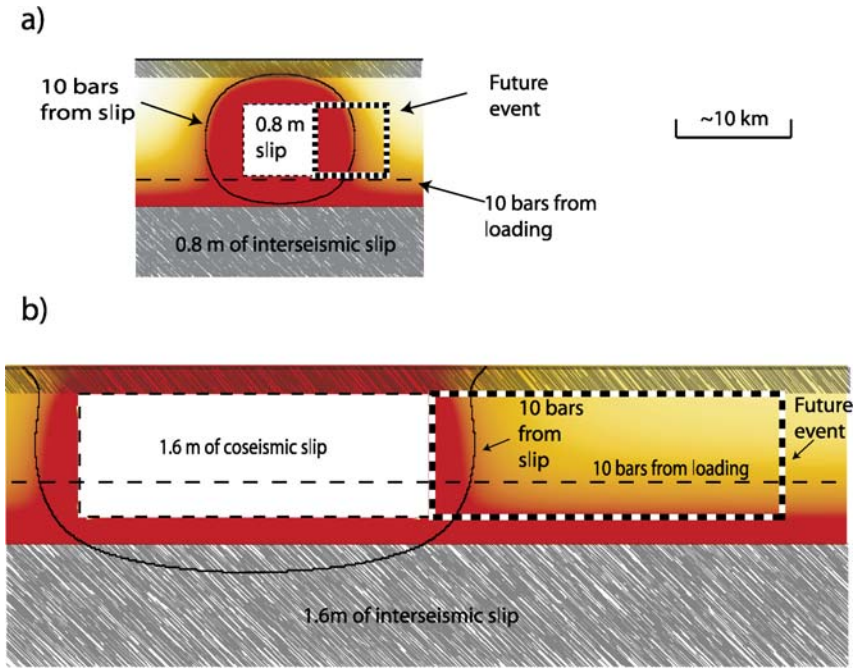


FIG. 2.6. Role of stress triggering and stress loading for faults of different L/W ratios. The colouring indicates the stress level. In the models, a surface that has already ruptured (white rectangle) increases stress around it. Under the seismogenic zone, the fault (shaded grey) moves aseismically causing loading. The relative contribution of each effect can be assessed by looking at the 10 bar contour indicated for each separately. (a) For small events, $L/W \sim 1$, the stress field on the fault plane may be weakly influenced by loading and greatly influenced by adjacent events; (b) For a large event ($L/W > 1$), loading predominates the stress field on the fault plane of the future event.

work is not practical for small earthquakes since the focal mechanisms for these events are not well constrained in most earthquake catalogs. As a result, investigations of precursory activity for small-to-intermediate magnitude seismicity are more suitable using statistical techniques such as the Pattern Informatics method (e.g. Tiampo et al., 2006).

This simple model for the relative importance of stress triggering versus fault loading also suggests that the approach of Bowman et al. (1998) can only be reliably used to observe AMR before large earthquakes. Furthermore, our results suggest that models of regional seismicity that are driven primarily by stress transfer and fail to account for loading by aseismic creep at depth (e.g. Robinson et al., 2005) will not produce systematic accelerating moment release before model events. Because the loading mechanism for events with $L/W > 1$ is dominated by aseismic creep on the fault in the lower crust, seismicity rate changes before large

earthquakes are more accurately represented by the model of King and Bowman (2003), as supported by the results of this study.

2.5. Appendix

Declustering Algorithm: Aftershocks represent a local perturbation of the background seismicity rate as the stress field is reorganised following an earthquake. This process is independent of the slow regional seismicity rate change represented by AMR. However, large aftershock sequences produce step-function changes in the regional seismicity rate. This is evident in the cumulative Benioff strain plot for the entire California catalog shown in Fig. 2.A1a. The c-value analysis described in the main text could incorrectly identify these aftershock clusters as AMR. To alleviate this problem, the seismicity catalog in this study is declustered before the c-value analysis is run.

The average seismicity rate is defined here as the number of events $M \geqslant 3.5$ in a two-week window. For the duration of the catalog (Fig. 2.A1), the average seismicity rate defined in this way is 17 events per two weeks. We classify an event as part of an aftershock cluster if it:

(1) occurs when the seismicity rate is greater than the long-term average;
(2) is preceded by an event greater than some arbitrary mainshock threshold, M_{main};
(3) is within some distance, R_{cluster}, of a mainshock.

For this study, $R_{\text{cluster}} = 50$ km. The mainshock threshold, M_{main}, is a parameter that is adjusted to achieve the desired level of declustering for the catalog. For instance, the dashed lines in Fig. 2.A1b indicate the occurrence of $M \geqslant 6.4$ events in the catalog used here. By inspection, it is clear that most of the bursts in seismicity rate coincide with events $M \geqslant 6.4$. We therefore select $M_{\text{main}} = 6.4$ as the mainshock threshold for the declustering. The declustering algorithm works by deleting all aftershocks until the seismicity rate is equal to the long-term average.

The results of this declustering algorithm are shown in Fig. 2.A2. In the new catalogue, 4069 events have been removed (28 of $m \geqslant 6.4$) for a total of 31 415 events. Figure 2.A2a shows the mapped distribution of "independent" earthquakes (black dots) and deleted aftershocks (white diamonds). Figure 2.A2b shows the seismicity rate of the declustered catalog. Note that not all of the clusters were removed. Our goal is not to identify all aftershocks, but to minimize their impact on the cumulative seismicity curves. An inspection of Fig. 2.A2c shows that this is the case; the steps in the curve caused by aftershocks have been removed, resulting in a smooth seismicity curve for the entire catalog. The same effect can be achieved with a wide range of declustering parameters.

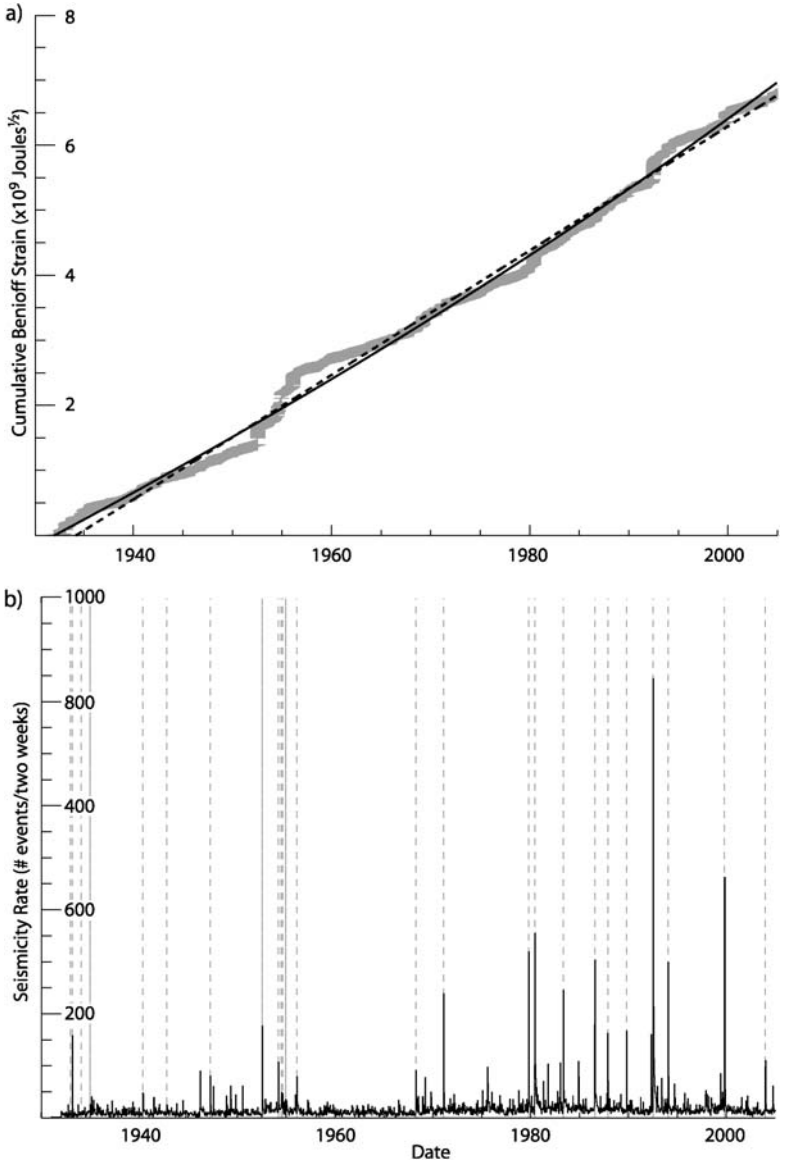


FIG. 2.A1. Seismicity catalogue for Central and Southern California (from 1930 to 2005) before declustering. (a) Evolution of the cumulative Benioff strain for the entire catalog. Note the irregularity of the curve, corresponding to the presence of large events and their associated aftershocks. (b) Seismicity rate through the duration of the catalog. The averaged number of events ($n \sim 17$) corresponds to the background seismicity for a period of 2 weeks (average background rate). The majority of the peaks in the seismicity rate can be related to the occurrence of $M \geqslant 6.4$ earthquakes.

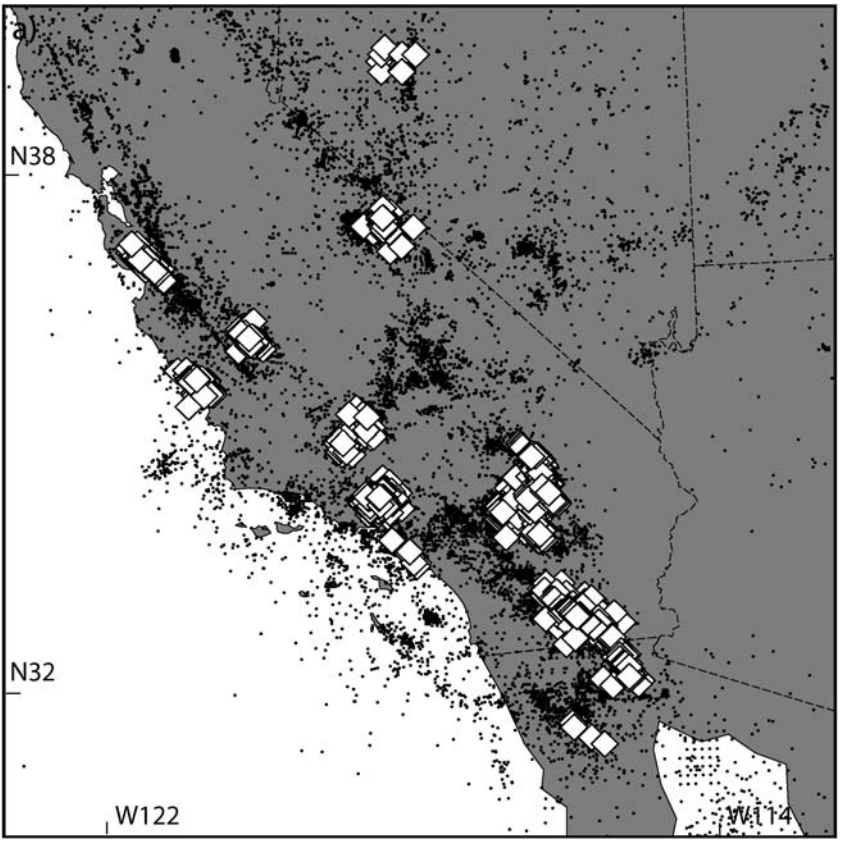


FIG. 2.A2. Declustered seismicity catalog. (a) Map of the region. White diamonds correspond to events that have been removed by the declustering procedure (see text for details). Each cluster is composed of a large event ($m \geqslant 6.4$) and its aftershocks. (b) Seismicity rate of the declustered catalog. Some high seismicity rate peaks are still present. These clusters are either swarms or are preceded a $M < 6.4$ mainshock, and are left in the catalog. (c) Evolution of the cumulative Benioff strain in the entire declustered catalog. Note the relatively linear trend of the curve after declustering.

2.6. Supplemental Information

Validity of the declustering method: If the methodology is correct, events removed from the catalogue should be aftershocks of large earthquakes. The modified Omori relation describes the decay in aftershock rate with time after a mainshock (Utsu, 1961) (extension of the Omori law (Omori, 1894)). An aftershock sequence will decay, on average, with the reciprocal of time ($1/t$) as following:

$$\lambda_0(t) = K(t + c)^{-p} \tag{2.4}$$

with K, c and p constants. The p-value is usually near the unity.

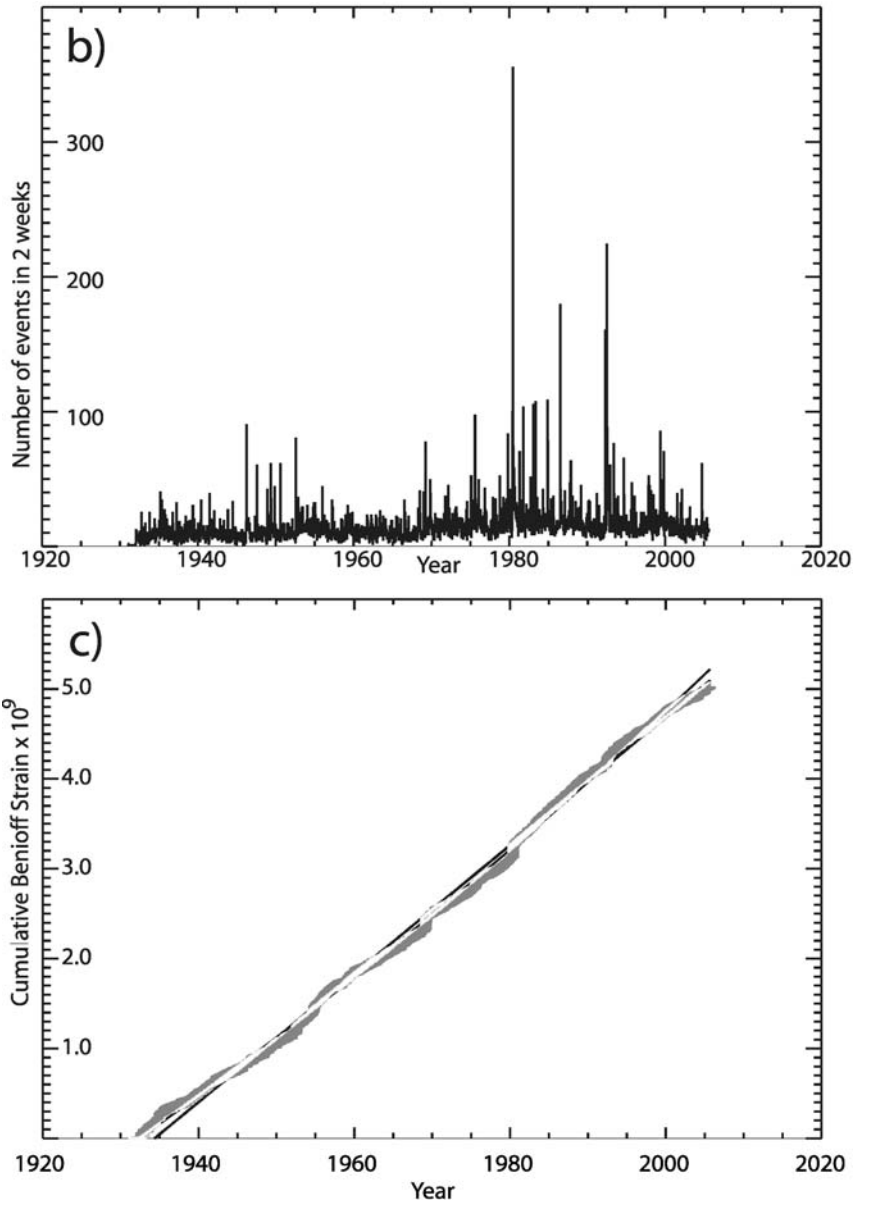


FIG. 2.A2. Continued.

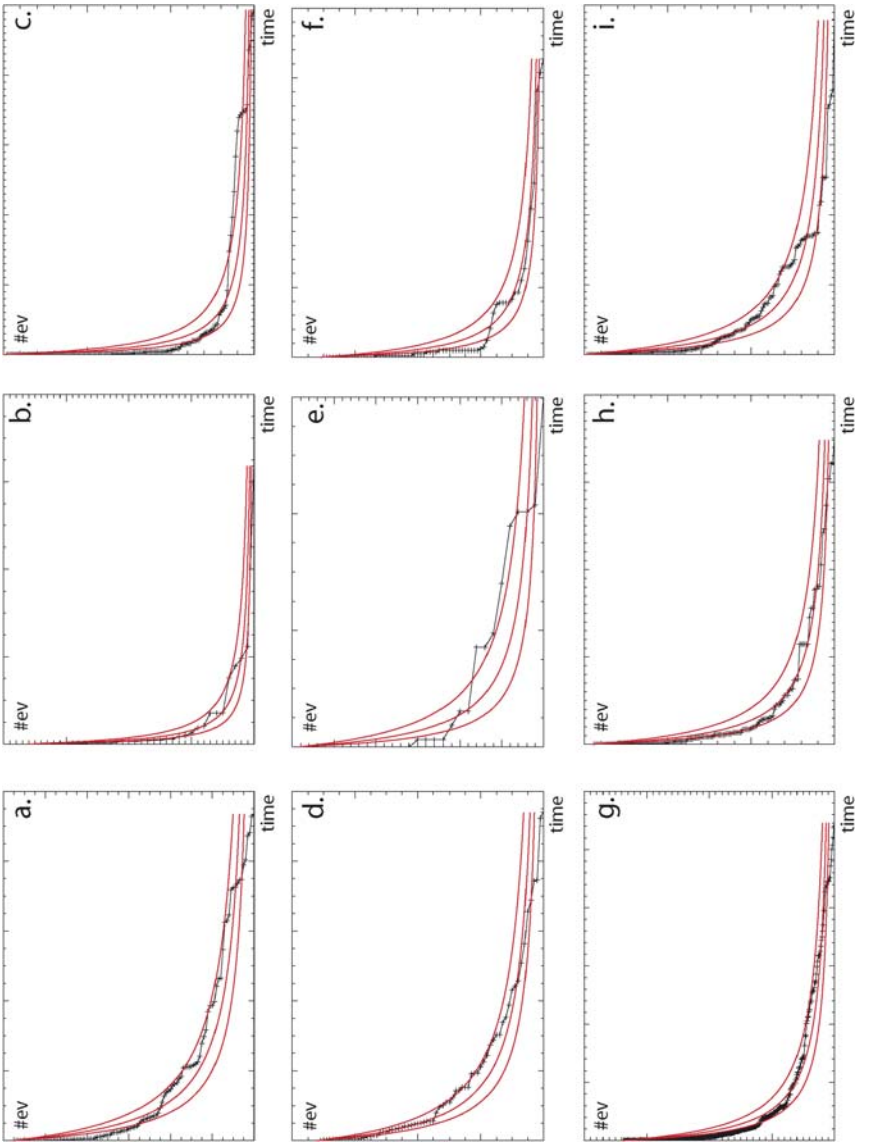

a.
b.
c.
d.
e.
f.
g.
h.
i.
#ev
time

FIG. 2.7. Validation of the declustering method by verification of the Omori law for clusters of large earthquakes $M \geqslant 6.5$ in Central and Southern California between 1950 and 2000. Curves correspond to the modified Omori law with three different p-values, $p = 0.95$, $p = 1.00$, $p = 1.05$ (up to down). The Omori decay model was fit by trial and error to best match the data; (a) Kern County earthquake; (b) Borrego Mountain earthquake; (c) San Fernando earthquake; (d) Coalinga earthquake; (e) Superstition Hills earthquake; (f) Loma Prieta earthquake; (g) Landers earthquake; (h) Northridge earthquake; (i) Hector Mine earthquake.

The Omori decay model was fit by trial and error to best match the data. It has been verified for all earthquakes of $M \geqslant 6.5$ in Central and Southern California from 1950 to 2000, corresponding to events used for the study of AMR. For the 9 earthquakes, the corresponding cluster follows a decay, which can be represented by the modified Omori with p-values between 0.95 and 1.05 (Fig. 2.7). This confirms that the clusters are composed of aftershocks.

The declustering procedure adopted does not remove all aftershocks associated with events of magnitude larger to M_c but it removes a sufficient number to prevent the cluster from disturbing the stability of the calculation of the c-value.

3. TEMPORAL DISTRIBUTION OF ACCELERATING MOMENT RELEASE

> "[...] a self-organised critical state is charactered by power law distributions of event sizes and long-range spatial correlation of fluctuations around the stationary state. Because earthquakes are characterised by several power laws, such as the Gutenberg–Richter frequency-magnitude relation, Omori's Law for aftershocks, and the fractal distribution of hypocenters, the application of this concept of self-organised criticality to earthquakes is now often taken for granted in the seismological community."

C.G. Sammis and D. Sornette, Positive Feedback, memory, and the predictability of earthquakes, Colloquium Proc. Natl. Acad. Sci. USA (2002)

A number of complex natural phenomena (i.e. non-linear dynamic systems) exhibit power-law frequency-size statistics in a very robust manner. This has been observed from Physics to Biology, from Languages to Economics. It is also observed in natural hazards, such as earthquake populations, forest fires and landslides (e.g. Turcotte et al., 2002).

It is commonly thought that power-law distributions are a "signature" of self-organised criticality systems (Bak et al., 1987). In that state, small perturbations result in chain reactions (avalanches), which can affect any number of elements within the system. Proponents of the theory see multiple manifestations of self-organised criticality in earthquake populations, from the Gutenberg–Richter law to accelerating moment release. In their view, accelerating moment release represents a cascade of events that leads to the largest one. This concept is not the same

as the Stress Accumulation model described in Sections 1 and 2 where accelerating moment release only represents the "signature" of the final step of aseismic loading on the major fault that is going to rupture (Elastic Rebound theory).

There has been a wide interest in comparing earthquake populations to critical systems as it can be seen in many reviews or special issues concerned with earthquake interactions and accelerating moment release (e.g. Jaumé and Sykes, 1999; Sammis and Sornette, 2002; Papazachos, 2003; Tiampo and Anghel, 2006). However this apparent consensus can lead to misunderstanding and the application of criticality to earthquake populations should not be taken for granted. For example, the Stress Accumulation model is understood by Tiampo and Anghel (2006) as a method based on the concept of stress transfer, but with a critical point physical origin. This is not correct. Sammis and Sornette (2002) also associate the Stress Accumulation model with a finite-time singularity process based on the concept of critical point, but this is not a necessary condition.

Since the Stress Accumulation model, which does not draw on the concept of criticality has been proposed as an explanation for accelerating moment release, the role of critical processes becomes questionable. However, it can be difficult to argue for or against criticality concepts since, by definition, these processes can only be caught by looking at the system globally (e.g. Self-Organised Criticality systems (Bak et al., 1987, 1988)) making it difficult to verify the reality of such underlying processes. It may be more useful to ask if such approaches have any practical utility.

The purpose of this chapter is to show that accelerating moment release can be formulated in terms of stress transfer with no need of criticality. First, critical processes are reviewed and the power-law time-to-failure equation explained in these terms (Section 3.1). Then a new mathematical formulation is demonstrated based on the Stress Accumulation model that draws on none of these concepts (Section 3.2).

3.1. Origin of Accelerating Moment Release from Critical Processes

3.1.1. Power-Laws: "Signature" for Criticality?

Self-organised criticality: The sandpile model (Bak et al., 1987, 1988; Bak and Tang, 1989; Bak and Paczuski, 1995) was first proposed to explain the concept of Self-Organised Criticality and is defined as follows.

The height (i.e. slope in real cases) of the "pile" at the point (i, j) of a grid is denoted $Z(i, j)$. Each height corresponds to an integer value (counter). At each time step, the height at some random point is increased by 1 unit, $Z(i, j) \rightarrow Z(i, j) + 1$. If the height at that site now exceeds an arbitrary critical height Z_{cr}, then a toppling event occurs, where the height of the unstable site is reduced by 4 units and the height at each of the four neighbours on the grid is increased by 1

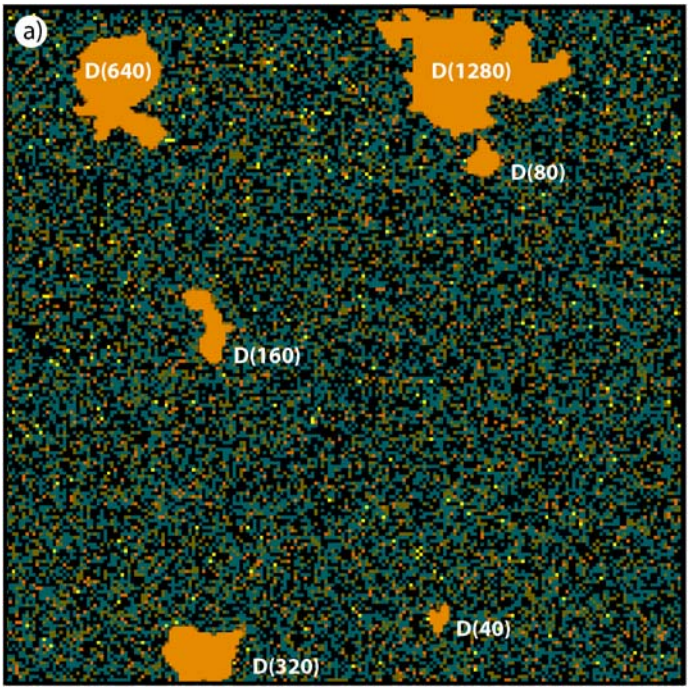


FIG. 3.1. Sandpile model; (a) Domain (areal extent) for several different avalanches in a sandpile model (SOC). Each avalanche (light grey colour) was triggered by the addition of a single increment (at a random point of the grid). The various shadings indicate heights 0–4, with 4 being the critical height. The light grey colour indicates sites that have toppled. White sites are active, toppling sites; (b) Scaling-law obtained by a sand pile model following (Bak et al., 1987) (see algorithm in Appendix A). Log–log plot of the frequency of occurrence of given avalanche size $D(s)$ versus the size of the avalanche s.

unit.

$$\begin{cases} Z(i, j) \to Z(i, j) - 4, \\ Z(i \pm 1, j) \to Z(i \pm 1, j) + 1, \\ Z(i, j \pm 1) \to Z(i, j \pm 1) + 1. \end{cases}$$

This is true except along the boundaries where "sand" is thrown out of the "pile". If any of the neighbouring sites are now unstable ($Z > Z_{\mathrm{cr}}$), the process continues until none of the Z values in the grid exceeds the critical value. Then an avalanche is over, and a new avalanche can be started by adding another increment to the grid.

One can imagine that for each point in the grid, there is a domain D which is the set of all points whose Z value was changed as a result of the initial perturbation. Figure 3.1a plots D for several different single perturbations that resulted in an avalanche. Each domain D has a finite size s, which is the areal extent of an

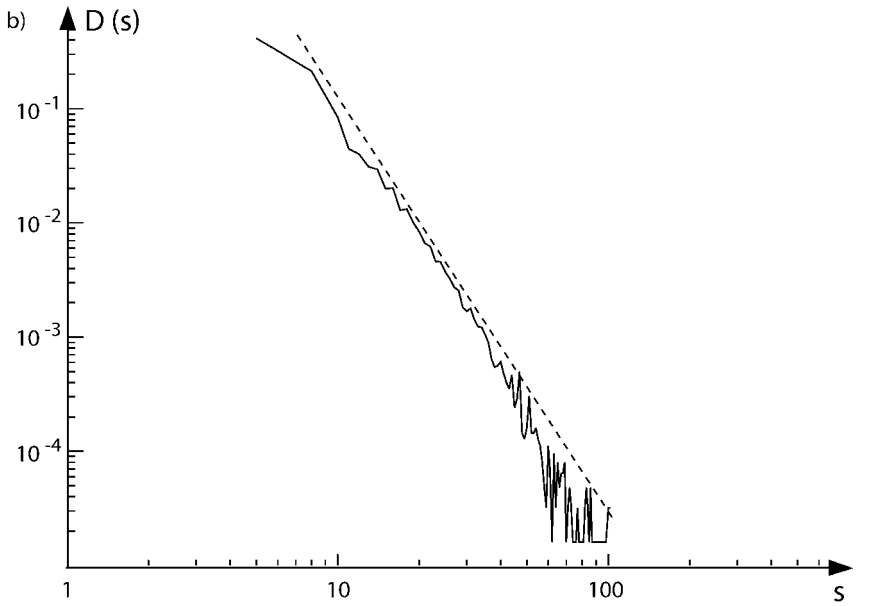


FIG. 3.1. Continued.

avalanche (number of topplings). Figure 3.1b plots D as a function of s. The figure clearly shows a power-law distribution, implying the system has self-organised to a critical state. The relation is

$$\log D(s) \propto s \tag{3.1}$$

where $D(s)$ is the frequency of occurrence of avalanches of size s. Each domain D (i.e. avalanche) was triggered by the addition of a single perturbation at a random point of the system. Note that avalanches also exhibit a power law distribution through time (Bak et al., 1987). Other models than the sandpile model have also been proposed, such as forest fire models (e.g. Turcotte et al., 2002; Papazachos, 2003) and slider blocks models (e.g. Carlson and Langer, 1989).

Critical point theory: In general, self-organised criticality formulations suffer from the problem that there should be no upper limit to the size of the largest event and they also generally fail to address the observations that major earthquakes recur on major recognisable faults. This can be resolved if it is explained as a finite-size effect (e.g. Sammis and Smith, 1999). In the critical point view, a large earthquake plays a special role and signals the end of a cycle on its fault network (see, e.g. Sornette and Helmstetter, 2002).

Singularities such as phase transitions characterise critical phenomena and critical points. The simplest representative dynamic evolution equation leading to a

finite-time singularity is

$$\frac{dE}{dt} = E^m, \quad \text{with } m > 1. \tag{3.2}$$

Its solution is

$$E(t) = E(0)\left(\frac{t_c - t}{t_c}\right)^{-1/(m-1)} \tag{3.3}$$

where the critical time $t_c = (m-1)/[E(0)]^{m-1}$ is determined by the initial condition $E(0)$. This kind of equation can be used for example in crack growth models, damage models or percolation models (see Sammis and Sornette, 2002, for more details).

3.1.2. Accelerating Moment Release as a Critical Phenomenon

In a self-organised criticality system, accelerating moment release is thought to be the consequence of the growth of the spatial correlation length. A small rupture can only grow into a large earthquake if a critical value of the spatial correlation length of the stress is reached. In the preparatory process before a mainshock, long-range correlations are established by the redistribution of stress from smaller to larger scales. In other words, small events cascade into larger ones to finally lead to the largest one (mainshock).

Some studies explain the accelerating moment release power-law using cascade triggering (e.g. Sornette and Helmstetter, 2002) or change of scaling in hierarchical models (e.g. Sornette and Sammis, 1995; Saleur et al., 1996). Other studies explain the origin of accelerating moment release using rock fracture mechanics, such as fiber-bundle models (e.g. Turcotte et al., 2003) or continuum damage models (e.g. Ben-Zion and Lyakhovsky, 2002).

Following existing theories and models based on self-organised criticality, an important number of studies that search for accelerating moment release in seismicity catalogues assume that these seismicity patterns are due to critical processes (e.g. Bowman et al., 1998; Zoller et al., 2001; Papazachos et al., 2002a). Some other studies test the critical point hypothesis by searching for indicators in earthquake populations, such as the growth of the spatial correlation length (Zoller and Hainzl, 2002) or seismic quiescence (Hainzl et al., 2000).

3.1.3. Previous Formulations of Accelerating Moment Release

Epidemic-Type Aftershock Sequence model: Sornette and Helmstetter (2002) show that for certain conditions, the rate of seismic activity in the Epidemic-Type Aftershock Sequence model diverges in finite time according to

$$N(t) \propto 1/(t_c - t)^m \tag{3.4}$$

where $N(t)$ is the seismicity rate. These specific conditions are obtained when the system is "supercritical" (as opposed to subcritical) (Helmstetter and Sornette, 2002). In that case, the average rate of seismicity $N(t)$ is no longer linear due to the role of large earthquakes. The nonlinearity is expressed as a positive feedback and self-amplifies leading to the finite-time singularity of Eq. (3.4).

This can be verified by stacking ETAS-derived synthetic catalogues to estimate the probability of events before "mainshocks". By relaxing the requirement that aftershocks are smaller than mainshocks, "foreshocks" can be produced. Because of the Omori decay (proportional to $1/t$), "foreshocks" are more likely to occur immediately before an event in a ETAS system. In other words, foreshock statistics and aftershock statistics are symmetric about the time of the mainshock, the only differences are that aftershocks have a much higher a-value than foreshocks and that foreshocks follow a power-law acceleration (i.e. accelerating moment release) whereas aftershocks follow a power-law deceleration. For more details, see Sornette and Helmstetter (2002), Helmstetter and Sornette (2002), Helmstetter et al. (2003), Helmstetter (2003).

Hierarchical models and Renormalisation Group Theory: Sornette and Sammis (1995) and Saleur et al. (1996) proposed that accelerating moment release can be formulated using the Renormalisation Group Theory, viewing accelerating moment release as a critical phenomenon in a hierarchical fault system. This theory is the transformation of a set of equations from one scale to another, by changing of variables. Sornette and Sammis (1995) postulated a one-parameter (time to failure) discrete renormalisation group, expressing how the (cumulative) seismic release ϵ at given time t can be related to that at another time t'. The times t and t' have a relation of the form $x' = \phi(x)$ where $x = t_c - t$, t_c is the failure time, and ϕ is called the renormalisation group flow map. Then

$$f(x) = g(x) + \frac{1}{\mu} f\big(\phi(x)\big) \tag{3.5}$$

where $f(x) = \epsilon$ and μ is a scaling constant. $g(x)$ represents the nonsingular part of the function $f(x)$ (see Saleur et al., 1996, for details). Saleur et al. (1996) show that the critical point x^* is the time at which $f(x)$ becomes singular and that a necessary condition is $\phi(x^*) = x^*$. Using the conditions $\phi(0) = 0$ and $\phi(x) = \lambda x + \cdots$ (linearised transformation), the solution of Eq. (3.5) close to $x = 0$ is $f(x) \sim x^m$ with $m = \log \mu / \log \lambda$. Equation (1.9) can be derived from this last equation.

Sornette and Sammis (1995) add an imaginary part to the exponent m and find

$$\epsilon(t) = A + B(t_f - t)^m \left[1 + C \cos\left(2\pi \frac{\log(t_f - t)}{\log \lambda} + \phi\right)\right] \tag{3.6}$$

with the term added to Eq. (1.9) being the logarithm periodic correction for the change of scale associated with the incoming mainshock. Several other studies

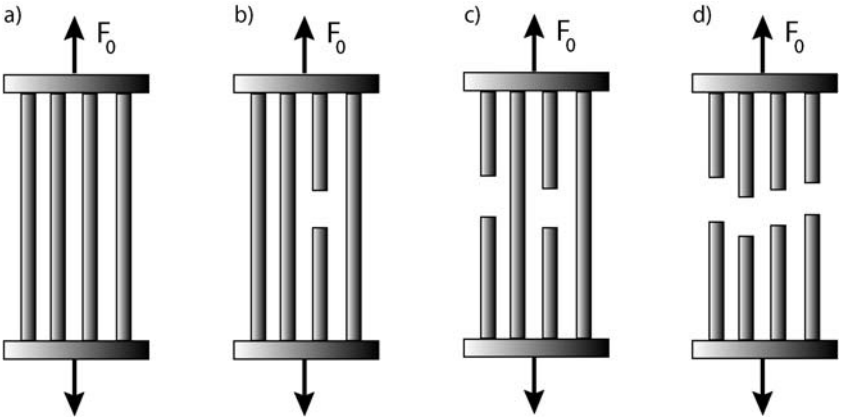


FIG. 3.2. Schematic illustration of the failure of a fiber-bundle with uniform load sharing; (a) Each of the fibers carries one-quarter of the load F_0; (b) One fiber has failed and each remaining fiber carries one-third of the load F_0; (c) Two fibers have failed and each remaining fiber carries one-half of the load F_0; (d) All four fibers have failed and no load is carried. Redrawn after Turcotte et al. (2003).

associate hierarchical models of faults with the Renormalisation Group Theory to identify accelerating moment release (e.g. Huang et al., 1998). However Eq. (3.6) has not proved to be a more accurate predictor of accelerating moment release than Eq. (1.9) (Gross and Rundle, 1998).

Fiber-bundle models: An approach to brittle failure is the fiber-bundle model (e.g. Newman and Phoenix, 2001) which takes into account a composite material made up of strong fiber embedded in a relatively weak matrix. The failure statistics of the individual fibers that make up the fiber-bundle are specified and can be either static or dynamic. In general, experiments favour dynamic-failure fiber-bundle models (Turcotte et al., 2003). When stress is applied to a fiber-bundle, the fibers begin to fail. The fiber-bundle model can also be used as a simple model for friction where the fibers represent the asperities on a surface and thus can be applied to fault rupture.

The failure of a simple fiber-bundle under uniform load sharing is illustrated in Fig. 3.2. Initially the load on the bundle F_0 is carried equally by the four fibers with $F = \frac{1}{4}F_0$. The weakest fiber fails and the load on that fiber is now carried by the surviving fibers with $F = \frac{1}{3}F_0$. The stress on each fiber increases from the original value σ_0 to $\frac{4}{3}\sigma_0$. The process of failure followed by stress redistribution continues until all fibers fail and no load can be carried.

As fibers fail, the stress on the remaining fibers increases leading to a catastrophic failure of the bundle and are equivalent to the microcracks that occur in a uniform brittle material as it is stressed to failure. A characteristic of material failure are the emergence of acoustic emission events, generated by microcracks

(Scholz, 1998). The microscopic fiber-bundle model can be used to obtain the predicted rate of the acoustic emission events which is a power-law time-to-failure equation (Turcotte et al., 2003). First it is necessary to prescribe the dependence of the hazard rate ν on the stress σ. For engineering materials it is standard practice (Newman and Phoenix, 2001) to empirically assume the power-law relation

$$\nu(t) = \nu_0 \left[\frac{\sigma(t)}{\sigma_0} \right]^{\rho} \tag{3.7}$$

where ν_0 is the hazard rate corresponding to the initial fiber stress σ_0 (values of ρ found experimentally in a range 2 to 5).

Equation (3.8) is the rate at which energy is radiated in acoustic emission events during the failure of a fiber-bundle

$$\frac{de_a(t)}{dt} = \frac{\eta_a \nu_0 \sigma_0^2}{2E_0} \frac{1}{(1 - t/t_f)^{(\rho+1)/\rho}} \tag{3.8}$$

where e_a is the energy associated with the acoustic emission event, η_a the acoustic emission efficiency (i.e. fraction of lost energy), ν_0 the hazard rate corresponding to the initial fiber stress σ_0, E_0 the Young's modulus of fibers and t_f the failure time. Details of the demonstration can be found in Turcotte et al. (2003).

By analogy between the energy associated with the acoustic emission event and the seismic energy release by a seismic event, Turcotte et al. (2003) find

$$\epsilon_B(t) = \epsilon_{Bf} - \frac{\sigma_0(\rho - 1)}{2\rho} \sqrt{\frac{\eta_a \nu_0}{2E_0}} \left(1 - \frac{t}{t_f} \right)^{(\rho-1)/(2\rho)} \tag{3.9}$$

where ϵ_B is the cumulative Benioff strain and ϵ_{Bf} the Benioff strain when the large earthquake occurs. Equation (3.9) is used in some AMR studies (Ben-Zion and Lyakhovsky, 2002; Turcotte et al., 2003).

Continuum damage models: The anelastic deformation of solids is often treated using continuum damage mechanics. Turcotte et al. (2003) show that continuum damage models can give exactly the same solution for material failure as the fiber-bundle model. The power-law scaling obtained by both models is consistent with the power-law seismic activity observed prior to some earthquakes.

In linear elasticity, $\sigma = E_0 \epsilon$ where E_0 is the Young's modulus of the undamaged material, a constant. Once microcracking is occurring, there is a deviation from linear elasticity. These microcracks weaken the material and for a given stress σ, the strain ϵ is greater than the value predicted by linear elasticity. The relation becomes $\sigma = E_{\text{eff}} \epsilon$ where E_{eff} is the effective Young's modulus, no longer assumed to be a constant. A continuum approach to this process is to introduce a damage variable α (e.g. Ben-Zion and Lyakhovsky, 2002) so that

$$E_{\text{eff}} = E_0(1 - \alpha). \tag{3.10}$$

The damage variable α quantifies the deviation from linear elasticity. In general $0 \leqslant \alpha \leqslant 1$. With $\alpha = 0$, linear elasticity is obtained but when $\alpha = 1$, failure occurs.

Using the relation (see details in Lyakhovsky et al., 1997)

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = C\epsilon^2 \tag{3.11}$$

(Ben-Zion and Lyakhovsky, 2002) integrate Eq. (3.10), which gives

$$\alpha = 1 - \left[1 - \left(3C\sigma^2/E_0^2\right)t\right]^{1/3}. \tag{3.12}$$

Substituting Eq. (3.12) back into Eq. (3.10) leads to strain accumulation in the power law form

$$\epsilon = \sigma/E_0\left[1 - \left(3C\sigma^2/E_0^2\right)t\right]^{-1/3}. \tag{3.13}$$

Using $t_f = E_0^2/3C\sigma^2$ in Eq. (3.13), defined by setting $\alpha = 1$ in Eq. (3.12), and changing constants gives

$$\epsilon(t) = \sigma/E_0(\Delta t/t_f)^{-1/3} \tag{3.14}$$

with $\Delta t = t_f - t$. Finally, Ben-Zion and Lyakhovsky (2002) obtain for the cumulative Benioff strain the form

$$\sum M_0^{1/2}(t) = A_1 + A_2 t + A_3(\Delta t/t)^{1/3}. \tag{3.15}$$

Equation (3.15) is equivalent to Eq. (1.9) with $m = 1/3$ with the background seismicity taken into account (i.e. the linear part of the equation).

3.2. Origin of Accelerating Moment Release from Stress Loading

The previous formulations of accelerating moment release have all been proposed based on the idea that accelerating moment release corresponds to some cascade triggering leading to the mainshock. However, another formulation based on the Stress Accumulation model can be proposed where accelerating moment release is formed by a different process.

3.2.1. Introduction

Here a new mathematical formulation of accelerating moment release based on the Stress Accumulation Model is reported, following Mignan *et al.* (2007): Large earthquakes can be preceded by a period of accelerating seismic activity of moderate-sized earthquakes. This phenomenon, usually termed accelerating moment release, has yet to be clearly understood. A new mathematical formulation of accelerating moment release is obtained from simple stress transfer considerations, following the recently proposed Stress Accumulation model. This model,

based on the concept of elastic rebound, simulates accelerating seismicity from theoretical stress changes during an idealised seismic cycle. In this view, accelerating moment release is simply the consequence of the decrease, due to loading, of the size of a stress shadow due to a previous earthquake. We show that a power-law time-to-failure equation can be expressed as a function of the loading rate on the fault that is going to rupture. We also show that the m-value which is the power-law exponent can be defined as $m = D/3$ with D a parameter that takes into account the geometrical shape of the stress lobes and the distribution of active faults. In the Stress Accumulation model, the power-law is not due to critical processes.

Accelerating moment release has been identified for a substantial number of earthquakes and observed for years to tens of years before a mainshock over tens to hundreds of kilometres from the future epicentre (e.g. Sykes and Jaumé, 1990; Bufe and Varnes, 1993; Knopoff et al., 1996; Bowman et al., 1998; Brehm and Braile, 1998; Jaumé and Sykes, 1999; Robinson, 2000; Bowman and King, 2001a; Zoller et al., 2001; Papazachos et al., 2002a; Bowman and Sammis, 2004). Although accelerating moment release might be fitted by any power-law, it is typically modelled by a simple power-law time-to-failure equation, following Bufe and Varnes (1993). This is a relation of the form

$$\epsilon(t) = A + B(t_f - t)^m \tag{3.16}$$

where t_f is the time of the large event, B is negative and m is usually about 0.3. A is the value of $\epsilon(t)$ when $t = t_f$. For a convenient data analysis, $\epsilon(t)$ is chosen to be the cumulative Benioff strain at time t and is defined as

$$\epsilon(t) = \sum_{i=1}^{N(t)} \sqrt{E_i(t)} \tag{3.17}$$

where E_i is the energy of the ith event and $N(t)$ is the number of events at time t. The cumulative Benioff strain is preferred over the cumulative number of earthquakes because the multitude of smaller earthquakes would dominate the data, whereas the moderate to large events would dominate if the cumulative seismic energy alone was used. Therefore the choice of Eq. (3.17) has no physical meaning and is only an observational tool. In this work, we keep the term "accelerating moment release" as it is referred in numerous publications. Equation (3.16) is equivalent to the rate-dependent failure equation of Voight (1989) as noted by Bufe and Varnes (1993) but the physical relation of this expression to accelerating moment release is not verified and other power-laws based on different physical processes can fit the pre-event seismicity rate changes.

Since accelerating seismic activity is a promising tool for earthquake forecasting, it is important to determine the origin of accelerating moment release. The majority of studies consider that accelerating moment release is due to critical processes (i.e. self-organised criticality or critical point theory). Studies which

give an explanation to the power-law behaviour of accelerating moment release are based on different critical concepts, such as the Epidemic-Type Aftershock Sequence Model (e.g. Sornette and Helmstetter, 2002), the Renormalisation Group theory for hierarchical systems (e.g. Sornette and Sammis, 1995; Saleur et al., 1996), fiber-bundle models (e.g. Newman and Phoenix, 2001; Turcotte et al., 2003), continuum damage mechanics (e.g. Ben-Zion and Lyakhovsky, 2002) or percolation models (e.g. Sammis and Sornette, 2002).

The Stress Accumulation model recently proposed by King and Bowman (2003) also explains accelerating moment release. In this model, accelerating moment release is due to the decrease of the size of a stress shadow left from one or more previous events. Consequently the main cause of increasing seismicity is loading by creep at depth on the fault that is going to fail. This view is not the same as other explanations based on critical processes where the acceleration is due to cascade triggering. Moreover, Mignan et al. (2006b) showed that the spatial distribution of accelerating moment release is in agreement with the predictions of the Stress Accumulation but not with stress triggering. The purpose of this work is to determine a mathematical formulation in agreement with Eq. (3.16) based on the principles of the Stress Accumulation model, which could lead to a better understanding of how to use accelerating moment release for forecasting.

3.2.2. Mathematical Formulation

In the Stress Accumulation model (King and Bowman, 2003), events occur at a constant rate (background seismicity) outside the stress shadow formed by the last mainshock. In the simplest form of the model, the stress is not sufficient for failure in the stress shadow and corresponds to a region of quiescence. The central point of the model is that aseismic slip on the fault at depth loads the upper part of the crust slowly removing the stress shadow prior to the next major event (Fig. 3.3). King and Bowman (2003) add a stress noise to the stress associated to the seismic cycle to simulate background seismicity. A new event occurs when the stress σ exceeds the failure stress σ_f.

Spatiotemporal evolution of a stress shadow: Let us consider the spatiotemporal evolution at the surface of a stress shadow during the seismic cycle of a theoretical fault (simplified to a point source at a given depth).

The time evolution of the stress σ at $r = 0$ (epicenter) is

$$\sigma(r = 0, t) = \sigma_0 + \frac{\mathrm{d}\sigma_l}{\mathrm{d}t} t, \qquad \sigma_0 = \sigma(r = 0, t = 0) \tag{3.18}$$

where σ_0 is the stress drop from the last mainshock (< 0, value at the surface), $\mathrm{d}\sigma_l/\mathrm{d}t$ is the loading rate and t is the time from the last mainshock. The loading rate is considered constant (Fig. 3.3) on a fault in a purely elastic medium. This idealised fault is localised and extends through both the crust and upper mantle, which are both assumed to retain long-term strength. Over the timescale of

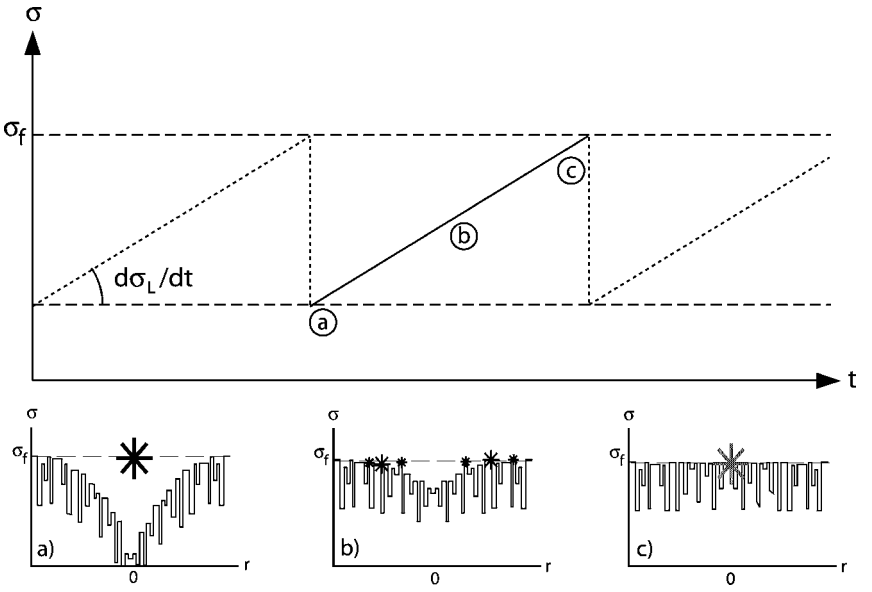


FIG. 3.3. Schematic representation of the spatiotemporal evolution of regional seismicity, at the origin of accelerating moment release, during the seismic cycle of a given fault. The loading rate $d\sigma_l/dt$ is constant on the fault and mainshocks are periodic with a constant stress drop (top). A background stress noise is added to simulate the background seismicity in space and time (bottom); (a) Stress shadow after the mainshock; (b) Decrease of the size of the stress shadow; (c) Stress shadow completely filled prior to the next mainshock. Large stars represent mainshocks and small black stars represent pre-mainshock seismicity. Pre-events occur when the stress σ exceeds the failure stress σ_f. Modified from King and Bowman (2003).

the seismic cycle, changes of the loading rate due to viscous relaxation in the lower crust or mantle are not required by either geological or recent geodetic data (see discussion in King and Bowman, 2003). Other processes such as time-dependent fault healing are not taken into account. Possible changes of the loading rate following a major earthquake are considered negligible since only precursory seismicity to the next major earthquake is studied.

The spatial evolution of the stress σ from the epicenter to r is given as

$$\sigma(r) = B + \frac{A}{(r^2 + h^2)^{3/2}} \tag{3.19}$$

where h is the depth of the point source and A and B are defined using limit conditions. Equation (3.19) is obtained from the simple relation

$$\sigma(R) \propto \frac{1}{R^3} \tag{3.20}$$

where $R = \sqrt{r^2 + h^2}$. Although Eq. (3.20) is an over-simplification, it permits a dimensional study of the evolution of stresses in space and is in agreement with the conceptual aspect of this demonstration which is based on a simple point source. Note that our demonstration is done for a tridimensional medium but because $r \gg h$ in practice, only stress evolutions on a horizontal plane will be shown.

Let us consider the following conditions

$$\begin{cases} \sigma(r \to \infty) = \sigma_f, \\ \sigma(r = 0) = \sigma(r = 0, t) \end{cases} \tag{3.21}$$

which mean that the loading only affects the stress shadow region through time whereas more distant stress is constant, tending to the failure stress σ_f (i.e. background stress region). Using the conditions from Eq. (3.21) in Eqs. (3.18) and (3.19), we find

$$\sigma(r, t) = \frac{h^3}{(r^2 + h^2)^{3/2}} \left(\sigma_0 - \sigma_f + \frac{\mathrm{d}\sigma_l}{\mathrm{d}t} t \right) + \sigma_f. \tag{3.22}$$

Equation (3.22) gives the evolution at the surface of the stress shadow ($\sigma < 0$) as a function of the distance from the source point (i.e. epicenter) and of the time from the last mainshock ($t_0 = 0$) to the next one (t_f) and is shown in Fig. 3.4. Note that except close to the source, the evolution is the same at any depth that is small compared to r.

Background stress noise: We now consider the background stress noise $\Delta\sigma_b$ as follows

$$\Delta\sigma_b = \sigma_f - \sigma^* \tag{3.23}$$

where σ^* is approaching σ_f. The peaks of the background stress noise (defined as each different value of σ^* in space) are random uniform in space and amplitude. The background stress noise is considered random to determine the only effects of the spatiotemporal evolution of the stress shadow on accelerating moment release. The use of a more complex but more realistic background stress field would only obscure the result of this demonstration.

Fluctuations of the background stress noise $\Delta\sigma_b$ must be such that the events that result from relaxing stress peaks ($\sigma^* \to \sigma_f$) result in a set of events obeying the Gutenberg–Richter frequency-magnitude relation (Fig. 3.5). The appropriate scaling law can be implemented in various ways as explained by King and Bowman (2003) and the easiest one is to adjust the size of the stress peaks and earthquake magnitude for any irregular function. A simple random stress distribution σ^* can be created by assigning random numbers to each element of a matrix. All points of the matrix that equal or exceed the failure stress σ_f are associated with an earthquake. The areas A of these regions are calculated to provide a measure of events size. These areas are then scaled such that the resulting value is

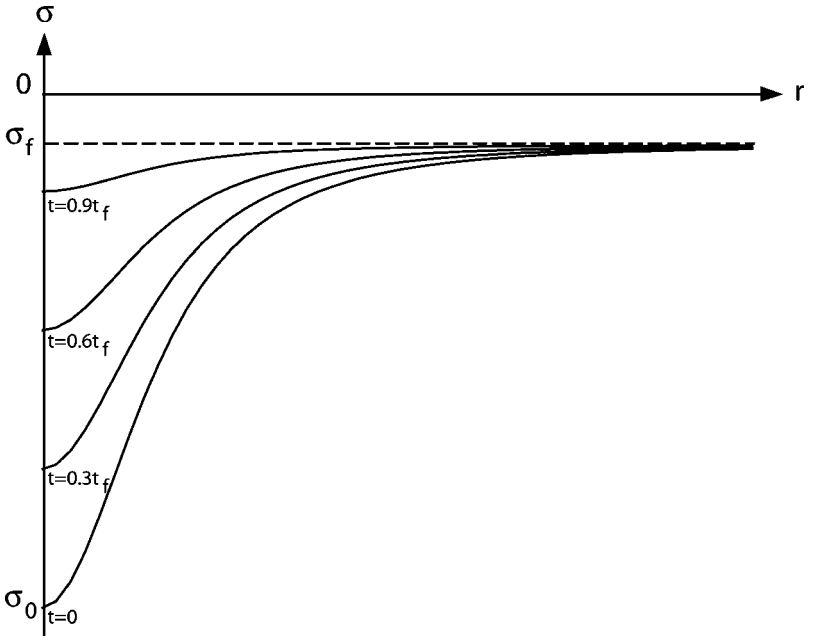


FIG. 3.4. Evolution at the surface of the stress shadow ($\sigma < 0$) as a function of the distance from the source point (i.e. epicenter) for different times t to the final time t_f: $t = 0$, $t = 0.3t_f$, $t = 0.6t_f$, and $t = 0.9t_f$. Far from the stress shadow, the background stress is constant and tends to the failure stress σ_f. Note that except close to the source, the evolution is the same at any depth that is small compared to r.

equivalent to an earthquake of magnitude M ($M = A^s$, with s a scaling parameter which depends of the range of areas A and of the range of magnitudes M chosen). With the appropriate scaling coefficients, this provides a Gutenberg–Richter relation with a b-value of 1.

We then define the region of background stress as the region where the stress shadow has no influence, at $r > r^*$. Also, we define a constant background stress noise that corresponds to the maximum amplitude $\sigma_f - \sigma^*_{\min}$. In the following demonstration, we will consider the stress noise $\Delta\sigma_b$ constant.

Decrease of the size of the stress shadow: We define the following conditions

$$\begin{cases} \sigma(r^*, t) = \sigma^*, \\ \sigma(r = 0, t_f) = \sigma_f. \end{cases} \tag{3.24}$$

The first condition from Eq. (3.24) means that the spatial limit of the stress shadow is at r^* when the stress σ is indistinguishable from the background stress σ^*.

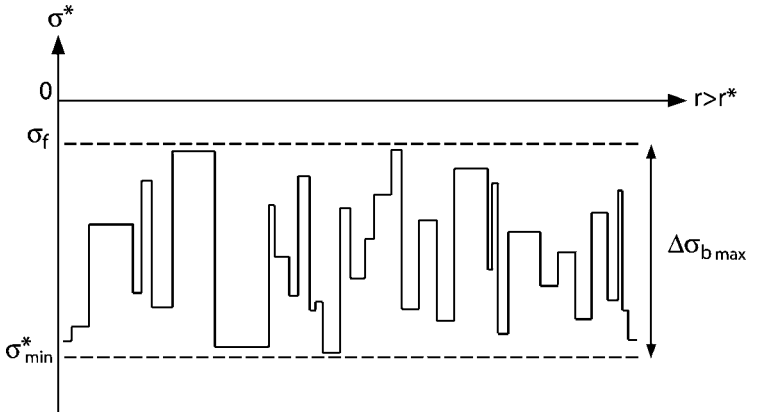


FIG. 3.5. Background stress noise $\Delta\sigma_b$. Stress fluctuations σ^* are approaching the failure stress σ_f. The peaks of the background stress noise (defined as each different value of σ^* in space) are random uniform in space and amplitude. The stress distribution is such that the events that result from relaxing stress peaks ($\sigma^* \to \sigma_f$) result in a set of events obeying the Gutenberg–Richter frequency-magnitude relation, following King and Bowman (2003) (see text for details). Note that $r > r^*$ corresponds to a distance from the source point (mainshock) where the associated stress shadow has no influence on the background stress.

Then using conditions from Eq. (3.24) in Eq. (3.22), we find

$$\begin{cases} \dfrac{h^3}{(r^{*2}+h^2)^{3/2}} A\left(\sigma_0 - \sigma_f + \dfrac{\mathrm{d}\sigma_l}{\mathrm{d}t} t\right) + \sigma_f = \sigma^*, \\ \sigma_0 + \frac{\mathrm{d}\sigma_l}{\mathrm{d}t} t_f = \sigma_f. \end{cases} \tag{3.25}$$

By substitution of σ_0, we obtain

$$\frac{h^3}{(r^{*2}+h^2)^{3/2}} \left(\frac{\mathrm{d}\sigma_l}{\mathrm{d}t}(t_f - t)\right) = \Delta\sigma_b \tag{3.26}$$

which leads to

$$r^*(t) = h\left[\left(\frac{\mathrm{d}\sigma_l}{\mathrm{d}t}\frac{(t_f - t)}{\Delta\sigma_b}\right)^{2/3} - 1\right]^{1/2}. \tag{3.27}$$

Equation (3.27) gives the decrease of the size of the stress shadow as a function of time t from $t = 0$ to t_f and is represented in Fig. 3.6. For a point at a distance from the source point $r < r^*$, it is located in the stress shadow ($\sigma(r, t) < \sigma^*$). For a point at a distance from the source point $r \geqslant r^*$, it is located outside the stress shadow, in the background stress noise ($\sigma(r, t) = \sigma^*$). It is important to note that the stress shadow does not disappear at the time t_f but at

$$t_{f'} = t_f - \Delta t, \quad \Delta t = \frac{\Delta\sigma_b}{\mathrm{d}\sigma_l/\mathrm{d}t} \tag{3.28}$$

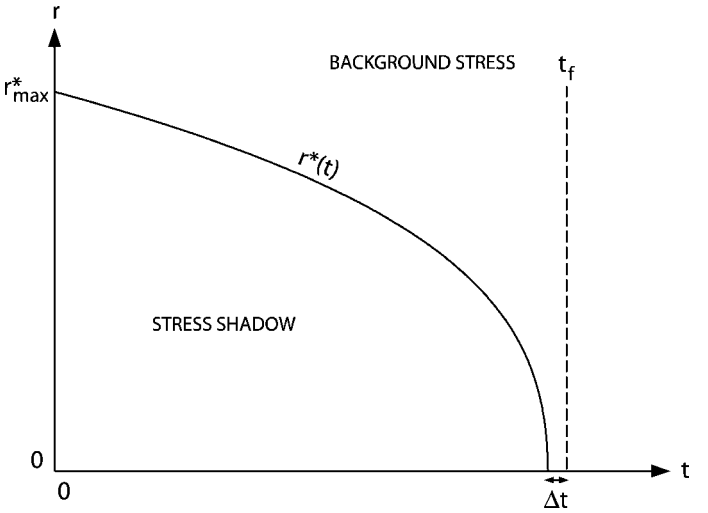


FIG. 3.6. Decrease of the size of the stress shadow r^* as a function of time t from $t = 0$ to t_f. For a point at a distance from the source point $r < r^*$, it is located in the stress shadow ($\sigma(r, t) < \sigma^*$). For a point at a distance from the source point $r \geqslant r^*$, it is located outside the stress shadow, in the background stress noise ($\sigma(r, t) = \sigma^*$). It is important to note that the stress shadow does not disappear at the mainshock time t_f but at $t_f - \Delta t$ (see text for details).

when the stress shadow is undistinguishable from the background stress σ^* at $r^* = 0$. Therefore the mainshock can occur between $t_{f'}$ and t_f due to the stress noise perturbation (with $t_{f'} < t_f$).

Time-to-failure power-law equation: We now use Eq. (3.27) to describe the evolution of seismicity λ as a function of time t.

We first consider the simple form

$$\lambda(t) = c_\lambda A(t) \tag{3.29}$$

with A the area where events occur as a function of time t and c_λ a constant that corresponds to the density of events per unit area (i.e. background density).

In the Stress Accumulation model (King and Bowman, 2003), accelerating moment release corresponds to events that occur where the stress shadow is filled by loading at depth ($\mathrm{d}\sigma_l/\mathrm{d}t$). Thus we can define the evolution of accelerating moment release λ_{AMR} through time t as

$$\lambda_{\mathrm{AMR}}(t) = c_\lambda \big[A_{\mathrm{sh}}(t_0) - A_{\mathrm{sh}}(t)\big], \quad t_0 = 0 \tag{3.30}$$

with A_{sh} the surface of the stress shadow and $A_{\mathrm{sh}}(t_0)$ the initial surface of the stress shadow (when $r = r^*_{\mathrm{max}}$). We can also define the evolution of the background seismicity λ_b where the stress shadow has no influence (where $r \geqslant r^*$)

as

$$\lambda_b(t) = c_\lambda A_b \delta_b t \tag{3.31}$$

with A_b a pre-defined surface (minus the stress shadow surface $A_{\mathrm{sh}}(t_0)$) and δ_b the background seismicity rate per unit area, assuming that the background seismicity is constant through time.

We now introduce Eq. (3.27) in

$$A_{\mathrm{sh}}(t) = k_A r^{*D}(t) \tag{3.32}$$

with k_A a constant that is linked to the geometrical shape of the surface A_{sh} and D the dimension of the geometrical shape. In a simple case where the stress shadow is circular and seismicity uniformly distributed in space, parameters of Eq. (3.32) would be $k_A = \pi$ and $D = 2$. However, the geometrical shape of stress lobes cannot be simply determined analytically only numerically. Also, seismicity is not uniformly distributed in space but is self-similar due to the fact that events can be associated with fault systems. This must be taken into account in Eq. (3.32) since the area A_{sh} is assumed to correspond to the spatial distribution of seismicity. Thus D is linked to the shape of the stress lobes and to the distribution of active faults. For an homogeneous distribution of events, D only depends of the geometrical shape of the stress shadow. By analogy to simple geometrical patterns such as a straight line ($D = 1$) and a circle ($D = 2$), a stress lobe corresponds to $1 \leqslant D \leqslant 2$ because a change of its size is not homogeneous in all directions of space (Fig. 3.7a). The spatial distribution of faults can be characterised (in theory at least) by an appropriate spatial fractal. In this case, a fractal generator with a dimension of 1 is equivalent to a straight line, a generator with a dimension of 2 can completely fill an area, and a generator with a dimension of 3 can fill space. For two-dimensional faulting, the surface will be partially filled with $1 \leqslant D \leqslant 2$ (e.g. King, 1983, Fig. 3.7b). Thus D corresponds to the lowest value obtained from the geometrical shape or from the fractal dimension of the fault network. For example, if the geometrical shape is a circle, D should be equal to 2 but if seismicity occurs on a fault network of fractal dimension 1, D will be equal to 1 (Fig. 3.7c).

With these geometrical considerations in mind, we can now study the shape of the accelerating moment release time-to-failure equation predicted by the Stress Accumulation model. From Eqs. (3.27), (3.30) and (3.32), we find

$$\lambda_{\mathrm{AMR}} \propto -\left(\frac{\mathrm{d}\sigma_l}{\mathrm{d}t}\frac{t_f - t}{\Delta\sigma_b}\right)^{D/3}, \quad 1 \leqslant D \leqslant 2. \tag{3.33}$$

By analogy with the accelerating moment release power-law time-to-failure Eq. (3.16), we obtain

$$\frac{1}{3} \leqslant m \leqslant \frac{2}{3}. \tag{3.34}$$

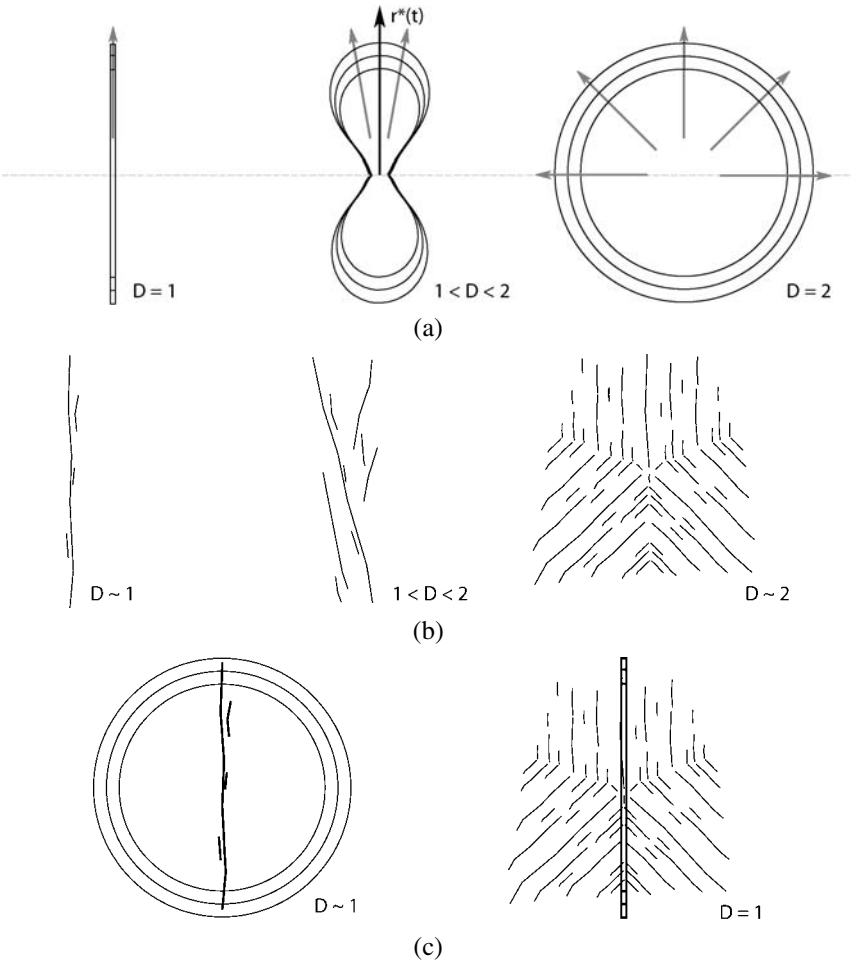


FIG. 3.7. Variability of the parameter D for different geometrical shapes of the stress shadow and for different distributions of active faults. These two criteria must be taken into account to determine the spatial distribution of the theoretical precursory seismicity; (a) Geometrical shape. $D = 1$ for a straight line, $D = 2$ for a circle and $1 \leqslant D \leqslant 2$ for a stress lobe. This can be explained by the fact that a change of the surface of a stress lobe is not homogeneous in all directions of space. Grey arrows show the directions of extension of the geometrical shapes; (b) Fractal dimension of the fault network. $D = 1$ for a straight line fault, $D = 2$ for a surface completely filled by faults, $1 \leqslant D \leqslant 2$ for a surface partially filled by faults; (c) D corresponds to the lowest value from the geometrical shape or from the fractal dimension.

This result is in agreement with m-values found for real data (Table 3.1) and is represented in Fig. 3.8 based on

TABLE 3.1. List of m-values determined from the study of accelerating moment release in different regions

m-value	Region	Reference
0.30 (0.10–0.39)	Northern California	Bufe and Varnes (1993)
0.26 (0.10–0.55)	California	Bowman et al. (1998)
0.24 (0.12–0.47)	New Madrid Seismic Zone	Brehm and Braile (1998)
0.47 (0.27–0.70)	Aegean region	Papazachos and Papazachos (2000)
0.36 (0.29–0.46)	New Zealand	Robinson (2000)
0.34 (0.29–0.40)	Northwest Anatolia	Papazachos et al. (2002a)

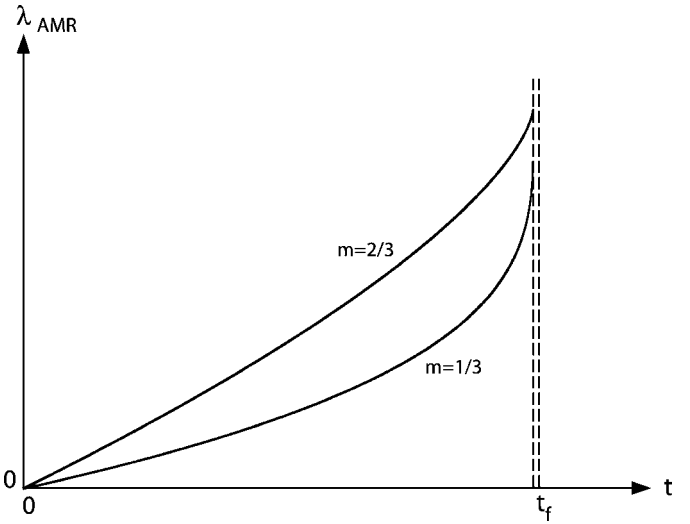


FIG. 3.8. Possible shapes for the time-to-failure power-law equation defined from the Stress Accumulation model. For two-dimensional faulting, $1/3 \leqslant m \leqslant 2/3$. Note that the two curves have been re-scaled for display reasons.

$$\lambda_{\mathrm{AMR}}(t) = c_\lambda k_A h^D \left[\left(\left(\frac{\mathrm{d}\sigma_l}{\mathrm{d}t} \frac{t_f}{\Delta\sigma_b} \right)^{2/3} - 1 \right)^{D/2} - \left(\left(\frac{\mathrm{d}\sigma_l}{\mathrm{d}t} \frac{t_f - t}{\Delta\sigma_b} \right)^{2/3} - 1 \right)^{D/2} \right] \tag{3.35}$$

which is the complete formulation of Eq. (3.33) that corresponds to the time-to-failure power-law equation defined from the Stress Accumulation model.

3.2.3. Discussion

A non-critical origin of the time-to-failure power-law: Other authors have considered that the Stress Accumulation model (King and Bowman, 2003) is a critical point process (e.g. Sammis and Sornette, 2002; Tiampo and Anghel, 2006). This is not the case (see below) although the Stress Accumulation model could be parametrised as a class of percolation critical point models if it is treated as a driven threshold with a mean field approximation (Sammis and Sornette, 2002). However this does not add to our understanding of the mechanical processes.

The fact that the Stress Accumulation model is not based on critical processes can be demonstrated by the following arguments:

(1) Loading processes are due to stable sliding at depth during the seismic cycle of a given fault.
(2) The mainshock is due to loading processes and not to pre-event seismicity, distant from the fault. In other words, the mainshock is not the final result of a cascading phenomenon, but the processes originally described by Reid as Elastic Rebound (Reid, 1910).
(3) Events that compose AMR are due to loading processes. Whereas aftershocks relax the excess stresses generated by a mainshock, AMR events relax excess stresses generated by loading. As a consequence, they delay (slightly) the time of occurrence of the future mainshock. This is opposite to the concept of cascade triggering from the smallest events to the largest one.
(4) The Gutenberg–Richter law, often considered as a signature of critical phenomena (e.g. Bak and Tang, 1989; Sornette et al., 1990; Blanter et al., 1997), is not a necessary condition for the emergence of a time-to-failure power-law pattern.

In our study, we do not claim that critical processes do not exist. Loading processes (Stress Accumulation model) and stress triggering (e.g. Epidemic-Type Aftershock Sequence model) should be seen as independent processes that may superimpose their effects but stress triggering is not the cause of AMR (Mignan et al., 2006b). It is also important to note that we do not consider that all critical processes involve cascade triggering. However other critical models such as percolation models (e.g. Sammis and Sornette, 2002), as explained earlier, do not add to our understanding of the mechanical processes. Rupture processes on the fault are commonly considered to be due to critical phenomena (e.g. Allegre et al., 1982) but the Stress Accumulation model is not concerned with these processes which are therefore not discussed here.

Acceleration of the number of events versus accelerating moment release: It can be noted that the quantity $\lambda_{\mathrm{AMR}}(t)$ corresponds to the evolution of the cumulative number of events through time whereas accelerating moment release is defined by

the evolution of the cumulative Benioff strain through time (Eq. (3.16)). However, the choice of the cumulative Benioff strain to fit accelerating seismicity has never been shown to be the most appropriate parameter, only a convenient observational tool (as explained previously).

In the Stress Accumulation model, the time power-law is due to the decrease of the size of the stress shadow which is similar to an increase of the size of the region of background seismicity. By consequence, the studied parameter $\lambda_{\mathrm{AMR}}(t)$ is linked to the density of events per unit of space and not to their magnitude. This is in agreement with the fact that accelerating moment release usually corresponds to an increase of the a-value and not of the b-value (e.g. Sammis et al., 2004).

We now add to Eq. (3.35) an appropriate noise to simulate the Gutenberg–Richter statistics to compare the power-law time-to-failure equation with real data. Figure 3.9 shows two examples of seismicity distribution (random uniform in space, random in magnitude in respect to the Gutenberg–Richter law). The first one corresponds to the expected seismicity evolution in the region of accelerating seismicity (power-law) following Eq. (3.35) (Fig. 3.9a–c) and the second one corresponds to an association of accelerating seismicity (power-law) following Eq. (3.35) and of background seismicity (linear-law) following Eq. (3.31) (Fig. 3.9d–f). This second example simulates a case where accelerating seismicity is searched in such a way that the optimised region contains events that are part of the acceleration but also some independent events from the background, where $r > r^*_{\mathrm{max}}$ (see Fig. 3.6).

Figure 3.9 also shows the variability of the power-law time-to-failure curve for different criteria (cumulative number of events, cumulative Benioff strain and cumulative seismic energy). From our formulation, it is evident that the best acceleration is found for the cumulative number of events (Fig. 3.9a). Acceleration is also clear for the cumulative Benioff strain and accelerating moment release simulated using Eq. (3.35) is similar to real patterns (Fig. 3.9b). We can note that no acceleration is visible for the cumulative seismic energy due to the role of the largest events of the fractal noise (Fig. 3.9c). In the case that a linear seismicity contribution is added, depending on the random noise, accelerating moment release can become difficult to observe (Fig. 3.9e).

Duration of accelerating seismicity: The mathematical formulation of accelerating seismicity shows that acceleration should be observed throughout the seismic cycle. However, if the signal is perturbed by background noise, the acceleration appears later during the seismic cycle, as observed in simulations by King and Bowman (2003) and in real data by Bowman and Sammis (2004). This phenomenon is illustrated schematically in Fig. 3.10 and can explain why no correlation

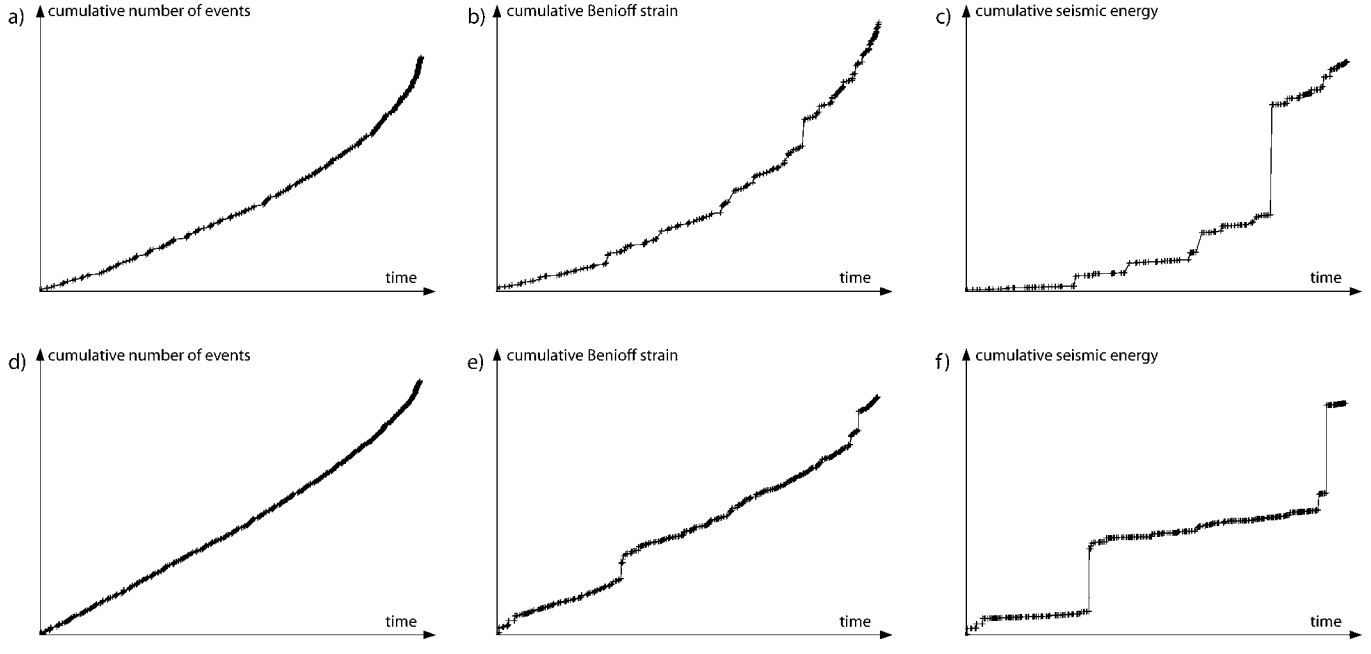


FIG. 3.9. Two examples of seismicity distribution (random uniform in space, random in magnitude in respect to the Gutenberg–Richter law). The first one corresponds to the expected seismicity evolution in the region of accelerating seismicity (power-law) following Eq. (3.35) (a), (b) and (c) and the second one corresponds to an association of accelerating seismicity (power-law) following Eq. (3.35) and of background seismicity (linear-law) following Eq. (3.31) (d), (e) and (f); (a)–(d): cumulative number of events through time; (b)–(e): cumulative Benioff strain through time (i.e. accelerating moment release); (c)–(f): cumulative seismic energy through time.

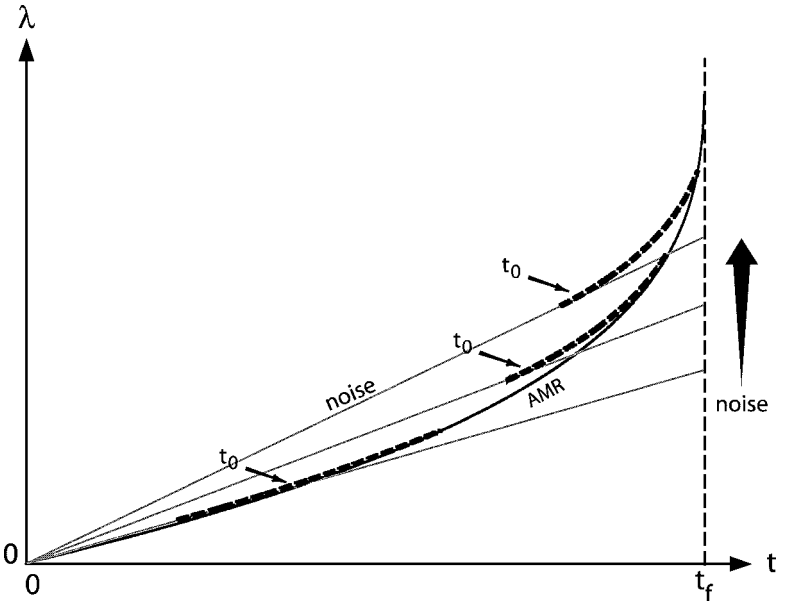


FIG. 3.10. Schematic representation of the observed duration of accelerating precursory seismicity. Accelerating moment release is masked by background noise (constant rate). The greater the noise, the later in the seismic cycle that acceleration is observable. t_0 indicates the approximate starting time determined from a temporal search.

has yet been made between the duration of acceleration moment release and the recurrence time of large earthquakes on a same fault. Because of the power-law behaviour, the curve is at first undistinguishable from a straight line.

A better determination of the spatial distribution of precursory seismicity, such as done by Bowman and King (2001a) and Mignan et al. (2006b), may permit some background seismicity noise to be removed and to extract accelerating patterns over longer time periods.

3.2.4. Conclusion

The mathematical formulation of accelerating seismicity based on the Stress Accumulation model permits low level regional seismic activity to be related to the loading rate of the fault that is going to fail. Contrary to critical processes, this view links directly the parameters of the acceleration to the behaviour of the main fault (loading rate, duration of the seismic cycle). Moreover, it gives a new explanation for the m-value which depends on the geometrical shape of the stress shadow and of the fractal dimension of the regional fault network.

The proposed formulation should in the future be used for real data sets, with real fault characteristics, to compare with current time-to-failure analysis employed in earthquake forecasts. The complementary role of the Stress Accumulation model to the concept of seismic gap and time recurrence of large earthquakes has already been discussed by Mignan et al. (2006a) and at present the loading rate is directly integrated in the power-law time-to-failure equation of accelerating moment release. Although at present we do not yet know how well the parameters in the equation can be determined and the degree to which this will improve the identification of AMR, nonetheless having a clear mathematical description of the process is the best approach to improve AMR related seismic hazard determination.

4. Earthquake Forecasts Using Accelerating Moment Release

> "No scientific prediction is possible without exact definition of the anticipated phenomenon and the rules, which define clearly in advance of it whether the prediction is confirmed or not."
>
> V.G. Kossobokov, Testing earthquake prediction models, Tectonophysics (2006)

Various phenomena have been suggested as short-term earthquake precursors. They can be based on changes in ground water characteristics (Silver and Wakita, 1996; Silver and Valette-Silver, 1992), on seismically-induced electric signal (Varotsos and Lazaridou, 1991), on changes in the geomagnetic field (Main and Meredith, 1989). Other potential precursors include seismicity patterns such as foreshocks (Main and Meredith, 1989; Gerstenberger et al., 2005) or changes in microearthquake activity (Joswig, 2001) and theoretical efforts have been made to identify their causative mechanisms using nucleation (Das and Scholz, 1981) and dilatancy–diffusion models (Scholz, 1990b). Unfortunately, none of these precursors has been systematically observed before large earthquakes (Geller et al., 1997), calling into question their utility for earthquake prediction.

With the stress-triggering theory, the attention of the seismological community has focused on medium- to long-term precursors, based on the study of earthquake populations (e.g. accelerating moment release, aftershocks). Many forecasting methods have been proposed based on an important diversity of approaches and looking at different precursory seismicity patterns. Some approaches are empirical, some others are based on physical processes and all of them have shown different degrees of success. The purpose of this section is to show what the Stress Accumulation model can provide as a forecasting method, independent but possibly complementary to existing forecasting methods. Contrary to approaches which are based on statistical principles, the Stress Accumulation model is based

on simple stress transfer relations and relates seismicity patterns (e.g. accelerating moment release) directly to the seismic cycle of a given fault.

In this section, forecasting methods are first described (Section 4.1). Then the Stress Accumulation model, via the study of accelerating moment release, is applied to the case of the Sumatra–Java Arc where a major earthquake caused a catastrophic tsunami in December 2004. If accelerating moment release has already been used for forecasting plans, this is the first time it is directly linked to the concept of a seismic gap (Section 4.2). Because the reliability of accelerating moment release is an important aspect of earthquake forecasts, a new statistical study is performed (Section 4.3). Finally, the role of the Stress Accumulation model and of accelerating moment release are discussed for the problem of earthquake prediction showing its possible impact on time-, space- and magnitude-predictable models (Section 4.4).

4.1. Existing Forecasting Methods

4.1.1. Recurrence of Large Earthquakes

Earthquake periodicity and aperiodicity: The recurrence intervals between past earthquakes are used to forecast the probability of future events for long-term hazard planning (e.g. Sieh, 1978; Wesnousky, 1986; Nishenko and Bollinger, 1990).

The recurrence interval μ and the coefficient of variation α (or aperiodicity) are calculated from seismicity data sets (paleoseismic and historical records) and used in a probability density function $f(t)$. Two commonly applied probability density functions (PDF) are the lognormal (e.g. Nishenko and Buland, 1987)

$$f(t, \mu, \alpha) = \frac{1}{t\alpha\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\alpha^2}\right) \tag{4.1}$$

and the Brownian passage time (e.g. Kagan and Knopoff, 1987)

$$f(t, \mu, \alpha) = \sqrt{\frac{\mu}{2\pi\alpha^2 t^3}} \exp\left(-\frac{(t - \mu)^2}{2\mu\alpha^2 t}\right) \tag{4.2}$$

which is now usually preferred for earthquake hazard calculations (Matthews et al., 2002). Then a time-dependent probability is obtained by summing the probability density function $f(t)$ as

$$P(t \leqslant T \leqslant t + \Delta t) = \int_t^{t+\Delta t} f(t)\, \mathrm{d}t. \tag{4.3}$$

Paleoseismological dating of large prehistoric earthquakes show that the timing between individual events can be highly erratic and even the longest paleoseismic records are short in the context of building a well-determined distribution. For example, the average recurrence interval for the last 10 large earthquakes on the

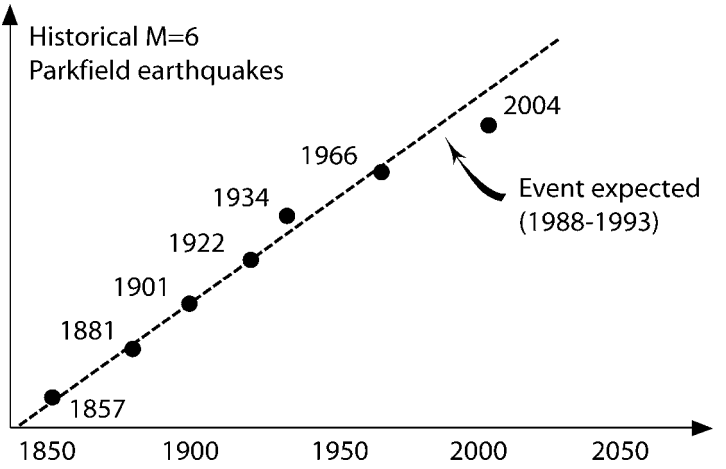


FIG. 4.1. Almost periodic recurrence of moderate ($M \sim 6$) earthquakes in Parkfield, California. From 1857 to 1966, 6 events occurred on the same segment of the San Andreas fault with a regular average time interval of 21 years (e.g. Bakun and McEvilly, 1984; Kerr, 1993). The next event was expected between 1988 and 1993 but only occurred in 2004.

San Andreas fault north of Los Angeles (Pallet Creek) is about 132 years but with a range from 44 to 332 years (Sieh et al., 1989). Thus it is possible that the mean recurrence and the coefficient of variation from a small sample do not represent the only distribution that satisfy the observations. Different approaches have been proposed to find a full array of solutions, such as Bayesian analysis (e.g. Ogata, 1999) or data-dependent normalisation (e.g. Nishenko and Buland, 1987).

However, forward extrapolation hinges on the assumption that there is no stress transfer from other earthquakes. One prime example of the failure of this type of prediction is the forecasting of the last earthquake in Parkfield, California. From 1857 to 1966, 6 earthquakes of moderate size ($M \sim 6$) occurred on the same segment of the San Andreas fault with a regular average time interval of 21 years (e.g. Bakun and McEvilly, 1984; Kerr, 1993). The next event was expected between 1988 and 1993 but it did not occur at the predicted time (Fig. 4.1). Only in 2004 did a new earthquake occur and with the delay possibly explained by a stress transfer from the 1983 Coalinga earthquake (e.g. Toda and Stein, 2002b; Tiampo et al., 2006).

Earthquake interactions: It is widely held that the time of occurrence of an earthquake on a fault undergoing tectonic loading is controlled both by the stress and frictional properties on that fault and by earthquakes on other faults nearby (Stein, 1999). Earthquakes change future seismicity rates in a volume roughly proportional to mainshock magnitude. Some events are advanced in parts of the volume while seismicity might be suppressed in others (see Section 1).

"*With stress transfer leading to changes in the probability of earthquake occurrence, the challenge is to quantify the probability of failure under tectonic loading alone and then to assess how probabilities change when loading is perturbed. Our understanding of stress transfer and triggering has advanced sufficiently that it has begun to be put into practical use, in ways that affect public policy and business decisions*" (quoted from Steacy et al., 2005; Gomberg et al., 2005).

To apply Coulomb stress change models to seismic hazard programs, one approach was first suggested by Stein et al. (1997) and has since been applied in several places (e.g. Toda et al., 1998; Parsons, 2004; Toda et al., 2005). Another approach has been suggested by Hardebeck (2004). In these studies, results from stress transfer models are converted into earthquake probability estimates. Models from Stein et al. (1997) and Hardebeck (2004) use time-dependent probabilities, based on rate-and-state functions (Dieterich, 1994) (see Section 1). The transient change in expected earthquake rate $R(t)$ after a stress step can be related to the probability of an earthquake of a given size over the time interval Δt through a nonstationary Poisson process as

$$P(t, \Delta t) = 1 - \exp\left[-\int_{t}^{t+\Delta t} R(t)\,\mathrm{d}t\right] \tag{4.4}$$

after Dieterich and Kilgore (1996).

Establishing the credibility of probability estimates requires studies of their robustness. Parsons (2005) examines theoretically the sensitivity of estimates of single-fault probability changes due to stress transfer. He shows that we should be confident in these estimates only when a perturbing event is close to the fault in question and/or the perturbing stress change is large relative to the ambient stressing rate and when the fault is well characterised. This is consistent with the results of Kagan et al. (2005), who find that for southern California seismicity, the impact of stress transfer from large earthquakes on the recurrence rate of other faults is weak, although they question whether this may be an observational limitation rather than a physical consequence with a model sensitive to arbitrary choices.

The Working Group on California Earthquake Probabilities and Virtual California: Some of the best forecasting plans are realised in California because of the large amount of data available. The Working Group on California Earthquake Probabilities (WGCEP) lead by the United States Geological Survey (USGS) involve a broad group of geophysicists and earth scientists who create a forecast algorithm for the San Francisco Bay region (http://quake.wr.usgs.gov/research/seismology/wg02). In the WGCEP method, the models that are used to compute a forecast from the data are statistical, meaning that the physics is encoded in (1) the form of the statistical distribution (lognormal or Brownian passage time) and (2) the parameters of the distribution (usually the mean and variance). WGCEP then uses the statistical distribution

for all the fault segments in a region, with parameters set using geological and geophysical data, to compute the conditional probability that a large ($m \geqslant 6.7$) earthquake will occur in the San Francisco Bay region during the period 2002–2031.

Rundle et al. (2006) compare the methods used by the WGCEP ("official forecasts") with their *Virtual California* ("research-quality forecasts"). This is a three-dimensional, interacting, laboratory friction-based, backslip-loaded model for the earthquake activity on the major strike-slip faults of California. The *Virtual California* model shows the importance of stress interactions and how these lead to the emergence of correlations in earthquake activity. Adding stress transfer in earthquake forecasting is complementary to the methods used by the WGCEP, but can improve them in important ways (see details in Rundle et al., 2006).

4.1.2. Pattern Recognition in the Background Seismicity

Many current earthquake prediction models are based on the study of seismicity patterns that lead up to a large event. Contrary to large earthquakes that can interact and which correspond to different sequences (meaning different fault segments), specific patterns that can be found in the background seismicity could be part of the same sequence (meaning the same fault segment). Some earthquake prediction methods are purely empirical but others are based on physical processes. Below are presented some of the most current search techniques for precursory seismicity (the list is far from exhaustive).

Algorithm M8 and CN: One interesting model for earthquake prediction based on seismicity is purely empirical and does not involve any physical causative mechanism. The algorithms CN and M8 (Keilis-Borok and Rotwain, 1990; Keilis-Borok and Kossobokov, 1990) analyse data from past "earthquake flow" in order to determine if a region is in a time of increased probability of a moderate to large earthquake. Earthquake flow includes the amount of seismic activity, its temporal variation, temporal and spatial clustering of events, spatial concentration of events, and long-range interaction between events (Keilis-Borok and Rotwain, 1990).

These algorithms (also the algorithm MSc, Kossobokov et al., 1990) have had varied success rates in predicting large magnitude events (e.g. Kossobokov et al., 1999; Keilis-Borok et al., 2001). However, many of the "predictions" were made in retrospect, using data after a large event had occurred to determine if the algorithms would predict the earthquake. This is an uniformitarian method for predicting a future event, with an empirical false alarm rate that has not been theoretically justified. Nevertheless, Keilis-Borok (1996) explains that prediction algorithms based on some seismicity patterns (magnitude range about $(M - 3)$ in an area about $5L$–$10L$ before an earthquake of magnitude M with a source di-

mension of L) may allow one to predict $\sim 80\%$ of strong earthquakes with alarms occupying altogether 20–30% of the time–space considered.

Reverse Tracing of Precursors (RTP): The Reverse Tracing of Precursors (RTP) described by Shebalin et al. (2006) is a methodology for medium-term prediction (i.e. months in advance). It is based on an ensemble of premonitory seismicity patterns which are considered to occur in a hierarchical system (e.g. Gabrielov et al., 2000; Shebalin et al., 2006). The important aspects of the Reverse Tracing of Precursors are (1) "precursory chains" that reflect the increase of the correlation length (e.g. Gabrielov et al., 2000; Zoller et al., 2001; Zoller and Hainzl, 2002), (2) "intermediate-term patterns" that correspond for example to accelerating seismicity (e.g. Gabrielov et al., 2000) and (3) "pattern recognition of infrequent events" that consider several "opinions" to decide the validity of the calculated chain of events. If sufficient number of votes are accumulated, the chain is considered precursory (see Shebalin et al., 2006, for details).

The Reverse Tracing of Precursors is currently tested by advance prediction in several seismicity active regions and its performance is yet to be validated (Shebalin et al., 2006). It can be noted that the 2003, $M_w = 6.5$ San Simeon earthquake in Central California and the 2003, $M_w = 8.1$ Tokachi-Oki earthquake in Northern Japan were predicted 6 and 7 months in advance, respectively, but some false alarms are also obtained (Shebalin et al., 2006). The statistical significance of these results for forecasting plans is discussed but not the size of the predicted "area of alarm" which seems very large for a realistic application (see Fig. 4.2).

Pattern Informatics (PI): The Pattern Informatics technique (PI) is a methodology for quantifying the spatiotemporal seismicity rate changes in historic seismicity (e.g. Rundle et al., 2002; Tiampo et al., 2002a, 2002b). By extension, this technique studies the temporal evolution of stress changes between earthquakes since changes in seismicity rates are directly related to changes in stress (e.g. Dieterich, 1994). Tiampo et al. (2006) derive a relationship between the "PI index" and stress change, based upon crack propagation theory (previous work about the relationship between seismicity rate changes and crack propagation, e.g. Bufe and Varnes, 1993). Practically, the Pattern Informatics method measures the change in seismicity rate at each box of a pre-defined grid, relative to the background seismicity rate, through the division of the average rate by the spatial variance over all boxes. As such, it identifies the characteristic patterns associated with the shifting of small earthquakes from one location to another through time prior to the occurrence of large earthquakes (Tiampo et al., 2006). Results correspond to a mapping of "PI anomalies" which are located where a new large earthquake can be expected. Figure 4.3 shows regions of anomalous seismicity found by Rundle et al. (2002) based on the Pattern Informatics technique. Darker colours indicate regions of increasingly anomalous activity and triangles

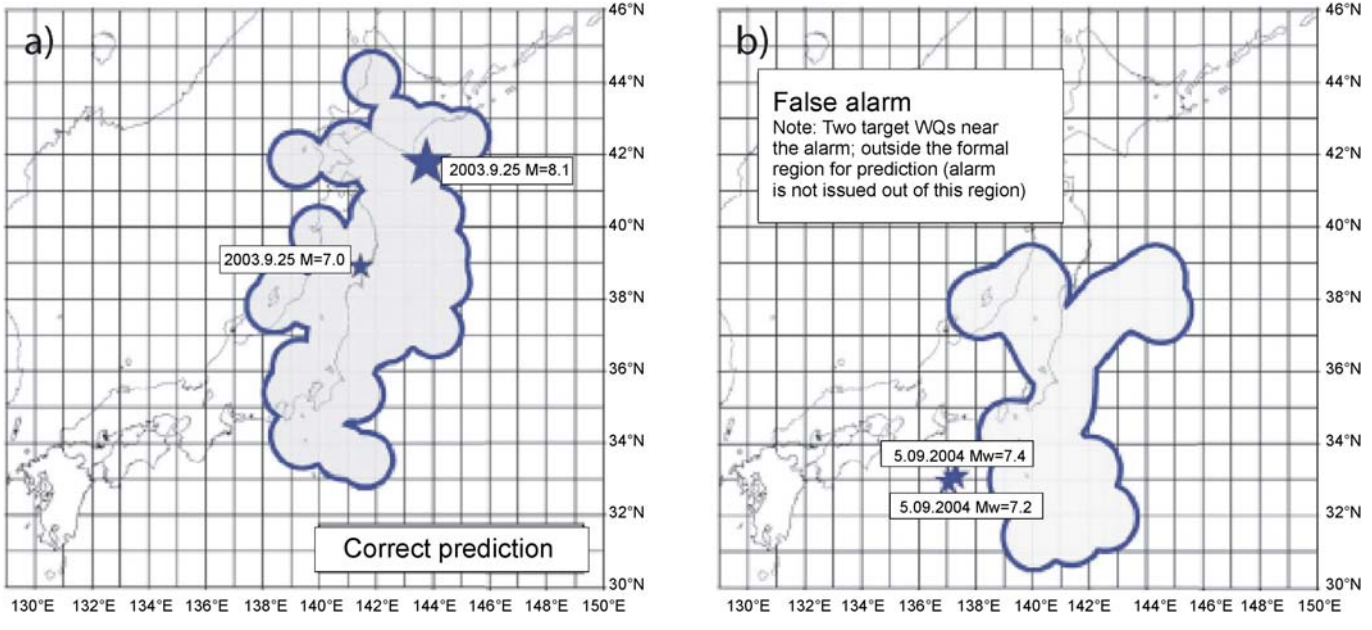


FIG. 4.2. Results of advance prediction in Japan using the Reverse Tracing of Precursors methodology. Two alarms were issued: one correct (panel a), and one false (panel b). It can be noted that the time window is several months and the space window hundreds of kilometres. Redrawn from Shebalin et al. (2006).

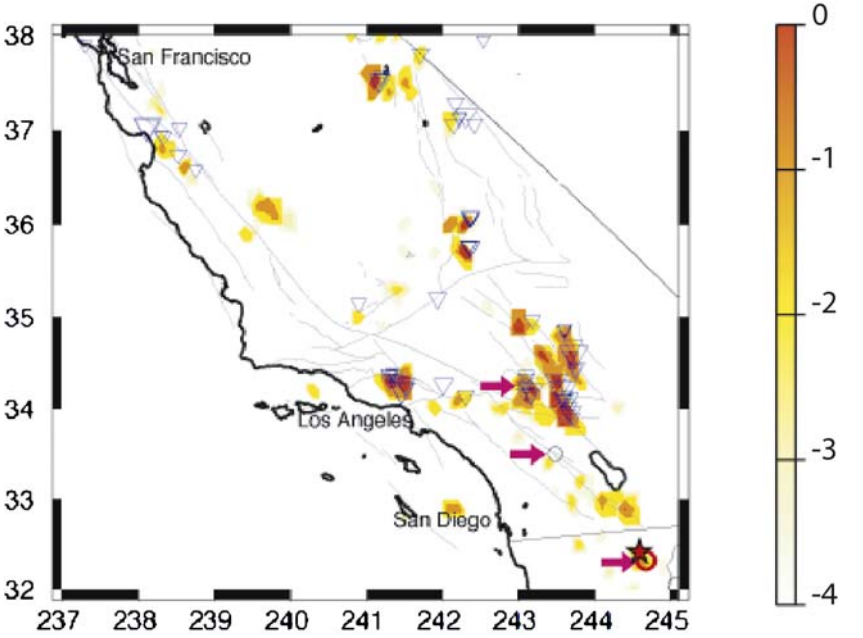


FIG. 4.3. Regions of anomalous seismicity found by Rundle et al. (2002) based on the Pattern Informatics technique. Darker colours indicate regions of increasingly anomalous activity and triangles are events that occurred during the learning phase of their algorithm (see e.g. Tiampo et al. (2002b) for details). Circles are events that occurred after the learning phase. The thick circle is the 2001, $M_w = 5.7$ event near Calexico, Mexico and the star indicates the location of a $M_w = 5.6$ event that happened in 2002, after the paper's submission (Sammis et al., 2004). Redrawn from Sammis et al. (2004).

are events that occurred during the learning phase of their algorithm (see e.g. Tiampo et al., 2002b, for details).

The Pattern Informatics technique has shown that significant anomalous seismicity changes occur during the mid-1970s in the region of the future 1983 Coalinga earthquake (Tiampo et al., 2006). Sammis et al. (2004) note that this technique has also predicted a $M_w = 5.6$ earthquake near Calexico, Mexico. This event occurred after the submission of the forecasting map of California of Rundle et al. (2002) (Fig. 4.3).

It can be noted that Holliday et al. (2006) propose to modify the Pattern Informatics method for forecasting large earthquake events by using complex eigenfactors. This is based on theoretical analysis which indicates that earthquake stresses may obey a wave-like equation. The new method uses the Principal Component Analysis (PCA) which is a mathematical procedure that transforms a set of correlated variables into a smaller set of uncorrelated variables called princi-

pal components (see Holliday et al., 2006 for details). New results show a better apparent correlation with future large earthquakes (Holliday et al., 2006).

4.1.3. Accelerating Moment Release Analysis

> "There is a controversy about earthquake prediction, supported, on one side, with a surprisingly small number of basic systematic studies and, on the other side, with a multitude of post-the-fact numerological exercises. As a result, although hundreds, if not thousands of observed phenomena have been claimed "to precede systematically" large earthquakes, there are almost no reproducible quantitative definitions of precursors."

V.G. Kossobokov, Testing earthquake prediction models, *Tectonophysics* (2006)

It is important to note that the different methods presented previously can indirectly search for the same precursory seismicity patterns. The different techniques are based on some statistics around the changes of seismicity rate and some of these aspects can possibly be linked to accelerating moment release. Indeed, the increase in seismic activity phenomenon is utilised as one of several indicators in the algorithm M8 (Keilis-Borok and Rotwain, 1990) as well as in the Reverse Tracing of Precursors (Shebalin et al., 2006). Also some attempts have been made to compare accelerating moment release and Pattern Informatics (Bowman et al., 2006). However contrary to the different forecasting techniques proposed, accelerating moment release has the advantage of being simply defined and thus more easily understandable.

Robustness of accelerating moment release: Accelerating moment release seems to be one of the most promising precursors of large earthquakes since it seems to be systematically observed when seismicity catalogues of sufficient quality are available. It is for example observed before all $M_w > 6.5$ earthquakes in California from 1950 to 2000 (Bowman et al., 1998; Bowman and King, 2001a). It is observed in many regions and for different fault mechanisms: normal, reverse, strike-slip (e.g. Jaumé and Estabrook, 1992; Bufe and Varnes, 1993; Bufe et al., 1994; Knopoff et al., 1996; Brehm and Braile, 1998; Papazachos and Papazachos, 2000; Robinson, 2000; Karakaisis et al., 2002; Papazachos et al., 2002a, 2002b, 2002c; Karakaisis et al., 2004; Sammis et al., 2004; Jiang and Wu, 2005; Mignan et al., 2006a).

Jaumé and Sykes (1999) reviewed some large earthquakes which were not preceded by accelerating moment release. For example, in a systematic review of moment release rates in the Alaska–Aleutian subduction zone, Bufe et al. (1994) found no accelerating moment release in the vicinity of the 1986 $M = 8.1$ Andreanof Islands earthquake. However this can possibly be explained by the poor quality of seismicity catalogues of this region for this time period. Also it is

difficult to verify if some accelerations of seismicity can be observed with no following earthquake. A systematic search would need many parameters to be taken into account (central point, region size, starting time, end time, magnitude cut-off). Brehm and Braile (1998) tested the "false-positive" alarm by randomly selecting 20 locations within the New Madrid Seismic Zone. The time window was free and the region size tested for 5 different radii. Brehm and Braile (1998) showed that some false-positive acceleration curves occurred over very short time intervals (less than 3 years) and that the false-positive occurrence rate was 2%. However the stability of this result is questionable.

Another important question is about the stability and false alarm rates associated with accelerating moment release patterns. Vere-Jones et al. (2001) pointed out that the least-squares technique used to fit and quantify accelerating moment release leads to instabilities in the determination of the parameters of the time-to-failure power-law equation. However, they also note that the least squares method will produce a "reasonable visual fit to the data". Also the quantification of the acceleration by the c-value (Bowman et al., 1998) is a simple method and possible instabilities (i.e. jumps of seismicity) can be removed with proper declustering (see Section 2). Accelerating moment release can then be quantified in terms of reliability to real processes thanks to the simulation of synthetic catalogues (Bowman et al., 1998; Brehm and Braile, 1998; Gross and Rundle, 1998; Zoller et al., 2001; Mignan et al., 2006a) and a c-value probability curve can be obtained for forecasting plans (see Section 4.2.7). Note that the statistical behaviour of accelerating moment release is studied in detail in Section 4.3.

Role of accelerating moment release in previous forecasting plans: In recent years, the accelerating moment release concept has caused much attention in the study of earthquake forecasting. Clark et al. (2003) quantify accelerating moment release observed at present time around several faults in California to determine which faults could fail in a near future, and this work is still in progress (Bowman et al., 2006, see Fig. 4.4). Robinson (2000) used a grid-search procedure to test if previous earthquakes could have been forecast using the accelerating moment release model. However, their Time of Increased Probability model (TIP) still needs to be validated. A similar study has been conducted by Papazachos et al. (2002a) for Turkey. They also search for accelerating moment release systematically in a grid and add a new quantification parameter, the quality index q_a (see in Papazachos et al., 2002a). However, these studies do not permit the time of occurrence of the incoming large earthquake to be forecast.

Some attempts have been made to predict large earthquake occurrence time using the time-to-failure equation which defines accelerating moment release patterns. Time-to-failure analysis is a technique for predicting earthquakes in which a failure function is fit to a time-series of accumulated Benioff strain (Gross and Rundle, 1998). The time-to-failure method as applied to earthquakes has first been described by Varnes (1989). The determination of the time of a future large event

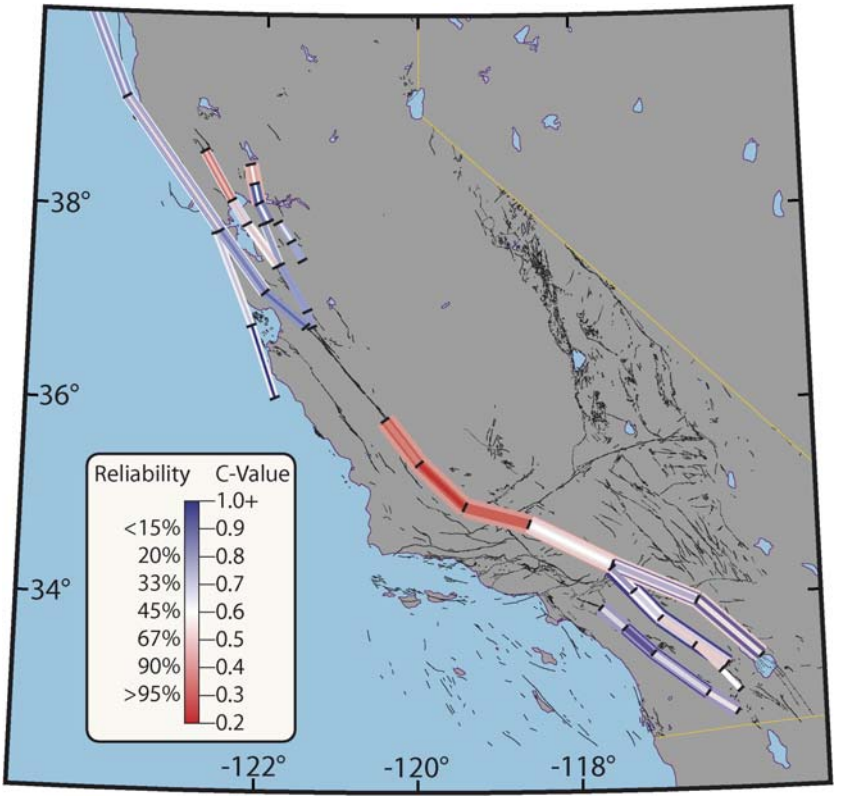


FIG. 4.4. Summary of the reliability of ongoing accelerating moment release for scenario earthquakes in California. The overlapping lines in the map above reflect the observed level of accelerating moment release for different combinations of scenario events. Redrawn from Bowman et al. (2006).

or mainshock, can be estimated from data on the time and magnitude of precursory shocks (Varnes, 1989). Back analyses indicate that, with adequate data, estimates of mainshock time may in some instances be made "with sufficient accuracy and far enough in advance to be useful in decreasing loss of life and property" (Varnes, 1989). This has also been tested for earthquakes in the New Madrid Seismic Zone by Brehm and Braile (1998) who explain that modelling precursory accelerating energy release is a "viable method" (Fig. 4.5). Bufe et al. (1994) also successfully forecasted the 1996 $M = 7.9$ Delarof Islands earthquake by using accelerating moment release patterns. However, other studies explain that "mainshock times are only loosely constrained" when looking at the AMR signal (Robinson, 2000).

Because the time-to-failure analysis is not perfect yet, the best solution seems to seek to understand better the origin of AMR. It has been shown in Section 3

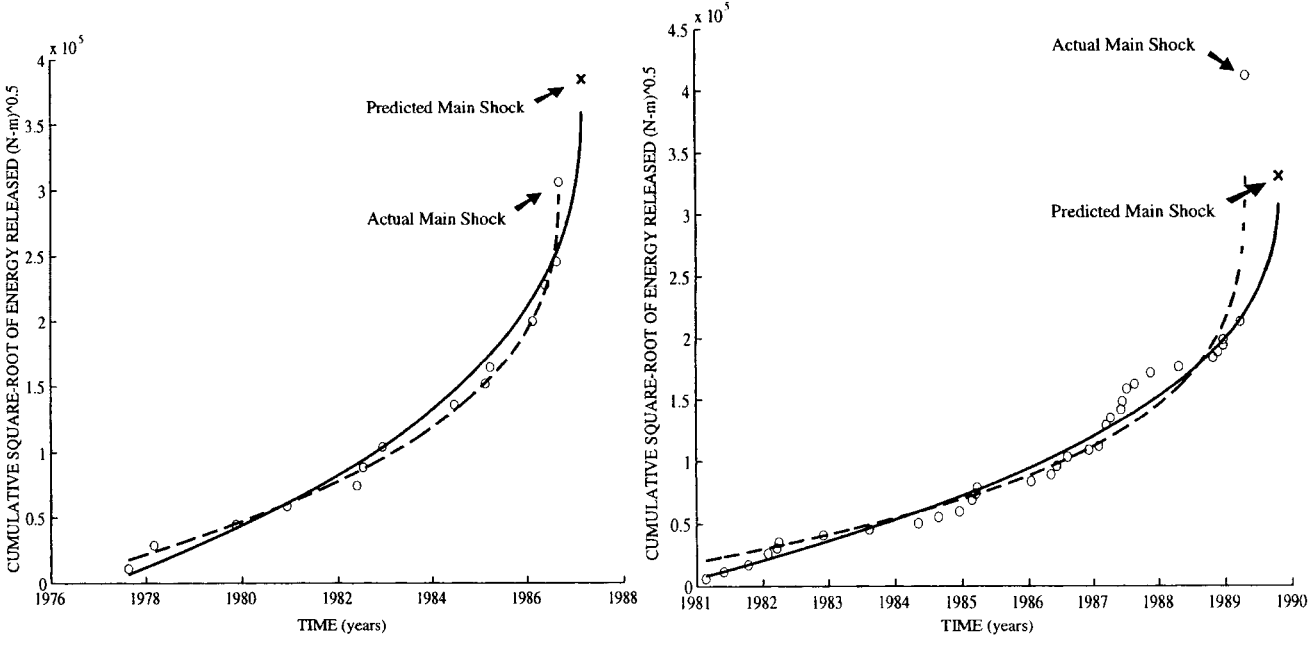


FIG. 4.5. Time-to-failure power-law fit of AMR observed before two earthquakes in the New Madrid Seismic Zone; (a) Precursory sequence for the August 1986 $M = 3.6$ earthquake; (b) Precursory sequence for the April 1989 $M = 4.3$ earthquake. The solid curve represents the model of AMR with the predicted mainshock time. The dashed curve represents the model of AMR with the actual mainshock time. Redrawn from Brehm and Braile (1998).

that a mathematical formulation for the time-to-failure power-law can be obtained from the Stress Accumulation model. This promising result should, in the future, permit the fit of accelerating moment release to be improved by determining some of the equation parameters from geological and geodetic data (see Section 3 for more detail).

4.2. Accelerating Moment Release and the Sumatra–Java Arc Seismic Hazard

Accelerating moment release is here tested as a precursor of large earthquakes in the view of the Stress Accumulation model. The proposed forecasting method is used in the case of the Sumatra–Java fault system because of the recent devastating consequences of the 2004, $M_w = 9.0$–9.3 Sumatra earthquake.

Other forecasting methods could also be used, such as algorithm M8 (Keilis-Borok and Rotwain, 1990) or Reverse Tracing of Precursors (Shebalin et al., 2006) that also search for an increase of the seismic activity (see Section 4.1). Pattern Informatics (Tiampo et al., 2006) could also be useful because it catches changes of the correlation length, which is implicitly observed in the Stress Accumulation model (Sammis and Sornette, 2002). However the physics underlying such methods remains unclear and it is preferable to base the forecasting plan on accelerating moment release alone.

This work associates the study of precursory seismicity patterns to the seismic cycle of faults and by consequence to the concept of seismic gap. Indeed the Stress Accumulation model explains that accelerating moment release is a consequence of stable sliding on the faults at depth (see Section 1). Simulations of King and Bowman (2003) and the mathematical formulation of accelerating moment release proposed in Section 3 show that this precursor should be observed in the last stages of loading on the fault, before rupture.

The relative role of the Stress Accumulation model and associated accelerating moment release in earthquake forecast is illustrated in Fig. 4.6. It shows that the Stress Accumulation model provides a complementary information to the study of recurrence intervals, independently of possible stress triggering from other earthquakes on other faults. However, as it will be shown in the Sumatra–Java Arc example, accelerating moment release does not yet determine the time of occurrence of the next earthquake. Thus time probabilities cannot yet be determined from this new forecasting method. In this study, a time-to-failure analysis (see Section 4.1) is not included.

4.2.1. Introduction

The study of the seismic activity in the Sumatra–Java region prior to the December 26, 2004 (Mw = 9.0–9.3) and March 28, 2005 (Mw = 8.7) earthquakes is presented, following Mignan *et al*. (2006a): A promising approach to assessing

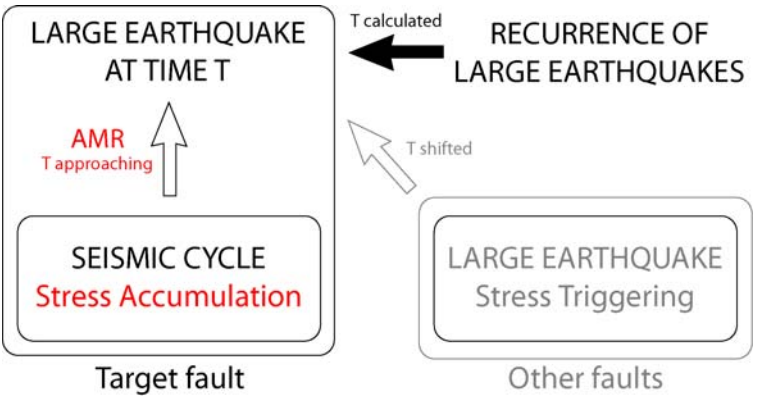


FIG. 4.6. Schematic representation of the role of the Stress Accumulation model and associated accelerating moment release in large earthquake forecasts. The time of occurrence T of a new earthquake on a target fault is first determined from the recurrence of past earthquakes on this fault. The recurrence is linked to the seismic cycle of the target fault and the Stress Accumulation model gives a complementary information about T. Independently, stress triggering from other large earthquakes (on other faults) is possible.

seismic hazards has been to combine the concept of seismic gaps with Coulomb-stress change modelling to refine short-term earthquake probability estimates. However, in practice the large uncertainties in the seismic histories of most tectonically active regions limit this approach since a stress increase is only important when a fault is already close to failure. In contrast, recent work has suggested that Accelerated Moment Release (AMR) can help to identify when a stretch of fault is approaching failure without any knowledge of the seismic history of the region. AMR can be identified in the regions around the Sumatra Subduction system that must have been stressed before the 26 December 2004 and 28 March 2005 earthquakes. The effect is clearest for the epicentral regions with less than a 2% probability that it could occur in a random catalogue. Less clear AMR is associated with the regions north of Sumatra around the Nicobar and Andaman islands where rupture in the December 2004 earthquake was less vigorous. No AMR is found for the region of the 1833 Sumatran earthquake suggesting that an event in this region in the near future is unlikely. AMR similar to that before the December 2004 and March 2005 events is found for a 750 km stretch of the southeastern Sumatra and western Java subduction system suggesting that it is close to failure. Should the whole of this stretch break in a single event the magnitude could be similar to the December 2004 earthquake.

The earthquake that resulted from slip along about 1600 km of the Sumatra-Andaman subduction system on 26 December 2004 has focused attention on the tsunami danger posed by such massive submarine events (Ammon et al., 2005; Bilham, 2005; Lay et al., 2005; McCloskey et al., 2005). Although it was known

that events along the Sumatra–Java subduction zone could generate tidal waves, such a massive earthquake was not anticipated since the relative paucity of large events along the arc (e.g. Ortiz and Bilham, 2003; Zachariasen and Sieh, 1999) led to the conclusion that substantial parts of the arc were uncoupled (aseismic) (Prawirodirdjo, 1997; Simoes et al., 2004) (Fig. 4.7). Prior to the recent events, known historical earthquakes had released little of the slip that had accumulated along the arc in the preceding 2–300 years. In retrospect, it seems that the historical record was too short to characterise the behaviour of the arc (Simoes et al., 2004).

Examining space–time patterns of regional seismicity is increasingly being used as the basis for assessing seismic hazard. In regions where good historical information is available, and taking Eastern Turkey and the Aegean as an example, it has been shown that the concept of seismic gaps can identify faults likely to host substantial events over the coming years or decades (Hubert-Ferrari et al., 2000; Stein, 1999; Toksoz et al., 1979). The reliability of this approach rests on the assumption that earthquakes repeat regularly, an assumption that many workers regard as unreliable (e.g. Benedetti et al., 2002; Rundle et al., 2001; Weldon et al., 2004). The identification of seismic gaps has been complemented in such areas by studies of earthquake stress interactions (e.g. Nalbant et al., 1998; King and Cocco, 2000). As a result of tertiary creep, the abrupt change of stress due to an earthquake enhances the probability of events over time periods of years in the regions where Coulomb-stress has been increased (Scholz, 1990a). This has been quantified using rate and state friction to estimate earthquake probabilities for the period shortly following large events (e.g. Stein, 1999; Scholz, 1998; Freed, 2005; Gerstenberger et al., 2005). However, because these studies involve the calculation of stress changes rather than absolute stress, they are of limited use unless a reliable recurrence time can be used to estimate where the region lies in its seismic cycle.

In this paper, we discuss a technique that allows proximity to failure to be assessed without knowledge of the seismic history of the region. The approach is to examine seismic activity in the region that must become loaded prior to an event. Numerous studies have suggested that seismic activity increases over a wide region around a future epicenter, an effect that has been labelled Accelerated Moment Release (AMR). Many explanations for these observations have been offered (Sornette and Sammis, 1995; Bowman et al., 1998; Grasso and Sornette, 1998; Sammis and Smith, 1999) but none established a clear relation between the fault that would fail and the evolving stress field. For events in California however, it has been shown that the increase in activity is found in the regions where stress must accumulate before motion on the future fault can occur and that these can be identified after the earthquake has occurred (Bowman and King, 2001a, 2001b). They also propose that AMR can be identified before an earthquake if geological data allows the parameters of a possible future event to be identified before it happens. Here, we apply this Stress Accumulation Model (King and Bowman,

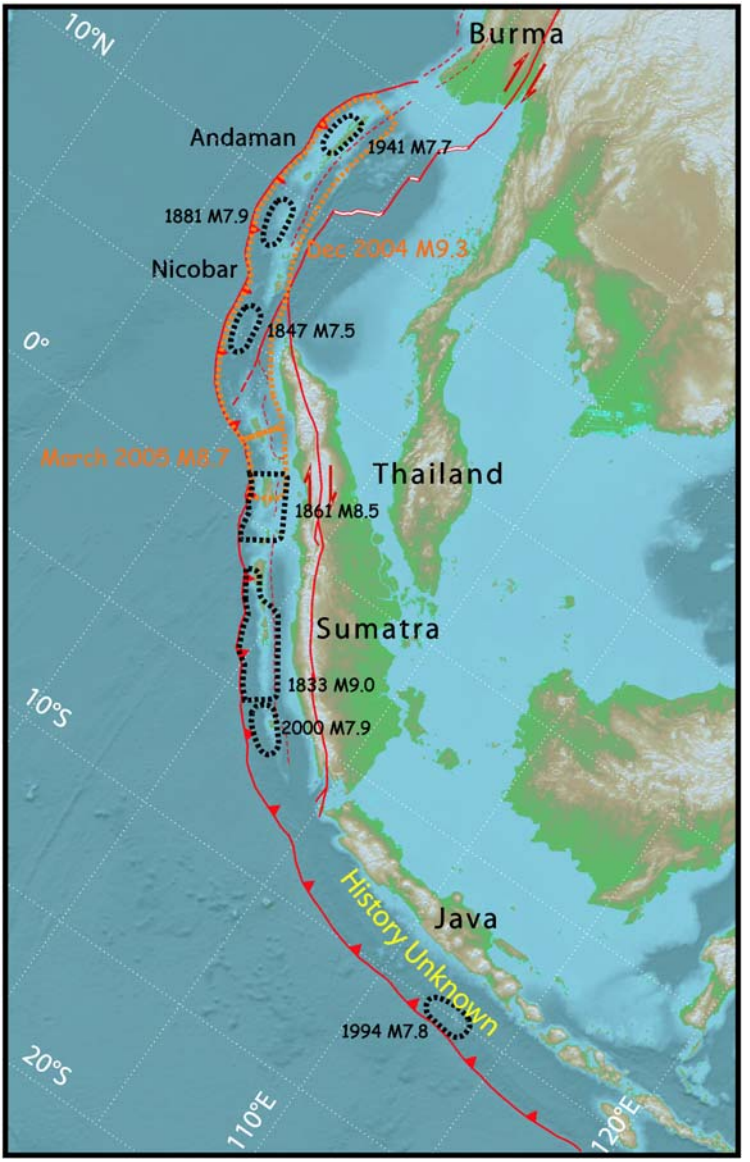


FIG. 4.7. Large earthquakes ($M > 7.5$) along the Sumatra–Java arc. Estimated rupture regions for historical events are shown in black and the rupture regions for the 26 December 2004 and 28 March 2005 are shown in grey.

2003) to the Sumatra–Java arc and demonstrate that the slip regions of both the 26 December 2004 and the 28 March 2005 events showed clear evidence that they were approaching failure prior to the two events. We further show that the regions of highest slip in the two events correspond to the sections of the arc that exhibited the most pronounced AMR. In contrast, the section of the subduction zone where the 1833 earthquake occurred shows no reliable evidence of accelerating activity at present. However, a region of accelerating seismicity associated with a 750 km stretch of the mapped plate boundary along southeastern Sumatra and western Java suggests that this region may be approaching failure.

4.2.2. The Stress Accumulation Model

The mechanical processes of the Stress Accumulation Model (SAM) are discussed in detail by King and Bowman (2003). They describe a complete seismic cycle for an isolated sticking patch on an otherwise creeping strike-slip fault. The model allows synthetic seismic catalogues to be generated that are closely comparable to those observed. An idealised representation of the evolution of stress and seismicity for the last 10% of the earthquake cycle is shown in Fig. 4.8. This is similar to the model described by King and Bowman (2003), but calculated for a reverse fault. The faults adjacent to the impending earthquake are considered to be early in their seismic cycle and are locked. The fault at depth creeps with slip increasing from 90% to 100% of the slip required to bring the fault in the seismic gap to failure. Figures 4.8a–d represent the last 10, 3.5, 1 and 0% of the seismic cycle. During the last 10% of the cycle, it can be seen that seismic activity (represented by small black dots) rapidly surrounds the future epicentre. This pattern of activity mimics the so-called "Mogi Doughnut", a halo of precursory activity that is sometimes observed around subduction zone earthquakes (Mogi, 1969). Following the earthquake, the stress pattern changes and the location of the seismicity switches to the aftershock lobes. King and Bowman (2003) show that realistic values of earthquake stress drops and stress field inhomogeneities can produce artificial catalogues with AMR similar to those observed.

The SAM method (Bowman and King, 2001a, 2001b) defines the region where stress must accumulate and pre-event seismicity must occur during the last part of the earthquake cycle. The technique is the same as that used to calculate post-earthquake Coulomb-stress (Stein, 1999; King and Cocco, 2000), but the fault is slipped with the opposite sense (the "backslip" model first described by Savage (1983)) to identify the form of the stress field before the earthquake. The form of the stress field is the same for half-space and plate calculations, only the decay with distance is changed (Stein et al., 1994). In common with most other Coulomb stress change studies a half-space calculation is adopted. Figure 4.9 shows the method applied to the segment of the subduction zone that experienced the largest and most rapid slip in the 26 December 2004 earthquake. The inset shows the surface projection of the fault as a black line. The outlined region cor-

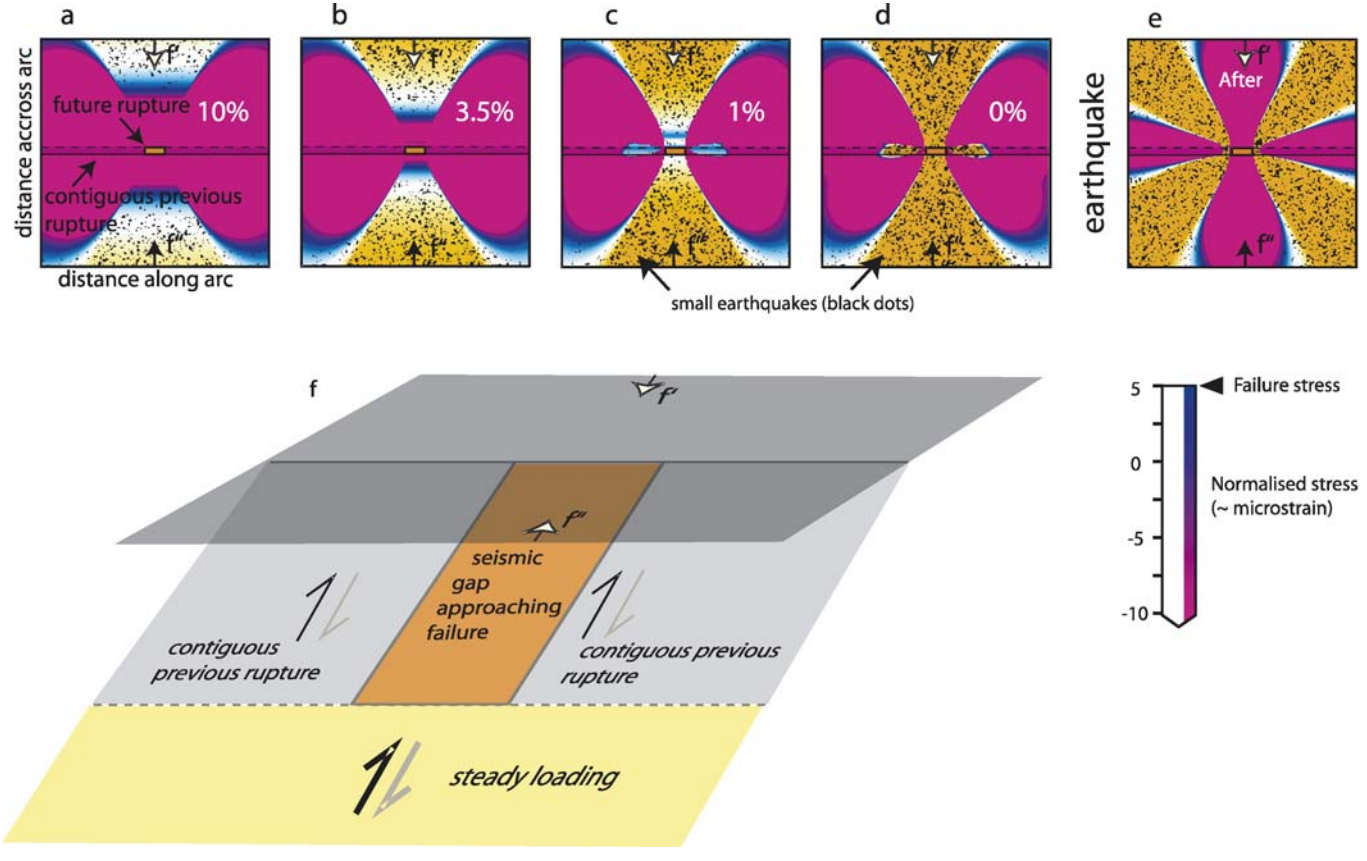

a
b
c
d
e
distance accross arc
future rupture
10%
contiguous previous rupture
distance along arc
3.5%
1%
0%
earthquake
After
small earthquakes (black dots)
f
seismic gap approaching failure
contiguous previous rupture
contiguous previous rupture
steady loading
5
0
-5
-10
Failure stress
Normalised stress (~ microstrain)

FIG. 4.8. The evolution of stress and seismicity for the last 10% of the earthquake cycle for a reverse fault. The fault geometry (not to scale) is shown in f. The faults adjacent to the impending earthquake are considered to be early in their seismic cycle and are locked, while the creeping fault at depth increases in slip from 90% to 100% of the slip required to bring the fault in the seismic gap to failure. Since only slip on the fault at depth increases below a locking depth until the earthquake occurs, this model is identical to those used to model geodetic data. The values used for the model are 7.5 metres slip for the adjacent faults, and the slip on the creeping fault below the seismogenic zone increases from 9 to 10 metres. Figures 2(a)–(d) represent the last 10, 3.5, 1 and 0% of the seismic cycle. It can be seen that during the last 10% of the cycle, activity (represented by small black dots) rapidly approaches the future epicentre. Following the earthquake the lobe patterns change and the black dots represent aftershocks. The mask to provide the inheritance stress (see King and Bowman (2003)) and the colour range are chosen so that failure stress reaches the future fault when deep slip reaches 10 metres. The colour bar shows normalised stress. The depth of the seismogenic zone is taken to be 15 km (locking depth) and the length of the fault is 30 km. To produce the artificial earthquakes, random stress variations (referred to as "noise" in King and Bowman (2003)) are added to the evolving stress level calculated for realistic earthquake parameters as described above. For this figure the random variations are exaggerated to make the events easily visible. It has previously been shown that, using realistic random variations, a model like this can create artificial seismic catalogues exhibiting AMR that is similar to real earthquake sequences when analysed by the same methods as real catalogs (King and Bowman, 2003).

responds to the 0.02 bar coulomb pre-stress contour for the fault parameters given in the figure caption. The choice of this value is discussed in the next section and in earlier papers (Bowman and King, 2001a, 2001b; Sammis et al., 2004), where it is pointed out that the form and not the absolute value of stress is of importance to identify the region that must be loaded prior to an earthquake. The earthquakes (shown by black dots) within this contour demonstrate AMR. The AMR shown in the figure is first identifiable in 1977 with activity steadily increasing until immediately prior to the December earthquake. It should be noted that the increase in activity results not from a few larger events near the end of the period but an overall increase of activity. This provides some support for the view that the AMR is a consequence of a general increase of activity (i.e. an increase of a-value and not a change of b-value in the Gutenburg–Richter relation) (Stein et al., 1994; Sammis et al., 2004). In this example the backslip earthquake model location and parameters are defined by the reported parameters of the event (Ammon et al., 2005).

To examine AMR along the whole arc we divide it into segments. The quality of earthquake catalogues in the Sumatra–Java region places a practical constraint on the size of the segments. In California, where catalogues are complete down to a magnitude of 3.5, it is possible to retrospectively demonstrate AMR before events for $M = 6.5$ events whose rupture parameters have been established (Bowman and King, 2001a, 2001b). However, catalogues in the Sumatra–Java region are not homogeneous below magnitude 4.5, so that it is unreasonable to expect to find AMR even for known events smaller than $M \sim 7.5$ (fault length $\sim$ 150 km).

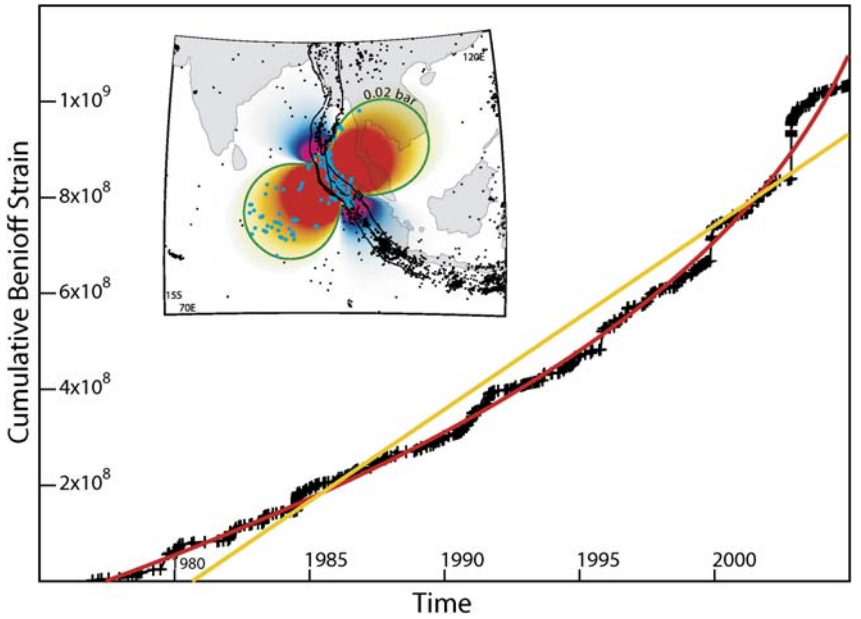


FIG. 4.9. Accelerating Moment Release (AMR) for the great Sumatra earthquake of December (12/26/2004). The Cumulative Benioff Strain and the time (in years) are respectively the y-axis and x-axis. The power law fit (dark grey curve) and the line fit (light grey line) used for the calculation of the c-value are shown. The c-value is 0.28, which corresponds to a 2% probability that the acceleration could occur randomly. The pre-stress field calculated by the backslip earthquake method for the high slip part of the 26 December 2004 event is shown in the inset. The grey contour (0.02 bars) outlines the area of events that provides the best c-value. Other contours also result in AMR; the one displayed is the optimum. The fault used corresponds to segments 5, 6, and 7 in Table 1. It has a dip of 11°, a down-dip width of 200 km and each segment is pure dip-slip.

With this constraint in mind, the subduction zone is divided into segments 250 km long to search for AMR before future events. This segmentation length means that only future events significantly greater than $M = 8.0$ will be detected.

4.2.3. Quantification of Accelerating Moment Release (AMR)

A number of studies have suggested that AMR can be modelled by fitting a simple power-law time-to-failure equation of the form

$$\sum \epsilon(t) = A + B(t_f - t)^m \tag{4.5}$$

where t_f is the time of the large event, B is negative and m is usually about 0.3. A is the value of $\epsilon(t)$ when $t = t_f$ (Bowman and Sammis, 2004; Bufe and Varnes,

1993). The cumulative Benioff strain at time t is defined as

$$\epsilon(t) = \sum_{i=1}^{N(t)} \sqrt{E_i(t)} \tag{4.6}$$

where E_i is the energy of the ith event and $N(t)$ is the number of events at time t. To quantify the AMR, we use the c-value (Bowman et al., 1998) defined to be the ratio between the root mean square of the power-law time-to-failure fit (Eq. (4.5)) and the root mean square of a linear fit to the cumulative energy of events. If the c-value tends to zero, a power law best approximates the curve and AMR is observed. If the c-value tends to unity then no AMR is seen. For a specified source (known or proposed) the program Nutcracker (http://geology.fullerton.edu/dbowman/downloads/NutcrackerX.1.zip) automatically searches for low c-values within a range of Coulomb-stress contours and a range of catalog start times. From this matrix, start times and stress levels can be identified that exhibit AMR, together with sensitivity of the AMR to parameter changes.

The stress levels identified in this procedure are used to define the shape of the stressed region (see Fig. 4.9). For any given azimuth, the extent of the pre-stress region depends on the fault slip. If known fault slip is used following an earthquake, the stress contours correspond to stress change at the time of the event. Where events have not yet happened, a nominal fault slip is selected and the values of the stress contours should be taken only as a rough guide. Clearly variations in modulus such as that between ocean and continental crust will modify stress contours, but such an effect is minor compared to other limitations of the approach addressed in the discussion.

Since the algorithm automatically seeks regions of low c-value, some degree of AMR will nearly always be found. To determine if the acceleration is due to a 'real' physical process or is merely a result of random clustering, we have used the Monte Carlo technique to examine the likelihood of AMR in random catalogues. Because any AMR found in these random catalogs cannot be due to a physical process, they provide a way of estimating the reliability of any give AMR sequence (Ikeda, 2004) (further information can be found in the Appendix). For the example in Fig. 4.9, the technique suggests that there is less than a 2% possibility that the observed AMR is due to random processes. In the study of the Sumatra–Java arc that follows, the reliability of the observed AMR is assessed by direct observation, examining the start-time versus stress matrix and the c-value (see Appendix). The total time over which the acceleration is observed is also an important guide. The length of the AMR is related to the length of the seismic cycle (King and Bowman, 2003) and consequently a short total time suggests that the AMR is incomplete and the region not yet fully re-stressed. This is consistent with such regions rupturing only as a consequence of propagation of slip from more stressed regions (as discussed below). However, an AMR with a short total

time also contains fewer events and is of less statistical significance than a longer time period.

4.2.4. Accelerating Moment Release along the Sumatra–Java Arc

The method outlined above has been applied along the length of the Sumatra–Java Arc (Fig. 4.10). The mechanisms are assumed to be pure reverse faulting with a dip of 11° and a width of 200 km. The regional stress for the coulomb calculation (King and Cocco, 2000) is chosen so that the greatest compressive stress is approximately perpendicular to the arc for each segment. The segments are shaded according to their c-value. The total time over which AMR can be identified is indicated beside each segment. The segments are numbered and the corresponding c-values, total time and the probability that the AMR is non-random (reliability, see Appendix) are shown in Table 4.1.

Figure 4.11 shows the slip distributions for the 26 December 2004 and the 28 March 2005 events (adapted from Ammon et al., 2005), which can be compared to the AMR for the event. The maximum slip in the December 2004 event corresponds to three segments 5, 6, and 7, while the maximum slip in the March 2004 event is in segment 8. Each segment individually had low c-values corresponding to less than 4% probability that they resulted from random processes. Segments 1 to 4 also exhibited AMR before the 26 December event, but with lower significance. In each of these cases, the time period over which AMR is observed is also relatively short. The higher c-value, and shorter time period, together with an inspection of the plots of Benioff strain (see discussion in the Appendix), suggests that the AMR for these parts of the fault are less convincing. Nonetheless, the AMR is consistent with segments 1 to 4 being loaded, albeit less dramatically than segments 5, 6, and 7. This is consistent with the lower slip on the northern segments during the earthquake. The 26 December earthquake initiated where the AMR was clearest and where the subsequent slip was greatest. The 28 March event also initiated where the AMR was clear and propagated to segment 9 where AMR is less clear. The evidence for AMR for segment 10 is almost absent and there is little evidence of AMR for segment 11 and none for segment 12. Further south, for segments 13,14 and 15 significant AMR is observed with values similar the zones that experienced the highest slip in December 2004 and March 2005. Segments 16 to 18 have no significant AMR but the results may be contaminated by stress systems and seismicity associated with the Banda and Phillipine arc.

4.2.5. Discussion

The cartoon evolution shown in Fig. 4.8 shows a rapid evolution with seismicity steadily approaching the future epicentre during the last 10% of the earthquake cycle and is calculated in a similar way to that used by King and Bowman (2003) (see the figure caption). For real situations, effects that are not considered for the

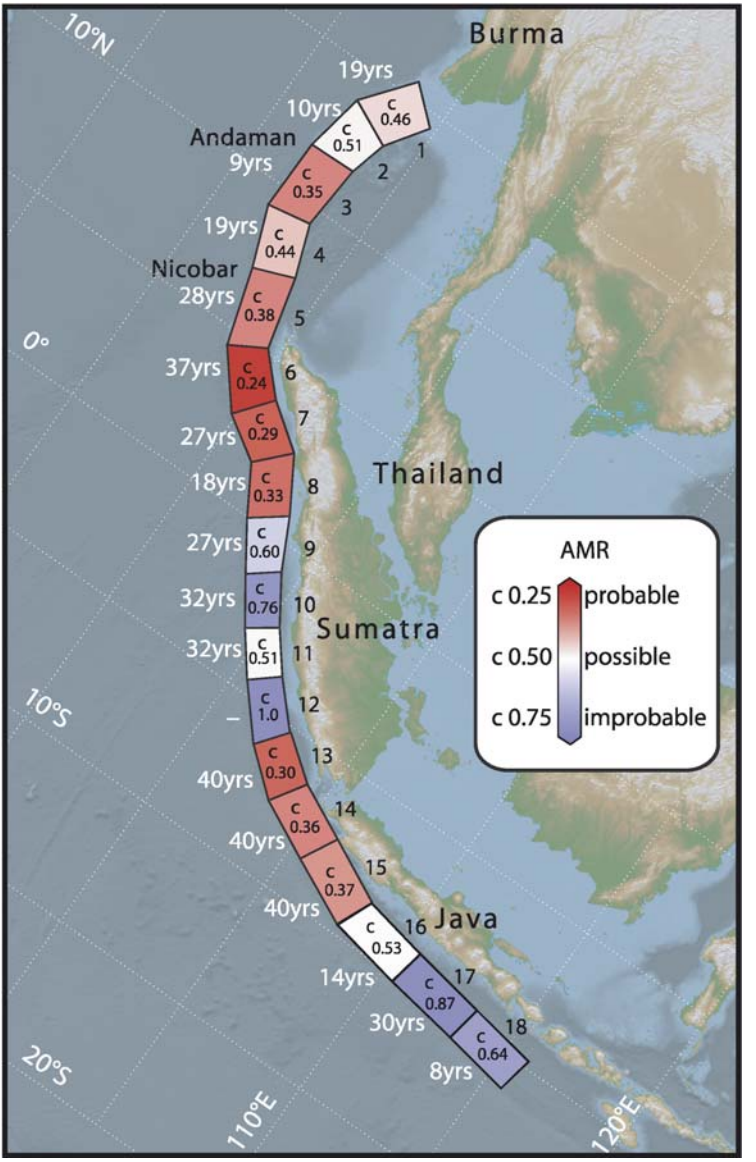


FIG. 4.10. AMR for the Sumatra–Java arc. For each segment the c-value is indicated and shaded in accordance with the colour bar in the inset. Beside each segment the length of time over which AMR is identified is indicated in years. The segments are numbered and the corresponding c-values, duration of AMR and probability that the AMR is non-random is shown in Table 4.1. The co-ordinates of the ends of the segments are also provided. Each segment has a dip of 11°, a width of 200 km and is pure dip slip with a nominal displacement of 20 metres.

TABLE 4.1. Segment characteristics. The fault segments all dip at 11°, a width of 200 km and are pure dip-slip. The regional stress field is chosen accordingly (see text). Since AMR is a far field effect it is not very sensitive to the fault parameters. The reliability is determined from the *c*-value alone using random catalogues (see Appendix). A high reliability and a long total time suggest that the AMR indicates that the segment is loaded. A long total time has no significance if the reliability is low

Segment	Seg start Lat°	Seg start Lon°	Seg end Lat°	Seg end Lon°	*c*-value	Reliability (%)	Total time (yr)
1 (slow rupture associated with the 26 December 2004 event)	93.48	15.58	92.03	13.85	0.46	84	19
2 (slow rupture associated with the 26 December 2004 event)	92.04	13.85	91.60	11.69	0.51	76	10
3 (slow rupture associated with the 26 December 2004 event)	91.60	11.69	91.53	9.17	0.35	98	9
4 (slow rupture associated with the 26 December 2004 event)	91.53	9.17	92.37	7.01	0.44	88	19
5 (26 December 2004 event)	92.37	7.01	92.86	4.28	0.38	96	28
6 (26 December 2004 event)	92.86	4.28	94.55	2.92	0.24	99	37
7 (26 December 2004 event)	95.55	2.92	96.48	1.33	0.29	98	27
8 (28 March 2005 event)	96.48	1.33	97.30	−0.13	0.33	97	18
9 (1861 event)	97.30	−0.13	98.26	−1.59	0.60	50	27
10 (1833 event)	98.26	−1.59	99.36	−3.19	0.76	23	32
11 (1833 event)	99.36	−3.19	100.53	−4.99	0.51	70	32
12 (1833 event)	100.53	−4.99	101.97	−6.63	1.0	0	–
13 (history unknown)	101.98	−6.63	103.76	−7.93	0.30	98	40
14 (history unknown)	103.76	−7.93	105.68	−8.74	0.36	97	40
15 (history unknown)	105.68	−8.74	108.28	−9.86	0.37	96	40
16 (history unknown)	108.28	−9.86	111.18	−10.21	0.53	76	14
17 (history unknown)	111.19	−10.21	114.81	−10.61	0.87	20	30
18 (history unknown)	114.81	−10.61	118.38	−10.56	0.64	60	8

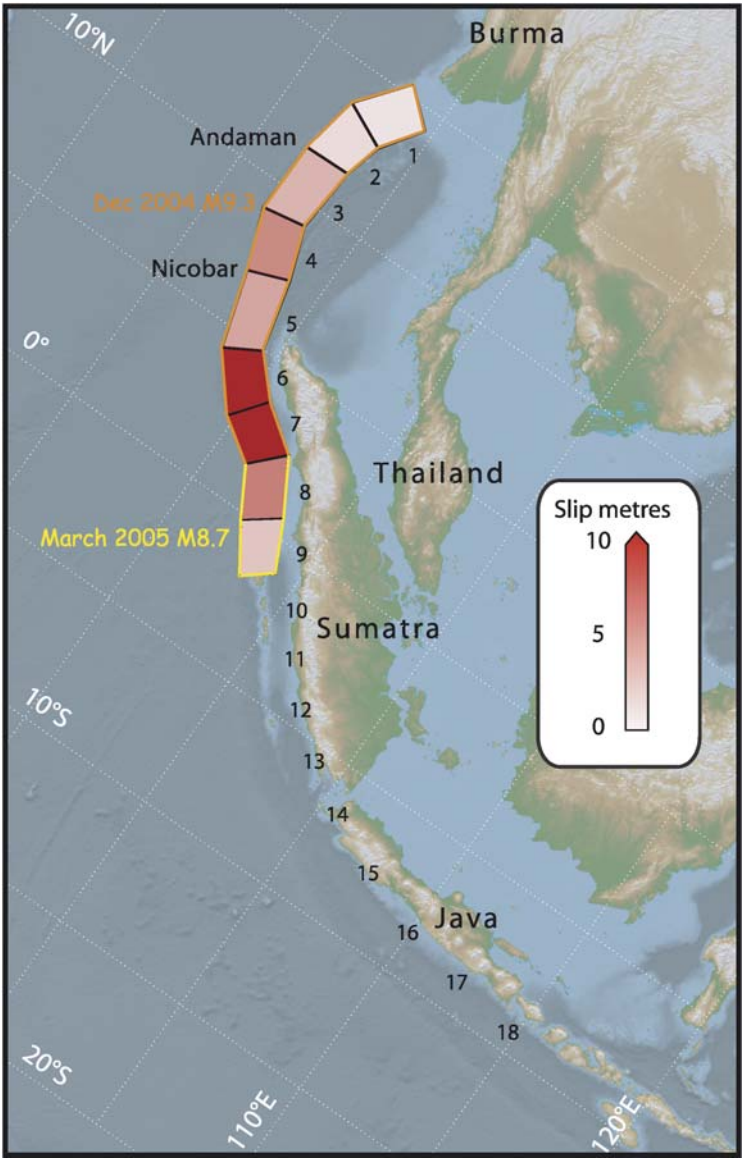


FIG. 4.11. Simplified slip distribution for the ruptures of the 26 December 2004 and 28 March 2005 earthquakes (modified from Ammon et al. (2005)). The displacements are shown in meters. Grey shading also represents slip with a range from 0 m – white to 10 m – grey.

simple model can intervene. The model supposes that the earthquake cycle consists of one simple loading system, where the only elements are creep at depth and the seismic motion of adjacent faults. At the end of the cycle, the stress conditions are identical to those at the beginning; the assumptions of a simple elastic rebound model. Elastic rebound concerns only horizontal stresses and displacements, however some permanent deformation (topography) is created at the end of the earthquake cycle for a dip-slip fault (King et al., 1988; Stein et al., 1988; Cattin and Avouac, 2000). The flexure, uplift, and subsidence associated with dip-slip earthquakes are associated with their own seismicity, which perturbs that expected for an idealised earthquake cycle. In more general terms, complex fault geometries result in the superposition of the stress fields and associated seismicity of multiple faults at different stages in their loading cycles. It is important to appreciate that the strain energy involved in these processes is very small and localised near to the earthquake epicentral region. The strain energy stored prior to, and then released in the main shock, is much greater. For Sumatra geodetic displacements of 1 or 2 centimetres occurred in India and Indonesia indicating the release of stress in a region larger than those in which we identify AMR.

Other major faults or plate boundaries can be associated with stress fields that are not responsible for the future earthquake. For example, the loading of the strike-slip fault behind the arc and the associated seismicity involves stress fields that cannot result in the observed dip-slip motion. Since at present such seismicity cannot be separated, events associated with loading of the strike-slip fault could be a source of error in the AMR plot for reverse faulting events. A similar situation occurs for western Java where the Banda and Philippine arcs and associated seismicity are sufficiently close that they overlap with regions that must be stressed prior to an eastern Java event.

While these sources of error reduce our ability to identify AMR they also offer future approaches to improving signal to noise ratios. The most straightforward would be to use only precursory events with identified focal mechanisms. Unfortunately, at present there are few parts of the world where seismic catalogues both include mechanisms and are homogeneous. The stress field prior to an event also varies with depth in a manner that can readily be calculated to improve the results. Unfortunately, most earthquake catalogs have poor depth resolution, rendering such an approach ineffective. While the limitations of existing catalogues must be accepted, future improvements may allow much improved AMR studies.

Prior to the work of King and Bowman (2003), it was assumed that AMR was correctly modelled by Eq. (4.5), based on the assumption that the events prior to the earthquake were part of either a failure process or a process analogous to a phase transition. The stress accumulation model (King and Bowman, 2003), in which creep at depth progressively fills a stress shadow, requires that events are a symptom of the loading process and are not a part of the failure leading to the main event. There is consequently no reason to suppose that the main event should occur at the asymptote of the failure function. The use of a power law

time-to-failure function (Eq. (4.5)) is used in this study since the filling of the stress shadow cannot be reduced to a simple analytic expression. In the future it may be possible to generate more appropriate empirical expressions for the AMR. However, in the absence of a clear alternative, Eq. (4.5) is retained on the understanding that the asymptote has no physical significance. It follows that since we do not know at present how much acceleration must be observed before an earthquake occurs, a clear time prediction is not possible.

The results presented above suggest that part of the southern Sumatra and western Java arc is now subject to similar loading conditions to those that prevailed prior to the Sumatra earthquakes of 26 December 2004 and 28 March 2005. The data presented is very suggestive, but would be more certain if we had more examples of AMR before large subduction earthquakes. Unfortunately although very large events have occurred (e.g Chile 1960, Alaska 1965) the NEIC catalogue is not long enough to provide the clear accelerations that span 10's of years. Where acceleration is identified over shorter time spans, it is more equivocal.

In the absence of many examples it is uncertain when the southern Sumatra and western Java arc might fail. Models suggest that the most pronounced AMR is within the last 10% of the earthquake cycle (which is perhaps between 200 and 500 years for the Sumatra–Java subduction system). AMR should therefore be identifiable in the last 20 to 50 years before the event. Thus, an AMR total time of 40 years might suggest that the event will occur in the next few years, but this is far from certain.

Various workers have proposed that some subduction systems are coupled and seismic, while some are uncoupled and largely aseismic. This has been explained by mechanical models that invoke parameters including (but not limited to) subduction rate, sediment load, and motion of the trench relative to the mantle as determining the seismic coupling (Ruff and Kanamori, 1983; Pacheco et al., 1993; Taylor et al., 1996). However, the seismic observations on which these models are based cover about 100 years, which may not be a sufficiently long time period for the characterisation of slow subduction systems (Taylor et al., 1996). For both the Sumatra (north and central) and Cascadia subduction systems, hitherto regarded as uncoupled, geological investigations have already demonstrated that this is incorrect (Simoes et al., 2004; McCaffrey and Goldfinger, 1995; Clague, 1997; Chlieh et al., 2004). It is therefore pertinent to question the implications of AMR in a subduction setting that has previously been considered to be uncoupled, such as southern Sumatra and western Java. Certainly the progressive build-up of stress implied by the stress accumulation model (King and Bowman, 2003) is not consistent with steadily creeping subduction. However, recent observations of transient strain events in subduction settings (e.g. Miller et al., 2002) and theoretical models (Liu and Rice, 2005) although concerning much smaller scales might suggest that the current period of AMR could be precursory to a non-destructive, slow earthquake.

4.2.6. Conclusion

Parts of the Sumatra–Java arc exhibited clear Accelerated Moment Release (AMR) prior to the 26 December 2004 and 28 March 2005 earthquakes. On the basis of a simple loading model of the type widely accepted to explain geodetic observations (Fig. 4.8 and caption), we propose a model in which the AMR provides evidence that sections of the arc were approaching failure prior to the catastrophic events. The method that we use has the potential to directly estimate absolute stress (relative to failure stress), a significant advance on coulomb stress change models (King and Cocco, 2000). If demonstrated to be robust, an ability to independently identify the latter part of the seismic cycle will enhance seismic gap methods of establishing earthquake risk.

The epicentres of both events occurred where AMR, and hence loading, was greater and propagated into regions where loading was less. To the south of the slip regions of the recent events, the epicentral area of the 1833 earthquake shows no evidence for an impending earthquake. Recent studies of corals suggest that this region is nearing the end of the earthquake cycle (Natawidjaja et al., submitted for publication), but the event could still not be close enough for AMR to be observable yet. However further to the south and east, parts of the Java–Sumatra arc appear to be loaded and may be capable of hosting a devastating earthquake (Fig. 4.12). In the absence of further studies in different regions, a possibility limited by the length of the available seismic catalogues, we cannot be certain that our results truly indicate an imminent earthquake danger for southeastern Sumatra and western Java. However, our conclusions rest on a simple physical/mechanical model. Bearing in mind the consequences, such an event should be taken seriously.

4.2.7. Appendix

Seismic data: The seismicity catalogs required for examining AMR must be homogeneous. Except for a few places, such as California, this can only be satisfied using the global catalogue based on the NEIC (ANSS) data set. The catalogue used here was taken from the Advanced National Seismic System (ANSS) Worldwide Earthquake Catalog, accessible from the Northern California Earthquake Data Center (http://quake.geo.berkeley.edu/anss), and extended from 01/01/1965 to 24/12/04. The minimum magnitude used was 4.5. This was chosen by examining the frequency-magnitude statistics of the catalog to define the minimum magnitude where a stable Gutenberg–Richter relation is observed (Wiemer and Wyss, 2000) (e.g. Fig. 4.A1). Because progressive improvements in the global seismic network have resulted in a progressive lowering of the detection threshold of small events, this step is necessary to remove apparent AMR caused by changing catalog completeness.

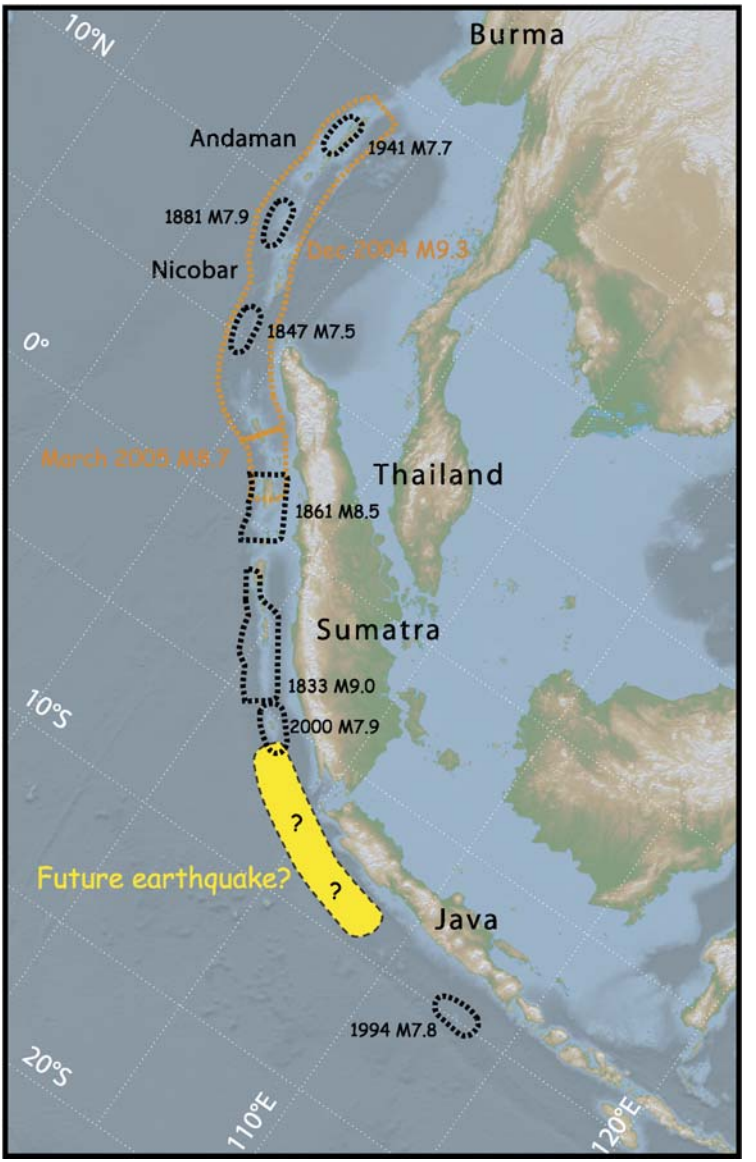


FIG. 4.12. The region where observed AMR suggests that the subduction system is approaching the end of the seismic cycle (light grey).

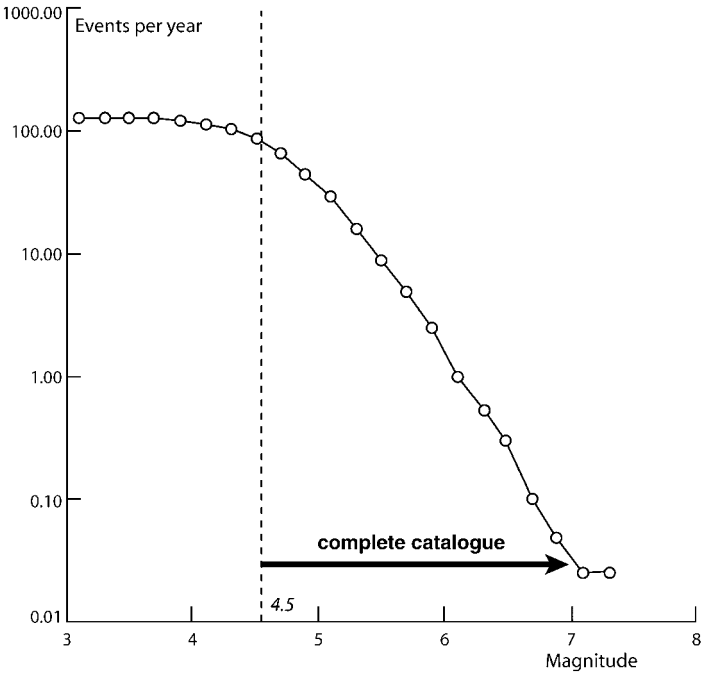


FIG. 4.A1. The Gutenberg–Richter plot for the data used to calculate AMR along the Sumatra–Java arc. The catalogue is complete above M 4.5. Subsets of the catalogue in time and space have also been examined to establish this throughout the region.

Only events with depths from the surface to 40 km were used. This is chosen to include those events with depths fixed at 33 km. The stress calculation was carried out for a depth of 8 km. Since the AMR is associated with seismicity distant from the future epicentre this is sufficiently accurate for the range of earthquake depths considered. Deeper events, in particular those associated with the lower part of the lithosphere and the subducting slab, may exhibit changes associated with the build up of load on the future fault (Astiz et al., 1988; Dmowska et al., 1988; Lay et al., 1989; Dmowska et al., 1996) but would require a three dimensional stress model and a more complicated search of the stress-time grid. This will be investigated in future work.

c-value and statistics: A measure of the reliability of the acceleration in a certain spatiotemporal area in a catalogue of seismicity is carried out by calculating the *c*-value, which is the ratio between the root mean square of a power law fit and the root mean square of a linear fit (Bowman et al., 1998). If the *c*-value tends to zero, the curve is best approximated by a power law and if the *c*-value tends to the unity, the curve can be best approximated by a straight line. Because the

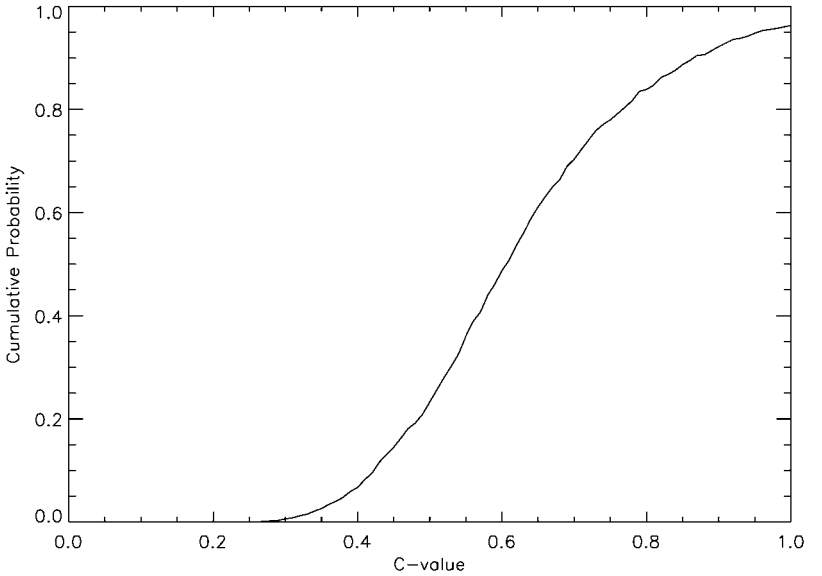


FIG. 4.A2. The probability of different c-values occurring in a random catalogues.

method searches for acceleration data sets may exhibit AMR that has no physical origin. To help determine whether an acceleration is likely to be "real" we have examined random catalogues of synthetic seismicity.

The Monte Carlo method is used to test the probability that AMR can be found due to random patterns in seismicity (the "false-alarm rate"). 1000 catalogues each composed of 500 events placed randomly in space and time are created. Each point is indexed to a magnitude that respects the Gutenberg–Richter law. Energy for each point is calculated using the magnitude-energy scaling law ($E = 10^{4.8+1.5M}$) (Kanamori and Anderson, 1975). This is then used to generate the cumulative Benioff strain curve. When real data is examined, the shape of the spatial region examined depends on the fault mechanism of the earthquake being examined. While there is no obvious reason why shape should affect the results, we test the false-alarm rate for circular, square and triangular patterns.

For each geometrical pattern, the search is done using a spatiotemporal matrix (c-value $= f$ [pattern, distance to the center of the pattern, starting time]). In detail, for the circle example, events are localised in a unit square [0, 0; 0, 1; 1, 0; 1, 1] and the centre of the circle corresponds to the center of the synthetic catalog. The calculation of c-value is first done by looking at events located in a circle of minimum radius (fixed parameter). For these chosen points, another search is done in time by increasing the value of the starting time of the cumulative energy curve (from 0.00 to 0.90 – fixed parameters). This temporal search is repeated several times for the spatial search from the minimum radius to the maximum radius (from 0.05 to 0.50 – fixed parameters). When the c-value is

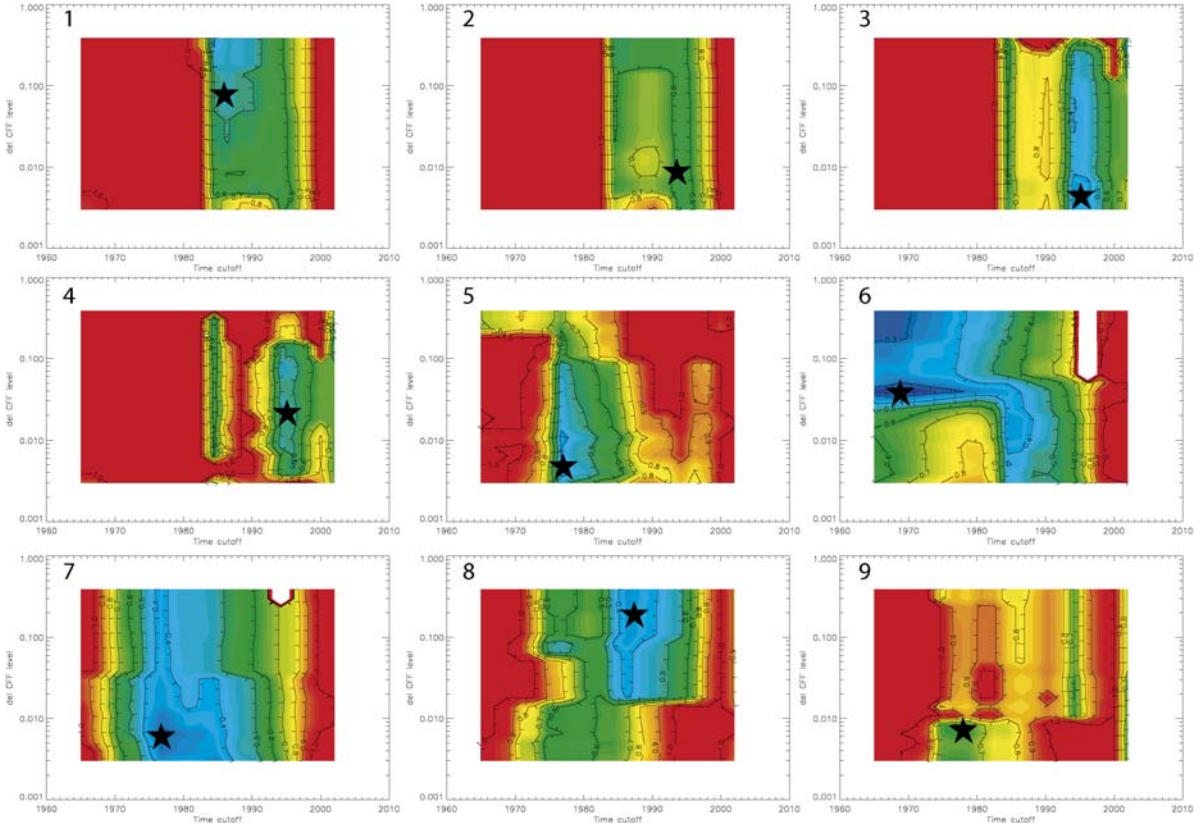


FIG. 4.A3. Start-time versus stress grids for the correspondingly numbered segments shown in the figures in the main text.

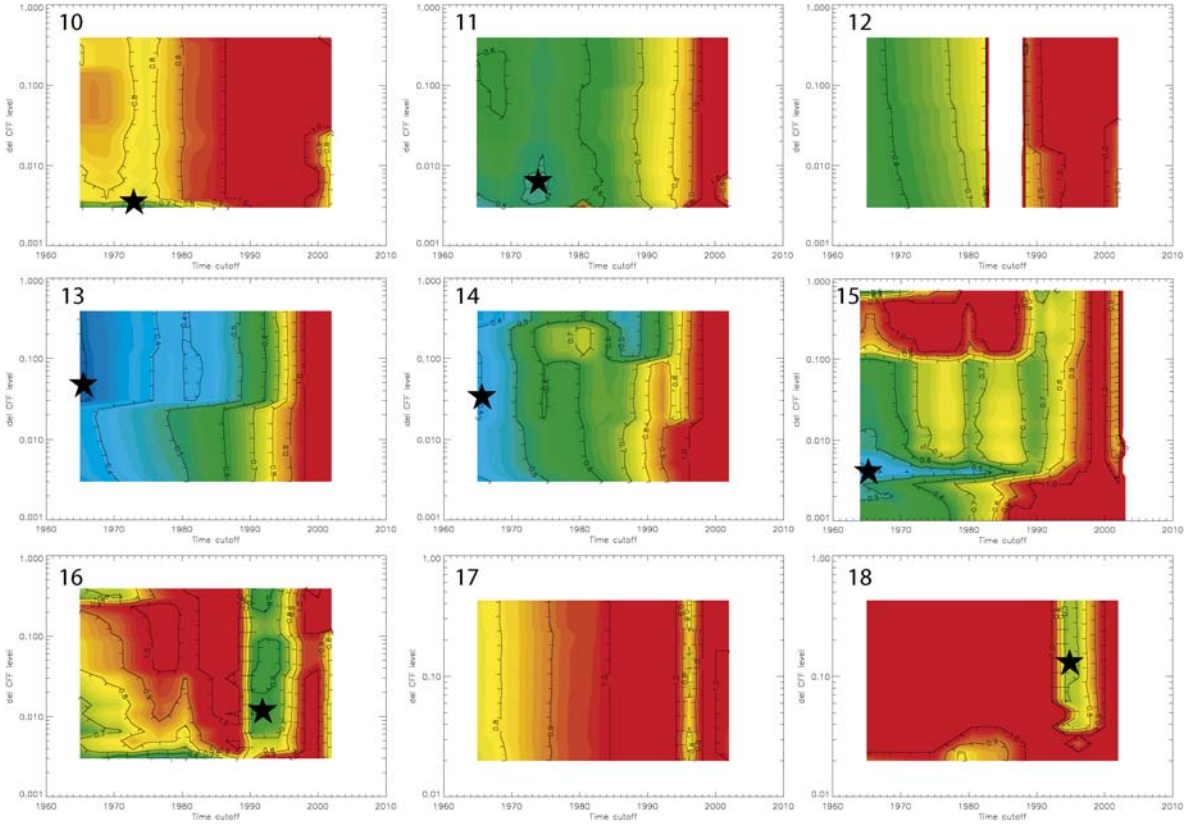


FIG. 4.A3. Continued.

calculated for all increments, the best one is kept (minimum c-value). This is repeated for each synthetic catalogue and a distribution of minimum c-values is then obtained. The search using other geometrical patterns adopts the same method by systematically changing the size of the geometrical figure. The relation between the probability of occurrence and c-value is the same for the three geometries tested, and is shown in Fig. 4.A2.

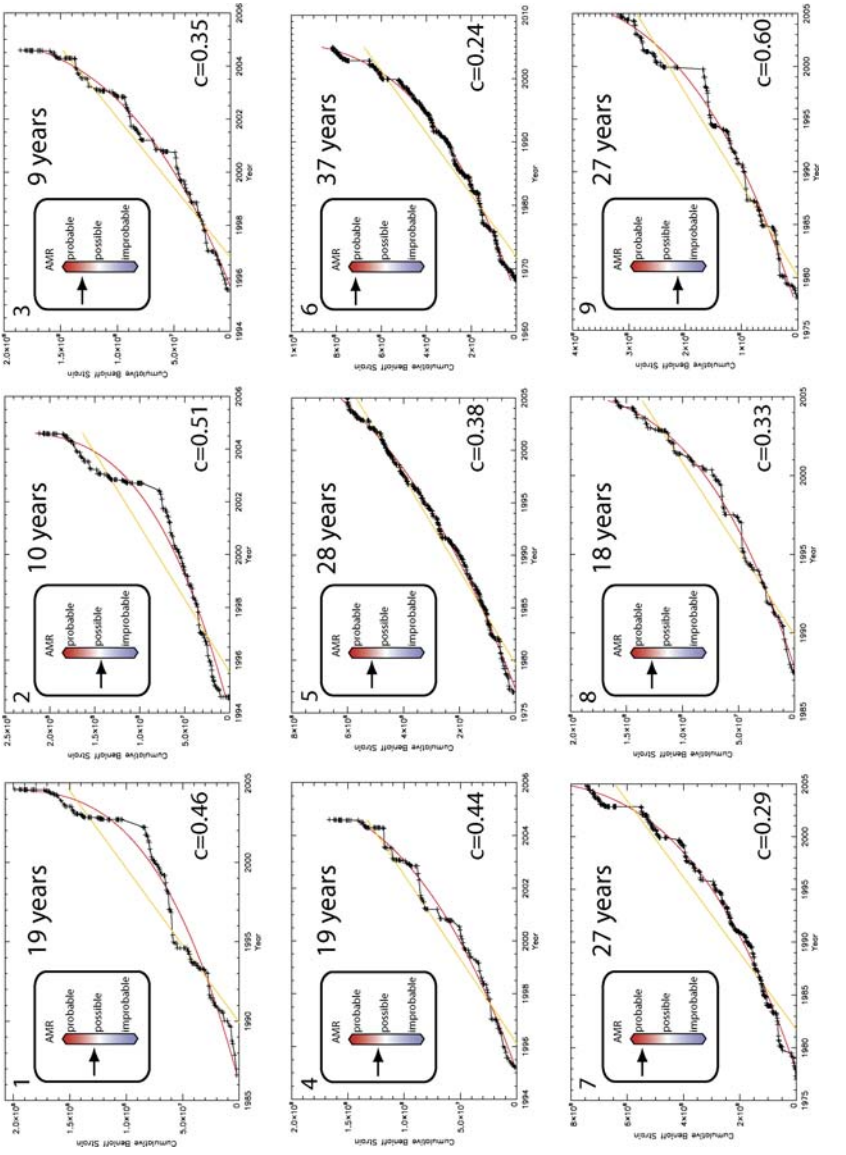
From the figure it can be concluded that low c-values (< 0.35) are very unlikely to occur by chance. However, once a c-value of 0.6 is reached the probability that it could result from random processes is about 50%. In a separate study, different ways of generating random catalogues have been adopted (Ikeda, 2004). These include randomising the times of real catalogues, using real epicentral distributions and assigning random earthquakes to them, and establishing catalogues that adopt the ETAS method to simulate aftershocks (Ogata, 1988, 1992). For c-values of 0.6 the probability that the observed AMR can be a random effect can increase to 55%. However for c-values less than 0.35 the probabilities are approximately equal to those in Fig. 4.A2. We consequently conclude that c-values less than 0.35 have a high probability of being real and those of 0.6 have little significance. As discussed in the text, a low c-value is a necessary condition for determining whether an AMR is real. However, it is not a sufficient condition.

Results for all segments: The start-time versus stress grid for all of the segments is shown in Fig. 4.A3. The vertical axis shows the Coulomb Stress level and the horizontal axis the time period. For each combination of Coulomb stress and time range, the plots indicate the c-value of the associated AMR plot. Light grey indicate high c-values and dark grey low c-values. The c-values are also contoured. The best c-value is provided by choosing a dark grey region. In each start-time versus stress grid, stars indicate the point chosen for the corresponding AMR plot in Fig. 4.A4. In some cases an apparently low c-value can be of little significance. In Fig. 4.A3 for example, the dark grey regions marked bad for segment 1 result from a search very close to the future fault slip; too few events are included to produce a convincing plot. It is important to appreciate that the reliable low c-values for segments 6, 7, 8, 14 and 15 are unambiguous.

4.3. Stability and Reliability of Accelerating Moment Release

The quantification of accelerating moment release by the c-value and its reliability have been defined previously (see Section 2 and the previous Section). However different studies show a certain variability of the c-value probability curve

FIG. 4.A4. The AMR plots chosen to be representative for the correspondingly numbered segments shown in the figures in the main text. Information for each segment is provided in Table 1 of the main text. The insets indicate the reliability of each plot. See the main text.


1
19 years
c=0.46
2
10 years
c=0.51
3
9 years
c=0.35
4
19 years
c=0.44
5
28 years
c=0.38
6
37 years
c=0.24
7
27 years
c=0.29
8
18 years
c=0.33
9
27 years
c=0.60
AMR
probable
possible
improbable
Cumulative Benioff Strain
Year

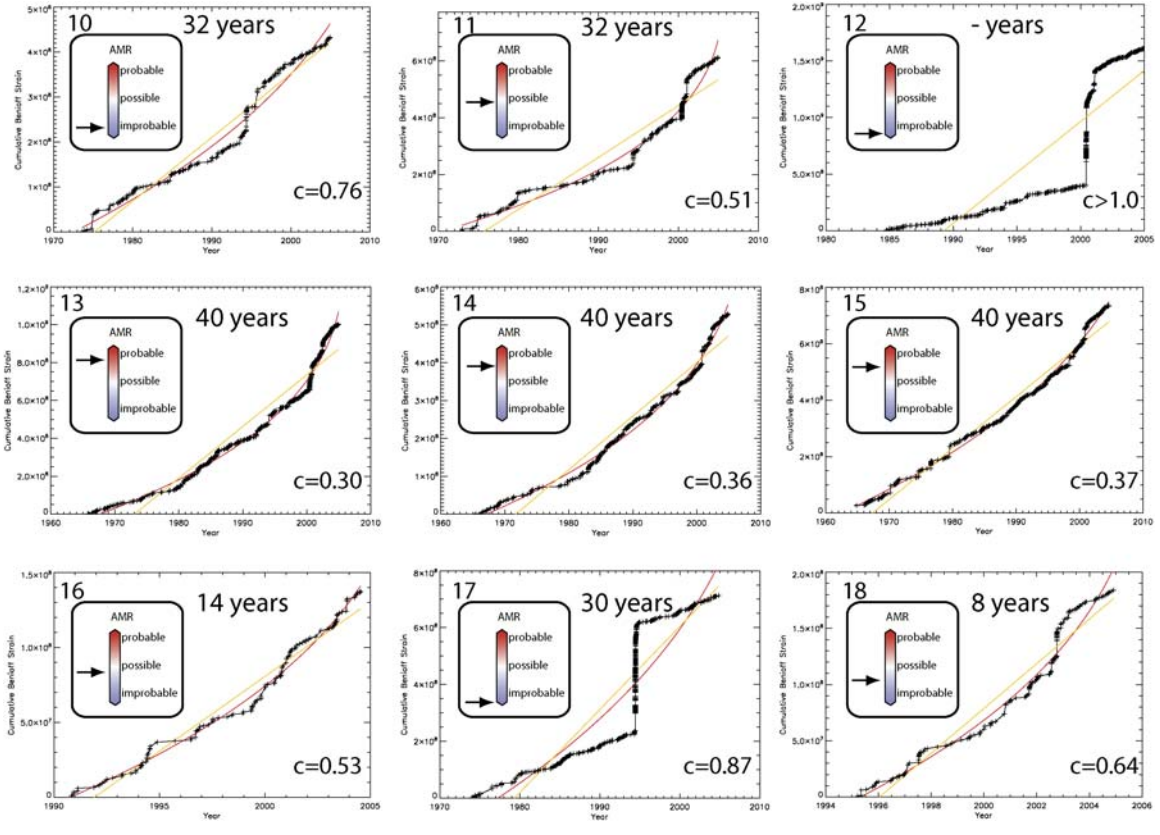


FIG. 4.A4. Continued.

depending of the method employed (e.g. Bowman et al., 1998; Zoller et al., 2001; Hardebeck et al., in preparation; Mignan et al., 2006a). In this section, a variety of approaches is first presented and compared to determine the sensitivity of the parameters employed. Because some differences cannot be explained, a new parameter (signal size) is taken into account in the calculation of the reliability of accelerating moment release. This gives a better understanding of the differences observed in the *c*-value probability curves. Moreover it gives a more accurate determination of the reliability for future forecasting studies.

4.3.1. Previous Studies

A region exhibiting accelerating seismicity has an increased likelihood of a large earthquake which can be termed an "alarm". The identification of such alarms can then be used in forecasting large earthquakes. However, there always exists a finite probability that the observed accelerating moment release in a region is due to random fluctuations in the background seismicity rate, leading to a false alarm. Therefore a *c*-value probability curve must be determined, which needs to define (1) the behaviour of synthetic catalogues and (2) the optimisation procedure.

Simulation of seismicity catalogues: Ordinary seismic activity corresponds to periods when there are no peaks of activity perturbing the regional seismicity. Once large earthquakes and associated events (e.g. aftershocks) are removed, ordinary seismicity is composed of random events and clusters in both space and time (e.g. Zhuang et al., 2005). A possible representation is the Epidemic-Type Aftershock Sequence model (e.g. Ogata, 1989; Ogata, 1998, see Section 1). However some studies pose the null hypothesis of spatially clustered, temporally random seismicity (e.g. McKernon and Main, 2005). Some others describe seismicity as uniformly random in space and time (e.g. Bowman et al., 1998; Mignan et al., 2006a). The distribution of events on a fractal fault network is also sometimes considered (e.g. Helmstetter et al., 2005).

Table 4.2 summarises the different random event distributions that can be used in time and space. The temporal distribution can be random uniform (Poisson

TABLE 4.2. Possible random event distributions in time and space in synthetic seismicity catalogues

Temporal distribution	Spatial distribution
random uniform	random uniform
ETAS	fractal
	cluster-linked
	real cluster

process) or aftershocks can be added (e.g. Zoller et al., 2001; Hardebeck et al., in preparation). Figure 4.13 represents possible spatial distribution of seismicity events. In the uniform distribution, events are randomly uniformly distributed in the space window (Fig. 4.13a). In the fractal distribution, events are localised in a Sierpinski triangle (e.g. Saleur et al., 1996, Fig. 4.13b). In the linked-cluster distribution, inter-event distances are normally distributed (Fig. 4.13c). It corresponds to a modified "random walk" that will produce pseudo-Brownian motion. It can be noted that some studies use a real cluster distribution of events from real seismicity catalogues (e.g. Gross and Rundle, 1998; Zoller et al., 2001; Hardebeck et al., in preparation).

The choice of the spatiotemporal distribution of events that compose synthetic catalogues has consequences for the calculation of the c-value probability curve. However, Ikeda (2004) shows that differences are small (see Fig. 4.14).

Optimisation procedure: The Monte Carlo method is used to calculate the c-value probability curve from a large number of synthetic seismicity catalogues. The higher the number of used catalogues, the more stable is the result. The optimisation procedure employed must be the same as the one used for real data searches. This can explain some important differences in c-value probability curves. For example, probabilities are very low in the study made by Bowman et al. (1998) compared to other ones (e.g. Zoller et al., 2001; Mignan et al., 2006a) because the time window was fixed whereas more recent studies let the starting time be free.

4.3.2. Signal Size and Stability of Accelerating Moment Release

The differences observed in c-value probability curves cannot be explained only by the choice of the behaviour of simulated seismicity and of the optimisation procedure. Indeed, some important differences are observed in different methods that use the same kind of spatiotemporal optimisation (e.g. Zoller et al., 2001; Hardebeck et al., in preparation; Mignan et al., 2006a).

Minimum number of events in the searched pattern: The Monte Carlo method is used for 1000 catalogues composed of 1000 events. Events are random uniform in space and time and respect the Gutenberg–Richter relation. The space window is optimised for 10 radii and the time window for 10 starting times. For each case, the c-value is calculated only if the number of events in the selected window is greater than a fixed value (5, 20 or 100 events) otherwise the c-value is fixed at 2.

c-value probability curves are represented in Fig. 4.15. They show that the probabilities change dramatically depending on the minimum number of events that compose the signal. This can explain why Hardebeck et al. (in preparation) suggest that accelerating moment release is an artifact of data-fitting and does not represent a real precursory process. The c-value probability curve that Hardebeck et al. (in preparation) propose can be reproduced if accelerating patterns can be

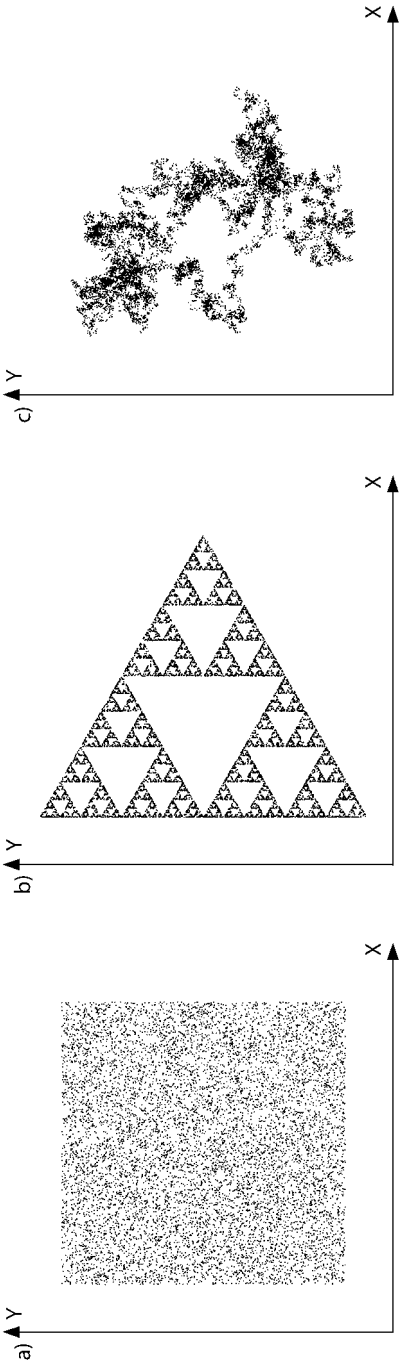


FIG. 4.13. Possible spatial distributions for a synthetic catalogue composed of 10000 events: (a) uniform distribution; (b) fractal distribution; (c) cluster-linked distribution.

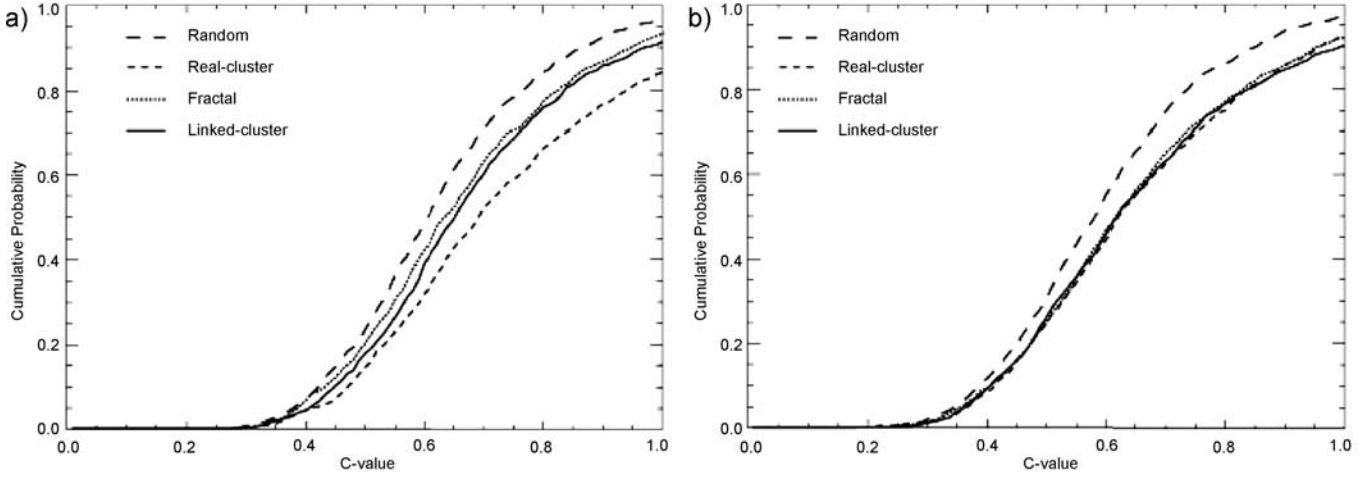


FIG. 4.14. *c*-value probability curves for different spatiotemporal distributions of events; (a) for random uniform time; (b) for ETAS time. The different distributions essentially produce similar results. Redrawn from Ikeda (2004).

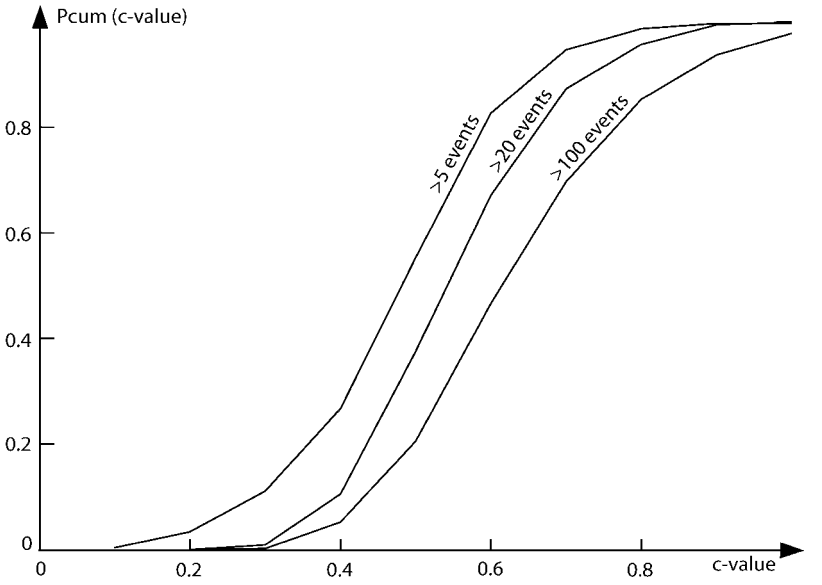


FIG. 4.15. *c*-value probability curves for 1000 catalogues composed of 1000 events. The minimum signal-size authorised is 5, 20 and 100 events. Probabilities change dramatically depending of the minimum signal size authorised (i.e. number of events that compose the signal).

composed of as few as 5 events. In the work of Mignan et al. (2006a) and Mignan et al. (2006b), the minimum number of events authorised is 15, which shifts the *c*-value probability curve to higher *c*-values.

Signal-size distribution: The differences observed in *c*-value probability curves can be easily explained by the fact that 5 random points are more likely to form a power-law curve than 15 points or more. Figure 4.16 shows the signal-size distribution for 1000 catalogues composed of 1000 events. Examples of accelerating pattern (c-value $= 0.4$) are shown for signals composed of 5, 50 and 400 events. The highest probability ($P \sim 0.4$) is obtained for small patterns (composed of $\sim$5 events) and decreases very fast to stabilise at a very low probability ($P < 0.01$) for signals composed of more than 200 events.

For reasons of stability, the search for accelerating moment release should define at first the minimum size authorised and then calculate the probabilities in consequence. It can be noted that accelerating moment release found in real cases have variable "sizes" depending of the methodology employed (see examples in Table 4.3). In the studies of Mignan et al. (2006a) and Mignan et al. (2006b), the largest signals are preferred to smaller ones for reasons of stability. This is not the case in other studies (e.g. Bufe et al., 1994; Bowman et al., 1998; Robinson, 2000) where the number of events that compose the signal is not a parameter taken into account in the reliability. The differences observed for Cali-

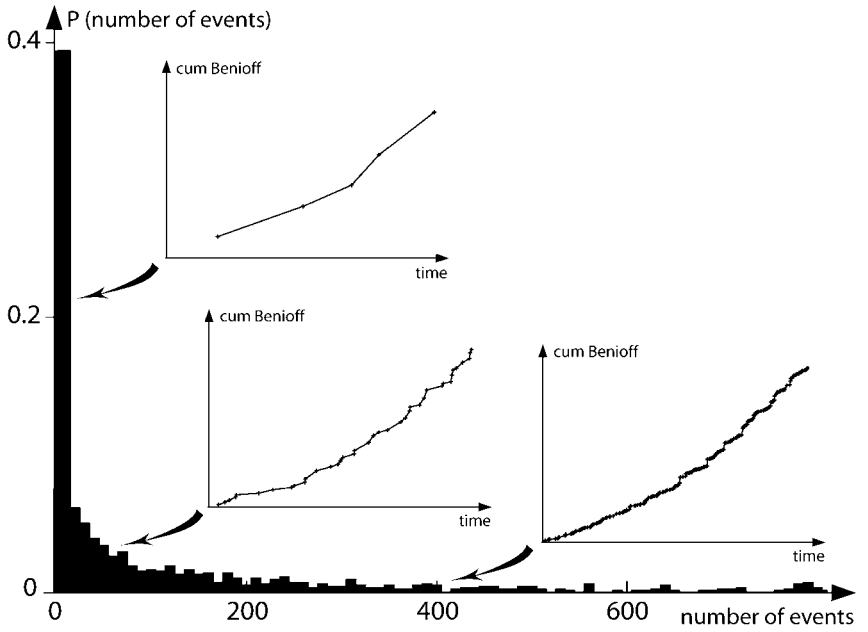


FIG. 4.16. Signal-size distribution for 1000 catalogues composed of 1000 events. Examples of accelerating pattern (c-value $= 0.4$) are shown for signals composed of 5, 50 and 400 events. The highest probability is obtained for small patterns (composed of $\sim$ 5 events) and decreases very fast to stabilise at a very low probability for signals composed of more than 200 events.

TABLE 4.3. Examples of signal-size for accelerating moment release observed in real seismicity catalogues

Region	Signal-size (# events)	Reference
California	30*	Bufe and Varnes (1993)
Alaska	20* (6–25)	Bufe et al. (1994)
California	40* (10–100)	Bowman et al. (1998)
New Madrid Seismic Zone	25* (15–40)	Brehm and Braile (1998)
Aegean region	70* (40–90)	Papazachos and Papazachos (2000)
New Zealand	20* (15–30)	Robinson (2000)
Sumatra	350	Mignan et al. (2006a)
California	250 (20–800)	Mignan et al. (2006b)

*Number of events approximated from figures.

fornia data between Bowman et al. (1998) and Mignan et al. (2006b) can also be explained by different magnitude cutoffs.

To summarise, this study shows that the number of events that compose the signal is of a major importance and should therefore be discussed in future studies. Also it shows that accelerating moment release composed of hundreds of events

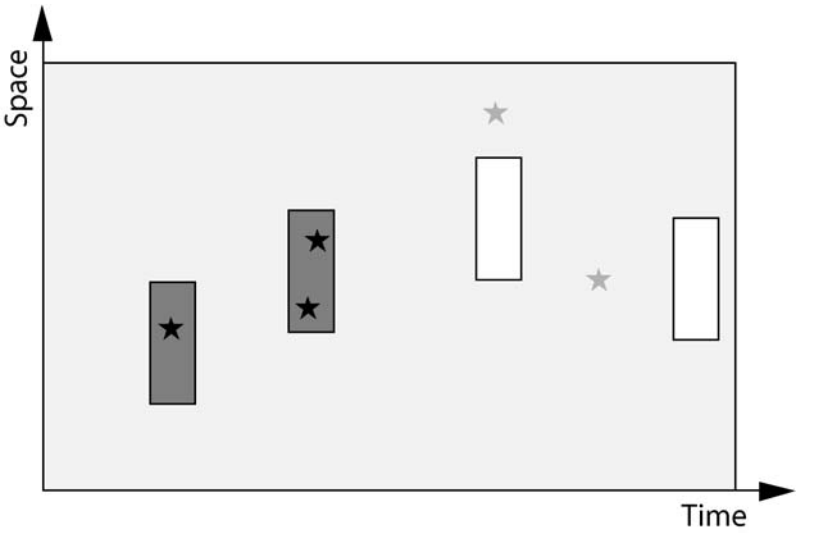


FIG. 4.17. Possible outcomes of prediction. For simplicity, the territory where the prediction is made is represented by a 1D space axis. Rectangles are space–time areas covered by correct (grey) and false (white) alarms. Black stars are predicted earthquakes and grey stars, failures to predict. Modified from Shebalin et al. (2006).

is very unlikely to occur randomly ($> 99\%$ of chance that it is a real precursory process).

4.4. Predictability of Earthquakes Using Accelerating Moment Release

Earthquake prediction is commonly defined as specifying the location, time, and magnitude of an impending earthquake within specified ranges (Fig. 4.17). It is important to distinguish between deterministic earthquake predictions and statistical forecasts. Predictions which are usually specified as deterministic or quasi-deterministic statements, are much more specific, implying considerably higher probability and more certainty than a forecast (Kagan and Jackson, 2006).

The "seismic risk map" presented in the case of the Sumatra–Java arc (Section 4.2) can be considered as a forecasting map. However, it does not directly determine the time window of the future large earthquake, but just compares the relative risk on different fault segments and gives an idea of the possible magnitude. The Stress Accumulation model only gives information about the degree of loading of a fault that is at the end part of its seismic cycle but is it possible in the future to use the Stress Accumulation model to improve forecasting techniques? A discussion about the space-, time-, and magnitude-prediction is made below to try to answer this question.

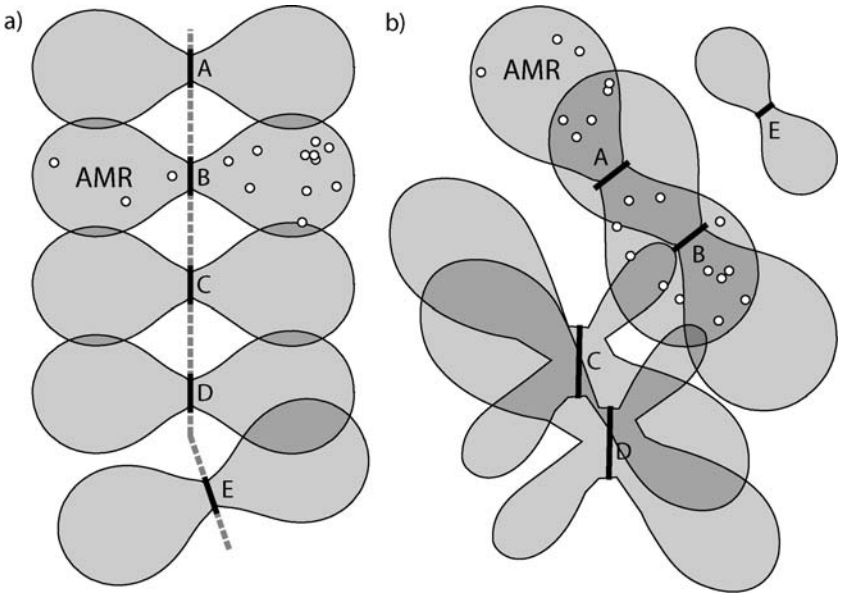


FIG. 4.18. Overlapping of stress lobes in different fault networks; (a) Schematic representation of the Sumatra–Java Arc subduction zone. If a fault segment is approaching failure (B), accelerating moment release will be observed essentially in the stress lobe associated with this fault segment; (b) Schematic representation of the southern California fault network. If a fault is approaching failure (A), accelerating moment release will be possibly observed in several stress lobes associated with different faults (approaching failure, A or not, B).

4.4.1. Space-Predictable?

The importance of the Stress Accumulation model is that it explains precursory seismicity by the behaviour of the stress field around the fault that is going to fail. By consequence, forecasting methods should focus on each fault that is susceptible to be the origin of a new large earthquake (i.e. space window = fault segment).

However the spatial distribution of accelerating moment release predicted by the Stress Accumulation model (see Section 2, Mignan et al., 2006b) shows that the stability of results depends on the geometry of active faults. A forecasting map of the Sumatra–Java Arc (Mignan et al., 2006a) can be proposed because the stress lobe associated with each segment of the fault system only overlaps with the nearby segments (Fig. 4.18a). This is also true for some parts of the San Andreas fault (see Fig. 4.4). In contrast, in the case of Southern California, it is difficult to create a forecasting map only based on accelerating moment release. Stress lobes of different faults can overlap and it is not possible to determine which fault will be the one to fail (Fig. 4.18b).

A solution to determine which fault is more likely to approach failure is to take into account other information: other forecasting techniques or geodetic data. Figure 4.19 shows an example of possible present seismic risk in the region of the Black Water fault in California (c-value = 0.47, Fig. 4.19a). It shows that several faults (e.g. Garlock, Panamint, Black Water, Harper, Helendale, Lenwood, Lockhart, Emerson, Calico faults) could be at the origin of accelerating seismicity because they are located in a large region of precursory seismicity (Fig. 4.19b). However, radar interferometry (InSAR) shows clear interseismic displacements between both sides of the Black Water fault (Peltzer et al. (2001), Fig. 4.19c). This complementary information suggests that this is the Black Water fault which is at the origin of accelerating seismicity.

4.4.2. Time-Predictable?

The mathematical formulation of the time-to-failure power-law obtained in Section 3 (Mignan et al., 2007) should permit the accelerating moment release fit analysis to be improved by determining the loading rate. However, the time-to-failure analysis (see Section 4.1) shows variable results and parameters for the accelerating moment release equation can have large uncertainties (e.g. loading rate and stress noise).

Mignan et al. (2007) show that the spatial distribution of precursory seismicity must be well established. This will allow the signal to be identified earlier by eliminating more of the background seismicity (see Fig. 3.10). New forecasting methods should discuss these different problems and determine a statistical method to determine time probabilities.

4.4.3. Magnitude-Predictable?

The magnitude expected for a future event can be determined from the length of the fault segment. It can also be noted that the size of the region where accelerating moment release is observed is proportional to the magnitude of the future large earthquake (which can be explained by the relation to the size of the stress lobes).

Bowman et al. (1998) showed that, for all 12 California earthquakes having magnitude $M \geqslant 6.2$, a region size could be found which optimised the accelerating moment release signal. Moreover, they have found that the logarithm of the radius R of these optimal critical regions scaled as the magnitude M

$$\log R \propto KM \tag{4.7}$$

with $K = 0.44$ the scaling parameter. Similar studies have been carried out in California (Zoller et al., 2001) and in the New Madrid Seismic Zone (NMSZ) (Brehm and Braile, 1998), and a compilation of these studies has been done by Jaumé and Sykes (1999). Other studies have been made outside the USA, such as

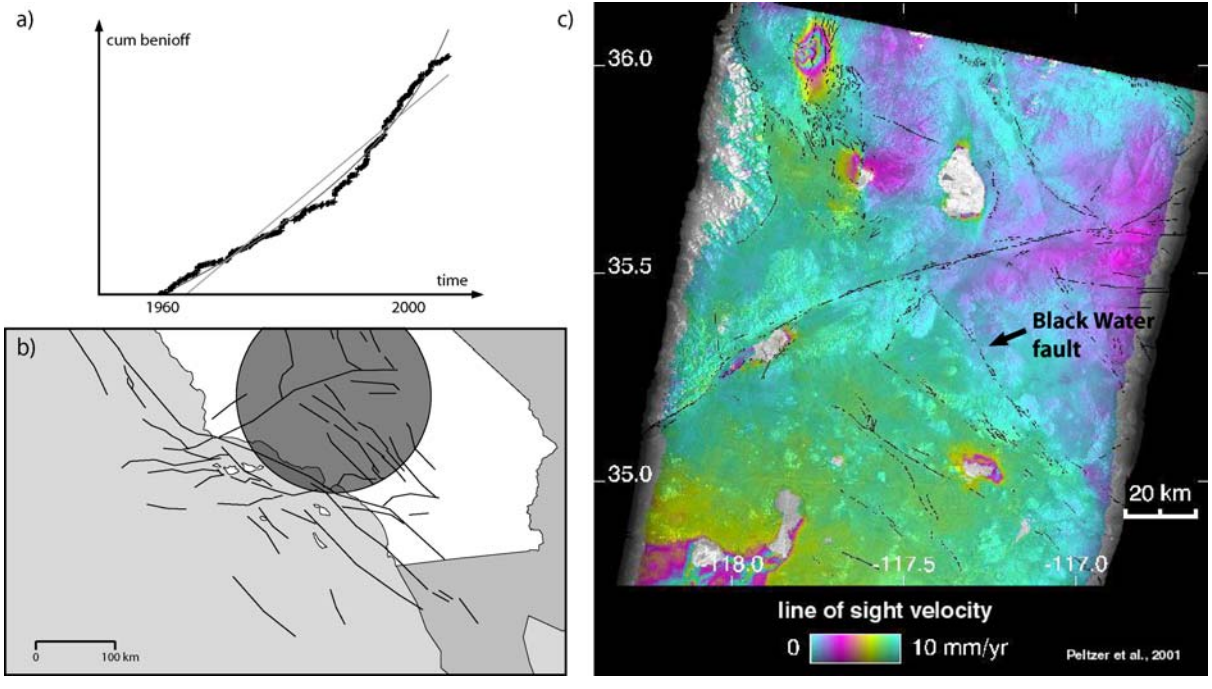


FIG. 4.19. Example of possible present seismic risk in the region of the Black Water fault in California; (a) accelerating moment release from 1960 to 2006 (c-value = 0.47); (b) Spatial distribution of accelerating moment release simplified by a circle (radius > 100 km). Stress lobes of different active faults can overlap the region of possible precursory seismicity (e.g. Garlock, Panamint, Black Water, Harper, Helendale, Lenwood, Lockhart, Emerson, Calico faults); (c) Interseismic displacement between the two sides of the Black Water fault, determined from InSAR, redrawn from Peltzer et al. (2001).

TABLE 4.4. Scaling relation between accelerating moment release critical radii and mainshocks magnitude

Scaling K	Location	Reference
0.44	California (USA)	Bowman et al. (1998)
0.75	NMSZ (USA)	Brehm and Braile (1998)
0.36	California & NMSZ (USA)	Jaumé and Sykes (1999)
0.41	Aegean area	Papazachos and Papazachos (2000)
0.70	California (USA)	Zoller et al. (2001)

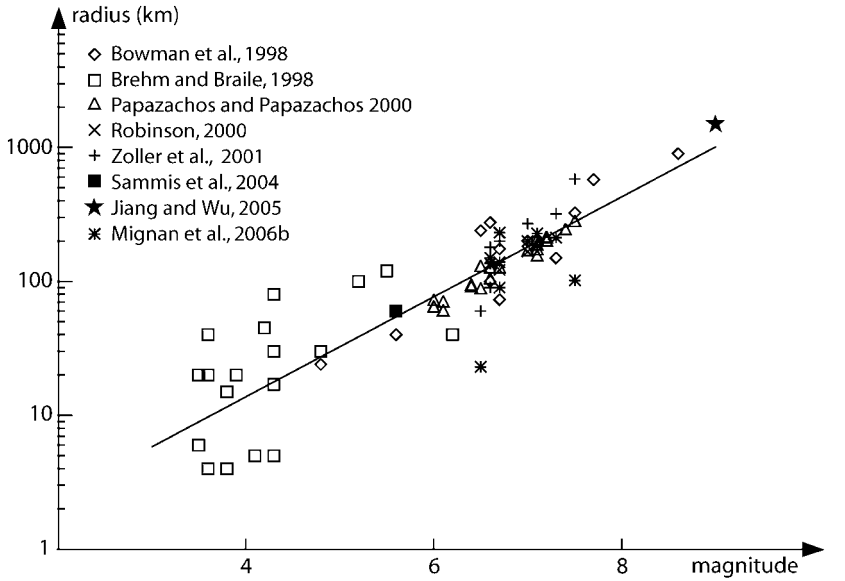


FIG. 4.20. Radius-magnitude scaling law. Compilation from different studies (see Table 4.5).

in the Aegean region (Papazachos and Papazachos, 2000). They are presented in Table 4.4.

Sammis and Smith (1999) note that the scaling relation found by Bowman et al. (1998) is compatible with the scaling of preparatory region with magnitude found by Keilis-Borok and Rotwain (1990) using the algorithm M8. Using the compilation of previous results and new ones obtained in Section 2 (Table 4.5), the new scaling relation is

$$\log R \propto 0.37M \tag{4.8}$$

TABLE 4.5. List of earthquakes with the magnitude/radius relation from different studies

Date yr/mo/day	Location	M_w	R km	Reference
1980-02-14	18°43′N,°39′W	$M_w = 4.8$	52	Varnes and Bufe (1996)
1995-10-31	37.00°N, 89.40°W	$M_w = 6.2$	40	Brehm and Braile (1998)
1968-11-09	38.00°N, 88.50°W	$M_w = 5.5$	120	Brehm and Braile (1998)
1975-06-13	36.54°N, 89.68°W	$M_w = 4.3$	30	Brehm and Braile (1998)
1980-07-05	36.60°N, 89.58°W	$M_w = 3.6$	4	Brehm and Braile (1998)
1980-12-02	36.21°N, 89.43°W	$M_w = 3.8$	4	Brehm and Braile (1998)
1983-05-15	38.77°N, 89.57°W	$M_w = 4.3$	80	Brehm and Braile (1998)
1984-02-14	37.21°N, 89.00°W	$M_w = 3.6$	20	Brehm and Braile (1998)
1985-12-05	35.90°N, 90.00°W	$M_w = 3.5$	20	Brehm and Braile (1998)
1986-08-26	38.32°N, 89.79°W	$M_w = 3.6$	40	Brehm and Braile (1998)
1987-06-10	38.71°N, 87.95°W	$M_w = 5.2$	100	Brehm and Braile (1998)
1987-06-13	36.58°N, 89.74°W	$M_w = 4.1$	5	Brehm and Braile (1998)
1987-09-29	36.95°N, 89.16°W	$M_w = 4.3$	17	Brehm and Braile (1998)
1987-10-14	37.05°N, 88.78°W	$M_w = 3.8$	15	Brehm and Braile (1998)
1989-04-27	36.09°N, 89.78°W	$M_w = 4.3$	5	Brehm and Braile (1998)
1990-09-26	37.15°N, 89.61°W	$M_w = 4.8$	30	Brehm and Braile (1998)
1991-07-07	36.69°N, 91.57°W	$M_w = 3.9$	20	Brehm and Braile (1998)
1993-01-08	35.93°N, 90.04°W	$M_w = 3.5$	6	Brehm and Braile (1998)
1994-02-05	37.37°N, 89.19°W	$M_w = 4.2$	45	Brehm and Braile (1998)
1994-08-20	36.14°N, 91.06°W	$M_w = 3.5$	20	Brehm and Braile (1998)
1952-07-21	Kern County	$M_w = 7.5$	325 ± 75	Bowman et al. (1998)
1992-06-28	Landers	$M_w = 7.3$	150 ± 15	Bowman et al. (1998)
1989-10-18	Loma Prieta	$M_w = 7.0$	200 ± 30	Bowman et al. (1998)
1994-01-17	Northridge	$M_w = 6.7$	73 ± 17	Bowman et al. (1998)
1983-05-02	Coalinga	$M_w = 6.7$	175 ± 10	Bowman et al. (1998)
1971-02-09	San Fernando	$M_w = 6.6$	100 ± 20	Bowman et al. (1998)
1987-11-24	Superstition Hills	$M_w = 6.6$	275 ± 95	Bowman et al. (1998)
1968-04-08	Borrego Mountain	$M_w = 6.5$	240 ± 60	Bowman et al. (1998)
1950-08-15	Assam	$M_w = 8.6$	900 ± 175	Bowman et al. (1998)
1906-04-18	San Fransisco	$M_w = 7.7$	575 ± 240	Bowman et al. (1998)
1986-07-08	Palm Spring	$M_w = 5.6$	40 ± 5	Bowman et al. (1998)
1980-02-14	Virgin Islands	$M_w = 4.8$	24 ± 2	Bowman et al. (1998)
1948-02-09	35.50°N, 27.20°E	$M_w = 7.1$	190	Papazachos and Papazachos (2000)
1952-12-17	34.40°N, 24.50°E	$M_w = 7.0$	171	Papazachos and Papazachos (2000)
1953-03-18	40.02°N, 27.53°E	$M_w = 7.4$	246	Papazachos and Papazachos (2000)
1953-08-12	38.30°N, 20.80°E	$M_w = 7.2$	216	Papazachos and Papazachos (2000)
1954-04-30	39.28°N, 22.29°E	$M_w = 7.0$	171	Papazachos and Papazachos (2000)
1956-07-09	36.64°N, 25.96°E	$M_w = 7.5$	284	Papazachos and Papazachos (2000)
1957-04-25	36.55°N, 28.80°E	$M_w = 7.2$	213	Papazachos and Papazachos (2000)
1967-03-04	39.20°N, 24.60°E	$M_w = 6.6$	106	Papazachos and Papazachos (2000)
1968-02-19	39.50°N, 25.00°E	$M_w = 7.1$	157	Papazachos and Papazachos (2000)
1970-03-28	39.20°N, 29.50°E	$M_w = 7.1$	175	Papazachos and Papazachos (2000)
1972-05-04	35.10°N, 23.60°E	$M_w = 6.5$	89	Papazachos and Papazachos (2000)
1976-05-11	37.40°N, 20.40°E	$M_w = 6.5$	131	Papazachos and Papazachos (2000)
1979-04-15	42.00°N, 19.00°E	$M_w = 7.1$	207	Papazachos and Papazachos (2000)

(*continued on next page*)

TABLE 4.5. Continued.

Date yr/mo/day	Location	M_w	R km	Reference
1981-02-24	38.07°N, 23.00°E	$M_w = 6.7$	128	Papazachos and Papazachos (2000)
1981-12-19	39.00°N, 25.26°E	$M_w = 7.2$	203	Papazachos and Papazachos (2000)
1983-01-17	38.10°N, 20.20°E	$M_w = 7.0$	171	Papazachos and Papazachos (2000)
1986-09-13	37.17°N, 22.09°E	$M_w = 6.0$	73	Papazachos and Papazachos (2000)
1992-04-30	34.99°N, 26.72°E	$M_w = 6.1$	71	Papazachos and Papazachos (2000)
1994-09-01	41.12°N, 21.25°E	$M_w = 6.1$	61	Papazachos and Papazachos (2000)
1995-05-13	40.15°N, 21.68°E	$M_w = 6.6$	127	Papazachos and Papazachos (2000)
1995-06-15	38.36°N, 22.28°E	$M_w = 6.4$	95	Papazachos and Papazachos (2000)
1997-10-13	36.33°N, 22.17°E	$M_w = 6.4$	92	Papazachos and Papazachos (2000)
1997-11-14	38.73°N, 25.91°E	$M_w = 6.0$	65	Papazachos and Papazachos (2000)
1997-11-18	37.58°N, 20.57°E	$M_w = 6.6$	106	Papazachos and Papazachos (2000)
1995-02-05	37.65°S, 179.49°E	$M_w = 7.0$	167	Robinson (2000)
1994-06-18	43.01°S, 166.71°E	$M_w = 6.7$	139	Robinson (2000)
1993-08-10	45.21°S, 177.46°E	$M_w = 6.7$	122	Robinson (2000)
1952-07-21	Kern County	$M_w = 7.5$	580	Zoller et al. (2001)
1968-04-08	Borrego Mountain	$M_w = 6.5$	60	Zoller et al. (2001)
1971-02-09	San Fernando	$M_w = 6.6$	90	Zoller et al. (2001)
1983-05-02	Coalinga	$M_w = 6.7$	200	Zoller et al. (2001)
1987-11-24	Superstition Hills	$M_w = 6.6$	180	Zoller et al. (2001)
1989-10-18	Loma Prieta	$M_w = 7.0$	270	Zoller et al. (2001)
1992-06-28	Landers	$M_w = 7.3$	320	Zoller et al. (2001)
1994-01-17	Northridge	$M_w = 6.7$	140	Zoller et al. (2001)
1999-10-16	Hector Mine	$M_w = 7.1$	180	Zoller et al. (2001)
2004-12-26	Sumatra	$M_w = 9.0$	1500	Jiang and Wu (2005)
1952-07-21	Kern County	$M_w = 7.5$	102	Mignan et al. (2006b)
1968-04-08	Borrego Mountain	$M_w = 6.5$	23	Mignan et al. (2006b)
1971-02-09	San Fernando	$M_w = 6.6$	150	Mignan et al. (2006b)
1983-05-02	Coalinga	$M_w = 6.7$	90	Mignan et al. (2006b)
1987-11-24	Superstition Hills	$M_w = 6.6$	136	Mignan et al. (2006b)
1989-10-18	Loma Prieta	$M_w = 7.0$	200	Mignan et al. (2006b)
1992-06-28	Landers	$M_w = 7.3$	212	Mignan et al. (2006b)
1994-01-17	Northridge	$M_w = 6.7$	231	Mignan et al. (2006b)
1999-10-16	Hector Mine	$M_w = 7.1$	228	Mignan et al. (2006b)

and is represented by Fig. 4.20. It can be simply explained in terms of stress transfer, by relating the magnitude of the future event to the maximum size of the stress lobes, which is in agreement with the Stress Accumulation model.

The Stress Accumulation model permits precursory seismicity to be better understood and thus forecasting may be improved. The spatial distribution of accelerating moment release can be specified, the time-to-failure power-law can be addressed in terms of fault loading and the magnitude can be evaluated from the size of the stress lobes. As a consequence, a discussion about the reliability of accelerating moment release and its relations to the future rupture segment is con-

cevable to determine if a forecast is possible, in which regions and under what conditions.

General Conclusions

> "Entities should not be multiplied beyond necessity – or – Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred."
>
> W. of Ockham, Occam's Razor principle (14th century).

The Stress Accumulation model seems to be the best explanation for the origin of accelerating moment release observed before large earthquakes.

(1) The spatial distribution of events that compose accelerating moment release fall mainly in regions of increased stress due to loading at depth on the fault that is going to fail (see Section 2).
(2) The power-law time-to-failure behaviour of accelerating moment release can be formulated in simple terms, with the loading rate on the fault (see Section 3).

This work has shown that the Stress Accumulation model can be defined with simple stress transfer considerations. Even if criticality concepts can match some aspects of the Stress Accumulation model, this is not needed. At present, a parallel can be made between accelerating moment release (debated and poorly understood) and aftershocks (accepted and quite well understood), which seem symmetrical about the mainshock time (acceleration/decceleration). Aftershocks are located in positive stress lobes of the stress field caused by the mainshock whereas accelerating moment release is located in positive stress lobes of the prestress field of the future mainshock. Aftershocks are due to the change of stress at the time of the mainshock and accelerating moment release is due to the change of stress during loading before the time of the mainshock. This view permits to better understand the possible behaviour of precursory seismicity by defining a simple pattern. This is in contrast with critical concepts where patterns emerge from chaos and are difficult to catch.

This has important consequences in earthquake forecasting. Indeed many precursors have been proposed, based on different mathematical or statistical tools around the criticality concepts. If accelerating moment release is simply defined by the Stress Accumulation model, forecasting is improved. Accelerating moment release gives a complementary information to the study of recurrence intervals by the determination of the degree of loading on the fault, as shown in the case of the Sumatra–Java Arc. However, the application of the Stress Accumulation model to earthquake forecasting also shows that a systematic use of accelerating moment

release is complex and depends of the studied region (e.g. fault network geometry). Also it shows that proper statistics are necessary to correctly determine the reliability of precursory patterns (see Section 4).

This work is a first step to a new understanding of precursory seismicity and to a new technique of large earthquake forecasting. However, time probabilities are not yet implemented in the Stress Accumulation model and future studies should concentrate on this problem.

ACKNOWLEDGMENTS

Sections 2.1 to 2.5 are the result of collaborative work with David D. Bowman (California State Univ., Fullerton), and Geoffrey C.P. King (Inst. Phys. Globe, Paris) (Mignan *et al*., 2006b). We are grateful for comments on the work by Russel Robinson, Kristy Tampo, and Steven Cohen, and for support by the French INSU-CNRS and the Southern California Earthquake Center, funded by USA NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008.

Sections 3.2.1 to 3.2.4 are the result of collaborative work with Geoffrey C.P. King and David D. Bowman (Mignan *et al*., 2007). We are grateful to Charles Sammis, John Rundle, Steven Cohen, and Renata Dmowska for helpful suggestions and critical comments, and for support by the same sources.

Sections 4.2.1 to 4.2.7 are the result of collaborative work with Geoffrey C.P. King, David D. Bowman, Robin Lacassin (Inst. Phys. Globe, Paris), and Renata Dmowska (Harvard Univ.) (Mignan *et al*., 2006a). We are grateful to Paul Tapponnier and James R. Rice for helpful suggestions and critical comments, to Roger Bilham for extensive commentary that substantially improved an early presentation of the results, and for support by the same sources.

The author acknowledges helpful reviews by Basil Papazachos and Charles Sammis. The author also thanks Geoffrey King, David Bowman, Renata Dmowska, Robert Muir-Wood and Raoul Madariaga for helpful suggestions and critical comments.

REFERENCES

Aki, K., Richards, P. (1980). Quantitative Seismology. W.H. Freeman and Co., San Francisco.

Allegre, C.J., Le Mouel, J.L., Provost, A. (1982). Scaling rules in rock fracture and possible implications for earthquake predictions. *Nature* **297**, 47–49.

Ammon, C.J., et al. (2005). Rupture process of the 2004 Sumatra–Andaman earthquake. *Science* **308**, 1133–1139.

Astiz, L., Lay, T., Kanamori, H. (1988). Large intermediate-depth earthquakes and the subduction process. *Phys. Earth Planet. Int.* **53**, 80–166, doi:10.1016/0031-9201(88)90138-0.

Bak, P., Paczuski, M. (1995). Complexity, contingency, and criticality. *Colloquium Proc. Natl. Acad. Sci. USA* **92**, 6689–6696.

Bak, P., Tang, C. (1989). Earthquakes as a self-organized critical phenomenon. *J. Geophys. Res.* **94**, 15635–15637.

Bak, P., Tang, C., Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the $1/f$ noise. *Phys. Rev. Lett.* **59**, 381–384.

Bak, P., Tang, C., Wiesenfeld, K. (1988). Self-organized criticality. *Phys. Rev. A* **38**, 364–374.

Bakun, W.H., McEvilly, T.V. (1984). Recurrence models and the Parkfield, California, earthquakes. *J. Geophys. Res.* **89**, 3051–3058.

Barka, A. (1999). The 17 August 1999 Izmit earthquake. *Science* **285**, 1858–1859.

Ben-Zion, Y., Lyakhovsky, V. (2002). Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. *Pure Appl. Geophys.* **159**, 2385–2412.

Benedetti, L., Finkel, R., Papanastassiou, D., King, G.C.P., Armijo, R., Ryerson, F.J., Farber, D., Flerit, F. (2002). Post-glacial slip history of the Sparta fault (Greece) determined by 36Cl cosmogenic dating: evidence for non-periodic earthquakes. *Geophys. Res. Lett.* **10**, doi:10.1029/2001GL014510.

Beroza, G.C., Zoback, M.D. (1993). Mechanism diversity of the Loma Prieta aftershocks and the mechanics of mainshock-aftershock interaction. *Science* **259**, 210–213.

Bilham, R. (2005). A flying start, then a slow slip. *Science* **305**, 1126–1127.

Blanter, E.M., Shnirman, M.G., Le Mouel, J.L. (1997). Hierarchical model of seismicity: scaling and predictability. *Phys. Earth Planet. Int.* **103**, 135–150.

Bowman, D.D., King, G.C.P. (2001a). Accelerating seismicity and stress accumulation before large earthquakes. *Geophys. Res. Lett.* **28**, 4039–4042, doi:10.1029/2001GL013022.

Bowman, D.D., King, G.C.P. (2001b). Stress transfer and seismicity changes before large earthquakes, Earth. *Planet. Sci. Lett.* **333**, 591–599.

Bowman, D., Sammis, C.G. (2004). Intermittent criticality and the Gutenberg–Richter distribution. *Pure Appl. Geophys.* **161**, 1945–1956.

Bowman, D., Ouillon, G., Sammis, C., Sornette, A., Sornette, D. (1998). An observational test of the critical earthquake concept. *J. Geophys. Res.* **103**, 24359–24372.

Bowman, D., Colella, H., Tiampo, K. (2006). Seismicity rate changes and earthquake forecasting beyond the San Andreas. 100th Anniversarry Earthquake Conference, USA.

Brehm, D.J., Braile, L.W. (1998). Intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone. *Bull. Seismol. Soc. Am.* **88**, 564–580.

Bufe, C.G., Varnes, D.J. (1993). Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *J. Geophys. Res.* **98**, 9871–9883.

Bufe, C.G., Nishenko, S.P., Varnes, D.J. (1994). Seismicity trends and potential for large earthquakes in the Alaska–Aleutian region. *Pure Appl. Geophys.* **142**, 83–99.

Carlson, J.M., Langer, J.S. (1989). Properties of earthquakes generated by fault dynamics. *Phys. Rev. Lett.* **62**, 2632–2635.

Cattin, R., Avouac, J.P. (2000). Modelling mountain building and the seismic cycle in the Himalaya of Nepal. *J. Geophys. Res.* **105**, 13389–14407.

Chlieh, M., Avouac, J., Sieh, K., Natawidjaja, D.H., Galetzka, J. (2004). Investigating lateral variations of interseismic strain along the Sumatran subduction zone. In: AGU Fall Meeting Abstracts, pp. C1385+.

Clague, J.J. (1997). Evidence for large earthquakes at the Cascadia subduction zone. *Rev. Geophys.* **35**, doi:10.1029/97RG00222.

Clark, J., Bowman, D., King, G., Sammis C. (2003). Accelerating moment release as a tool for seismic hazard assessment in southern California. In: EGS–AGU–EUG Joint Assembly, Abstracts from the meeting held in Nice, France, 6–11 April 2003, abstract #7422, pp. 7422-+.

Cocco, M., Nostro, C., Ekstrom, G. (2000). Static stress changes and fault interaction during the 1997 Umbria-Marche earthquake sequence. *J. Seismol.* **4**, 501–516.

Das, S., Scholz, C.H. (1981). Theory of time-dependent rupture in the earth. *J. Geophys. Res.* **86**, 6039–6051.

Deng, J., Sykes, L.R. (1997). Stress evolution in southern California and triggering of moderate-, small-, and micro-size earthquakes. *J. Geophys. Res.* **102**, 24411.

Dieterich, J. (1994). A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.* **99**, 2601–2618.

Dieterich, J., Kilgore, B. (1996). Implications of fault constitutive properties for earthquake prediction. *Proc. Natl. Acad. Sci. USA* **93**, 3787–3794.

Dmowska, R., Rice, J.R., Lovison, L.C., Josell, D. (1988). Stress transfer and seismic phenomena in coupled subduction zones during the earthquake cycle. *J. Geophys. Res.* **93**, 7869–7884.

Dmowska, R., Zheng, G., Rice, J.R. (1996). Seismicity and deformation at convergent margins due to heterogeneous coupling. *J. Geophys. Res.* **101**, 3015–3029.

Fedotov, S.A. (1968). The seismic cycle, quantitative seismic zoning, and long-term seismic forecasting. In: Medvedev, S. (Ed.), *Seismic Zoning of the USSR*, pp. 133–166.

Felzer, K.R., Brodsky, E.E. (2005). Testing the stress shadow hypothesis. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003277.

Freed, A.M. (2005). Earthquake triggering by static, dynamic and postseismic stress transfer. *Annu. Rev. Earth Planet. Sci.* **33**, 335–367.

Gabrielov, A., Zaliapin, I., Newman, W.I., Keilis-Borok, V.I. (2000). Colliding cascades model for earthquake prediction. *Geophys. J. Int.* **143**, 427–437.

Geller, R.J., Jackson, D.D., Kagan, Y., Mulargia, F. (1997). Earthquakes cannot be predicted. *Science* **275**, 1616–1617.

Gerstenberger, M.C., Wiemer, S., Jones, L.M., Reasenberg, P.A. (2005). Real-time forecasts of tomorrow's earthquakes in California. *Nature* **435**, 328–331.

Gomberg, J., Blanpied, M.L., Beeler, N.M. (1997). Transient triggering of near and distant earthquakes. *Bull. Seismol. Soc. Am.* **87**, 294–309.

Gomberg, J., Beeler, N.M., Blanpied, M.L., Bodin, P. (1998). Earthquake triggering by transient and static deformations. *J. Geophys. Res.* **103**, 24411–24426, doi:10.1029/98JB01125.

Gomberg, J., Belardinelli, M.E., Cocco, M., Reasenberg, P. (2005). Time-dependent earthquake probabilities. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003405.

Grasso, J.R., Sornette, D. (1998). Testing self-organized criticality by induced seismicity. *J. Geophys. Res.* **103**, 2965–2988.

Gross, S., Kisslinger, C. (1997). Estimating tectonic stress rate and state with Landers aftershocks. *J. Geophys. Res.* **102**, 7603.

Gross, S., Rundle, J. (1998). A systematic test of time-to-failure analysis. *Geophys. J. Int.* **133**, 57–64.

Gutenberg, B., Richter, C.F. (1956). In: *Seismicity of the Earth and Associated Phenomena*. Princeton University Press, p. 310.

Hainzl, S., Zoller, G., Kurths, J., Zshau, J. (2000). Seismic quiescence as an indicator for large earthquakes in a system of self-organized criticality. *Geophys. Res. Lett.* **27**, 597–600.

Hardebeck, J.L. (2004). Stress triggering and earthquake probability estimates. *J. Geophys. Res.* **109**, 4310, doi:10.1029/2003JB002437.

Hardebeck, J.L., Felzer, K.R, Michael, A.J. (in preparation). Rigorous observational tests contradict the accelerating moment release hypothesis.

Hardebeck, J.L., Nazareth, J.J., Hauksson, E. (1998). The static stress change triggering model: Constraints from two southern California aftershock sequences. *J. Geophys. Res.* **103**, 24427–24438, doi:10.1029/98JB00573.

Harris, R.A. (1998). Introduction to special section: Stress triggers, stress shadows, and implications for seismic hazard. *J. Geophys. Res.* **103**, 24347–24358.

Harris, R.A., Simpson, R.W. (1993). In the shadow of 1857; an evaluation of the static stress changes generated by the M8 Ft. Tejon, California earthquake. *EOS Trans. AGU* **74** (Fall Meet. Suppl.), 427.

Harris, R.A., Simpson, R.W. (1996). In the shadow of 1857; the effect of the great Ft. Tejon earthquake on subsequent earthquakes in Southern California. *Geophys. Res. Lett.* **23**, 229–232.

Harris, R.A., Simpson, R.W. (1998). Suppression of large earthquakes by stress shadows: A comparison of Coulomb and rate-and-state failure. *J. Geophys. Res.* **103**, 24439–24452, doi:10.1029/98JB00793.

Harris, R.A., Simpson, R.W. (2002). The 1999 Mw 7.1 Hector Mine, California, earthquake: A test for the stress shadow hypothesis? *Bull. Seismol. Soc. Am.* **92**, 1497–1512.

Harris, R.A., Simpson, R.W., Reasenberg, P. (1995). Influence of static stress changes on earthquake locations in Southern California. *Nature* **375**, 221–224.

Helmstetter, A. (2003). Is earthquake triggering driven by small earthquakes?. *Phys. Rev. Lett.* **91**, 058501, doi:10.1103/PhysRevLett.91.058501.

Helmstetter, A., Sornette, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J. Geophys. Res.* **107**, doi:10.1029/2001JB001580.

Helmstetter, A., Sornette, D., Grasso, J.R. (2003). Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws. *J. Geophys. Res.* **108**.

Helmstetter, A., Kagan, Y.Y., Jackson, D.D. (2005). Importance of small earthquakes for stress transfers and earthquake triggering. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003286.

Holliday, J.R., Rundle, J.B., Tiampo, K.F., Klein, W., Donnelan, A. (2006). Modification of the pattern informatics method for forecasting large earthquake events using complex eigenfactors. *Tectonophysics* **413**, 87–91.

Hough, S.E. (2002). Earthshaking Science. Princeton University Press.

Huang, Y., Saleur, H., Sammis, C., Sornette, D. (1998). Precursors, aftershocks, criticality and self-organized criticality. *Europhys. Lett.* **41**, 43–48.

Hubert-Ferrari, A., Barka, A., Jacques, E., Nalbant, S.S., Meyer, B., Armijo, R., Tapponnier, P., King, G.C.P. (2000). Seismic hazard in the Marmara Sea region following the 17 August 1999 Izmit earthquake. *Nature* **404**, 269–273.

Hubert-Ferrari, A., King, G., Manighetti, I., Armijo, R., Meyer, B., Tapponnier, P. (2003). Long-term elasticity in the continental lithosphere; Modelling the Aden Ridge propagation and the Anatolian extrusion process. *Geophys. J. Int.* **153**, doi:10.1046/j.1365-246X.2003.01872.x.

Ikeda, N. (2004). Testing the false alarm rate of accelerating moment release before large earthquakes. M.S. Thesis, Cal. State Univ. Fullerton.

Ishimoto, M., Ida, K. (1939). Observations of earthquakes registered with the microseismograph constructed recently. *Bull. Earthq. Res. Inst.* **17**, 443–478.

Jacques, E., King, G.C.P., Tapponnier, P., Ruegg, J.C., Manighetti, I. (1996). Seismic activity triggered by stress changes after the 1978 events in the Asal Rift, Djibouti. *Geophys. Res. Lett.* **88**, 2481–2484.

Jaumé, S.C., Estabrook, C.H. (1992). Accelerating seismic moment release and outer-rise compression: Possible precursors to the next great earthquake in the Alaska Peninsula region. *Geophys. Res. Lett.* **19**, 345–348.

Jaumé, S.C., Sykes, L.R. (1999). Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. *Pure Appl. Geophys.* **155**, 279–305.

Jiang, C.-S., Wu, Z.-L. (2005). The December 26, 2004, off the west coast of northern Sumatra, Indonesia, $Mw = 9.0$, earthquake and the critical-point-like model of earthquake preparation. *Acta Seismol. Sinica* **18**, 290–296.

Joswig, M. (2001). Mapping seismic quiescence in California. *Bull. Seismol. Soc. Am.* **91**, 64–81.

Kagan, Y.Y., Jackson, D.D. (2006). Comment on "Testing earthquake prediction methods: The West Pacific short-term forecast of earthquakes with magnitude MwHRV $\geqslant$ 5.8" by V.G. Kossobokov. *Tectonophysics* **413**, 33–38.

Kagan, Y.Y., Knopoff, L. (1987). Statistical short-term earthquake prediction. *Science* **236**, 1563–1567.

Kagan, Y.Y., Jackson, D.D., Liu, Z. (2005). Stress and earthquakes in southern California: 1850–2004. *J. Geophys. Res.* **110**, 1850–2004.

Kanamori, H., Anderson, D.L. (1975). Theoretical basis of some empirical relations in seismology. *Bull. Seismol. Soc. Am.* **65**, 1073–1095.

Karakaisis, G.F., Papazachos, C.B., Savvaidis, A.S., Papazachos, B.C. (2002). Accelerating seismic crustal deformation in the North Aegean Trough, Greece. *Geophys. J. Int.* **148**, 193–200.

Karakaisis, G.F., Papazachos, C.B., Scordilis, E.M., Papazachos, B.C. (2004). Current accelerating seismic excitation along the northern boundary of the Aegean Microplate. *Tectonophysics* **383**, 81.

Karakostas, V.G., Papadimitriou, E.E., Karakaisis, G.F., Papazachos, C.B., Scordilis, E.M., Vargemezis, G., Aidona, E. (2003). The 2001 Skyros, Northern Aegean, Greece, earthquake sequence: off-fault aftershocks, tectonic implications, and seismicity triggering. *Geophys. Res. Lett.* **30**, doi:10.1029/2002GL015814.

Karakostas, V.G., Papadimitriou, E.E., Papazachos, C.B. (2004). Properties of the 2003 Lefkada, Ionian Islands, Greece, earthquake seismic sequence and seismicity triggering. *Bull. Seismol. Soc. Am.* **94**, 1976–1981.

Katsumata, K., Kasahara, M. (1999). Precursory seismic quiescence before the 1994 Kurile earthquake ($Mw = 8.3$) revealed by three independent seismic catalogs. *Pure Appl. Geophys.* **155**, 443–470.

Keilis-Borok, V.I. (1996). Intermediate-term earthquake prediction. *Proc. Natl. Acad. Sci. USA* **93**, 3748–3755.

Keilis-Borok, V.I., Kossobokov, V.G. (1990). Premonitory activation of earthquake flow: algorithm M8. *Phys. Earth Planet. Int.* **61**, 73–83, doi:10.1016/0031-9201(90)90096-G.

Keilis-Borok, V.I., Rotwain, I.M. (1990). Diagnosis of time of increased probability of strong earthquakes in different regions of the world: algorithm CN. *Phys. Earth Planet. Int.* **61**, 57–72, doi:10.1016/0031-9201(90)90095-F.

Keilis-Borok, V.I., Ismail-Zadeh, A., Kossobokov, V., Shebalin, P. (2001). Non-linear dynamics of the lithosphere and intermediate-term earthquake prediction. *Tectonophysics* **338**, 247–260.

Kenner, S., Segall, P. (1999). Time-dependence of the stress shadowing effect and its relation to the structure of the lower crust. *Geology* **27**, 119–122.

Kerr, R.A. (1993). Parkfield quakes skip a beat. *Science* **259**, 1120–1123.

Kilb, D., Ellis, M., Gomberg, J., Davis, S. (1997). On the origin of diverse aftershock mechanisms following the 1989 Loma Prieta earthquake. *Geophys. J. Int.* **128**, 557–570.

King, G. (1983). The accommodation of large strains in the upper lithosphere of the earth and other solids by self-similar fault systems: the geometrical origin of b-value. *Pure Appl. Geophys.* **121**, 761–815.

King, G.C.P., Bowman, D. (2003). The evolution of regional seismicity between large earthquakes. *J. Geophys. Res.* **108**, doi:10.1029/2001JB000783.

King, G.C.P., Cocco, M. (2000). Fault interaction by elastic stress changes: new clues from earthquake sequences. *Adv. Geophys.* **44**, 1–38.

King, G.C.P., Stein, R.S., Rundle, J.B. (1988). The growth of geological structures by repeated earthquakes 1. Conceptual framework. *J. Geophys. Res.* **93**, 307–313.

King, G.C.P., Stein, R.S., Lin, J. (1994). Static stress changes and the triggering of earthquakes. *Bull. Seismol. Soc. Am.* **84**, 935–953.

Kisslinger, C. (1996). Aftershocks and fault-zone properties. *Adv. Geophys.* **38**, 1–36.

Knopoff, L., Levshina, T., Keilis-Borok, V., Mattoni, C. (1996). Increased long-range intermediate-magnitude earthquake activity prior to strong earthquakes in California. *J. Geophys. Res.* **101**, 5779–5796.

Kossobokov, V.G., Keilis-Borok, V.I., Smith, S.W. (1990). Localization of intermediate-term earthquake prediction. *J. Geophys. Res.* **95**, 19763–19772.

Kossobokov, V.G., Romashkova, L.L., Keilis-Borok, V.I., Healy, J.H. (1999). Testing earthquake prediction algorithms: statistically significant advance prediction of the largest earthquakes in the Circum-Pacific, 1992–1997. *Phys. Earth Planet. Int.* **111**, 187–196.

Lay, T., Astiz, L., Kanamori, H., Christensen, D.H. (1989). Temporal variation of large intraplate earthquakes in coupled subduction zones. *Phys. Earth Planet. Int.* **54**, 258–312, doi:10.1016/0031-9201(89)90247-1.

Lay, T., et al. (2005). The great Sumatra–Andaman earthquake of 26 December 2004. *Science* **308**, 1127–1133.

Levin, S.Z., Sammis, C.G., Bowman, D.D. (2006). An observational test of the stress accumulation model based on seismicity preceding the 1992 Landers, CA earthquake. *Tectonophysics* **413**, 39–52.

Lienkaemper, J.J., Galehouse, J.S., Simpson, R.W. (1997). Creep response of the Hayward fault to stress changes caused by the Loma Prieta earthquake. *Science* **276**, 2014–2016.

Lin, A., Stein, R.S. (2004). Stress triggering in thrust and subduction earthquakes and stress interactions between the southern San Andreas and nearby thrust and strike slip faults.

Liu, Y., Rice, J.R. (2005). Aseismic slip transients emerge spontaneously in 3D rate and state modeling of subduction earthquake sequences. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003424.

Lyakhovsky, V., Ben-Zion, Y., Agnon, A. (1997). Distributed damage, faulting, and friction. *J. Geophys. Res.* **102**, 27635–27650, doi:10.1029/97JB01896.

Ma, K.F., Chang, C.H., Stein, R.S. (2005). Response of seismicity of Coulomb stress triggers and shadows of the 1999 M (sub w) = 7.6 Chi-chi, Taiwan, earthquake. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003389.

Main, I.G., Meredith, P.G. (1989). Classification of earthquake precursors from a fracture mechanics model. *Tectonophysics* **167**, 273–283.

Matsuura, M., Jackson, D.D., Cheng, A. (1986). Dislocation model for aseismic crustal deformation at Hollister, California. *J. Geophys. Res.* **91**, 12661–12674.

Matthews, M.V., Ellsworth, W.L., Reasenberg, P.A. (2002). A Brownian model for recurrent earthquakes. *Bull. Seismol. Soc. Am.* **92**, 2233–2250.

McCaffrey, R., Goldfinger, C. (1995). Forearc deformation and great subduction earthquakes: Implications for cascadia offshore earthquake potential. *Science* **267**, 856–859.

McCloskey, J., Nalbant, S.S., Steacy, S. (2005). Earthquake risk from co-seismic stress. *Nature* **434**, 291.

McKernon, C., Main, I.G. (2005). Regional variations in the diffusion of triggered seismicity. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003387.

Meyer, B., Tapponnier, P., Bourjot, L., Métivier, F., Gaudemer, Y., Peltzer, G., Guo, S., Chen, Z. (1998). Crustal thickening in Gansu-Qinghai, lithospheric mantle subduction, and oblique, strike-slip controlled growth of the Tibet Plateau. *Geophys. J. Int.* **135**, 1–47.

Mignan, A., King, G.C.P., Bowman, D. (2007). A mathematical formulation of accelerating moment release based on the stress accumulation model. *J. Geophys. Res.* **112**, doi:10.1029/2006JB004671.

Mignan, A., King, G.C.P., Bowman, D., Lacassin, R., Dmowska, R. (2006a). Seismic activity in the Sumatra–Java region prior to the December 26, 2004 (Mw = 9.0–9.3) and March 28, 2005 (Mw = 8.7) earthquakes. *Earth. Planet. Sci. Lett.* **244**, 639–654.

Mignan, A., Bowman, D., King, G.C.P. (2006b). An observational test of the origin of accelerating moment release before large earthquakes. *J. Geophys. Res.* **111**, doi:10.1029/2006JB004374.

Miller, M., Melbourne, T., Johnson, D.J., Summer, W.Q. (2002). Periodic slow earthquakes from the Cascadia subduction zone. *Science* **295**, 2423.

Mogi, K. (1969). Some features of recent seismic activity in and near Japan (2), Activity before and after large earthquakes. *Bull. Earthquake Research Institute, University of Tokyo* **47**, 395–417.

Mogi, K. (1981). Earthquake prediction program in Japan. In: Simpson, D.W., Richards, P.G. (Eds.), *Earthquake Prediction: An International Review*. American Geophysical Union, Washington, DC, pp. 635–666.

Nalbant, S.S., Hubert, A., King, G.C.P. (1998). Stress coupling between earthquakes in northwest Turkey and the north Aegean Sea. *J. Geophys. Res.* **103**, 24469–24486, doi:10.1029/98JB01491.

Natawidjaja, D., Sieh, K., Galetzka, J., Suwargadi, B., Chlieh, M., Cheng, H., Edwards, R.L., Avouac, J.P., Ward, S. (submitted for publication). The giant Sumatran megathrust ruptures of 1797 and 1833: paleoseismic evidence from coral microatolls. *J. Geophys. Res*.

Newman, W.I., Phoenix, S.L. (2001). Time-dependent fiber-bundles with local load sharing. *Phys. Rev. E* **63**, 021507.

Nishenko, S.P., Bollinger, G.A. (1990). Forecasting damaging earthquakes in the Central and Eastern United States. *Science* **249**, 1412–1416.

Nishenko, S.P., Buland, R. (1987). A generic recurrence interval distribution for earthquake forecasting. *Bull. Seismol. Soc. Am.* **77**, 1382–1399.

Nostro, C., Chiaraluce, L., Cocco, M., Baumont, D., Scotti, O. (2005). Coulomb stress changes caused by repeated normal faulting earthquakes during the 1997 Umbria-Marche (central Italy) seismic sequence. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003386.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.* **83**, 9–27.

Ogata, Y. (1989). Statistical model for standard seismicity and detection of anomalies by residual analysis. *Tectonophysics* **169**, 159–174.

Ogata, Y. (1992). Detection of precursory relative quiescence before great earthquakes through a statistical model. *J. Geophys. Res.* **97**, 19845.

Ogata, Y. (1998). Space–time point-process models for earthquake occurrences. *Ann. Inst. Stat. Math.* **50**, 379–402.

Ogata, Y. (1999). Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. *J. Geophys. Res.* **104**, 17995–18014.

Ogata, Y. (2005). Detection of anomalous seismicity as a stress change sensor. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003245.

Ogata, Y., Zhuang, J. (2006). Space-time ETAS models and an improved extension. *Tectonophysics* **413**, 13–23.

Ogata, Y., Jones, L.M., Toda, S. (2003). When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *J. Geophys. Res.* **108**, 1–6, doi:10.1029/2002JB002009.

Omori, F. (1894). On the aftershocks of earthquakes. *Journal of Collegiate Science, Imperial University Tokyo* **7**, 111–200.

Ortiz, M., Bilham, R. (2003). Source area and rupture parameters of the 31 December 1881 $Mw = 7.9$ Car Nicobar earthquake estimated from tsunamis recorded in the Bay of Bengal. *J. Geophys. Res.* **108**, 2215–2230.

Pacheco, J.F., Sykes, L.R., Scholz, C.H. (1993). Nature of seismic coupling along simple plate boundaries of the subduction type. *J. Geophys. Res.* **98**, 14133.

Papazachos, B., Papazachos, C. (2000). Accelerated preshock deformation of broad regions in the Aegean area. *Pure Appl. Geophys.* **157**, 1663–1681.

Papazachos, B.C. (2003). Chaos in seismology and earthquake prediction.

Papazachos, B.C., Karakaisis, G.F., Savvaidis, A.S., Papazachos, C.B. (2002a). Accelerating seismic crustal deformation in the southern Aegean area. *Bull. Seismol. Soc. Am.* **92**, 570.

Papazachos, B.C., Savvaidis, A.S., Karakaisis, G.F., Papazachos, C.B. (2002b). Precursory accelerating seismic crustal deformation in the Northwestern Anatolia Fault Zone. *Tectonophysics* **347**, 217–230.

Papazachos, B.C., Savvaidis, A.S., Papazachos, C.B., Karakaisis, G.F. (2002c). Precursory seismic crustal deformation in the area of southern Albanides. *J. Seismology* **6**, 237.

Parsons, T. (2002). Global Omori law decay of triggered earthquakes: Large aftershocks outside the classical aftershock zone. *J. Geophys. Res. (Solid Earth)* **107**, doi:10.1029/2001JB000646.

Parsons, T. (2004). Recalculated probability of $M \geqslant 7$ earthquakes beneath the Sea of Marmara, Turkey. *J. Geophys. Res.* **109**, 5304, doi:10.1029/2003JB002667.

Parsons, T. (2005). Significance of stress transfer in time-dependent earthquake probability calculations. *J. Geophys. Res.* **110**, 5304, doi:10.1029/2004JB003190.

Parsons, T., Stein, R.S., Simpson, R.W., Reasenberg, P.A. (1999). Stress sensitivity of fault seismicity: A comparison between limited-offset oblique and major strike-slip faults. *J. Geophys. Res.* **104**, 20183–20202, doi:10.1029/1999JB900056.

Peltzer, G., Crampe, F., Hensley, S., Rosen, P. (2001). Transient strain accumulation and fault interaction in the Eastern California Shear zone. *Geology* **29**, 975–978.

Pollitz, F., Bakun, W.H., Nyst, M. (2004). A physical model for strain accumulation in the San Francisco Bay region: Stress evolution since 1838. *J. Geophys. Res.* **109**, 11408, doi:10.1029/2004JB003003.

Pondard, N., Armijo, R., King, C.P.K., Meyer, B., Flerit, F. (submitted for publication). Fault interactions in the Sea of Marmara pull-apart (North Anatolian Fault): Earthquake clustering and propagating earthquake sequences, *Geophys. J. Int*.

Prawirodirdjo, L., et al. (1997). Geodetic observations of interseismic strain segmentation at the Sumatra subduction zone. *Geophys. Res. Lett.* **24**, 2601–2604, doi:10.1029/97GL52691.

Reasenberg, P.A., Hanks, T.C., Bakun, W.H. (2003). An empirical model for earthquake probabilities in the San Francisco Bay region, California, 2002–2031. *Bull. Seismol. Soc. Am.* **93**, 1–13.

Reid, H.F. (1910). The mechanics of the earthquake, the California earthquake of April 18, 1906. In: *Report of the State Investigation Commission*, vol. 2. Carnegie Institution of Washington, Washington, DC.

Robinson, R. (2000). A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. *Geophys. J. Int.* **140**, 568–576.

Robinson, R. (2003). Fault interactions and subduction tectonics: a re-examination of the Weber, New Zealand, earthquake sequence of 1990. *Geophys. J. Int.* **154**, 745–753, doi:10.1046/j.1365-246X.2003.02007.x.

Robinson, R., Zhou, S., Johnston, S., Vere-Jones, D. (2005). Precursory Accelerating Seismic Moment Release (AMR) in a synthetic seismicity catalog; A preliminary study. *Geophys. Res. Lett.* **32**, doi:10.1029/2005GL022576.

Ruff, L., Kanamori, H. (1983). Seismic coupling and uncoupling at subduction zones. *Tectonophysics* **99**, 99–117.

Rundle, J.B., Tiampo, K.F., Klein, W., Sa Martins, J.S. (2002). Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobilogy, and forecasting. *Colloquium Proc. Natl. Acad. Sci. USA* **99**, 2514–2521.

Rundle, J.B., Rundle, P.B., Donnellan, A., Li, P., Klein, W., Morein, G., Turcotte, D.L., Grant, L. (2006). Stress transfer in earthquakes, hazard estimation and ensemble forecasting: Inferences from numerical simulations. *Tectonophysics* **413**, 109–125.

Rundle, P.B., Rundle, J.B., Tiampo, K.F., Sa Martins, J.S., McGinnis, S., Klein, W. (2001). Nonlinear network dynamics of earthquake fault systems. *Phys. Rev. Lett.* **87**, doi:10.1103/PRL87.148501.

Saleur, H., Sammis, C.G., Sornette, D. (1996). Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. *J. Geophys. Res.* **101**, 17661–17678, doi:10.1029/96JB00876.

Sammis, C.G., Smith, S.W. (1999). Seismic cycles and the evolution of stress correlation in cellular automaton models of finite fault networks. *Pure Appl. Geophys.* **155**, 307–334.

Sammis, C.G., Sornette, D. (2002). Positive feedback, memory, and the predictability of earthquakes. *Colloquium Proc. Natl. Acad. Sci. USA* **99**, 2501–2508.

Sammis, C.G., Bowman, D.D., King, G.C.P. (2004). Anomalous seismicity and accelerating moment release preceding the 2001 and 2002 earthquakes in Northern Baja California, Mexico. *Pure Appl. Geophys.* **161**, 2369–2378, doi:10.1007/s00024-004-2569-3.

Savage, J.C. (1983). A dislocation model of strain accumulation and release at a subduction zone. *J. Geophys. Res.* **88**, 4984–4996.

Scherbakov, R., Turcotte, D.L., Rundle, J.B. (2005). Aftershock statistics. *Pure Appl. Geophys.* **162**, 1051–1076.

Scholz, C.H. (1998). Earthquakes and friction laws. *Nature* **391**.

Scholz, C.H. (1990a). Geophysics; earthquakes as chaos. *Nature* **348**, 197–198.
Scholz, H.C. (1990b). In: *The Mechanics of Earthquakes and Faulting*. Cambridge University Press, New York, p. 439.
Shebalin, P., Keilis-Borok, V., Gabrielov, A., Zaliapin, I., Turcotte, D. (2006). Short-term earthquake prediction by reverse analysis of lithosphere dynamics. *Tectonophysics* **413**, 63–75.
Sieh, K.E. (1978). Prehistoric large earthquakes produced by slip on the San Andreas Fault at Pallett Creek, California. *J. Geophys. Res.* **83**, 3907–3939.
Sieh, K.E., Stuiver, M., Brillinger, D. (1989). A more precise chronology of earthquakes produced by the San Andreas fault in southern California. *J. Geophys. Res.* **94**, 603–623.
Silver, P.G., Valette-Silver, N.J. (1992). Detection of hydrothermal precursors to large Northern California earthquakes. *Science* **257**, 1363–1368.
Silver, P.G., Wakita, H. (1996). A search for earthquake precursors. *Science* **273**, 77–78.
Simoes, M., Avouac, J.P., Cattin, R., Henry, P. (2004). The Sumatra subduction zone: a case for a locked fault zone extending into the mantle. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002958.
Simpson, R.W., Reasenberg, P.A. (1994). Earthquake-induced static stress changes on Central California faults. In: Simpson, R.W. (Ed.), *The Loma Prieta, California Earthquake of Oct. 17, 1989 – Tectonic Processes and Models*, U.S. Geol. Surv. Prof. Pap., 1994.
Simpson, R.W., Schulz, L.D., Dietz, S.S., Burford, R.O. (1988). The response of creeping parts of the San Andreas fault to earthquakes on nearby faults: Two examples. *Pure Appl. Geophys.* **126**, 665–685.
Smalley, R.F., Turcotte, D.L., Solla, S.A. (1985). A renormalization group approach to the stick-slip behavior of faults. *J. Geophys. Res.* **90**, 1894–1900.
Sornette, D., Helmstetter, A. (2002). Occurrence of finite-time singularities in epidemic models of rupture, earthquakes, and starquakes. *Phys. Rev. Lett.* **89** (15), 158501.
Sornette, D., Sammis, C. (1995). Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. *J. Physique* **5**, 607–619.
Sornette, D., Sornette, A., Davy, P. (1990). Structuration of the lithosphere in plate tectonics as a self-organized critical phenomenon. *J. Geophys. Res.* **95**, 17353–17361.
Steacy, S., Gomberg, J., Cocco, M. (2005). Introduction to special section: stress transfer, earthquake triggering and time-dependent seismic hazard. *J. Geophys. Res.* **110**, doi:10.1029/2005JB003692.
Stein, R. (1999). The role of stress transfer in earthquake occurrence. *Nature* **402**, 605–609.
Stein, R., King, G., Rundle, J. (1988). The growth of geological structures by repeated earthquakes 2. Field examples of continental dip-slip faults. *J. Geophys. Res.* **93**, 13319–13331.
Stein, R., King, G., Lin, J. (1994). Stress triggering of the 1994 $M = 6.7$ Northridge, California, earthquake by its predecessors. *Science* **265**, 1432–1435.
Stein, R., Barka, A.A., Dieterich, J.H. (1997). Progressive failure on the North Anatolian fault since 1939 by earthquake stress triggering. *Geophys. J. Int.* **128**, 594–604.
Sykes, L., Jaumé, S. (1990). Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Fransisco Bay area. *Nature* **348**, 595–599.
Tajima, F., Kanamori, H. (1985). Global survey of aftershock area expansion patterns. *Phys. Earth Planet. Int.* **40**, 77–134.
Taylor, M., Zheng, G., Rice, J.R., Stuart, W., Dmowska, R. (1996). Cyclic stressing and seismicity at strongly coupled subduction zones. *J. Geophys. Res.* **101**, 8363–8382.
Taylor, M., Dmowska, R., Rice, J.R. (1998). Upper plate stressing and seismicity in the subduction earthquake cycle. *J. Geophys. Res.* **103**, 24523–24542.
Tiampo, K.F., Anghel, M. (2006). Introduction to special issue: Critical point theory and space–time pattern formation in precursory seismicity. *Tectonophysics* **413**, 1–3.
Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J. (2002a). Eigenpatterns in southern California seismicity. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000562.
Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J., Klein, W. (2002b). Mean-field threshold systems and phase dynamics: An application to earthquake fault systems. *Europhys. Lett.* **60**, 481–487.

Tiampo, K.F., Rundle, K.F., Klein, W. (2006). Premonitory seismicity changes prior to the Parkfield and Coalinga earthquakes in Southern California. *Tectonophysics* **413**, 77–86.

Toda, S., Stein, R.S. (2000). Did stress triggering cause the large off-fault aftershocks of the 25 March 1998 $Mw = 8.1$ Antarctic plate earthquake? *Geophys. Res. Lett.* **27**, 2301–2304, doi:10.1029/1999GL011129.

Toda, S., Stein, R.S. (2002a). In: Effect of multiple stress changes on aftershock decay; an example from 26 March (M 6.5) and 13 May (M 6.3) 1997 Kagoshima earthquakes. In: *Seismological Research Letters*, Seismological Society of America: El Cerrito, CA, United States, p. 261.

Toda, S., Stein, R.S. (2002b). Response of the San Andreas Fault to the 1983 Coalinga-Nunez earthquakes; an application of interaction-based probabilities for Parkfield. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000172.

Toda, S., Stein, R.S. (2003). Toggling of seismicity by the 1997 Kagoshima earthquake couplet: A demonstration of time-dependent stress transfer. *J. Geophys. Res.* **108**, 1–7, doi:10.1029/2003JB002527.

Toda, S., Stein, R.S., Reasenberg, P.A., Dieterich, J.H., Yoshida, A. (1998). Stress transferred by the 1995 $M_W = 6.9$ Kobe, Japan, shock: Effect on aftershocks and future earthquake probabilities. *J. Geophys. Res.* **103**, 24543–24566, doi:10.1029/98JB00765.

Toda, S., Stein, R.S., Richards-Dinger, K., Bozkurt, S.B. (2005). Forecasting the evolution of seismicity in southern California: Animations built on earthquake stress transfer. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003415.

Toksoz, Y.M.N., Shakal, A.F., Michael, A.J. (1979). Space–time migration of earthquakes along the North Anatolian fault zone and seismic gaps. *Pure Appl. Geophys.* **117**, 1258–1270.

Tse, S.T., Rice, J.R. (1986). Crustal earthquake instability in relation to the depth variation of frictional slip properties. *J. Geophys. Res.* **91**, 9452–9472.

Turcotte, D.L. (Ed.) (1992). *Fractals and Chaos in Geology and Geophysics*. Cambridge University Press, Cambridge.

Turcotte, D.L., Malamud, B.D., Guzzetti, F., Reichenbach, P. (2002). Self-organization, the cascade model, and natural hazards. *Colloquium Proc. Natl. Acad. Sci. USA* **99**, 2530–2537.

Turcotte, D.L., Newman, W.I., Scherbakov, R. (2003). Micro- and macro-scopic models for rock fracture. *Geophys. J. Int.* **152**, 718–728.

Utsu, T. (1961). Statistical study on the occurrence of aftershocks. *Geophys. Mag.* **30**, 521–605.

Varnes, D.J. (1989). Predicting earthquakes by analyzing accelerating precursory seismic activity. *Pure Appl. Geophys.* **130**, 661–686.

Varnes, D.J., Bufe, C.G. (1996). The cyclic and fractal seismic series preceding an mb 4.8 earthquake on 1980 February 14 near the Virgin Islands. *Geophys. J. Int.* **124**, 149–158.

Varotsos, P., Lazaridou, M. (1991). Latest aspects of earthquake prediction in Greece based on seismic electric signals. *Tectonophysics* **188**, 321–347.

Vere-Jones, D., Robinson, R., Yang, W. (2001). Remarks on accelerated moment release model: problems of model formulation, simulation and estimation. *Geophys. J. Int.* **144**, 517–531.

Voight, B. (1989). A relation to describe rate-dependent material failure. *Science* **243**, 200–203.

Wang, E., Chen, C.H. (2001). Static stress transferred by the 1999 Chi-Chi, Taiwan, earthquake: effects on the stability of the surrounding fault systems and aftershock triggering with a 3D fault-slip model. *Bull. Seismol. Soc. Am.* **91**, 1041–1052.

Weldon, R., Scharer, K.M., Fumal, T., Biasi, G. (2004). Wrightwood and the earthquake cycle: what a long recurrence tells us about how faults work. *GSA Today* **14**, 4–10.

Wesnousky, S.G. (1986). Earthquakes, quaternary faults, and seismic hazard in California. *J. Geophys. Res.* **91**, 12587–12632.

Wiemer, S., Wyss, M. (1994). Seismic quiescence before the Landers ($M = 7.5$) and Big Bear ($M = 6.5$) 1992 earthquakes. *Bull. Seismol. Soc. Am.* **84**, 900–916.

Wiemer, S., Wyss, M. (2000). Minimum magnitude of complete reporting in earthquake catalogs: examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Am.* **90**, 859–869.

Wiemer, S., Wyss, M. (2002). Mapping spatial variability of the frequency-magnitude distribution of earthquakes. *Adv. Geophys.* **45**, 259–302.

Zachariasen, J., Sieh, K. (1999). Submergence and uplift associated with the giant 1833 Sumatran subduction earthquake: evidence from coral microatolls. *J. Geophys. Res.* **104**, 895–919.

Zhuang, J., Chang, C.P., Ogata, Y., Chen, Y.-I. (2005). A study on the background and clustering seismicity in the Taiwan region by using point process models. *J. Geophys. Res.* **110**.

Zoller, G., Hainzl, S. (2002). A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL014856.

Zoller, G., Hainzl, S., Kurths, J. (2001). Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. *J. Geophys. Res.* **106**, 2167–2176.

This page intentionally left blank

ADVANCES IN GEOPHYSICS, VOL. 49, CHAPTER 3

SEISMIC RAY TRACING AND WAVEFRONT TRACKING IN LATERALLY HETEROGENEOUS MEDIA

N. RAWLINSON[†], J. HAUSER AND M. SAMBRIDGE

Research School of Earth Sciences, Australian National University, Canberra ACT 0200, Australia

1. INTRODUCTION

1.1. Motivation

One of the most common and challenging problems in seismology is the prediction of source-receiver paths taken by seismic energy in the presence of lateral variations in wavespeed. The solution to this problem is required in many applications that exploit the high frequency component of seismic records, such as body wave tomography, migration of reflection data and earthquake relocation. The process of tracking the kinematic evolution of seismic energy also brings with it the possibility of computing various other wave-related quantities such as traveltime, amplitude, attenuation, or even the high frequency waveform, which can then be compared to observations.

The difficulties associated with locating a two point path arise from the non-linear relationship between velocity and path geometry. Figure 1, which shows a fan of ray paths propagate from a point source in a strongly heterogeneous medium, provides useful insight into this non-linearity. If the medium had been homogeneous, then the paths would simply have been straight lines emitting at uniform angular distance from the source. However, the focusing and defocusing effects of the velocity heterogeneity have imposed strong and varying curvature to the paths. In addition to the extremely non-linear distribution of rays along the boundaries of the medium, the phenomenon of multi-pathing, which equates to wavefront self-intersection, can also be observed. Thus, the already difficult problem of locating a valid path which connects two points has been further complicated by the fact that there may well be more than one path.

Over the past few decades, the growing need for fast and accurate prediction of high frequency wave properties (most commonly traveltime) in complex 2-D

[†] Author thankse-mail: Nicholas.Rawlinson@anu.edu.au; jurg.hauser@anu.edu.au; Malcolm.Sambridge@anu.edu.au

ISSN: 0065-2687
DOI: 10.1016/S0065-2687(07)49003-3

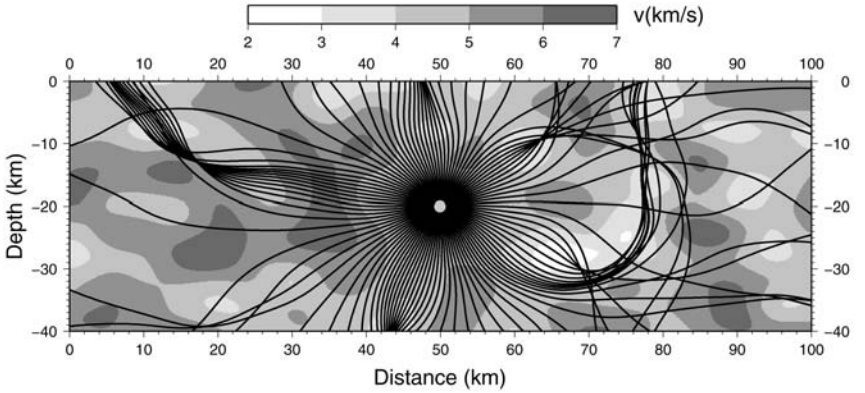


FIG. 1. Trajectories followed by a uniform fan of 100 rays emitted by a source point (grey dot) in a smoothly varying heterogeneous medium. Although the angular distance between all adjacent paths at the source point is identical, this relationship is not preserved as the rays track through the medium.

and 3-D media has spawned a prolific number of grid and ray based solvers. Traditionally, the method of choice has been ray tracing (Julian and Gubbins, 1977; Červený, 1987; Virieux and Farra, 1991; Červený, 2001), in which the trajectory of paths corresponding to wavefront normals are computed between two points. This approach is often highly accurate and efficient, and naturally lends itself to the prediction of various seismic wave properties. However, it is non-robust, and may fail to converge to a true two-point path even in mildly heterogeneous media. In addition, it usually provides no guarantee as to whether a located path corresponds to a first or later arrival.

Grid based schemes, which usually involve the calculation of traveltimes to all points of a regular grid that spans the velocity medium, have become increasingly popular in recent times. They are often based on finite difference solution of the eikonal equation (Vidale, 1988; Qin *et al.*, 1992; Hole and Zelt, 1995; Kim and Cook, 1999; Popovici and Sethian, 2002; Rawlinson and Sambridge, 2004a) or shortest path (network) methods (Nakanishi and Yamaguchi, 1986; Moser, 1991; Cheng and House, 1996), both of which tend to be computationally efficient and highly robust, a combination which makes them viable alternatives to ray tracing. Wavefronts and rays can be obtained *a posteriori* if required by either contouring the traveltime field or following the traveltime gradient from receiver to source, respectively. Disadvantages of these schemes include that in most cases they only compute the first-arrival traveltime, and their accuracy is generally not as high as ray tracing. In addition, it is difficult to compute quantities other than traveltime without first extracting ray paths.

The aim of this review paper is to describe a variety of ray and grid based solvers for locating ray paths, or implicitly or explicitly tracking wavefronts, in laterally heterogeneous media. The profusion of different schemes that can be

found in the literature means that it is not possible to provide a comprehensive review; instead, we focus on methods that have been successful in practical applications. In addition, we only look at the kinematic component of the problem i.e. ray trajectory and wavefront evolution, rather than the dynamic component, which is required for computing various quantities such as amplitudes or waveforms. After providing some introductory background material on asymptotic ray theory and geometric optics, the review proceeds with a description of various shooting and bending methods of ray tracing. This is followed by eikonal solvers and shortest path methods.

In recent years, there has been an increase in the development of new methods aimed at tracking all arrivals in heterogeneous media. These schemes can be used to predict more of the seismic wavefield when multi-pathing of seismic energy results in more complex wavetrains. One possible use of these schemes is in seismic tomography, where the exploitation of multi-arrivals may result in improved images. A detailed description of several recently developed grid and ray based schemes for predicting multi-arrivals is included in Section 4.

1.2. The Eikonal Equation

All ray and grid based methods we consider in this review are subject to the so-called "high frequency approximation"; that is, the wavelength of the propagating wave is substantially shorter than the seismic heterogeneities that characterise the medium through which they pass. Under this assumption, the full elastic wave equation can be greatly simplified, and the problem of computing the seismic wavefield made much more tractable. For a seismic P-wave in an isotropic medium, the elastic wave equation can be written (Chapman, 2004)

$$\nabla^2\Phi - \frac{1}{\alpha^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \tag{1}$$

where Φ represents the scalar potential of a P-wave, α is P-wavespeed and t is time. If we assume that the solution of Eq. (1) has the general form

$$\Phi = A\exp\big[-i\omega\big(T(\mathbf{x}) + t\big)\big],$$

where $A = A(\mathbf{x})$ is amplitude, ω is angular frequency and T is a surface of constant phase, then the Laplacian of the scalar potential is

$$\begin{aligned}\nabla^2\Phi = {} & \nabla^2 A\exp\big[-i\omega(T+t)\big] - i\omega\nabla T\cdot\nabla A\exp\big[-i\omega(T+t)\big] \\ & - i\omega\nabla A\cdot\nabla T\exp\big[-i\omega(T+t)\big] - i\omega A\nabla^2 T\exp\big[-i\omega(T+t)\big] \\ & - \omega^2 A\nabla T\cdot\nabla T\exp\big[-i\omega(T+t)\big]\end{aligned}$$

and the second derivative of Φ with respect to time is

$$\frac{\partial^2\Phi}{\partial t^2} = -\omega^2 A\exp\big[-i\omega(T+t)\big].$$

Substitution of the above two expressions into Eq. (1) yields

$$\nabla^2 A - \omega^2 A|\nabla T|^2 - i\left[2\omega \nabla A \cdot \nabla T + \omega A \nabla^2 T\right] = \frac{-A\omega^2}{\alpha^2}, \tag{2}$$

which can be divided into real and imaginary parts. If we take the real part and divide through by $A\omega^2$, then

$$\frac{\nabla^2 A}{A\omega^2} - |\nabla T|^2 = \frac{-1}{\alpha^2}.$$

Application of the high frequency approximation ($\omega \rightarrow \infty$) then yields the *eikonal equation*

$$|\nabla T| = s, \tag{3}$$

where $s = 1/\alpha$ is slowness. $T(\mathbf{x})$ is a time function (the eikonal) which describes surfaces of constant phase (wavefronts) when T is constant. If we now take the imaginary part of Eq. (2) and divide through by ω, we obtain the *transport equation*

$$2\nabla A \cdot \nabla T + A\nabla^2 T = 0 \tag{4}$$

which can be used to compute the amplitude of the propagating wave. Substitution of the appropriate general S-wave vector potential into the elastic wave equation for an S-wave leads to identical expressions for the eikonal and transport equations; thus, Eqs. (3) and (4) are valid for any high frequency body wave with slowness s. In fully anisotropic media (Červený, 2001), the eikonal and transport equations have a slightly more complex form due to the presence of the elastic tensor $\mathbf{c}$.

1.3. The Kinematic Ray Tracing Equations

Rather than directly solve the eikonal equation, one can instead consider its characteristics, which are the trajectories orthogonal (in isotropic media) to the wavefront. If $\mathbf{r}$ represents the position vector of a point on a wavefront, and l the pathlength of the curve traced out by this point as the wavefront evolves (see Fig. 2), then

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l} = \frac{\nabla T}{s} \tag{5}$$

since both $\mathrm{d}\mathbf{r}/\mathrm{d}l$ and $\nabla T/s$ are unit vectors parallel to the path. The rate of change of traveltime along the path is simply the slowness, so

$$\frac{\mathrm{d}T}{\mathrm{d}l} = s \tag{6}$$

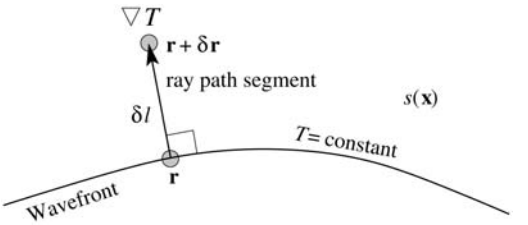


FIG. 2. Variables used to describe wavefronts and rays. T is traveltime, $\mathbf{r}$ is the position vector of a point on the wavefront, l is ray path length and $s(\mathbf{x})$ is slowness.

and by taking the gradient of both sides (noting the commutation of $\mathrm{d}/\mathrm{d}l$ and ∇)

$$\frac{\mathrm{d}\nabla T}{\mathrm{d}l} = \nabla s. \tag{7}$$

Equations (5) and (7) can be combined in order to remove ∇T which gives

$$\frac{\mathrm{d}}{\mathrm{d}l}\left[s\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l}\right] = \nabla s. \tag{8}$$

Equation (8) is the *kinematic ray equation* and describes the trajectory of ray paths in smoothly varying isotropic media. It will be shown later how Eq. (8) can be reduced to forms suitable for initial and boundary value ray tracing. The kinematic ray equation can also be derived using the calculus of variations, because Fermat's principle of stationary time states that ray paths correspond to extremal curves of the integral

$$T = \int_L s\,\mathrm{d}l \tag{9}$$

where L represents the path. In this case, Eq. (8) turns out to be the corresponding Euler–Lagrange equation.

In the presence of wavespeed discontinuities, Eq. (8) cannot be used because ∇s is not defined. Instead, Snell's law can be applied, which in its simplest form can be expressed:

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_o}{v_o} \tag{10}$$

where θ_i and v_i are the angle and wavespeed of the incoming ray, and θ_o and v_o are the angle and wavespeed of the outgoing ray. For reflected rays, $v_1 = v_2$ so $\theta_1 = \theta_2$, which in general will not be the case for transmitted rays provided a wavespeed discontinuity exists.

The derivation of Eq. (8) is relatively straightforward, due largely to the assumption of isotropic media. However, if we also wish to include anisotropy, then a more general approach is required. A commonly used treatment in this case is the Hamiltonian formalism of classical mechanics (Chapman, 2004;

Červený, 2001). In this case, rays are equivalent to the characteristic curves of the Hamiltonian, which may be expressed in various ways. In isotropic media, the Hamiltonian is often written (Chapman, 2004)

$$H(\mathbf{x}, \mathbf{p}) = \frac{\mathbf{p}^2}{2s^2} \tag{11}$$

or (Virieux and Farra, 1991)

$$H(\mathbf{x}, \mathbf{p}) = \frac{1}{2}\left[\mathbf{p}^2 - s^2\right], \tag{12}$$

where $\mathbf{p} = \nabla T$. Setting $H = 1/2$ in Eq. (11) or $H = 0$ in Eq. (12) results in characteristic curves which satisfy the eikonal equation. The Hamilton equations, which describe the characteristic curves of the Hamiltonian, can be written (Chapman, 2004)

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \frac{\partial H}{\partial \mathbf{p}} \quad \text{and} \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\nabla H \tag{13}$$

which is a coupled system of six ordinary differential equations. These equations can be integrated forward in time from given initial conditions using standard numerical solvers such as the Runge–Kutta method (e.g. Kreyszig, 1993). The Hamilton equations need not be written in the form of Eq. (13) with time as the independent variable; for example, one could also use path length (see Červený, 2001 for more details).

In anisotropic media, Eq. (13) remains valid, but an alternative form of the Hamiltonian, which takes into account the 21 independent elastic parameters $c_{i,j,k,l}$ (where $i, j, k, l = 1, \ldots, 3$) and density ρ, is required. The presence of anisotropy means that we can no longer treat P- and S-waves as equivalent, with only their propagation speeds being different. Instead, there will be three distinct wave types, a quasi-compressional wave qP, and two quasi shear waves qS1 and qS2. It turns out that the behaviour of these waves can be described by finding the eigenvectors and eigenvalues corresponding to the solution of the equation (see Červený and Firbas, 1984):

$$(\mathbf{\Gamma} - \mathbf{I})\mathbf{w} = \mathbf{0}, \tag{14}$$

where the 3×3 matrix $\mathbf{\Gamma} = \{\Gamma_{jk}\} = p_i p_l \mathbf{c}_{il}/\rho$ (note implied summation over i and l) is the so-called Christoffel matrix and $\mathbf{w}$ is the three component displacement vector. Equation (14) can be derived by seeking the asymptotic solution to the full elastic wave equation. The eigenvalues G_m ($m = 1, \ldots, 3$) satisfy:

$$\det[\mathbf{\Gamma} - \mathbf{I}G_m] = 0 \tag{15}$$

and the eigenvectors $\mathbf{g}_m$ satisfy:

$$[\mathbf{\Gamma} - \mathbf{I}G_m]\mathbf{g}_m = \mathbf{0} \tag{16}$$

noting that m is not used as an implied summation variable. Equation (14) is satisfied provided any of the eigenvalues $G_m = 1$ ($m = 1, \ldots, 3$). Each G_m corresponds to the eikonal equation for a different wave type, and the associated eigenvectors $\mathbf{g}_m$ describe the direction of particle motion imposed by the wave. It turns out (Červený and Firbas, 1984) that the eigenvalues can be written in terms of the eigenvectors as

$$G_m = \mathbf{g}_m^{\mathrm{T}} \boldsymbol{\Gamma} \mathbf{g}_m = \frac{p_i \, p_l \mathbf{g}_m^{\mathrm{T}} \mathbf{c}_{il} \mathbf{g}_m}{\rho}. \tag{17}$$

In isotropic media, $G_1 = \alpha^2 \mathbf{p} \cdot \mathbf{p}$ and $G_2 = \beta^2 \mathbf{p} \cdot \mathbf{p}$, where α and β are the P and S wavespeeds respectively.

For general anisotropic inhomogeneous media, the Hamiltonian can therefore be written as

$$H_m(\mathbf{x}, \mathbf{p}) = \frac{1}{2} G_m(\mathbf{x}, \mathbf{p}) = \frac{1}{2} \mathbf{g}_m^{\mathrm{T}} \boldsymbol{\Gamma} \mathbf{g}_m = 1. \tag{18}$$

Substitution of these expressions into the Hamilton equations (Eq. (13)) allows ray paths for any of the three different wave types to be traced using an appropriate numerical solver. For mildly anisotropic media, Červený and Firbas (1984) derive a linearisation procedure which allows anisotropic traveltimes to be computed using paths provided by an isotropic ray tracer. If an interface is located within an anisotropic media, then up to three reflected and three transmitted waves can be generated. Although Eq. (10) is still valid, it strictly applies to phase angle and phase velocity; thus, in order to correctly propagate rays in the presence of interfaces, additional constraints are required (see Slawinski *et al.*, 2000).

In this section, we have only touched on kinematic ray theory; later, we will explore the more practical aspects of implementation. For further details on the underlying theory, the interested reader is referred to the comprehensive texts of Červený (2001) and Chapman (2004), which cover various aspects of kinematic and dynamic ray theory, ray amplitudes and synthetic seismograms.

1.4. Common Model Parameterisations

When rays or wavefronts are tracked through 2-D or 3-D laterally heterogeneous media, a formal description or parameterisation of the spatial variations in seismic properties is required. There are many ways that this can be done, and the final choice can depend on a variety of factors including the types of seismic structures that are present, the prediction scheme that is to be applied, and the dataset that is being simulated. For example, if the dataset was very large, then a fast and robust prediction scheme would be desirable; if the dataset contained refraction and reflection phases, then a layered medium would be required and the prediction scheme would need to account for discontinuities. On the other

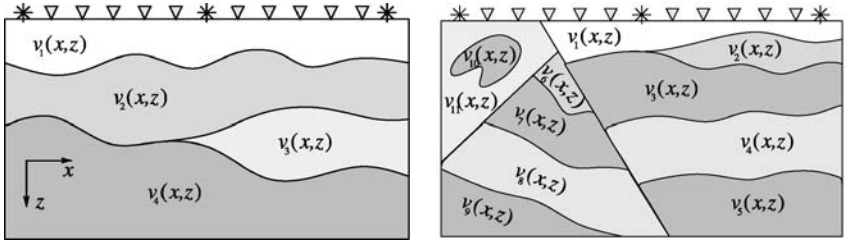


FIG. 3. General schemes for representing structure. (a) Laterally continuous interfaces separating layers within which wavespeed $v_i(x, z)$ varies smoothly, (b) more flexible framework based on an aggregate of irregular blocks within which wavespeed $v_i(x, z)$ varies smoothly.

hand, if the preferred prediction scheme was based purely on numerically solving Eq. (8), then only isotropic structures described by continuous variations in wavespeed would be permitted. However, if one was trying to predict phases in the presence of complex structures such as salt domes, subduction zones or heavily faulted sedimentary basins, then this class of prediction scheme would not be appropriate.

In laterally heterogeneous media, the most general type of parameterisation needs to allow for both velocity and interface variation. This variation would need to be almost arbitrary if one wanted to represent all possible types of Earth structure, but in practice, a number of acceptable assumptions can usually be made. In seismic tomography (Nolet, 1987; Iyer and Hirahara, 1993; Rawlinson and Sambridge, 2003a), for example, it is common practice, when interfaces are required, to represent the medium by layers, within which wavespeed varies continuously, separated by sub-horizontal interfaces which vary in depth (e.g. Chiu *et al.*, 1986; Farra and Madariaga, 1988; Williamson, 1990; Sambridge, 1990; Wang and Houseman, 1994; Zelt, 1999; Rawlinson *et al.*, 2001a) as shown in Fig. 3a. The relative simplicity of this representation makes it amenable to fast and robust data prediction, and it also allows a variety of later arriving phases to be computed. However, in exploration seismology, where data coverage is usually dense, and near surface complexities (particularly faults) often need to be accurately represented, this class of parameterisation can be too restrictive.

An alternative approach is to divide the model region up into an aggregate of irregularly shaped volume elements (see Fig. 3b), within which material property varies smoothly, but is discontinuous across element boundaries (e.g. Pereyra, 1996; Bulant, 1999). This allows most geological features such as faults, folds, lenses, overthrusts etc. to be faithfully represented. However, in the presence of such complexity, the data prediction problem becomes much more difficult to resolve, and reconciling data observations with these predictions (e.g. via seismic tomography) would be extremely challenging in the absence of accurate *a priori* information.

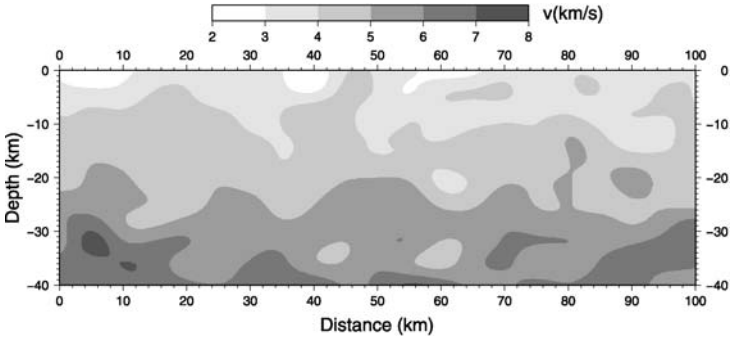
Common parameterisations used to describe wavespeed variations (or other seismic properties) in a continuum include constant velocity (or slowness) blocks (e.g. Aki *et al.*, 1977; Nakanishi, 1985; Williamson, 1990; Saltzer and Humphreys, 1997), triangular/tetrahedral meshes within which velocity is constant or constant in gradient (e.g. White, 1989; Sambridge and Faletic, 2003), and grids of velocity nodes which are interpolated using a predefined function (e.g. Thomson and Gubbins, 1982; Thurber, 1983; Červený *et al.*, 1984; Virieux and Farra, 1991; Zhao *et al.*, 1992; Neele *et al.*, 1993, etc.). Constant velocity blocks are conceptually simple, but require a fine discretisation in order to subdue the undesirable artifact of block boundaries. These discontinuities also have the potential to unrealisticly distort the wavefield and make the two-point ray tracing problem more unstable. Triangular/tetrahedral meshes are flexible and allow analytic ray tracing when velocity is constant or constant in gradient within a cell; however, like constant velocity blocks, they usually require a fine discretisation, and can also destabilise the data prediction problem.

Velocity grids which describe a continuum using an interpolant offer the possibility of smooth variations with relatively few parameters, but are generally more computationally intensive to evaluate. In addition, analytic solution of the ray tracing equations is usually not possible. However, for many practical applications, the benefits of smoothness outweigh these considerations. One of the simplest and most popular interpolants is pseudo-linear interpolation, which in 3-D Cartesian coordinates is:

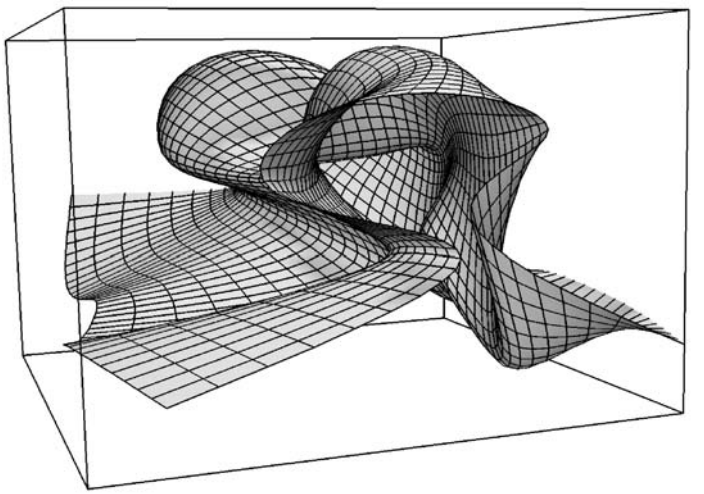
$$v(x, y, z) = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} V(x_i, y_j, z_k) \left(1 - \left|\frac{x - x_i}{x_2 - x_1}\right|\right) \times \left(1 - \left|\frac{y - y_j}{y_2 - y_1}\right|\right) \left(1 - \left|\frac{z - z_k}{z_2 - z_1}\right|\right) \tag{19}$$

where $V(x_i, y_j, z_k)$ are the velocity (or slowness) values at eight grid points surrounding (x, y, z). For Eq. (19), velocity is continuous, but the velocity gradient is not (i.e. C^0 continuity). Despite this feature, pseudo-linear interpolation has been frequently used in problems which require traveltime prediction (Eberhart-Phillips, 1986; Zhao *et al.*, 1992; Scott *et al.*, 1994; Steck *et al.*, 1998).

Higher order interpolation functions are required if the velocity field is to have continuous first and second derivatives, which is usually desirable for schemes which numerically solve the ray tracing or eikonal equations. There are many types of spline functions that can be used for interpolation, including Cardinal (Thomson and Gubbins, 1982; Sambridge, 1990), Bezier (Bartels *et al.*, 1987), B-splines (Farra and Madariaga, 1988; Virieux and Farra, 1991; Rawlinson *et al.*, 2001a) and splines under tension (Červený *et al.*, 1984; Smith and Wessel, 1990; VanDecar *et al.*, 1995). Cubic B-splines are particularly useful (Virieux and Farra, 1991; Rawlinson *et al.*, 2001a) as they offer C_2 continuity, local control and the potential for an irregular distribution of nodes. For a set of velocity values $V_{i,j,k}$



(a)



(b)

FIG. 4. Examples of cubic B-spline parameterisation for (a) velocity structure (b) interface structure.

on a 3-D grid of points $\mathbf{p}_{i,j,k} = (x_{i,j,k}, y_{i,j,k}, z_{i,j,k})$, the B-spline for the ijkth volume element is

$$\boldsymbol{B}_{i,j,k}(u, v, w) = \sum_{l=-1}^{2} \sum_{m=-1}^{2} \sum_{n=-1}^{2} b_l(u) b_m(v) b_n(w) \mathbf{q}_{i+l,j+m,k+n}, \qquad (20)$$

where $\mathbf{q}_{i,j,k} = (V_{i,j,k}, \mathbf{p}_{i,j,k})$. Thus, the three independent variables $0 \leqslant u, v, w \leqslant 1$ define the velocity distribution in each volume element. The weighting factors $\{b_i\}$ are the uniform cubic B-spline functions (Bartels *et al.*, 1987). Figure 4a shows a 2-D velocity field described by a mosaic of cubic B-spline elements.

Rather than use velocity grids in the spatial domain to describe smooth media, one could also exploit the wavenumber domain by employing a spectral parameterisation. These are often popular for global applications e.g. spherical harmonics (Dziewonski *et al.*, 1977), but can also be used for problems on a local or regional scale. For example, Wang and Pratt (1997) use the following Fourier series to describe a 2-D slowness distribution in their inversion of reflection amplitude and traveltimes

$$s(\mathbf{r}) = a_{00} + \sum_{m=1}^{N} \left[a_{m0} \cos(\mathbf{k} \cdot \mathbf{r}) + b_{m0} \sin(\mathbf{k} \cdot \mathbf{r}) \right] + \sum_{m=-N}^{N} \sum_{n=1}^{N} \left[a_{mn} \cos(\mathbf{k} \cdot \mathbf{r}) + b_{mn} \sin(\mathbf{k} \cdot \mathbf{r}) \right], \tag{21}$$

where $\mathbf{r} = x\mathbf{i} + z\mathbf{j}$ and $\mathbf{k} = m\pi k_0 \mathbf{i} + n\pi k_0 \mathbf{j}$ are the position and wavenumber vector respectively, and a_{mn} and b_{mn} are the amplitude coefficients of the (m, n)th harmonic term. Although Eq. (21) is infinitely differentiable, it is globally supported in that adjustment of any amplitude coefficient influences the entire model. Spectral parameterisations have been used in a number of studies which require traveltime prediction (Hildebrand *et al.*, 1989; Hammer *et al.*, 1994; Wiggins *et al.*, 1996).

Interfaces are often described using equivalent parameterisations to those used for velocity. For example, linear segments (2-D volume), triangular meshes (3-D volume) or nodes with a specified interpolant are common. In 2-D, piecewise linear segments have been used in several studies (e.g. Zelt and Smith, 1992; Williamson, 1990), but the discontinuities in gradient between adjacent segments can have a destabilising effect on traveltime prediction schemes – for example, two ray paths with similar trajectories which impinge on an interface at either side of a discontinuity may reflect at very different angles. Zelt and Smith (1992) overcome this problem by applying a smoothing filter to the interface normals which are required by Snell's law. Cubic B-splines in parametric form have been used by a number of authors (Farra and Madariaga, 1988; Virieux and Farra, 1991; Rawlinson *et al.*, 2001a) to describe interface structure for the data prediction problem. For interface surfaces, Eq. (20) becomes

$$\boldsymbol{B}_{i,j}(u, v) = \sum_{k=-1}^{2} \sum_{l=-1}^{2} b_k(u) b_l(v) \mathbf{p}_{i+k, j+l}, \tag{22}$$

where $\mathbf{p}_{i,j} = (x_{i,j}, y_{i,j}, z_{i,j})$ is a set of control vertices on a topologically regular grid. Equation (22) has the same desirable properties as its velocity counterpart (Eq. (20)), and thanks to its parametric representation, allows multi-valued surfaces to be represented. Figure 4b shows a complex interface surface that has been parameterised using cubic B-splines.

Finally, it is worth noting that irregular parameterisations have been used for both ray based and grid based schemes. In the case of grid based schemes which solve the eikonal equation, irregular grids have the potential to improve computational efficiency by varying grid resolution in response to wavefront curvature (e.g. Kimmel and Sethian, 1998; Qian and Symes, 2002). Another motivation for adopting irregular grids comes from seismic tomography; for many large seismic datasets, path distribution can be highly heterogeneous, resulting in a spatial variability in resolving power. The ability to "tune" a parameterisation to these variations using some form of irregular mesh has a range of potential benefits, including increased computational efficiency (fewer unknowns), improved stability of the inverse problem, and improved extraction of structural information (Michelini, 1995; Curtis and Snieder, 1997; Vesnaver *et al.*, 2000; Spakman and Bijwaard, 2001; Rawlinson and Sambridge, 2003a; Sambridge and Faletic, 2003; Sambridge *et al.*, 2005). Completely unstructured meshes, such as those that use Delaunay tetrahedra or Voronoi polyhedra (Sambridge *et al.*, 1995), offer high levels of adaptability, but have special book-keeping requirements when solving the forward problem of data prediction. Sambridge *et al.* (1995) and Sambridge and Gudmundsson (1998) describe techniques for locating points within these meshes, which allows ray tracing to be performed efficiently.

2. Ray Tracing Schemes

In this section, we describe a number of practical ray tracing schemes for solving the boundary value problem of locating source-receiver ray paths in various classes of media (2-D, 3-D, with and without discontinuity). There are two broad categories – shooting and bending – which exploit different formulations of the ray tracing equations (Eq. (8)).

2.1. Shooting Methods

Shooting methods of ray tracing are conceptually simple; they formulate Eq. (8) as an initial value problem which allows a complete ray path (with appropriate application of Snell's law in the presence of any interface) to be traced given an initial trajectory of the path. The two point problem of finding a source-receiver path then becomes an inverse problem in which the unknown is the initial direction vector of the ray, and the function to be minimised is a measure of the distance between the ray end point and receiver. The main challenge that faces this class of method is the non-linearity of the inverse problem, which tends to increase dramatically with the complexity of the medium (as Fig. 1 testifies).

2.1.1. The Initial Value Problem

The appropriate form of the equation required to solve the initial value problem depends largely on the choice of parameterisation. In a medium described by constant velocity (slowness) blocks, the ray path is simply described by a piecewise set of straight line segments; all that is required to solve the initial value problem is repeated application of Snell's law at cell boundaries. This can be accomplished with high computational efficiency (e.g. Williamson, 1990). Analytic ray tracing can also be applied to other parameterisations; for example, triangular or tetrahedral meshes in which the velocity gradient is constant (e.g. White, 1989). The expression for ray trajectory in a medium with a constant velocity gradient can be expressed in various ways, including parametrically as (Rawlinson *et al.*, 2001a)

$$\mathbf{x} = \frac{v(z_0)}{k}\left[\frac{a_0(c-c_0)}{1-c_0^2}, \frac{b_0(c-c_0)}{1-c_0^2}, 1-\sqrt{\frac{1-c^2}{1-c_0^2}}\right] + \mathbf{x}_0, \tag{23}$$

where $\mathbf{x}_0$ is the origin of the ray segment, $[a, b, c]$ is a unit vector tangent to the ray path, $[a_0, b_0, c_0]$ is a unit vector tangent to the ray path at x_0, k is the velocity gradient, and $v(z_0)$ is the velocity at z_0. The associated traveltime is then given by

$$T = \frac{1}{2k}\ln\left[\left(\frac{1+c}{1-c}\right)\left(\frac{1-c_0}{1+c_0}\right)\right] + T_0, \tag{24}$$

where T_0 is the traveltime from the source to $\mathbf{x}_0$. For application to tetrahedra (or triangles in 2-D), it is simply a matter of rotating the coordinate system so that the velocity gradient is in the direction of the z-axis. A number of other velocity functions yield analytic ray tracing solutions, such as the constant gradient of $\ln v$, and the constant gradient of the nth power of slowness $1/v^n$ (Červený, 2001).

Although analytic ray tracing is possible for a few special cases, in general one needs to solve Eq. (8) using numerical methods. This usually requires Eq. (8) to be reduced to a convenient first-order initial value system of equations, which can be done in a variety of ways. For example, by considering the following unit vector in the direction of the ray

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l} = [\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta], \tag{25}$$

where θ is the inclination of the ray with the vertical (z-axis), and ϕ is the azimuth of the ray (angle between ray and positive x-axis in xy plane). Substitution of this expression into Eq. (8) and application of the product rule yields:

$$\left.\begin{aligned}
\frac{\partial s}{\partial x} &= s\cos\theta\cos\phi\frac{\mathrm{d}\theta}{\mathrm{d}l} - s\sin\theta\sin\phi\frac{\mathrm{d}\phi}{\mathrm{d}l} + \sin\theta\cos\phi\frac{\mathrm{d}s}{\mathrm{d}l} \\
\frac{\partial s}{\partial y} &= s\cos\theta\sin\phi\frac{\mathrm{d}\theta}{\mathrm{d}l} + s\sin\theta\cos\phi\frac{\mathrm{d}\phi}{\mathrm{d}l} + \sin\theta\sin\phi\frac{\mathrm{d}s}{\mathrm{d}l} \\
\frac{\partial s}{\partial z} &= -s\sin\theta\frac{\mathrm{d}\theta}{\mathrm{d}l} + \cos\theta\frac{\mathrm{d}s}{\mathrm{d}l}
\end{aligned}\right\}. \tag{26}$$

These three equations can be rearranged to remove the $\mathrm{d}s/\mathrm{d}l$ term and produce expressions for $\mathrm{d}\theta/\mathrm{d}l$ and $\mathrm{d}\phi/\mathrm{d}l$, which together with Eq. (25) produce the following system of equations

$$\left.\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}l} &= \sin\theta\cos\phi \\
\frac{\mathrm{d}y}{\mathrm{d}l} &= \sin\theta\sin\phi \\
\frac{\mathrm{d}z}{\mathrm{d}l} &= \cos\theta \\
\frac{\mathrm{d}\theta}{\mathrm{d}l} &= \frac{\cos\theta}{s}\left[\cos\phi\frac{\partial s}{\partial x} + \sin\phi\frac{\partial s}{\partial y}\right] - \frac{\sin\theta}{s}\frac{\partial s}{\partial z} \\
\frac{\mathrm{d}\phi}{\mathrm{d}l} &= \frac{1}{s\sin\theta}\left[\cos\phi\frac{\partial s}{\partial y} - \sin\phi\frac{\partial s}{\partial x}\right]
\end{aligned}\right\}. \tag{27}$$

Thus, given some initial position and trajectory, a ray path can be obtained by solving this coupled system of equations e.g. using a fourth order Runge–Kutta scheme (e.g. Kreyszig, 1993).

The initial value formulation of the kinematic ray tracing equations given by Eq. (27) uses path length l as the independent variable. However, it is often more convenient to use traveltime t, since this parameter is usually required in addition to path geometry. Conversion of Eq. (27) into a form that uses t as the independent variable and velocity v instead of slowness s (the use of slowness or velocity is often a matter of convention, but it can have practical implications for certain classes of problem, e.g. Rawlinson and Sambridge, 2003b) can be achieved using the following simple set of relationships

$$\begin{aligned}
&\frac{\mathrm{d}\theta}{\mathrm{d}l} = s\frac{\mathrm{d}\theta}{\mathrm{d}t}, \\
&\frac{\mathrm{d}\phi}{\mathrm{d}l} = s\frac{\mathrm{d}\phi}{\mathrm{d}t}, \quad \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}l} = s\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}, \quad \frac{\partial s}{\partial \mathbf{x}} = -\frac{1}{v^2}\frac{\partial v}{\partial \mathbf{x}},
\end{aligned} \tag{28}$$

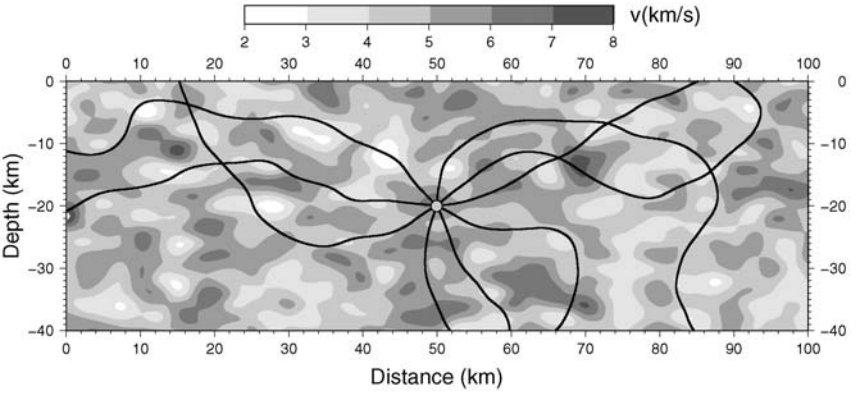


FIG. 5. A fan of nine ray paths traced from a source point in a complex 2-D medium described by a mesh of cubic B-spline functions. A fourth order Runge–Kutta scheme is used to solve Eq. (29) in this case.

which result in the following system of equations

$$\left.\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}t} &= v\sin\theta\cos\phi \\
\frac{\mathrm{d}y}{\mathrm{d}t} &= v\sin\theta\sin\phi \\
\frac{\mathrm{d}z}{\mathrm{d}t} &= v\cos\theta \\
\frac{\mathrm{d}\theta}{\mathrm{d}t} &= -\cos\theta\left[\cos\phi\frac{\partial v}{\partial x}+\sin\phi\frac{\mathrm{d}v}{\mathrm{d}y}\right]-\sin\theta\frac{\partial v}{\partial z} \\
\frac{\mathrm{d}\phi}{\mathrm{d}t} &= \frac{1}{\sin\theta}\left[\sin\phi\frac{\partial v}{\partial x}-\cos\phi\frac{\partial v}{\partial y}\right]
\end{aligned}\right\}. \tag{29}$$

This system of equations has a similar form to Eq. (27), and therefore can be solved using the same class of technique. The nine ray paths shown in Fig. 5 were computed by solving Eq. (29) using a 4th order Runge–Kutta method (with constant ϕ). This procedure is very efficient; for example, tracing 10,000 ray paths using a time step of 0.1 s from the same source point to the edge of the medium only takes 6 s on a 1.6 GHz Opteron PC running Linux.

Rather than use ray inclination and azimuth to describe ray trajectory, one could also use the components of the ray direction (or *slowness*) vector $\mathbf{p} = \nabla T$. A system of first-order equations can then be simply derived from Eq. (8) by using this vector as the substitution variable i.e. $\nabla T = s\,\mathrm{d}\mathbf{r}/\mathrm{d}l$. This results in

$$\left.\begin{aligned}\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l} &= v\mathbf{p} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}l} &= \nabla\frac{1}{v}\end{aligned}\right\}, \tag{30}$$

which is a system of six equations. The independent variable s can be replaced by traveltime t in the same way as before to produce

$$\left.\begin{aligned}\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} &= v^2\mathbf{p} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\frac{1}{v}\nabla v\end{aligned}\right\}. \tag{31}$$

This is the same set of equations that would result from substituting Eq. (11) into the Hamilton equations (13). Although Eq. (29) has one less dependent variable to compute than Eq. (31), it contains trigonometric functions which, from a computational point of view, are less desirable. The 6D vector ($\mathbf{r}$, $\mathbf{p}$), which uniquely describes the position and trajectory of a ray in 3-D space, is sometimes referred to as the *phase space* vector (Chapman, 2004), while its 5-D counterpart ($\mathbf{r}$, θ, ϕ) is sometimes referred to as the *reduced phase space* vector (e.g. Osher *et al.*, 2002).

It is worth emphasising that a variety of other first-order formulations of the kinematic ray tracing equations in isotropic media can be derived – for more details, see Červený (1987, 2001). In general anisotropic media, a system of six first-order equations can be obtained by substituting the Hamiltonian defined by Eq. (13) into the Hamilton equations, which results in (Červený and Firbas, 1984; Chapman, 2004)

$$\left.\begin{aligned}\frac{\mathrm{d}x_i}{\mathrm{d}t} &= \frac{\partial G_m}{\partial p_i} = \frac{c_{ijkl}\,p_k\,g^j_m\,g^l_m}{\rho} \\ \frac{\mathrm{d}p_i}{\mathrm{d}t} &= -\frac{\partial G_m}{\partial x_i} = -\frac{1}{2}\frac{\partial[c_{jklm}/\rho]}{\partial x_i}\,p_j\,p_m\,g^k_m\,g^l_m\end{aligned}\right\}. \tag{32}$$

Although much more computationally expensive to solve than its isotropic equivalent, initial value anisotropic ray tracing in 2-D or 3-D media does not pose a significant challenge for modern computers.

In the presence of smooth velocity variations, the kinematic ray tracing equations provide the required solution, but as soon as velocity discontinuities are introduced, two additional problems need to be solved: (1) locate ray-interface intersection point; (2) find trajectory of departing ray path. The first problem can be difficult to solve, particularly if both ray paths and interfaces are described by non-linear functions. If a medium is divided up into cells, then a first level of refinement can be achieved by simply knowing which cells contain an interface;

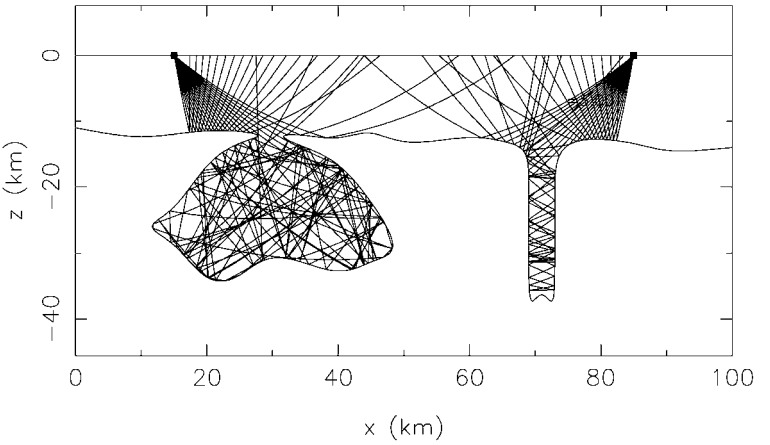


FIG. 6. Reflection paths from two source points in the presence of a complex multi-valued interface described by cubic B-splines. Path trajectories are obtained by solving an initial value problem until the ray impinges on the surface $z = 0$. Note that it is decided *a priori* that all rays impinging on the interface will reflect. Thus, rays trapped inside complex features of the model keep reflecting until they finally emerge and intersect with the surface.

thus when a ray enters a cell, one knows whether or not an intersection is possible. Sambridge and Kennett (1990) devise a boundary value ray tracing scheme for 3-D media which contain discontinuities; rather than exactly locate the ray-interface intersection point, the time step length used in Eq. (29) is iteratively adjusted in the vicinity of an interface based on a linear update procedure. The iteration process ceases when the distance between the ray end point and interface satisfies a tolerance criteria.

In an alternative approach, Virieux and Farra (1991) exploit the convex hull and subdivision properties of B-splines. The first property states that the B-spline surface must lie within the convex volume defined by the control points; the second property allows the same surface to be described by larger numbers of control points. Thus, the approximate location of a ray-interface intersection point can be found by determining whether the ray (locally linear in this case) intersects a parallelepiped containing the convex volume; if it does, then the convex volume can be subdivided and the simple ray-parallelepiped problem can be resolved for each of the new, smaller, parallelepipeds. Repetition of this process allows all possible intersection points to be found. This scheme is very robust, but can be computationally expensive; as a result, Virieux and Farra (1991) switch to a more efficient local search method once it is considered a unique intersection point is being targeted by the subdivision scheme.

Like Virieux and Farra (1991), Rawlinson *et al.* (2001a) also use cubic B-splines in parametric form to represent interface surfaces, but adopt a different approach to finding ray-interface intersection points. In this case, rays are de-

fined by circular arc segments. Rather than exploit the convex hull property of B-splines, an initial intersection point is found by approximating the surface by a mosaic of triangular patches, for which analytic solution of the ray-interface intersection problem is possible. Once a preliminary intersection point is found, an iterative non-linear scheme based on the generalised Newton method is used to target the true intersection point. Although computationally efficient, the procedure may fail to locate some intersection points, particularly if the trajectory of the ray is nearly parallel to the interface when intersection occurs. Another possible problem is that the local method converges to the incorrect intersection point. However, in practice the scheme is effective, as demonstrated in 2-D by Fig. 6, which shows reflection paths propagating through a highly complex interface structure. As well as demonstrating the robustness of the intersection scheme, this example also illustrates the power of initial value ray tracing: given virtually any structure, it is possible to trace ray paths of almost any specified initial trajectory.

The second problem that needs to be solved when a ray impinges on an interface is to find the trajectory of the departing ray path. In 2-D, a simple application of Snell's law (Eq. (10)) can produce up to four departing paths (transmission, reflection, converted transmission, converted reflection) for isotropic media. In 3-D, the same class of paths can be generated, but in addition to Snell's law, the continuity of projection of the incident ray must also be considered. This is equivalent to requiring that the incident path, departing path and normal to the interface at the intersection point, all lie in the same plane. By combining this constraint with Snell's law, the slowness vector $\mathbf{p}_r$ of the reflected or refracted ray path can be defined by (Červený, 1987; Sambridge and Kennett, 1990):

$$\mathbf{p}_r = \mathbf{p}_i + \left\{ \kappa \left[\frac{1}{v_r^2} - \frac{1}{v_i^2} + (\mathbf{p}_i \cdot \mathbf{n})^2 \right]^{1/2} - \mathbf{p}_i \cdot \mathbf{n} \right\} \mathbf{n}, \tag{33}$$

where $\mathbf{p}_i$ is the slowness vector of the incident path, $\mathbf{n}$ is a normal vector to the interface at the intersection point, and v_i and v_r are the wavespeeds of the incident and departing rays. $\kappa = \text{sign}(\mathbf{p}_i \cdot \mathbf{n})$ and equals $+1$ if $\mathbf{p}_i$ makes an acute angle with $\mathbf{n}$ and -1 otherwise. When unconverted reflected waves are required ($v_i = v_r$), Eq. (33) reduces to:

$$\mathbf{p}_r = \mathbf{p}_i - 2(\mathbf{p}_i \cdot \mathbf{n})\mathbf{n}. \tag{34}$$

For general anisotropic media, up to three reflected and three transmitted waves can be generated for a single incident path – see Slawinski *et al.* (2000) for details on how they can be calculated.

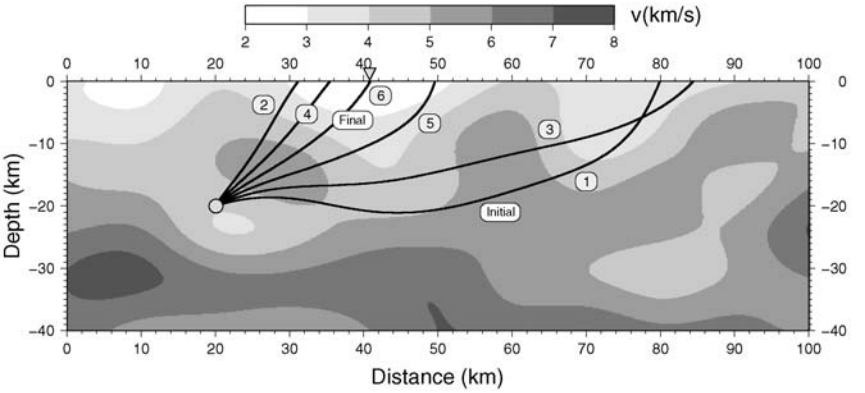


FIG. 7. Principle of the shooting method. In this case, an initial path trajectory is updated until it converges at the receiver.

2.1.2. *The Boundary Value Problem*

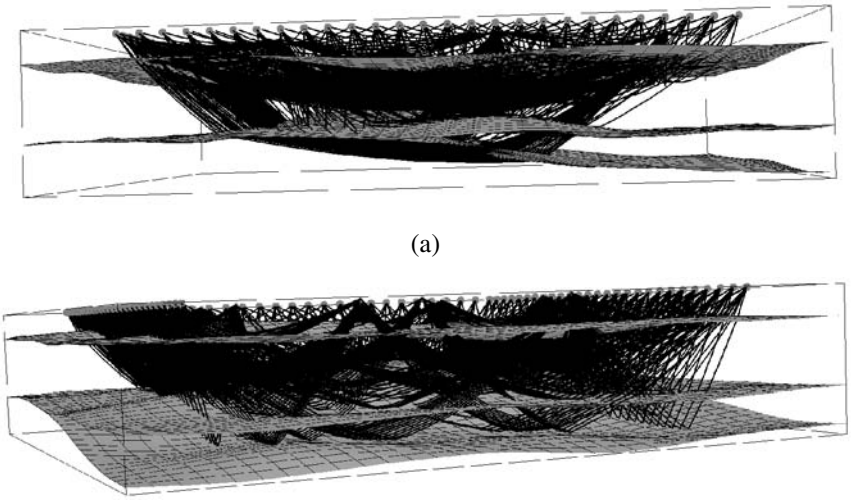
Shooting methods of ray tracing usually solve the boundary value problem by probing the medium with initial value ray paths and then exploiting information from the computed paths to better target the receiver. Figure 7 illustrates this basic concept in 2-D. If a ray emanates from a source point in a 3-D medium with take off angles θ_o and ϕ_o, and the aim is for the ray end point (x_e, y_e) on the receiver plane (z = constant) to coincide with the receiver location (X_r, Y_r), then the boundary value problem amounts to finding the (θ_o, ϕ_o) that solve the two non-linear simultaneous equations

$$\left.\begin{aligned} x_e(\theta_o, \phi_o) &= X_r \\ y_e(\theta_o, \phi_o) &= Y_r \end{aligned}\right\}. \tag{35}$$

Given that (x_e, y_e) cannot be expressed explicitly as a function of (θ_o, ϕ_o) for most velocity fields, it is usually the case that the boundary value problem is posed as an optimisation problem, with the misfit function to be minimised expressed as some measure of the distance between the ray end point and its intended target. Since the optimisation problem is non-linear, a range of iterative non-linear and fully non-linear schemes can be applied.

Julian and Gubbins (1977) propose two iterative non-linear schemes for solving Eq. (35). The first of these is Newton's method, which amounts to a simple linearisation:

$$\begin{bmatrix} \dfrac{\partial x_e}{\partial \theta_o} & \dfrac{\partial x_e}{\partial \phi_o} \\ \dfrac{\partial y_e}{\partial \theta_o} & \dfrac{\partial y_e}{\partial \phi_o} \end{bmatrix} \begin{bmatrix} \theta_o^{n+1} - \theta_o^n \\ \phi_o^{n+1} - \phi_o^n \end{bmatrix} = \begin{bmatrix} X_r - x_e(\theta_o^n, \phi_o^n) \\ Y_r - y_e(\theta_o^n, \phi_o^n) \end{bmatrix}. \tag{36}$$



(a)

(b)

FIG. 8. Two-point paths computed through a 3-D layered model using the shooting scheme of Rawlinson *et al.* (2001a). (a) Refracted paths, (b) reflected paths.

Thus, given some starting initial trajectory θ_o^0, ϕ_o^0, solution of Eq. (36) provides an updated initial trajectory θ_o^1, ϕ_o^1, and the process is repeated until an appropriate tolerance criterion is met. The success of this scheme depends largely on two factors: (1) accurate calculation of the partial derivative matrix, and (2) obtaining an initial guess ray that will converge to the correct minimum under the assumption of local linearity. Both of these requirements can be difficult to satisfy, particularly in complex media. One approach to estimating the partial derivatives involves fitting two planes through the end points of a cluster of three ray paths with different initial projection angles (i.e. one plane through all three $x_o(\theta_o, \phi_o)$ and the other plane through all three $y_o(\theta_o, \phi_o)$). The gradient of these planes in the θ_o and ϕ_o directions provide estimates of the four partial derivatives in Eq. (36). This approach is actually equivalent to the method of false position, which is the second iterative non-linear scheme proposed by Julian and Gubbins (1977). The method of false position is unlikely to be first-order accurate like a true Newton method, and therefore will converge more slowly and possibly with less stability. However, it will be faster at each iteration, and Rawlinson *et al.* (2001a, 2001b) found it to be sufficiently robust to use in solving the forward step of large 3-D tomographic inverse problems. Figure 8 shows two-point paths through a 3-D laterally heterogeneous structure that were computed using this scheme.

Sambridge and Kennett (1990) directly compute the partial derivatives in Eq. (36) by exploiting wavefront curvature information at the end point of the ray. This can be done by differentiating Eq. (29) with respect to the initial take-off an-

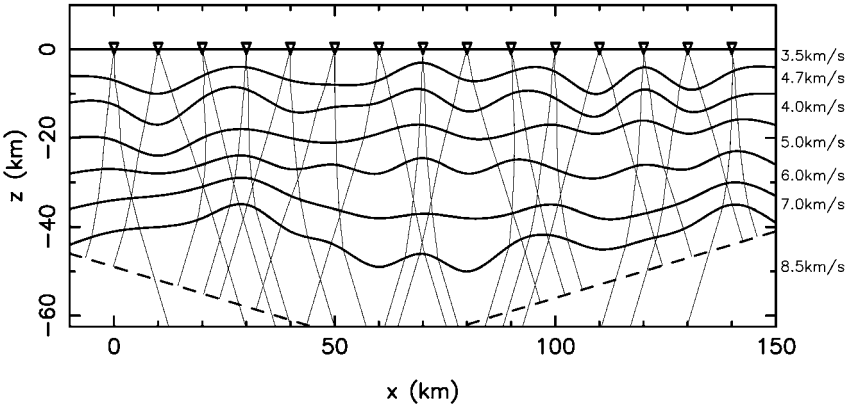


FIG. 9. A boundary value problem involving teleseismic wavefronts rather than point sources can also be solved using shooting methods of ray tracing.

gles (θ_o, ϕ_o) and reversing the order of differentiation, resulting in an additional set of 10 first-order equations which must be solved together with the original set of five. The computed variables are closely related to the derivative terms in Eq. (36), and allow the Newton scheme to be applied with first-order accuracy. Note that the additional equations now contain second-order derivatives of velocity, which means that the velocity field must have C_2 continuity.

The second requirement of a successful iterative non-linear shooting method is a sufficiently accurate initial guess ray. This can be obtained in various ways, including shooting a broad fan of rays in the general direction of the receiver array, and then (if necessary) shooting out increasingly targeted clusters of rays towards zones containing receivers until a suitably accurate initial ray is obtained (Virieux and Farra, 1991; Rawlinson *et al.*, 2001a). Another approach is to use the correct two-point ray for a laterally averaged version of the model as the initial guess ray (Thurber and Ellsworth, 1980; Sambridge, 1990). It is worth noting that the two-point problem in ray tracing is not the only type of boundary value problem; for example, one may wish to compute paths from an incident teleseismic wavefront below the crust or lithosphere to a receiver array on the surface. In this case, rays that end at a receiver begin at specific points along the wavefront surface. Figure 9 shows a solution to this class of boundary value problem obtained using the iterative non-linear shooting scheme of Rawlinson and Houseman (1998).

Shooting methods of ray tracing are widely used in seismology due to their conceptual simplicity, and potential for high accuracy and efficiency. One area in which they enjoy frequent application is seismic tomography, where 2-D or 3-D variations in seismic properties are imaged by matching data observations with data predictions using inversion techniques (e.g. Cassell, 1982; Benz and Smith, 1984; Langan *et al.*, 1985; Farra and Madariaga, 1988; White, 1989; Sambridge,

1990; Zelt and Smith, 1992; VanDecar *et al.*, 1995; McCaughey and Singh, 1997; Rawlinson *et al.*, 2001b).

2.1.3. Paraxial Ray Tracing

An important field of ray theory that is yet to be mentioned is the paraxial ray approximation (Červený and Pšenčik, 1983; Červený and Firbas, 1984; Červený, 1987; Farra and Madariaga, 1988; Virieux and Farra, 1991; Červený, 2001; Červený *et al.*, 2006), which is widely employed by the seismology community for various aspects of data prediction. It essentially involves using first order perturbation theory to deduce characteristics of the wavefield in the neighbourhood of a reference ray. Thus, given some reference path with position and slowness vector $\mathbf{y}_0(t) = [\mathbf{r}_0(t), \mathbf{p}_0(t)]^{\mathrm{T}}$, a paraxial ray can be defined by the first-order approximation

$$\mathbf{y}(t) = \mathbf{y}_0(t) + \delta\mathbf{y}(t) = \big[\mathbf{r}_0(t) + \delta\mathbf{x}(t), \mathbf{p}_0(t) + \delta\mathbf{p}(t)\big]. \tag{37}$$

To obtain the paraxial ray tracing equations, consider the linearisation of the kinematic ray tracing equations $\dot{\mathbf{r}} = \nabla_{\mathbf{p}} H$ and $\dot{\mathbf{p}} = -\nabla_{\mathbf{r}} H$ (equivalent to Eq. (13) with $(\dot{})$ denoting differentiation with respect to t), which can be written as

$$\left.\begin{aligned} \delta\dot{\mathbf{r}} &= \frac{\partial\dot{\mathbf{r}}}{\partial\mathbf{r}}\delta\mathbf{r} + \frac{\partial\dot{\mathbf{r}}}{\partial\mathbf{p}}\delta\mathbf{p} \\ \delta\dot{\mathbf{p}} &= \frac{\partial\dot{\mathbf{p}}}{\partial\mathbf{r}}\delta\mathbf{r} + \frac{\partial\dot{\mathbf{p}}}{\partial\mathbf{p}}\delta\mathbf{p} \end{aligned}\right\}, \tag{38}$$

or in more compact form as

$$\delta\dot{\mathbf{y}} = A\delta\mathbf{y} \quad \text{where } A = \begin{bmatrix} \nabla_{\mathbf{r}}\nabla_{\mathbf{p}} H & \nabla_{\mathbf{p}}\nabla_{\mathbf{p}} H \\ -\nabla_{\mathbf{r}}\nabla_{\mathbf{r}} H & -\nabla_{\mathbf{p}}\nabla_{\mathbf{r}} H \end{bmatrix}. \tag{39}$$

The paraxial ray tracing equation has six independent solutions in 3-D and four in 2-D. Since we can write that

$$\mathbf{y}(t) = \mathbf{y}(t_0) + \frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t}\delta t = \mathbf{y}(t_0) + \frac{\partial\mathbf{y}(t)}{\partial\mathbf{y}(t_0)}\frac{\mathrm{d}\mathbf{y}(t_0)}{\mathrm{d}t}\delta t, \tag{40}$$

it is common to express the solution to Eq. (39) in terms of a ray propagator matrix $P(t, t_0)$ (see Červený, 1987; Virieux and Farra, 1991)

$$\delta\mathbf{y}(t) = P(t, t_0)\delta\mathbf{y}(t_0) \quad \text{where } P(t, t_0) = \begin{bmatrix} \dfrac{\partial\mathbf{x}(t)}{\partial\mathbf{x}(t_0)} & \dfrac{\partial\mathbf{x}(t)}{\partial\mathbf{p}(t_0)} \\ \dfrac{\partial\mathbf{p}(t)}{\partial\mathbf{x}(t_0)} & \dfrac{\partial\mathbf{p}(t)}{\partial\mathbf{p}(t_0)} \end{bmatrix}, \tag{41}$$

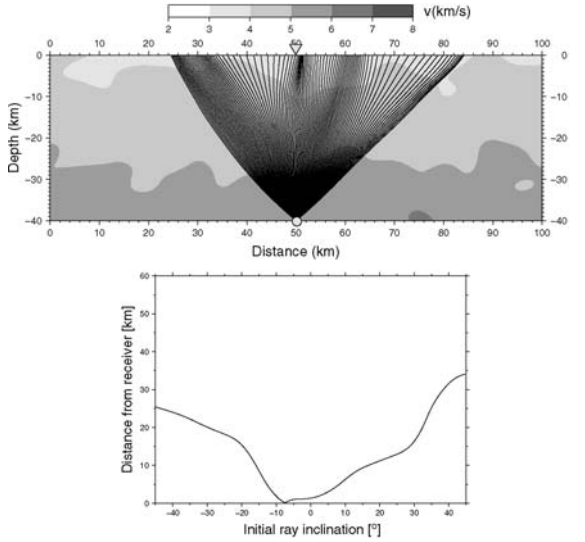
which has the initial condition that $P(t = t_0, t_0) = I$, the identity matrix. The power of paraxial rays is that they allow information about the wavefield in the vicinity of a reference ray to be used, for example to detect caustics in two-point ray tracing. In fact, the shooting method of Sambridge and Kennett (1990)

described in the previous section, which makes use of wavefront curvature information in the vicinity of a ray to compute partial derivatives for an iterative non-linear update scheme, is an example of using paraxial ray theory in two-point ray tracing. Virieux and Farra (1991) describe a similar scheme, and in fact there are many shooting schemes of ray tracing which exploit the paraxial ray approximation (e.g. Červený and Firbas, 1984; Farra and Madariaga, 1988; Bulant, 1996). For more details on paraxial ray theory and its many other potential applications, refer to Červený and Pšenčik (1983); Červený and Firbas (1984); Červený *et al.* (1984); Červený (1987); Farra and Madariaga (1987); Klimeš (1989); Bulant (1996); Červený (2001); Červený *et al.* (2006).

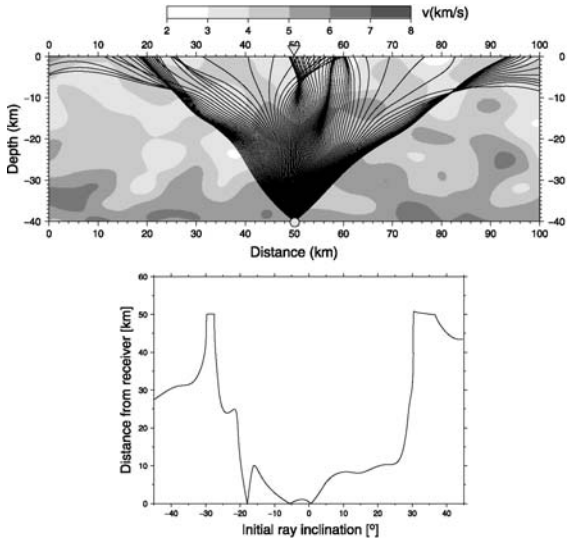
2.1.4. Fully Non-Linear Shooting Methods

Shooting methods which use an iterative non-linear approach to solve the boundary value problem are often very efficient in mildly heterogeneous media, but generally become less robust as the complexity of the medium increases. Figure 10 illustrates why this is the case by showing the relationship between ray inclination angle and distance from a receiver for two velocity models of different complexity. In the first model (Fig. 10a), velocity increases linearly with depth but has random velocity fluctuations with a standard deviation of 0.4 km/s superimposed. These velocity perturbations are sufficient to cause some mild focusing and defocusing of the wavefront, as reflected by the variable density of ray paths along the surface. This effect is also manifest in the corresponding plot of ray end point to receiver distance vs. initial ray inclination; the plot is asymmetrical about 0° (vertical initial inclination) and contains significant variations in curvature. Despite these non-linearities, it is likely that any iterative non-linear shooting scheme will converge to the correct solution using a starting ray with initial angle in the range $-45° \rightarrow 45°$, although the initiation of a triplication several kilometres to the right of the receiver (Fig. 10a) may cause some difficulties (it appears as a small region of near zero gradient in the ray inclination vs. distance plot at about $-4°$).

The second model (Fig. 10b) is identical to Fig. 10a except that the random velocity fluctuations now have a standard deviation of 1.4 km/s, which results in significant lateral heterogeneity. Note that the pattern of anomalies is identical to the first model; it is only their amplitude which has changed. The effect of the increased amplitude on the ray distribution is dramatic, with several triplications of the ray field now evident. The corresponding distance vs. ray inclination plot reveal these multi-pathing effects as extrema. Clearly, at least three arrivals are detected by the receiver. If one was to use an iterative non-linear shooting method in this situation with a starting ray in the initial inclination range of $-45° \rightarrow 45°$, then it is possible that a two point path will be located, but it is also possible that the scheme will become trapped in a local minimum. Shooting a broad fan of rays, and then increasingly more targeted clusters based on previous sampling of



(a)



(b)

FIG. 10. Demonstration of the non-linear relationship that exists between initial ray trajectory and distance from ray end point to receiver. (a) Mildly heterogeneous medium with lateral velocity standard deviation of 0.4 km/s; (b) strongly heterogeneous medium with lateral velocity standard deviation of 1.4 km/s. The source is denoted by a grey circle and the receiver by a grey triangle. Note that while 100 ray paths are shown in each of the ray tracing plots (top), the corresponding ray end point vs. initial ray inclination diagrams (bottom) were generated using 1,000 rays.

the wavefield, may provide sufficiently accurate initial rays (e.g. Rawlinson *et al.*, 2001a) to locate one or more global minima. Ultimately, though, there will be a trade-off between computing time and the number of two-point paths found, and one must always make a decision as to which point along this trade-off curve is adequate for the problem at hand.

Given the potential pitfalls of using an iterative non-linear solver in two point shooting methods, it would appear that fully non-linear solvers would be at least worthy of investigation. However, there are relatively few examples in the literature, perhaps due to the recent proliferation of grid based and wavefront construction type schemes that are designed to overcome these limitations (these schemes will be discussed later in some detail). Velis and Ulrych (1996) describe a fully non-linear shooting method of ray tracing that uses simulated annealing to locate the global minimum path. Simulated annealing (e.g. Kirkpatrick *et al.*, 1983) is based on an analogy with physical annealing in thermodynamic systems to guide variations to the model parameters, in this case the initial ray trajectories. Velis and Ulrych (2001) extend their method to heterogeneous 3-D velocity models which can include variable thickness layers, faults, and complex structures such as salt domes. The scheme exhibits several advantages compared to more conventional ray tracing schemes; in particular increased robustness for locating the global minimum solution. However, it does not appear to be practical for finding all multipaths, and tends to be more computationally expensive than iterative non-linear solvers.

2.2. Bending Methods

The principle of the bending method of ray tracing is to iteratively adjust the geometry of an initial arbitrary path that joins source and receiver until it becomes a true ray path (i.e. it satisfies Fermat's principle of stationary time) – see Fig. 11. A common approach to implementing the bending method is to derive a boundary value formulation of the kinematic ray tracing equations which can then be solved iteratively. There are many ways that this can be done; here, we describe a scheme that was first devised by Julian and Gubbins (1977). The traveltime T of a ray path between source S and receiver R can in general be expressed by the integral

$$T = \int_S^R s \, \mathrm{d}l, \tag{42}$$

where s is slowness and l is path length. The ray path can be described parametrically by a monotonic function λ, the normalised path length ($\lambda = l/L$, where L is the total path length of the ray), in which case $\mathbf{r} = \mathbf{r}(\lambda)$. A perturbation in path length can therefore be written as

$$\frac{\mathrm{d}l}{\mathrm{d}\lambda} = \sqrt{\dot{\mathbf{r}} \cdot \dot{\mathbf{r}}} = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} = F, \tag{43}$$

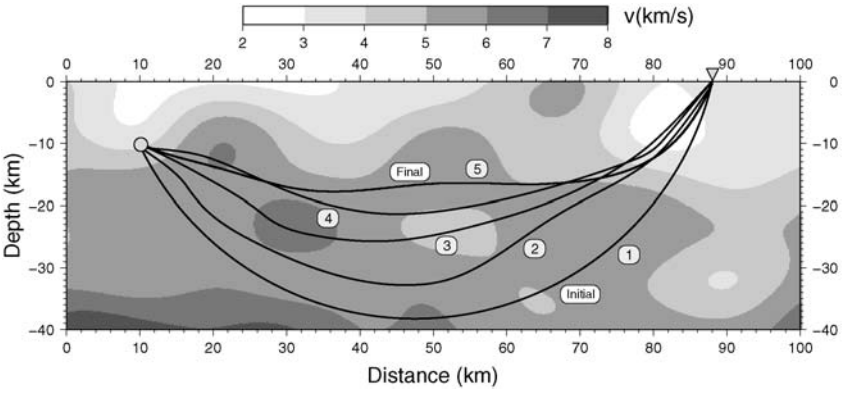


FIG. 11. Principle of the bending method. In this case, an initial two point path is perturbed until it satisfies Fermat's principle.

where $(\dot{})$ denotes differentiation with respect to λ, and the use of normalised path length means that $F = L$ and $\mathrm{d}F/\mathrm{d}\lambda = 0$. Using this expression, Eq. (42) can be rewritten as

$$T = \int_S^R sF \, \mathrm{d}\lambda. \tag{44}$$

The ray tracing equations can be obtained by extremizing this integral using the calculus of variations (e.g. Jeffreys and Swirles, 1966). For any integrand $G(\lambda, \mathbf{r}(\lambda), \dot{\mathbf{r}}(\lambda))$ the Euler–Lagrange equations are

$$\frac{\partial G}{\partial \mathbf{r}} - \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{\partial G}{\partial \dot{\mathbf{r}}}\right] = 0. \tag{45}$$

In our case, the integrand is $G = sF = s\sqrt{\dot{\mathbf{r}} \cdot \dot{\mathbf{r}}}$, and substitution into Eq. (45) yields

$$s\ddot{\mathbf{r}} + (\dot{\mathbf{r}} \cdot \nabla s)\dot{\mathbf{r}} - (\dot{\mathbf{r}} \cdot \dot{\mathbf{r}})\nabla s = 0. \tag{46}$$

It is easy to show that only two of these equations are independent (Julian and Gubbins, 1977), and that one may be ignored without loss of generality. This leaves two equations with three unknowns $\mathbf{r} = (x, y, z)$; a final constraint comes from $\mathrm{d}F/\mathrm{d}\lambda = 0$ (so $(\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}}) = 0$). Thus, a system of three independent non-linear second-order differential equations can be explicitly written (Julian and Gubbins, 1977)

$$\left.\begin{aligned} &s\ddot{x} + s_y\dot{y}\dot{x} + s_z\dot{z}\dot{x} - s_x\left(\dot{y}^2 + \dot{z}^2\right) = 0 \\ &s\ddot{y} + s_x\dot{x}\dot{y} + s_z\dot{z}\dot{y} - s_y\left(\dot{x}^2 + \dot{z}^2\right) = 0 \\ &\dot{x}\ddot{x} + \dot{y}\ddot{y} + \dot{z}\ddot{z} = 0 \end{aligned}\right\}, \tag{47}$$

where $\nabla s = (s_x, s_y, s_z)$. The boundary conditions for this problem are $\mathbf{r}(0) = \mathbf{r}_S$ and $\mathbf{r}(1) = \mathbf{r}_R$, and an iterative non-linear solution approach is possible given some initial estimate of the path $\mathbf{r}(\lambda)^0$, so that in general $\mathbf{r}(\lambda)^{n+1} = \mathbf{r}(\lambda)^n + \delta\mathbf{r}(\lambda)^n$. Substitution of this expression into Eq. (47) and linearising the resulting equations for $\delta\mathbf{r}(\lambda)^n$ allows solutions to be obtained using, for example, second order finite difference techniques (Julian and Gubbins, 1977). The iterative process can be continued until some suitable convergence criterion, based on the path perturbation integrated along the ray, is satisfied.

Pereyra *et al.* (1980) devise a similar technique to that described above for ray bending, but extend it to allow for the presence of interfaces. This can be achieved by considering a separate system of differential equations in each smooth region, and coupling them using the known discontinuity condition at each interface that is traversed by the ray path. The order in which the interfaces are intersected by the ray path needs to be known in advance, which may be a drawback in complex structures. This scheme is developed further in very complex 3-D models by Pereyra (1996), who also implements shooting to obtain an initial guess ray. This helps to overcome the problem of knowing *a priori* the correct path sequence.

2.2.1. *Pseudo-Bending Methods*

Pseudo-bending methods are similar in principle to the bending scheme described above, but avoid direct solution of the ray equations. One of the first of these schemes was developed by Um and Thurber (1987), and is based upon the ray path being represented by a set of linearly interpolated points. Given some initial arbitrary path, the aim is to sequentially adjust the location of each point so that the path better satisfies the ray equations. This can be accomplished quite efficiently by locating the direction of the ray path normal and then directly exploiting Fermat's principle of stationary time. If we denote $\mathbf{m}_t$ and $\mathbf{m}_n$ as vectors tangent and normal to the ray path respectively at some point $\mathbf{r}$ (see Fig. 12a), then

$$\mathbf{m}_t = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l} \quad \text{and} \quad \mathbf{m}_n = \frac{\mathrm{d}\mathbf{m}_t}{\mathrm{d}l}, \tag{48}$$

where $\mathbf{m}_t$ is a unit vector. By using Eq. (5), $\mathbf{m}_n$ can be written

$$\mathbf{m}_n = \frac{\mathrm{d}}{\mathrm{d}l}\left(\frac{\nabla T}{s}\right) = \frac{1}{s}\frac{\mathrm{d}\nabla T}{\mathrm{d}l} + s\nabla T\frac{\mathrm{d}}{\mathrm{d}l}\left(\frac{1}{s}\right). \tag{49}$$

$\mathbf{m}_n$ can be expressed completely in terms of s and $\mathbf{m}_t$ by making use of Eq. (7), so that

$$\mathbf{m}_n = \frac{\nabla s}{s} + \nabla T\frac{\mathrm{d}}{\mathrm{d}\mathbf{r}}\left(\frac{1}{s}\right)\cdot\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}l} = \frac{1}{s}\big(\nabla s - (\nabla s\cdot\mathbf{m}_t)\mathbf{m}_t\big). \tag{50}$$

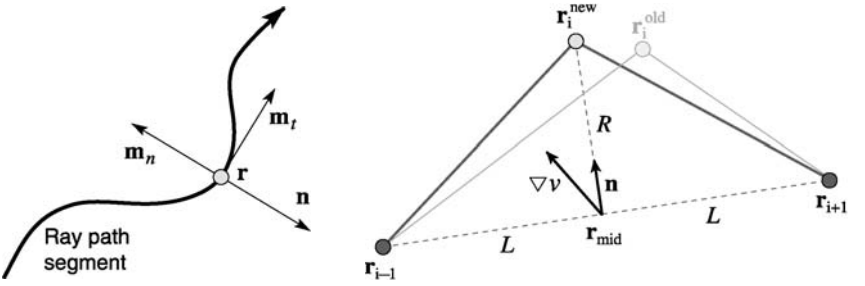


FIG. 12. Schematic representation of parameters used in the pseudo bending method. (a) Definition of ray tangent $\mathbf{m}_t$, ray normal $\mathbf{m}_n$ and antinormal unit vector $\mathbf{n}$ at a point $\mathbf{r}$ along a ray path; (b) stencil for the three point perturbation scheme.

If we now let $\mathbf{n}$ represent the anti-normal unit vector to the ray path at $\mathbf{r}$, and substitute velocity v for slowness s, then

$$\mathbf{n} = \frac{\nabla v - (\nabla v \cdot \mathbf{m}_t)\mathbf{m}_t}{|\nabla v - (\nabla v \cdot \mathbf{m}_t)\mathbf{m}_t|} \tag{51}$$

which is equivalent to the expression derived by Um and Thurber (1987). The vector $\mathbf{n}$ thus defines the direction of ray path curvature.

A three-point perturbation scheme is devised by Um and Thurber (1987) to sequentially update points along a path. Consider Fig. 12b, which shows three points $\mathbf{r}_{i-1}$, $\mathbf{r}_i$ and $\mathbf{r}_{i+1}$. The aim is to replace the initial guess point $\mathbf{r}_i^{\text{old}}$ with an improved estimate $\mathbf{r}_i^{\text{new}}$. The improved estimate is obtained by considering a perturbation to the point $\mathbf{r}_{\text{mid}}$, which lies at the midpoint between $\mathbf{r}_{i-1}$ and $\mathbf{r}_{i+1}$. The vector $\mathbf{m}_t$ can then be simply approximated by

$$\mathbf{m}_t = \frac{\mathbf{r}_{i+1} - \mathbf{r}_{i-1}}{|\mathbf{r}_{i+1} - \mathbf{r}_{i-1}|} \tag{52}$$

and the anti-normal unit vector $\mathbf{n}$, which specifies the bending direction, is computed from Eq. (51). The next step is to find the distance R in the direction $\mathbf{n}$ which results in an improved estimate of the path. An approximate analytic expression for the traveltime $T(R)$ between $\mathbf{r}_{i-1}$ and $\mathbf{r}_{i+1}$ can be obtained using the trapezoidal rule

$$T(R) = \sqrt{L^2 + R^2}\left[\frac{1}{v_i^{\text{new}}} + \frac{1}{2}\left(\frac{1}{v_{i-1}} + \frac{1}{v_{i+1}}\right)\right]. \tag{53}$$

The appropriate value for R can be obtained by appealing directly to Fermat's principle of stationary time, which in this case equates to setting $\mathrm{d}T/\mathrm{d}R = 0$.

Therefore,

$$\left.\begin{aligned}\frac{\mathrm{d}T}{\mathrm{d}R} &= \sqrt{L^2+R^2}\,\frac{\mathrm{d}}{\mathrm{d}R}\left(\frac{1}{v_i^{\mathrm{new}}}\right)+\left(\frac{1}{v_i^{\mathrm{new}}}+c\right)\frac{R}{\sqrt{L^2+R^2}}\\ &= \left(L^2+R^2\right)\frac{\mathrm{d}v_i^{\mathrm{new}}}{\mathrm{d}R}-Rv_i^{\mathrm{new}}\left(cv_i^{\mathrm{new}}+1\right)=0\end{aligned}\right\}, \tag{54}$$

where $c = (1/v_{i-1} + 1/v_{i+1})/2$. The quantity v_i^{new} is unknown, but a first-order accurate (in R) estimate based on the velocity and velocity gradient at $\mathbf{r}_{\mathrm{mid}}$ can be made

$$v_i^{\mathrm{new}} \approx v_{\mathrm{mid}} + (\mathbf{n}\cdot\nabla v_{\mathrm{mid}})R. \tag{55}$$

Substitution of this expression into Eq. (54) results in

$$\begin{aligned}&\left[c(\mathbf{n}\cdot\nabla v_{\mathrm{mid}})^2\right]R^3-\left[2cv_{\mathrm{mid}}(\mathbf{n}\cdot\nabla v_{\mathrm{mid}})\right]R^2-\left[v_{\mathrm{mid}}(cv_{\mathrm{mid}}+1)\right]R\\ &\quad+(\mathbf{n}\cdot\nabla v_{\mathrm{mid}})L^2=0.\end{aligned} \tag{56}$$

Um and Thurber (1987) ignore the cubic term of R, which results in a quadratic equation with two roots. The solution which produces real positive traveltimes is

$$R=\frac{-v_{\mathrm{mid}}(cv_{\mathrm{mid}}+1)+\sqrt{v_{\mathrm{mid}}^2(cv_{\mathrm{mid}}+1)^2+8cv_{\mathrm{mid}}(\mathbf{n}\cdot\nabla v_{\mathrm{mid}})^2L^2}}{4cv_{\mathrm{mid}}(\mathbf{n}\cdot\nabla v_{\mathrm{mid}})}. \tag{57}$$

The update procedure for pseudo bending as outlined above is very simple and computationally efficient, as it only requires two relatively simply equations to be solved (Eqs. (51) and (57)). In practice, Um and Thurber (1987) apply the update scheme simultaneously from both end points of the ray path to the central point. This process is repeated until a convergence criterion is met. An initial path can be approximated in various ways, but one simple option is to begin with a three point ray joining source and receiver. Once the central point has been perturbed, two new points are introduced that bisect each line segment. The central three points can then be relocated, before four new points are introduced in the same way as before (see Fig. 13). This process can be continued until a suitably accurate path is obtained.

Despite the relatively crude approximations made in pseudo bending, Um and Thurber (1987) find it to be much more computationally efficient than conventional bending schemes; consequently, it has become quite popular for problems which require large traveltime datasets to be predicted, e.g. 3-D local earthquake tomography (Eberhart-Phillips, 1990; Scott *et al.*, 1994; Eberhart-Phillips and Reyners, 1997; Graeber and Asch, 1999). Zhao *et al.* (1992) modify the three point perturbation scheme of Um and Thurber (1987) to allow for the presence of interfaces. In this case, the sequence of points which discretely defines the path includes points which lie on each interface traversed by the ray. When the update

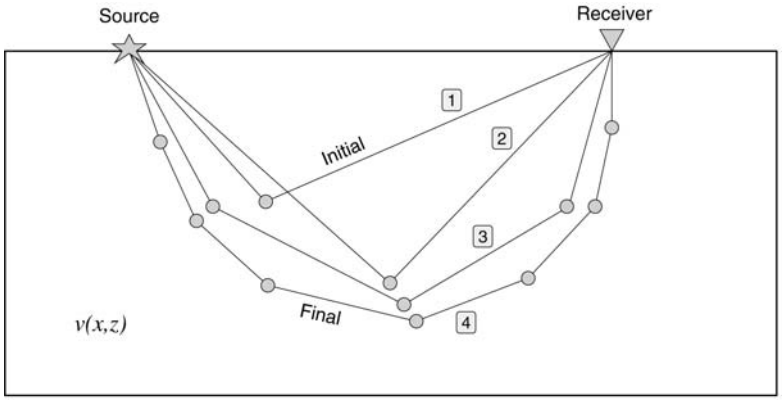


FIG. 13. Principle of the pseudo-bending method of Um and Thurber (1987). In this schematic example, an initial guess ray is defined by three points. The center point is perturbed to satisfy Fermat's principle of stationary time. The number of path segments is then doubled and the process repeated.

scheme reaches these points, they are perturbed along the interface (with adjacent points on either side of the interface held fixed) until Snell's law is satisfied. Koketsu and Sekine (1998) devise a similar scheme in a 3-D spherical coordinate system.

2.2.2. *Other Bending Schemes*

Two classes of bending method have been described above. In the first, the ray equation is linearised and iteratively solved as a boundary value problem with fixed end points. The second approach – pseudo bending – applies a simple perturbation scheme based on a direct application of Fermat's principle of stationary time, to a ray path described by a sequence of points. Although these are the two principal methods, several other schemes have been developed which warrant a brief mention.

Prothero *et al.* (1988) develop a 3-D bending scheme based on the simplex method of function minimisation. An initial path is obtained by using an exhaustive search method to find the minimum-time circular path between source and receiver. Perturbations to this path, described by a sum of sine-wave harmonics, are then made using the simplex method, which searches for the amplitude coefficients that produce the path of least time. Although the method appears to be more robust than pseudo-bending, it is significantly slower.

Like ray shooting, ray bending can also be carried out using fully non-linear update schemes. Sadeghi *et al.* (1999) develop a method which uses genetic algorithms (GAs) to globally search for the minimum time ray path between two fixed points. The scheme is similar to pseudo-bending in that paths are described by a set of linearly interpolated points, which are perturbed until a convergence criterion is satisfied. However, in this case, a population of multiple two-point

paths join source and receiver, and the GA drives the bending of paths until the traveltime converges to a minimum. When this occurs, all two point rays within the population should follow almost identical paths.

Dębski and Ando (2004) develop a so-called "spectral ray tracer", which bares some resemblance to the scheme of Prothero *et al.* (1988), except that ray paths are parameterised as a series of Chebyshev polynomials. The bending problem can then be formulated as one of function minimisation, in which the decomposition coefficients of the Chebyshev polynomials become the variables to be adjusted until the two point traveltime is minimised. Instead of adopting a linearised approach to the optimisation problem, Dębski and Ando (2004) use a genetic algorithm to generate and select the decomposition components.

Most schemes for solving the boundary value problem in ray tracing can usually be characterised as either shooting or bending. However, other schemes do exist, most notably those that are based on structural perturbation (Červený, 2001). In this type of scheme, a known two point path exists in a reference medium, and the aim is to locate the equivalent two point path in a medium that is slightly modified from the reference medium. Solution of this class of problem can be achieved using ray perturbation theory, which is described in various papers including Farra and Madariaga (1987); Snieder and Sambridge (1992); Pulliam and Snieder (1996). Snieder and Spencer (1993) show that ray bending and ray perturbation theories can in fact be combined into a single perturbation theory.

3. Grid Based Schemes

An alternative to tracing rays between source and receiver is to compute the traveltime of the evolving wavefront at all points of a grid which spans the medium. The complete traveltime field implicitly contains the wavefront location as a function of time (i.e. isochrons of $T(\mathbf{x})$) and all possible ray path trajectories (specified by ∇T). Compared to conventional shooting and bending methods of ray tracing (Julian and Gubbins, 1977; Cassell, 1982; Um and Thurber, 1987; Virieux and Farra, 1991; Koketsu and Sekine, 1998), grid based traveltime schemes have a number of clear advantages: (1) Most are capable of computing traveltimes to all points of a medium, and will locate diffractions in ray shadow zones; (2) the non-linearity of both ray shooting and bending means that they may fail to converge to a true two-point path, whereas most grid based schemes are highly stable and will find the correct solution even in strongly heterogeneous media; (3) grid based schemes can be very efficient in computing traveltime and path information to the level of accuracy required by practical problems. Ray tracing schemes can be inefficient if solution non-linearity is significant; (4) most grid-based schemes consistently find first-arrivals in continuous media. It is often difficult to ascertain with ray tracing whether the located path is a first or later arrival.

Despite these advantages, grid based schemes have a number of limitations which should be considered prior to application. These include: (1) accuracy is a function of grid spacing – in 3-D halving the spacing of a grid will increase computation time by at least a factor of 8. Thus, computation time may become unacceptable if highly accurate traveltimes are required; (2) most practical schemes compute first-arrivals only – thus, features such as wavefront triplications cannot be predicted; (3) quantities other than traveltime (such as amplitude) are difficult to compute accurately without first extracting path geometry and applying ray based techniques. Two grid-based schemes – finite difference solution of the eikonal equation and shortest path methods – have emerged in the last few decades as popular alternatives to conventional ray tracing, and are described below.

3.1. Eikonal Solvers

One of the first grid schemes based on finite difference solution of the eikonal equation was proposed by Vidale (1988). From a given source point, traveltimes are progressively computed outwards along an expanding square in 2-D (see Fig. 14). Traveltimes to points which lie in the initial square around the source point are computed as follows. The four points $(i \pm 1, j)$ and $(i, j \pm 1)$ have traveltimes estimated by the formulae $T_{i\pm1,j} = \delta x/2(s_{i\pm1,j} + s_{i,j})$ and $T_{i,j\pm1} = \delta z/2(s_{i,j\pm1} + s_{i,j})$, where s is slowness and δx and δz are grid spacing in x and z. The remaining four points are computed by appealing to the eikonal equation (Eq. (3)); in particular, the ∇T term can be approximated in the cell

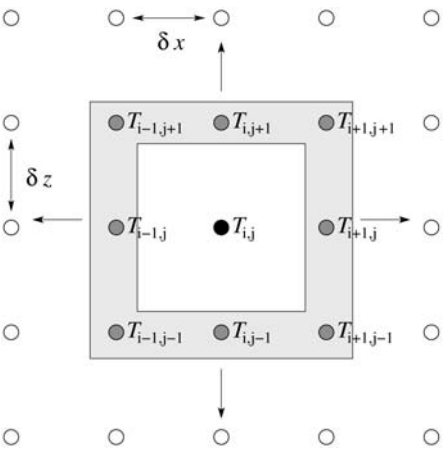


FIG. 14. The expanding square method for progressive calculation of traveltimes throughout a gridded velocity field.

defined by points (i, j), $(i+1, j)$, $(i, j+1)$ and $(i+1, j+1)$ by

$$\left.\begin{aligned}\frac{\partial T}{\partial x} &= \frac{T_{i,j} + T_{i,j+1} - T_{i+1,j} - T_{i+1,j+1}}{2\delta x} \\ \frac{\partial T}{\partial z} &= \frac{T_{i,j} + T_{i+1,j} - T_{i,j+1} - T_{i+1,j+1}}{2\delta z}\end{aligned}\right\}. \tag{58}$$

Substitution into Eq. (3) produces the following quadratic equation

$$\frac{(T_{i,j} + T_{i,j+1} - T_{i+1,j} - T_{i+1,j+1})^2}{\delta x^2} + \frac{(T_{i,j} + T_{i+1,j} - T_{i,j+1} - T_{i+1,j+1})^2}{\delta z^2} = 4\bar{s}^2, \tag{59}$$

where $\bar{s}$ is the average slowness of all four points defining the cell. This equation can easily be solved for $T_{i+1,j+1}$. If $h = \delta x = \delta z$ (Vidale, 1988), then the solution to Eq. (59) has the simple form $T_{i+1,j+1} = T_{i,j} + \sqrt{2(h\bar{s})^2 - (T_{i,j+1} - T_{i+1,j})^2}$. Vidale (1988) also defines a solution stencil for a locally circular wavefront. This allows a mixed scheme to be devised, which uses the locally circular assumption in regions of high wavefront curvature (e.g. in the source neighbourhood), and Eq. (59) when the wavefront is more planar.

As the computational front evolves outward from the source, points within the square band cannot have traveltimes computed in arbitrary order; causality at least requires that new traveltimes be computed using only those traveltimes from surrounding points of lesser value. Vidale (1988) devises a scheme which sweeps through each of the four sides of the square in order to locate the minimum time solution at each point. The resulting computational scheme is both fast and accurate, with CPU time being approximately proportional to the number of points defining the grid. The method is readily extendable to 3-D, as demonstrated by Vidale (1990).

The use of an expanding square formalism to define the shape of the computational front cannot always respect the direction of flow of traveltime information through the medium. This is demonstrated in Fig. 15, which shows how the expanding square can fail if the first-arriving path to a point inside the square needs to sample structure outside the square. As a consequence, first-arrivals are not always guaranteed, which can ultimately lead to instability. Nevertheless, the basic scheme proposed by Vidale (1988) remains popular, and its stability has been improved upon by introducing new features to the finite difference stencil and evolution of the computational front. Both Hole and Zelt (1995) and Afnimar and Koketsu (2000) introduce special headwave operators to better deal with the presence of strong velocity contrasts, and Hole and Zelt (1995) implement an iterative post-sweeping scheme to help account for the non-causal nature of the expanding square. van Trier and Symes (1991) use entropy-satisfying first-order upwind

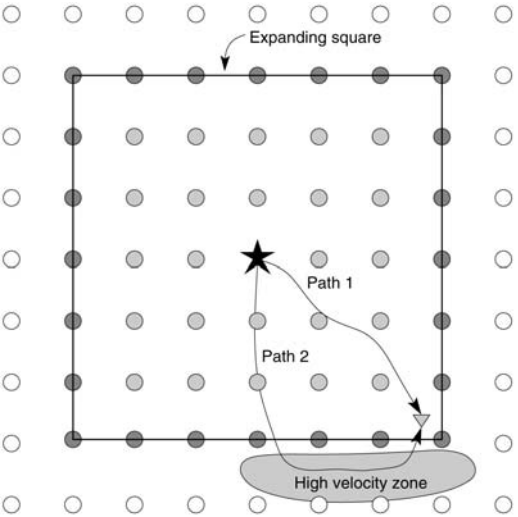


FIG. 15. Schematic illustration of how the expanding square method can fail. In this example, path 1 would be computed by the expanding square, but path 2 actually arrives first.

difference operators to improve the computational efficiency of the method and better deal with wavefront discontinuities. Comparable improvements are made by Podvin and Lecomte (1991), who systematically apply Huygen's principle in the finite difference approximation in order to account for all possible modes of propagation.

More recently, Essentially Non-Oscillatory (ENO) finite difference schemes developed in the field of computational mathematics (Shu and Osher, 1988, 1989) have been used to solve the eikonal equation within an expanding square framework. The attraction of ENO schemes is that they can be readily extended to very high orders of accuracy, yet remain stable. Kim and Cook (1999) present a scheme, which they describe as ENO-DNO-PS, to efficiently compute first-arrival traveltime fields. DNO (or *down 'n' out*) refers to the expanding box scheme first introduced by Vidale (1988), and PS refers to post-sweeping, which is an iterative correction strategy applied to the traveltime field computed by ENO-DNO in an attempt to deal with causality breaches introduced by the expanding box. This correction scheme is similar in principle to the one used by Hole and Zelt (1995). Fig. 16 demonstrates the importance of post-sweeping in complex models; the expanding box results in the overestimation of traveltimes in various regions of the model (dashed lines), which appear to be corrected by the post-sweeping step (solid line). A third-order WENO-DNO-PS solver is used in this example, which is slightly different to that used by Kim and Cook (1999). Weighted Essentially Non-Oscillatory or WENO schemes have a number of advantages over ENO in-

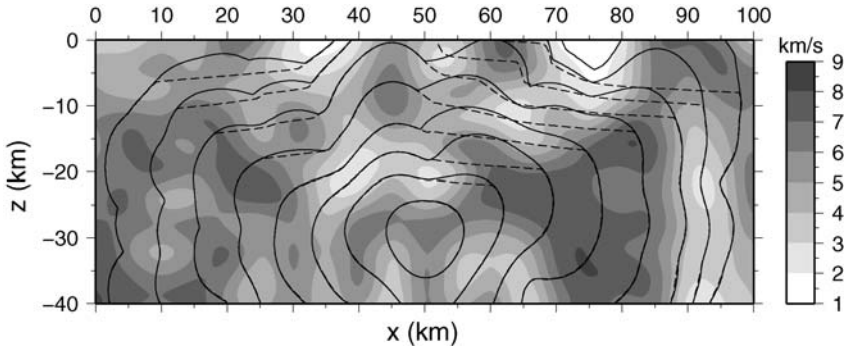


FIG. 16. Illustration of how post-sweeping can correct for errors due to the use of an expanding box strategy for computing traveltimes. Wavefronts for WENO-DNO denoted by dashed lines; wavefronts for WENO-DNO-PS denoted by solid lines.

cluding computation time and stability (Liu *et al.*, 1994; Jiang and Shu, 1996; Jiang and Peng, 2000).

Other authors to use ENO-type schemes for seismic traveltimes include Qian and Symes (2002), who implement a WENO scheme with adaptive gridding, and Buske and Kästner (2004), who use an ENO scheme in polar coordinates; in both cases, the computed traveltimes are sufficiently accurate to solve the transport equation (Eq. (4)) and obtain amplitudes.

Rather than use an expanding square as the computational front for traveltimes, it is also possible to use the expanding wavefront itself (Qin *et al.*, 1992; Cao and Greenhalgh, 1994). In the case of Qin *et al.* (1992), the finite difference stencil proposed by Vidale (1988) is retained, and traveltimes to the first band of points surrounding the source are computed in the same way. However, thereafter the point of global minimum traveltime along the perimeter of points computed so far is used as the next source to locally expand the solution region. This approach ensures that the shape of the computational front conforms to that of the first-arriving wavefront, which minimises the possibility of computing arrivals other than the first. The need to locate the global minimum traveltime along the computational front requires an ordered storage of traveltimes (e.g. heap sorting), which is less efficient than the expanding box regime.

Eikonal solvers have been used in a variety of seismological applications, and are particularly useful for those classes of problems that require large traveltime datasets to be predicted, such as seismic imaging. For example, they have been frequently used in 3-D refraction/reflection tomography (Hole, 1992; Zelt, 1996; Riahi *et al.*, 1997; Zelt and Barton, 1998; Zelt, 1999; Zelt *et al.*, 2001; Day *et al.*, 2001) and in the migration of coincident reflection sections (Gray and May, 1994; Bevc, 1997; Buske, 1999a, 1999b).

3.1.1. The Fast Marching Method

One of the more recently developed grid based eikonal solvers which is both highly robust and computationally efficient is the so called Fast Marching Method (Sethian, 1996, 1999; Popovici and Sethian, 2002) or FMM. It was originally developed in the field of computational mathematics for solving various types of interface evolution problems, and to date has been applied in numerous areas of the physical sciences including optimal path planning, medical imaging, geodesics, and photolithographic development (Sethian, 1999, 2001). In seismology, FMM has been used in the migration of coincident reflection profiles (Popovici and Sethian, 2002) and teleseismic tomography (Rawlinson *et al.*, 2006a, 2006b).

A common feature of first-arrival traveltime fields, particularly in complex media, is that they are not spatially differentiable at every point (i.e. the ∇T term in Eq. (3) is not defined everywhere). This can lead to instability in schemes that do not implicitly or explicitly recognise this behaviour (e.g. Vidale, 1988; Qin *et al.*, 1992). When a wavefront self-intersects (i.e. multipathing occurs), the first-arrival wavefront will contain a kink or discontinuity which would normally be spanned by the later arriving wavefront. However, since only first-arrival traveltimes are computed, this information is discarded, and to ensure stability, it becomes important not to use traveltimes from both sides of the discontinuity in the same finite difference stencil to compute new traveltimes. One way to overcome this problem is to solve the viscous version of the eikonal equation, which smooths out discontinuities; the limit of smooth solutions is a weak solution that corresponds to the first arriving wavefront. It turns out that the viscous limit solution can also be obtained by solving Eq. (3) using upwind entropy-satisfying operators which take into account the direction of flow of information when evaluating ∇T.

A commonly used entropy-satisfying upwind scheme (e.g. Sethian and Popovici, 1999; Chopp, 2001; Sethian and Popovici, 2002) may be expressed as follows

$$\begin{aligned}\Big[&\max\big(D_a^{-x}T, -D_b^{+x}T, 0\big)^2 + \max\big(D_c^{-y}T, -D_d^{+y}T, 0\big)^2 \\ &+ \max\big(D_e^{-z}T, -D_f^{+z}T, 0\big)^2\Big]_{ijk}^{1/2} = s_{i,j,k},\end{aligned} \tag{60}$$

where (i, j, k) are grid increment variables in any orthogonal coordinate system (x, y, z), and the integer variables a, b, c, d, e, f define the order of accuracy of the upwind finite difference operator used in each of the six cases. For example, in a Cartesian coordinate system, the first and second-order operators for $D^{-x}T_i$ are

$$\begin{aligned}D_1^{-x}T_{i,j,k} &= \frac{T_{i,j,k} - T_{i-1,j,k}}{\delta x}, \\ D_2^{-x}T_{i,j,k} &= \frac{3T_{i,j,k} - 4T_{i-1,j,k} + T_{i-2,j,k}}{2\delta x},\end{aligned} \tag{61}$$

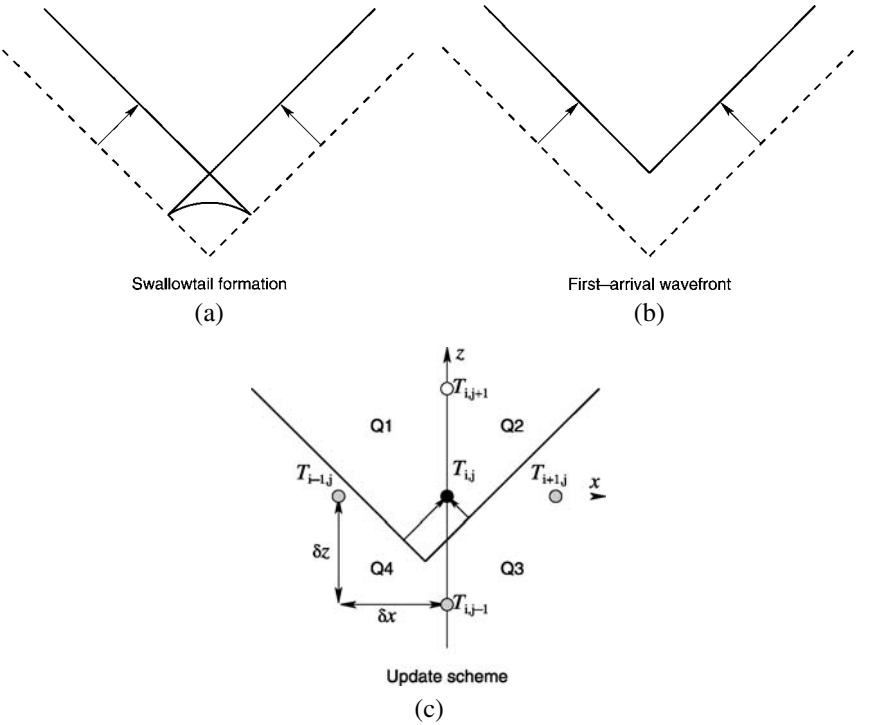


FIG. 17. Discontinuities in first-arrival wavefronts usually arise from discarding later-arriving information. (a) Formation of a swallowtail from an initial wavefront (dashed line); (b) the first-arrival wavefront contains a discontinuity; (c) the upwind entropy-satisfying finite difference stencil properly respects the flow of information by considering solutions from each quadrant.

where δx is the grid spacing in x. First-order accurate schemes only use D_1 operators and second-order accurate schemes preferentially use D_2 operators. Strictly speaking, the second-order method is really mixed-order because it will use first order approximations when causality does not permit the use of the required operator. For example, if we implement a second-order method and $T_{i-1} > T_{i-2}$, then the operator used would be $D_2^{-x} T_i$; while if $T_{i-1} < T_{i-2}$ we would have to resort to $D_1^{-x} T_i$. Mixed schemes which use D_3 or even higher order operators could also be devised, but in practice, the need to resort to D_1 operators on occasion means that improvements in overall accuracy are usually not substantial (Rawlinson and Sambridge, 2004b). It should also be noted that the first-order FMM scheme has been proven to be unconditionally stable (Sethian, 2001), but no such proof exists for higher order schemes.

Figure 17 schematically illustrates why the entropy-satisfying upwind finite difference stencil of Eq. (60) preserves stability in the presence of a wavefront discontinuity. In this case, the complete wavefront includes a swallowtail, but this

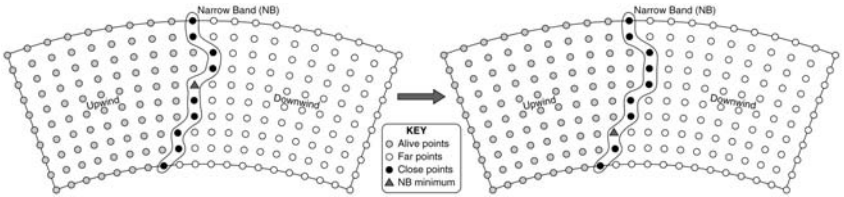


FIG. 18. Narrow band (NB) evolution scheme used for the ordered update of grid points. *Close* points have trial traveltimes computed using Eq. (60) (using appropriate operators for polar coordinates in this case). The narrow band locally advances from the *close* point with minimum traveltime.

is discarded and the first-arrival wavefront propagates with a discontinuity. The solution of Eq. (60), in 2-D media with first-order operators, can be obtained by considering each quadrant of Fig. 17c, where the aim is to compute $T_{i,j}$ from known traveltimes at surrounding points. In Q1 and Q2, the only approximations that can be made are $T_{i,j} = T_{i-1,j} + \delta x s_{i,j}$ and $T_{i,j} = T_{i+1,j} + \delta x s_{i,j}$ respectively, which are not even first-order accurate. However, in Q3 and Q4, which both have two known traveltimes, true first-order accurate traveltimes can be obtained. In the case of Q3, the quadratic equation is

$$\left(\frac{T_{i+1,j} - T_{i,j}}{\delta x}\right)^2 + \left(\frac{T_{i,j} - T_{i,j-1}}{\delta z}\right)^2 = (s_{i,j})^2. \tag{62}$$

Of the two possible solutions, the larger one corresponds to the correct local plane wave approximation. Similarly in quadrant Q4, the quadratic equation is

$$\left(\frac{T_{i,j} - T_{i-1,j}}{\delta x}\right)^2 + \left(\frac{T_{i,j} - T_{i,j-1}}{\delta z}\right)^2 = (s_{i,j})^2, \tag{63}$$

where again the larger solution corresponds to the correct local plane wave approximation. Of the four possible solutions that have been identified, the one with minimum traveltime is correct; in this case, it is the plane wave solution from Q3. The extension to 3-D and higher-order operators is straightforward, although the number of candidate solutions obviously increases.

The FMM stencil encapsulated in Eq. (60) describes how to calculate new traveltimes using known traveltimes from neighbouring grid points, but in order to populate a grid of nodes with traveltime values, the order in which nodes are updated must be consistent with the direction of flow of information; that is, from smaller values of T to large values of T. FMM achieves this by systematically constructing traveltimes in a downwind fashion from known values upwind by employing a *narrow band* approach. This is illustrated in Fig. 18. *Alive* points have their values correctly calculated, *close* points lie within the narrow band and have trial values computed using Eq. (60) with *alive* points only, and *far* points have no values calculated. The narrow band is evolved by identifying the *close*

point with minimum traveltime, tagging it as *alive* and then updating any adjacent *close* or *far* point, the latter being re-tagged as *close*. Using this approach, the shape of the narrow band approximates the shape of the first-arrival wavefront, and the idea is to propagate the band through the grid until all points become alive. This basic approach is similar to that advocated by Qin *et al.* (1992). In fact, looking back at the literature produced by the seismology community on grid-based eikonal solvers, it is evident that all the ideas on which FMM is based have been considered at one time or another, but had not previously been combined in this particular way.

The use of a heap sort algorithm to identify the node with global minimum traveltime within the narrow band means that FMM has an operation count of $O(M \log M)$, where M is the total number of grid points. Thus, in order to achieve greatest efficiency in practical applications (particularly in 3-D), it is important to choose the largest grid spacing permitted by the uncertainty in the data.

The stability of FMM is demonstrated by Fig. 19, which shows the evolution of a wavefront in a highly complex medium. In this case, FMM has been solved in spherical shell coordinates using the appropriate spherical form of ∇T in Eq. (3) with constant radius r. This particular implementation would be useful for tracking the evolution of higher frequency surface waves e.g. for surface wave tomography using ambient seismic noise (Shapiro *et al.*, 2005; Yao *et al.*, 2006). The velocity variations shown in Fig. 19 are somewhat pathological, but the example seeks to demonstrate the robustness of FMM. Discontinuities in the first-arriving wavefront clearly develop, but are successfully evolved without the introduction of instabilities. Further FMM examples and detailed analysis of computational efficiency and accuracy can be found in Rawlinson and Sambridge (2004b, 2004a); de Kool *et al.* (2006).

3.1.2. Improving Accuracy in the Source Neighbourhood

A point source is an upwind singularity of the traveltime field and can be a major contributor to the overall error of finite difference eikonal solvers due to high wavefront curvature in the source neighbourhood. Schemes that use regular traveltime grids often poorly approximate this curvature, and the resulting error is carried forward in subsequent calculations. This problem has been recognised and addressed in a variety of ways. As alluded to earlier, Vidale (1988) formulates two finite difference stencils, one that is most accurate for locally plane wavefronts, and one that is most accurate for locally circular wavefronts. Using the circular finite difference scheme in the source vicinity would largely address this problem, although it is difficult to gauge the stability of this approach.

Other schemes for addressing near-source error include using spherical grids centered on the source point (Alkhalifah and Fomel, 2001), and local grid refinement in the source neighbourhood (Kim and Cook, 1999). In the latter case, grid spacing is progressively increased away from the source as wavefront curvature

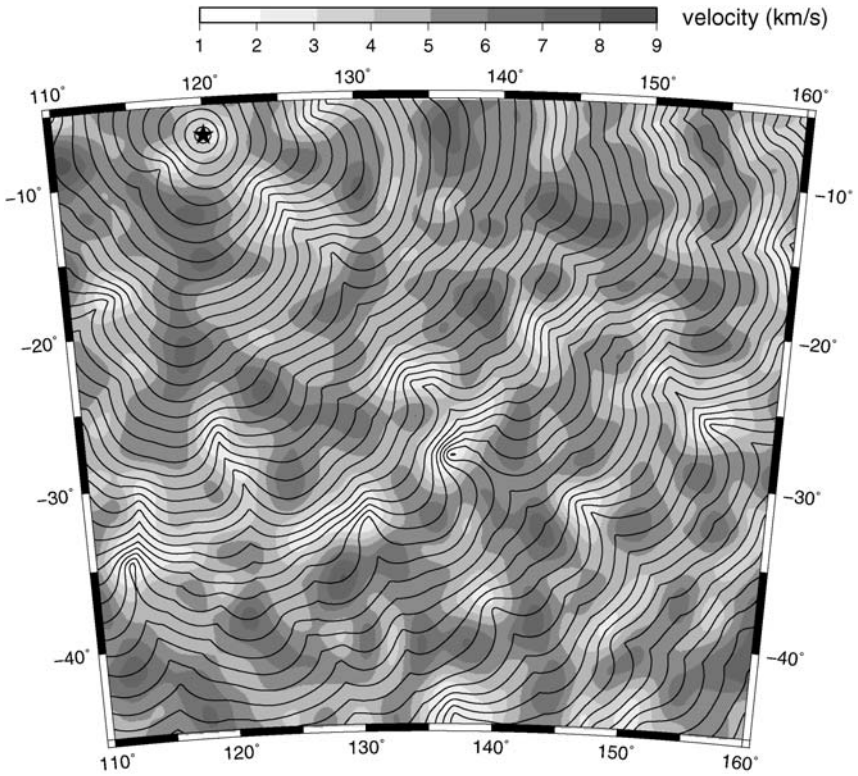


FIG. 19. Wavefronts implicitly computed by FMM in a complex velocity field. Note the formation and propagation of wavefront discontinuities. Wavefronts are contoured at 20 s intervals, and the synthetic model is in spherical shell coordinates.

decreases. Adaptive gridding with grid refinement and coarsening based on *a posteriori* error estimation (Qian and Symes, 2002) has also been successfully used to minimise near source error. Another approach is to solve the eikonal equation in the so-called celerity domain. Celerity is defined as the distance from the source divided by the traveltime from the source, and transformation of the problem into this domain can significantly reduce source error (Zhang *et al.*, 2005).

In the context of FMM, Rawlinson and Sambridge (2004a) develop a grid refinement strategy that significantly improves the CPU time versus accuracy trade-off. The scheme is similar to that of Kim and Cook (1999), except that only a single level of refinement is used, and the edge of the refined grid ultimately conforms to the shape of the first-arrival wavefront. This latter property ensures that the stability of the scheme is not compromised. Figure 20 schematically illustrates how the computational front evolves through the refined grid, then is mapped back onto the coarser grid as soon as the wavefront reaches the edge of

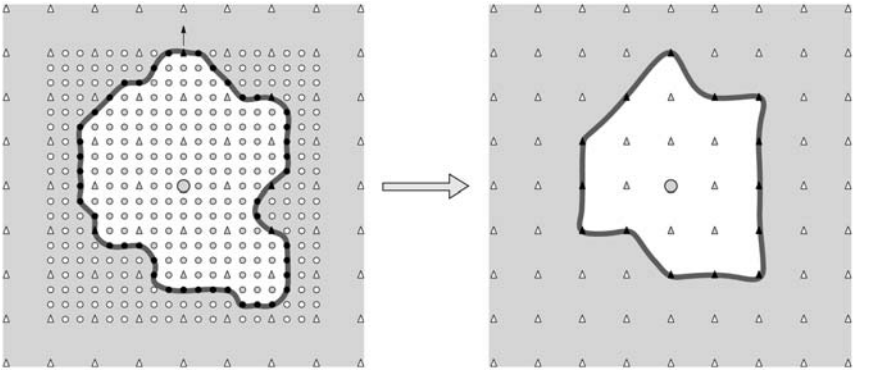


FIG. 20. Implementation of source-grid refinement when the computational front approximates the first-arrival wavefront. When the narrow band impinges on the edge of the refined grid, it is mapped onto the underlying coarse grid before continuing to evolve.

the refined region. Rawlinson and Sambridge (2004a) show that in 2-D, accuracy can be improved by an order of magnitude with little increase in computing time.

3.1.3. Reflections and Transmissions

Grid-based eikonal solvers have largely been used to compute first-arrival traveltime fields in continuous media, and in fact, most methods cannot be naturally extended to track later arrivals, as they achieve their efficiency by only computing a single valued traveltime field. However, it is possible to compute reflection and transmission phases, thereby allowing much more of the complete wavefield to be simulated. Podvin and Lecomte (1991) and Riahi and Juhlin (1994) find reflections by tracking first-arrival traveltime fields from both source and receiver to the entire interface. Fermat's principle of stationary time can then be applied to locate reflection points along the interface. Although this scheme has the advantage that multiple reflection paths can be found for a single interface, traveltime fields need to be computed for all sources and receivers. Hole and Zelt (1995) overcome this problem by explicitly applying Snell's law in a local plane wave approximation in the neighbourhood of an interface.

In an alternative approach, Li and Ulrych (1993) compute reflected and refracted traveltimes in two dimensions by using a local regridding technique to decompose a cell containing an interface into several rectangular and triangular cells so that the true interface shape is better represented by the computational grid. The incident traveltime field is computed using the scheme of Vidale (1988), and the reflected traveltime field is obtained by reinitialising the computational front from the point of minimum traveltime on the interface. The multi-stage FMM scheme of Rawlinson and Sambridge (2004b, 2004a) is similar in principle to the Li and Ulrych (1993) approach, but has been successfully applied to

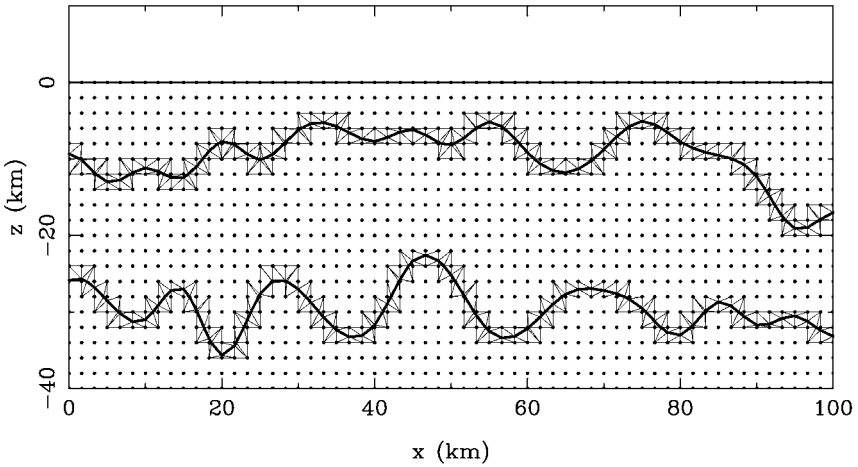


FIG. 21. An adaptive triangular mesh can be used to locally suture the irregular interface nodes to the regular nodes of the velocity grid (grey dots). To aid visualisation, the grid is coarser than would be used in practice. Refer to Rawlinson and Sambridge (2004b, 2004a) for more details.

much more complex velocity models, and allows paths composed of any number of reflection and transmission segments to be tracked. Two main difficulties need to be overcome when reflection and refraction phases are computed: (1) accurate description of the embedded interface; (2) correct propagation of refracted and reflected wavefronts. In a heterogeneous layered media, interfaces vary with depth, and therefore do not conform to the regular distribution of grid nodes used by the velocity field (of course, completely irregular meshes could be used, but ease of implementation and computational cost can suffer). This can be overcome by using a locally adaptive mesh of triangles which sutures regular velocity nodes to irregular interface nodes, as illustrated in Fig. 21. Within the irregular mesh, a first-order accurate scheme for triangular elements is used to update traveltimes; elsewhere, the usual regular scheme (e.g. Eq. (60)) can be applied.

To understand how more than one arrival can be tracked with FMM, consider Fig. 22a, which shows a wavefront propagating through a layer from a point source and impinging on an interface. Rather than continuing to evolve the narrow band, the wavefront is tracked only as far as the interface, which is treated as one of the four boundaries of the computational domain. Once all points on the interface become *alive*, the FMM process is terminated. The next step is to track a reflected or transmitted wavefront into the top or bottom layer respectively; it turns out that this can be done with first-order accuracy by using only the traveltime values of nodes which lie on the interface, with no explicit application of Snell's law. Thus, in order to track a reflection, all nodes in the top layer are set to *far* and all interface nodes are set to *close*. Therefore, the starting narrow band conforms to the shape of the interface, and FMM can be reinitialised. Us-

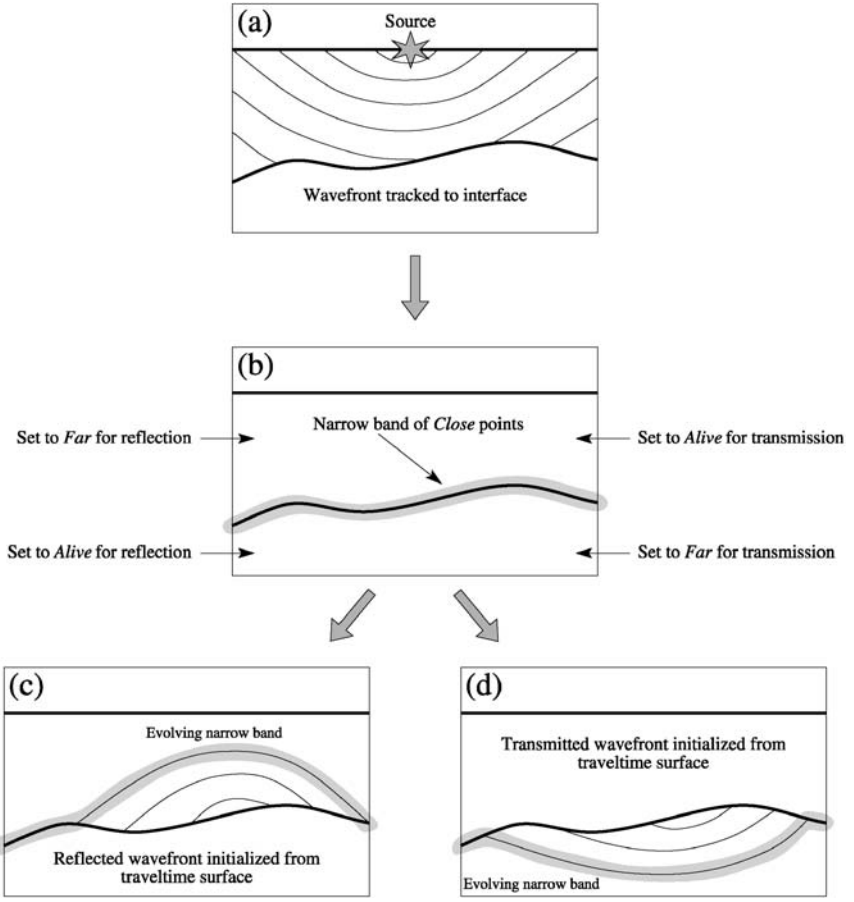


FIG. 22. Principle behind the multi-stage FMM scheme. Each layer represents a separate computational domain in which FMM is initialised either from a source or a layer boundary. (a) Incident wavefront generated from a point source; (b) narrow band defined by a set of interface nodes; (c) reflected wavefront tracked; (d) refracted wavefront tracked.

ing a starting narrow band that does not conform to the shape of the first-arriving wavefront is acceptable provided it is understood that the wavefront cannot reflect from or transmit through the interface more than once. A refracted wavefront can be tracked in the same way, although in this case all nodes in the bottom layer are set to *far*. This basic procedure can be applied iteratively to track phases comprising any number of reflected and transmitted phases in media of virtually any complexity. It should be noted, however, that the arrival found in each case is the absolute first-arrival of the specified phase, and is therefore likely to have small

amplitude, particularly if it has experienced multiple interactions with complex interfaces and velocity gradients.

Figure 23 shows an example of a reflection multiple tracked in the presence of a highly heterogeneous velocity model. As in Fig. 19, the velocity variations are pathological relative to real Earth structure, but the purpose of the example is to demonstrate the robustness of the method. The two-point ray-path (Fig. 23a) favours fast regions of the model and preferentially reflects from convex regions of the interface; this is due to the fact that the first-arrival of this class of phase is tracked. The wavefront becomes more complex as it evolves (Fig. 23b, c) in response to strong velocity contrasts and interface geometry, but the multi-stage FMM remains stable. This stability is highly desirable in many applications including seismic tomography and non-linear earthquake relocation, which typically require many data prediction steps in order to obtain a solution.

In a recent paper, de Kool *et al.* (2006) extend the multi-stage FMM scheme to 3-D spherical coordinates. The new implementation shares many of the basic principles used in the 2-D Cartesian version of Rawlinson and Sambridge (2004a), but several improvements have been implemented. These include discarding the explicit irregular mesh used to suture interface nodes to neighbouring velocity nodes; in 3-D one would need to use an irregular tetrahedral mesh, which can get complicated, particularly for interfaces with high curvature. Instead, a procedure based on testing all combinations of nodes (up to three at a time) adjacent to the node to be updated (interface node or velocity node adjacent to the interface) is used. Each combination of nodes allows a plane wave approximation to be made, and the one that yields the minimum traveltime is chosen. Another procedure built into the new 3-D method allows later-arriving reflections such as the global *PP* phase to be tracked. This is done by initiating FMM from both source and receiver and matching reflection points using Fermat's principle of stationary time. The flexibility of the new scheme is demonstrated in Fig. 24, which shows ray paths from teleseismic and local earthquake sources (extracted from traveltime fields computed using the multi-stage FMM) in a subduction zone setting. [The 3-D FMM software is freely available to download from http://rses.anu.edu.au/seismology/fmmcode.]

3.2. Shortest Path Ray Tracing

Shortest path ray tracing or SPR is another popular method for determining first-arrival traveltimes at all points of a gridded velocity field (Nakanishi and Yamaguchi, 1986; Moser, 1991; Fischer and Lees, 1993; Cheng and House, 1996; Zhao *et al.*, 2004; Zhou and Greenhalgh, 2005). Rather than solve a differential equation, a network or graph is formed by connecting neighbouring nodes with traveltime path segments. Dijkstra-like algorithms can then be used to find the shortest path between a given point and all other points in the network. According

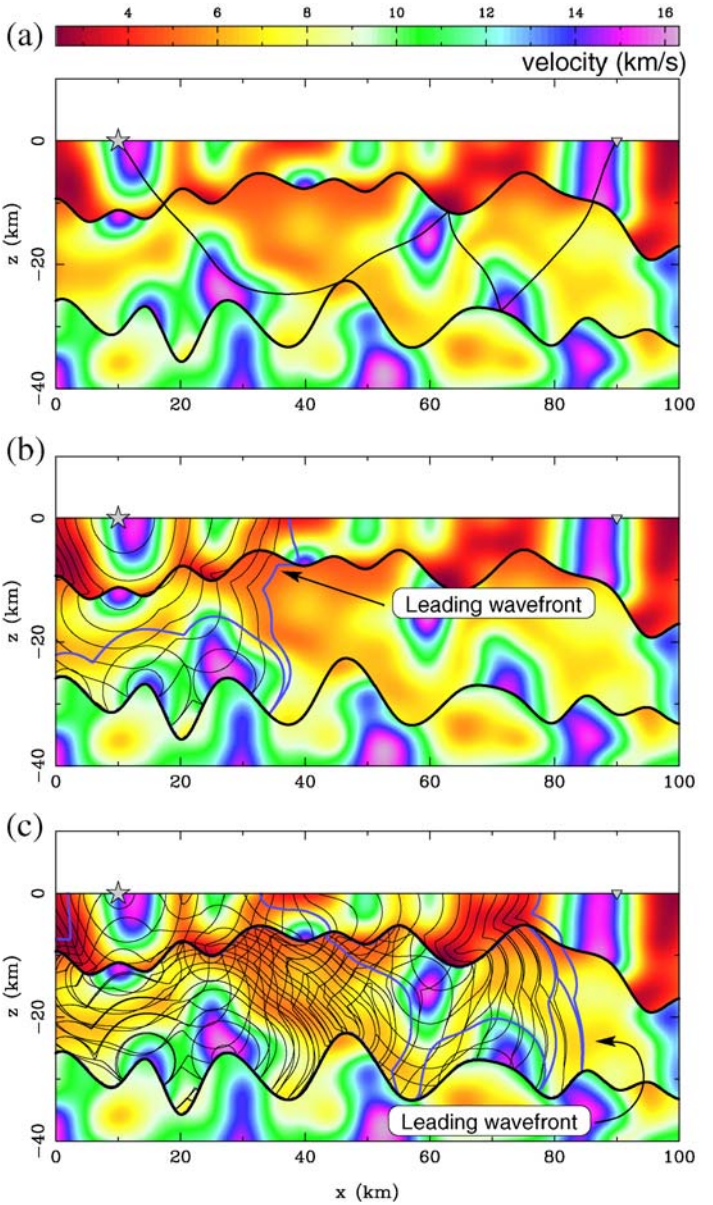


FIG. 23. A reflection multiple computed using the multi-stage FMM scheme of Rawlinson and Sambridge (2004a). (a) A two point ray path corresponding to the reflection multiple, computed after the completion of the multi-stage FMM; (b) snapshot of the evolving wavefront. Isochrons are contoured at 0.4 s intervals, and the leading wavefront is shown in blue; (c) same as (b) but at a later point in time.

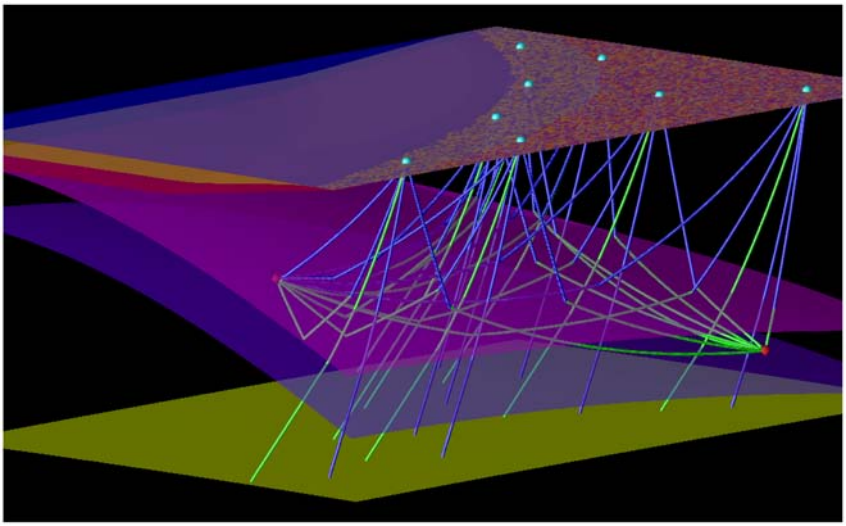


FIG. 24. Ray paths from local earthquake and teleseismic sources, computed using the multi-stage FMM scheme of de Kool *et al.* (2006), in the presence of a subduction zone. Receivers are denoted by blue spheres and local earthquake sources by red diamonds. P-wave paths are blue, and S-wave paths are green. S to P conversions occur at the upper interface of the dipping slab for the local earthquake sources.

to Fermat's principle of stationary time, the shortest time path between two points corresponds to a true ray path.

Shortest path networks are commonly defined in terms of either a cell or a grid centered framework. In a seminal paper by Nakanishi and Yamaguchi (1986), the velocity field is defined by a set of constant velocity cells with network nodes placed on the interface between each cell (see Fig. 25a). The advantage of this approach is that traveltimes between each node pair can be easily evaluated as $t = ds$ where d is the distance between the two nodes and s is the slowness of the cell containing the ray segment. Accuracy may be increased by either reducing cell size or increasing the number of nodes on cell edges. Figure 25b illustrates a selection of minimum time ray paths from one of the nodes to several other nodes in a constant velocity medium. This example provides some intuitive understanding of the sources of error inherent to this approach, as the true paths would be straight lines. Paths such as OC and OD are relatively straight, but OB and in particular OA are poor approximations due to the absence of connections between nodes on the same cell surface.

Another way of creating a network is to use a regular grid of velocity nodes and form linear connections between adjacent velocity nodes (Moser, 1991), as shown in Fig. 26a. The traveltime between two connected nodes A and B can be simply approximated by $t = d(s_A + s_B)/2$ where s_A and s_B are the slowness

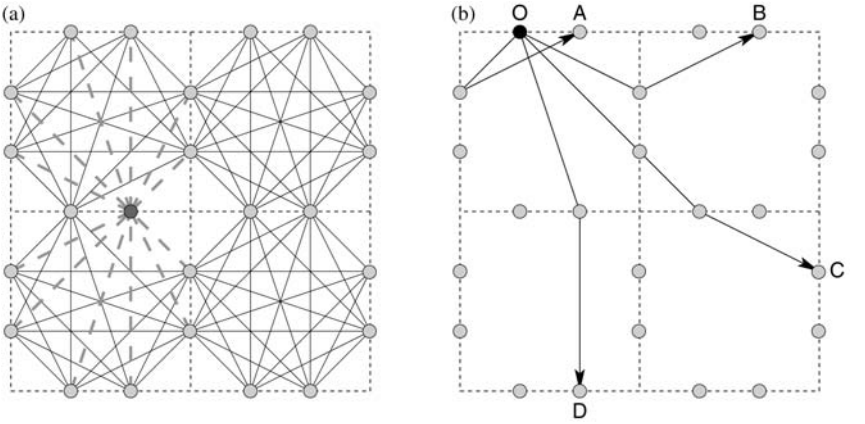


FIG. 25. Shortest path network built on a grid of constant velocity cells (Nakanishi and Yamaguchi, 1986; Moser, 1991). In this example, two network nodes are placed on the edge of each cell boundary. (a) All allowable path connections between nodes. Thick dashed grey lines highlight the connections from a single node; (b) shortest paths between node O and a selection of surrounding nodes in a homogeneous medium.

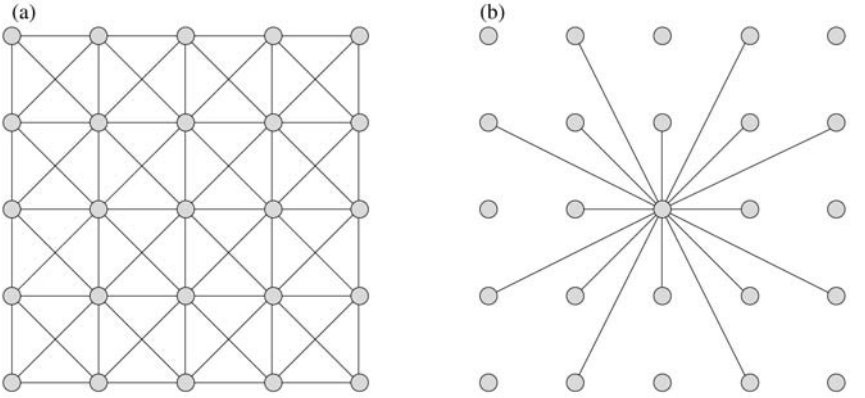


FIG. 26. Shortest path network built on a grid of velocity nodes (Moser, 1991). (a) A grid of 25 nodes with at most 8 connections per node; (b) increasing the number of connections allows smaller path deviations to be more accurately represented.

at nodes A and B respectively. The angular distance between branches emanating from a single node in Fig. 26b is 45°, which is relatively large. This can be reduced by increasing the number of node connections, as demonstrated in Fig. 26b. The connection stencil for a particular node is sometimes referred to as the "forward star" (Klimeš and Kvasnička, 1994). In the case of Fig. 26a, the forward star has 8 connections, while in Fig. 26b the forward star has 16 connections. The advantage

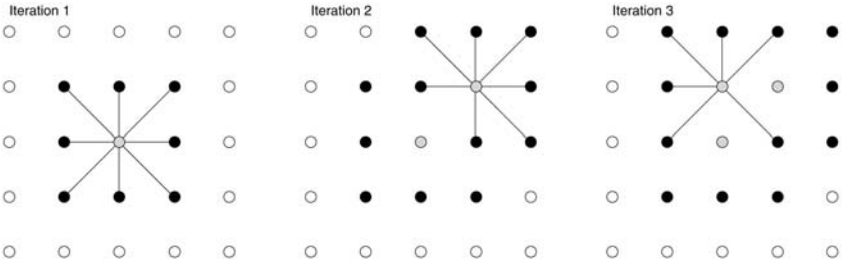


FIG. 27. Three iterations of a simple shortest path scheme using a forward star with 8 connections. Grey dots have known traveltimes, black dots have trial traveltimes, and white dots are yet to have traveltime computed.

of a grid formalism for the shortest path network is that continuously varying velocity fields can be more accurately represented, and interfaces are more easily inserted.

Once a network structure and method of traveltime determination between node pairs has been chosen, the next step is to use a shortest path algorithm to compute the full traveltime field and associated ray paths. The original network theory algorithm was developed by Dijkstra (1959), for which computation time scales as $O(M^2)$, where M is the number of nodes. The basic approach is conceptually simple, with nodes divided into two groups: a set of P nodes with known traveltimes, and a set of Q nodes with unknown traveltimes (Moser, 1991); initially, Q contains M elements and P is empty, with the traveltimes of nodes in Q set to an arbitrarily large value. The scheme is initiated from a source node by adding it to P, and then calculating traveltimes to all neighbouring nodes as defined by the forward star (see Fig. 27). These constitute trial traveltime values, and the method proceeds by choosing the trial node with minimum traveltime, adding it to P and then computing trial times to all neighbouring nodes in Q. If the node in Q already has a trial value from a previous iteration, then the one with minimum traveltime is selected. The complete traveltime field is found in exactly M iterations, and ray paths are obtained by simply recording the update order of nodes.

The most time consuming part of the Dijkstra algorithm is locating the minimum traveltime node, because in its traditional usage (Moser, 1991), the entire set of nodes in Q is tested, even those with arbitrarily large values. Thus, the total operation count is $\sum_{i=1}^{M} i = \frac{1}{2}M(M+1) \rightarrow O(M^2)$. However, it is possible to use much more efficient schemes such as the heap sort, which has an operation count of $O(M \log M)$, due to the ordered storage of trial traveltimes in a binary tree. Much work has been done to increase the computational speed of shortest path algorithms by introducing increasingly efficient sorting algorithms (Moser, 1991; Klimeš and Kvasnička, 1994; Cheng and House, 1996; Zhang and Toksöz, 1998).

It is worth noting the similarity between SPR and eikonal solvers such as FMM, which use the shape of the first-arrival wavefront as the computational front (cf.

Figs. 27 and 18). The only real difference is in the way traveltimes are updated to neighbouring nodes. Errors in SPR are due to finite node spacing and the angular distribution of node connectors. A coarse grid of nodes may poorly approximate the true velocity variations, while a limited range of angles between adjacent connectors may not allow for an accurate representation of the path. The accuracy of eikonal solvers is also a function of grid spacing, and parallels can be drawn between the complexity of the forward star used in SPR and the finite difference stencil used to solve the eikonal equation.

Although SPR methods can only be used to locate absolute first-arrivals in continuous media, they can be adapted to track refracted and reflected waves in much the same way as eikonal methods. Moser (1991) describes a constrained shortest path approach to computing this class of phase in layered media, which requires that paths visit a specified set of nodes that lie on an interface. In principle, the scheme is equivalent to the multi-stage FMM described earlier, which separates the medium into different computational domains. Thus, shortest paths are computed from the source node to all other nodes in the layer; the traveltimes at interface nodes are recorded and ordered in a heap; and then the layer into which the wave propagates has all nodes placed in Q and P set to empty. The evolution of the new time field will therefore be initiated from the interface node with minimum traveltime. Like the multi-stage FMM, this scheme could be extended to compute paths comprising any number of reflection and refraction branches.

SPR has proven to be effective in a number of practical seismic applications that require large datasets to be predicted in the presence of significant lateral heterogeneity. In their original implementation of SPR, Nakanishi and Yamaguchi (1986) invert traveltimes from local earthquakes for 2-D velocity structure. Zhang and Toksöz (1998) use an updated version of the SPR scheme, which samples the wavefront with uniform angular coverage and removes unnecessary nodes from the network, in a 2-D refraction tomography algorithm. In 3-D, Toomey *et al.* (1994) invert refraction traveltimes for crustal structure using a similar SPR scheme to that proposed by Moser (1991). More recently, Bai (2005) use SPR in the inversion of local and regional earthquake traveltimes for the 3-D velocity structure of Rabaul volcano in Papua New Guinea.

4. Multi-Arrival Wavefront Tracking

Most of the ray and grid-based schemes described previously are only suitable for tracking a single or limited number of arrivals between two points. In many cases, the presence of velocity heterogeneity results in an evolving wavefront self-intersecting, a phenomenon commonly referred to as *multipathing*, because it results in more than one ray path passing through a given point. This behaviour is

readily observed in Fig. 1. Eikonal and shortest path methods will only provide the first-arrival in such cases; ray tracing schemes can potentially provide more than one arrival, but the severe non-linearity of the two-point problem makes reliability a major issue. The heterogeneity of Earth structure means that multipathing commonly contributes to the complexity of a recorded waveform, and having a scheme that can efficiently predict all arrivals of significant amplitude has important implications in various areas of seismology, such as seismic imaging. In particular, it has already been noted that first-arrival traveltimes, which are commonly exploited in seismic tomography, tend to avoid regions of low wavespeed. Later arrivals, which do not necessarily exhibit such behaviour, could therefore provide better path coverage and consequently result in improved imaging.

To date, a number of methods have been developed to solve the multi-arrival problem. These include grid-based (Benamou, 1999; Steinhoff *et al.*, 2000; Engquist *et al.*, 2002; Fomel and Sethian, 2002; Osher *et al.*, 2002; Symes and Qian, 2003), ray-based (Vinje *et al.*, 1993; Lambaré *et al.*, 1996; Vinje *et al.*, 1996, 1999) and hybrid schemes (Benamou, 1996). Due to the relative infancy of this branch of seismology, many of the methods that have been proposed so far are not sufficiently developed for practical application. Here, several ray and grid based schemes are discussed.

4.1. Ray Based Schemes

One of the first ray-based schemes for computing multi-paths was proposed by Vinje *et al.* (1993), and is commonly referred to as wavefront construction. The basic principle underlying the method is that a wavefront can be evolved by repeated application of local ray tracing from a set of points lying on the wavefront surface. Thus, for some given time step δt, a new wavefront is defined by the end points of the rays traced from the old wavefront, as illustrated schematically in Fig. 28. In order to avoid undersampling of the wavefront as it evolves, new points are interpolated based on a distance criterion between adjacent rays. Thus, in order to apply this scheme, all that is required is some initial value formulation of the ray equation (e.g. Eq. (29)), an interpolation criterion and an interpolation method. From a practical implementation point of view, an efficient indexing scheme for storing the path end points and their connectivity is also required.

The basic 2-D scheme of Vinje *et al.* (1993) was subsequently extended to 3-D (Vinje *et al.*, 1996), and modified for media involving complex interface structures (Vinje *et al.*, 1999). The solution of the initial value ray tracing equations can be computed with high accuracy, so the main source of error in the wavefront construction method comes from interpolation. To decide whether a point needs to be interpolated, Vinje *et al.* (1993) uses a distance criterion between adjacent nodes i.e. $|\mathbf{r}_{i+1} - \mathbf{r}_i| \geqslant d_n$, where $\mathbf{r}_{i+1}$ and $\mathbf{r}_i$ are two adjacent points on the same wavefront and d_n is the distance tolerance above which a new point is interpolated. The problem with this interpolation criterion is that it does not account for

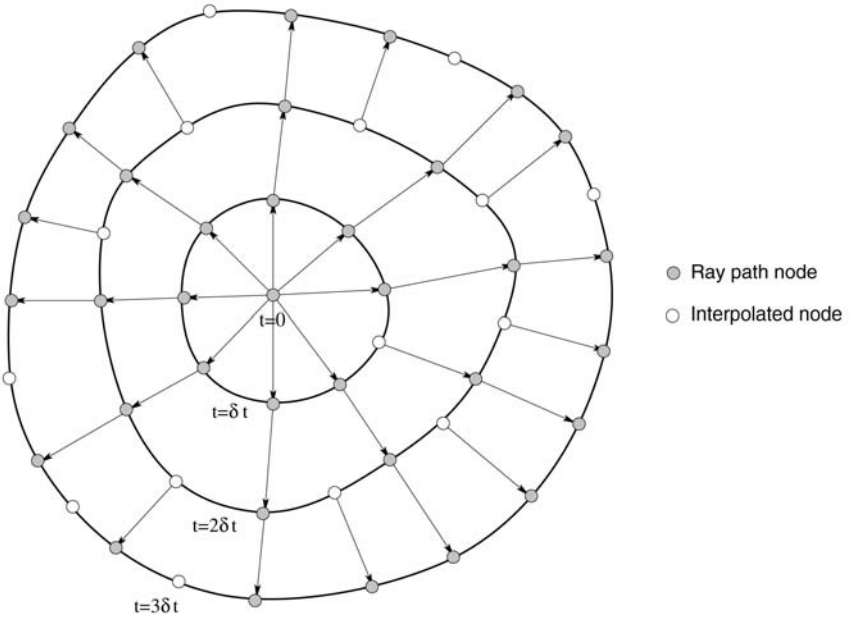


FIG. 28. Schematic diagram illustrating the principle behind the wavefront construction scheme proposed by Vinje *et al.* (1993). At each iteration, a new wavefront is constructed from the previous wavefront using local ray tracing and interpolation.

variations in wavefront curvature, so regions of high detail are likely to be undersampled. This can be ameliorated to some extent by using a criterion based on angular distance as well, but difficulties can still arise due to the presence of sharp corners, which usually occur when a wavefront triplicates. Since most interpolation schemes are best applied to smooth, differentiable functions, they may not produce satisfactory results.

An alternative approach to interpolation in physical (or "normal") space is to instead perform it in phase space (Lambaré *et al.*, 1996; Hauser *et al.*, 2006). As noted earlier, phase space is defined as $(\mathbf{r}, \nabla T)$, where $\mathbf{r}$ is a point on the wavefront and ∇T is the traveltime gradient, equivalent to the ray trajectory. If normal space is extended to higher dimensional phase space by using ∇T as the other coordinate directions, then the wavefront is "unfolded" into a smooth curve sometimes referred to as a "bicharacteristic strip" or curve (Osher *et al.*, 2002). Fig. 29 shows a circular wavefront in 2-D real space unfolded into a spiral curve in 3-D reduced phase space.

Reduced phase space makes use of the relationship $\nabla T = f(\theta, \phi)$, where θ and ϕ are ray inclination and azimuth respectively (see Eq. (29)); thus, it is possible to transform 6-D phase space into 5-D reduced phase space for problems in three dimensional normal space, and 4-D phase space into 3-D reduced phase space for

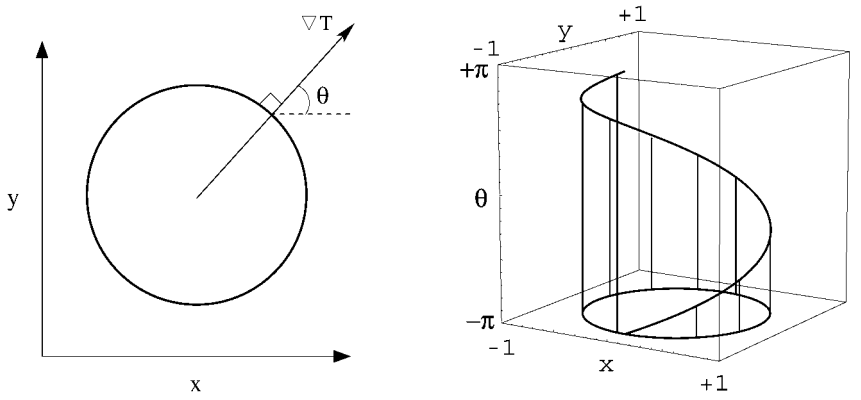


FIG. 29. Representation of a circular wavefront in 3-D reduced phase space; the ray path inclination θ is used as the third coordinate.

problems in two dimensional normal space. The advantage of a reduced phase space representation is that when the wavefront self intersects and contains sharp corners, its corresponding bicharacteristic curve in phase space will be locally smooth and single valued, as demonstrated in Fig. 30.

In order to compute the distance between neighbouring points in reduced phase space, it is necessary to scale the normal space coordinates so that they lie in the interval $[-\pi, \pi]$; for example, in 2-D Cartesian coordinates, $-\pi \leqslant x, y \leqslant \pi$. This scaling correlates the metric distance with the angular distance. The criterion for adding a new point is then given by $|\mathbf{x}_{i+1} - \mathbf{x}_i| \geqslant d_p$, where $\mathbf{x}_{i+1}$ and $\mathbf{x}_i$ are adjacent points on the wavefront and d_p is the tolerance. The important difference compared to the criterion used by Vinje *et al.* (1993) is that here $\mathbf{x} = (\mathbf{r}, \boldsymbol{\theta})$, rather than $\mathbf{x} = \mathbf{r}$. The smoothness of the bicharacteristic curve in reduced phase space means that linear interpolation of $\mathbf{x}$ (five variables in 3-D, three variables in 2-D) is often sufficient. However, it is relatively straightforward to use higher-order interpolators. For example, in 3-D phase space, one could insert a new point $\mathbf{x}_m$ midway between $\mathbf{x}_{i+1}$ and $\mathbf{x}_i$ using a weighted average defined by (Dyn *et al.*, 1990)

$$\mathbf{x}_m = \frac{-\mathbf{x}_{i-1} + 9\mathbf{x}_i + 9\mathbf{x}_{i+1} - \mathbf{x}_{i+2}}{16}, \tag{64}$$

which provides a smoother interpolation. As well as using interpolation to introduce new points to the wavefront, it also makes sense to have a minimum distance criterion to allow the removal of points which become too close together as a result of wavefront focusing. Properly applied, this can significantly improve computational efficiency without degrading accuracy.

The robustness of wavefront construction i n reduced phase space is demonstrated in Fig. 31, which shows several snapshots of an evolving wavefront

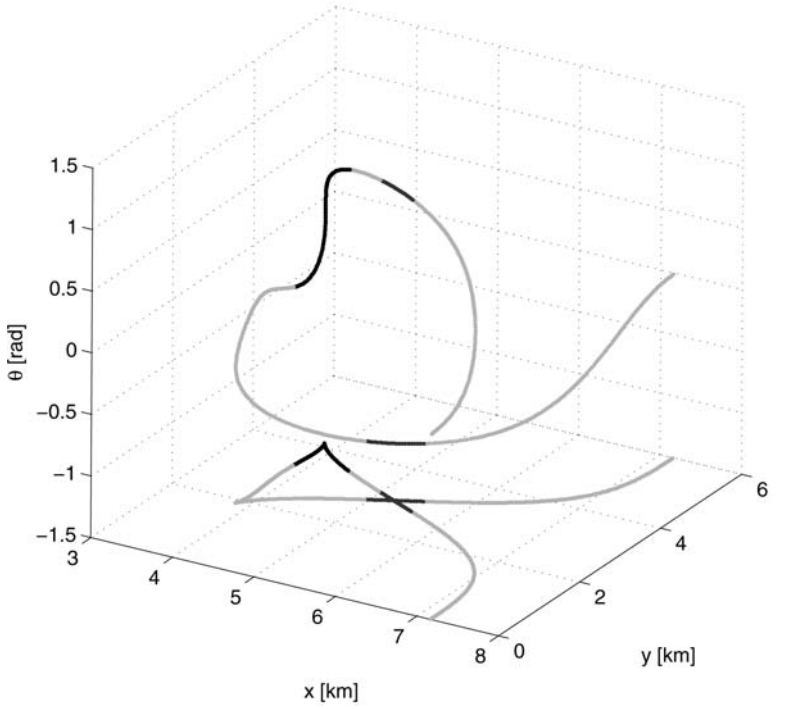


FIG. 30. A triplicating wavefront in normal space is represented by a smooth bicharacteristic curve in reduced phase space. Note that the intersecting segments in normal space no longer intersect in phase space, and the sharp corner is unfolded into a smooth segment of the bicharacteristic curve.

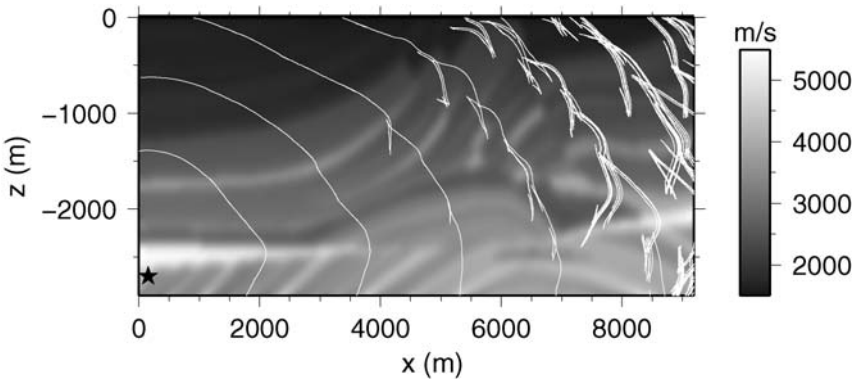


FIG. 31. Wavefronts tracked through the smoothed Marmousi model using the wavefront construction method of Hauser *et al.* (2006).

in the smooth version of the Marmousi model. The Marmousi model is one of the most challenging benchmarks for ray tracing and grid-based traveltime schemes, and is frequently used to test new methods (e.g. Lambaré *et al.*, 1996; Alkhalifah and Fomel, 2001). In the case of Fig. 31, which uses the scheme of Hauser *et al.* (2006), the wavefront gradually develops multiple self-intersections as it propagates, which is testament to the complexity of the underlying model. However, the smoothness of the corresponding bicharacteristic curve in reduced phase space ensures that the scheme remains well behaved. In this example, over 60 arrivals from a single source are detected in some regions of the model.

A second wavefront construction example, which tracks a surface wavefront from an earthquake in northern Papua New Guinea through the southwest Pacific region, is shown in Fig. 32. The phase velocity model (Fig. 32a) was constructed on the basis of five dispersion curves for fundamental mode Rayleigh waves with a 15 s period, with interpolation using average local topography to create the complete 2-D velocity field. At very low frequencies, great circle path propagation is a reasonable approximation, but at higher frequencies, the effects of velocity variations on path geometry should be considered. Fig. 32b shows the result of shooting a large fan of equally spaced rays (in azimuth) from the source. The strong velocity gradients present in the model leads to significant multipathing, and ray coverage is highly non-uniform. Using this approach, it is not possible to extract the shape of the evolving wavefront with much fidelity. In contrast, the phase space wavefront construction method of Hauser *et al.* (2006) appears to capture the full detail of the wavefront (Fig. 32c), with many levels of triplications evident. All ray paths (41 in total) from the source to a station located in the South Island of New Zealand are shown in Fig. 32d. Clearly, the energy received by the station will have sampled a broad region of the velocity model, which at the very least suggests that a great circle path approximation for higher frequency surface waves would not be valid in this case.

Extension of the phase space wavefront reconstruction scheme to 3-D is not trivial, as the connectivity between neighbouring points on a wavefront surface is not unique (unlike the 2-D case). Points could be connected using a triangular mesh, but a carefully considered interpolation strategy would be required to properly target regions of the wavefront that become undersampled. In addition, the use of higher order interpolation would be desirable to minimise the number of points that are required to accurately represent the evolving wavefront. One possibility is to use an edge splitting strategy that subdivides adjacent triangles that have a common edge which exceeds some distance tolerance. The butterfly subdivision scheme of Dyn *et al.* (1990) works along these lines, and allows for linear as well as higher order interpolation.

Several wavefront construction schemes have already been devised for 3-D media, most notably those of Vinje *et al.* (1996, 1999) and Lucio *et al.* (1996). Both use triangular meshes to define the wavefront surface, although they use different interpolation strategies. In the case of Vinje *et al.* (1999), wavefronts can also

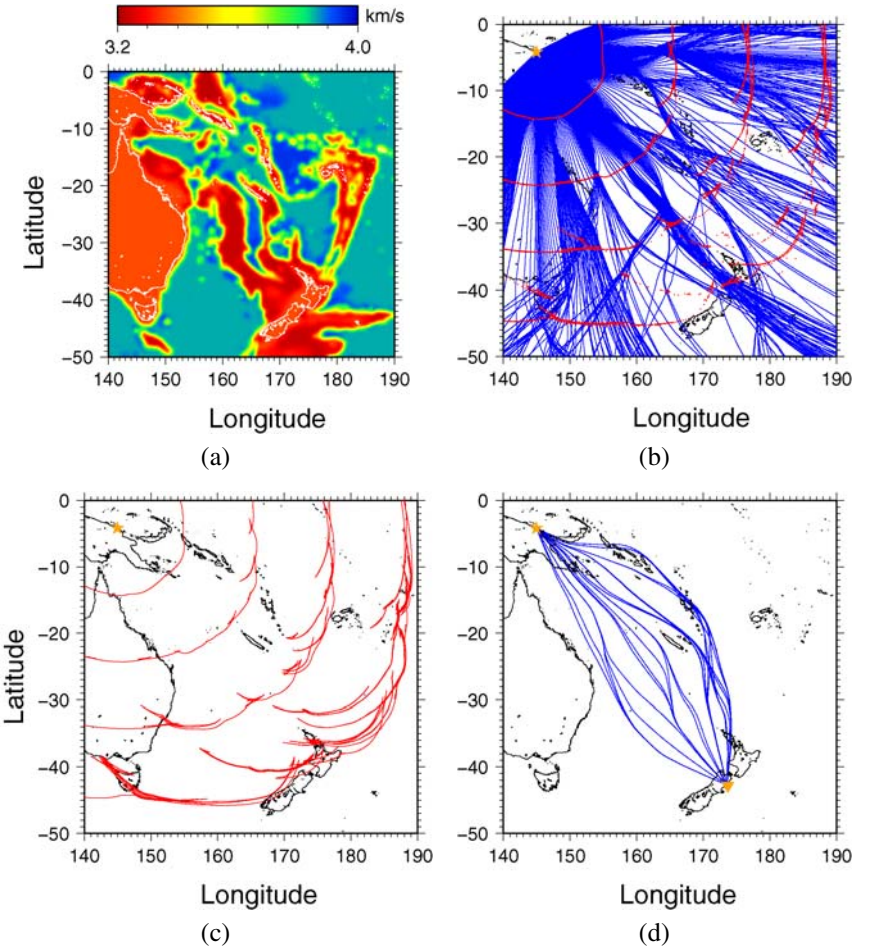


FIG. 32. Surface wavefront propagation in the southwest Pacific. (a) Phase velocity model for fundamental mode Rayleigh waves with a period of 15 s; (b) rays and wavefronts computed using a fan shooting scheme. Wavefronts are visualised by placing red dots along ray paths at predetermined intervals in time. Note that there are more red dots (10,000 per time step) than rays (500) because not all rays are plotted; (c) wavefronts plotted at the same time increments as (b), but computed using the wavefront construction scheme of Hauser *et al.* (2006); (d) ray paths between the source and a receiver located in New Zealand, obtained from the wavefront construction solution.

be tracked in so called "open models" which may contain interfaces with holes. This allows multiple reflection and refraction arrivals to be computed. The use of local ray tracing to evolve the propagating wavefront means that other ray-related variables such as amplitude may also be computed (Vinje *et al.*, 1993; Lambaré *et al.*, 1996). Amplitude is a potentially important quantity when multi-

ple paths are considered, because it offers a means of discriminating between the relative importance of each arrival.

Of all the multi-arrival schemes proposed so far, wavefront construction is by far the most developed, and is perhaps the only one that is sufficiently robust and efficient for practical problems. It has become a standard tool in the exploration industry via the NORSAR-3D modelling package, which is based on the scheme of Vinje *et al.* (1999), and has been used in various applications including coincident reflection migration (Xu and Lambaré, 2004; Xu *et al.*, 2004).

Another scheme for tracking multiple arrivals is "big ray-tracing" (Benamou, 1996; Abgrall, 1999), which is actually a hybrid method that attempts to combine initial value ray tracing and finite difference solution of the eikonal equation in order to compute the multi-valued traveltime field. The first step of the procedure is to use initial value ray tracing to partition the traveltime field into single-valued regions via a fan shoot. The second step involves computing traveltimes on an irregular grid of points that span each adjacent pair of ray paths using an eikonal solver. This results in the complete traveltime field spanned by the ray paths being found, and allows wavefronts and intermediate ray paths to be extracted from the solution. However, the principal drawback of this scheme is that the fan shoot, even with a large number of initial rays, is unlikely to properly sample the evolving wavefront (as demonstrated previously in Fig. 32b); for example, the detail contained in an evolving triplication is unlikely to be retained. Rietdijk (1999) discusses the limitations of big ray-tracing in more detail.

4.2. Grid Based Schemes

Compared to a ray-based technique such as wavefront construction, grid based schemes are potentially attractive as they avoid the need for complex data structures to track irregular ray geometries, and the use of interpolation to retain a uniform sampling of the propagating wavefront. Over the last decade, a number of new grid-based methods for tracking multi-valued wavefronts have been proposed, and new developments appear regularly in the current literature.

One possibility for tracking multiple arrivals is to exploit the stability and efficiency of first-arrival eikonal solvers by partitioning the multi-valued solution into a series of single valued solutions in a way that is analogous to big ray-tracing. A detailed examination of this approach is given by Benamou (1999); other papers on domain decomposition methods include those by Symes and Qian (2003) and Bevc (1997). The main challenge faced by these schemes is devising a robust splitting strategy, as the structure of the multi-valued solution is not known *a priori*. Given that many levels of triplication can occur (e.g. Figs. 31 and 32), it is difficult to envisage this approach becoming generally applicable. Benamou (2003) provides an overview of a number of different splitting methods.

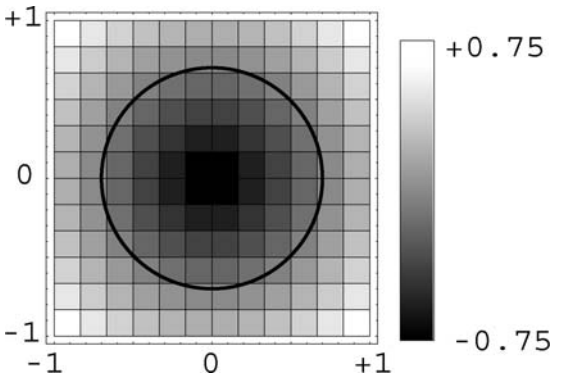


FIG. 33. Level set representation of an interface. In this example, the zero level set of the signed distance function is a circle.

Recently, a number of schemes have been proposed for tracking multi-valued wavefronts using the level set method. Level set methodology was pioneered by Osher and Sethian (1988) for the tracking of interfaces, and has been applied to various problems in the physical sciences (Sethian, 1999; Osher and Fedkiw, 2003), perhaps most notably in fluid dynamics for the evolution of interfaces between fluids and gasses (e.g. Mulder *et al.*, 1992; Sussman *et al.*, 1994; Chang *et al.*, 1996). The basic idea behind a level set formulation is to represent an interface as the zero level set, or zero contour, of a function defined as the signed distance from each grid point to the closest point on the interface (see Fig. 33). This implicitly represented interface can then be evolved by finite difference solution of a set of partial differential equations that explicitly describe the behaviour of the signed distance function over time. In general, the level set method describes the evolution of an $m-1$ dimensional manifold in m dimensional space; thus, a curve is tracked on a 2-D grid of points, and a surface is tracked on a 3-D grid of points. Note that while the FMM scheme presented earlier is also an interface tracking algorithm, it deals strictly with monotonically advancing interfaces, which allows for a boundary value formulation of the problem. Level set methods do not have this restriction, and consequently require an initial value formulation.

If we denote the signed distance function by $\phi(\mathbf{x})$, then at some given point $\mathbf{x} = \mathbf{x}_A$, $\phi(\mathbf{x}_A)$ is the distance from $\mathbf{x}_A$ to the closest point on the interface, and is negative if $\mathbf{x}_A$ is inside the interface, and positive if $\mathbf{x}_A$ is outside the interface (Fig. 33). The propagation of the surface implicitly described by $\phi(\mathbf{x})$ can be defined in space and time by $\phi(\mathbf{r}(t), t) = 0$, where $\mathbf{r}(t)$ is the location of the zero level set in space at time t. The time derivative of this expression is given by the chain rule as

$$\phi_t + \nabla\phi\big(\mathbf{r}(t), t\big) \cdot \mathbf{r}_t(t) = 0, \tag{65}$$

where ϕ_t and $\mathbf{r}_t$ are time derivatives of ϕ and $\mathbf{r}$ respectively. If F is defined as the scalar speed function in the outward normal direction to the evolving surface, then $F = \mathbf{r}_t(t) \cdot \mathbf{n}$, where the unit normal direction $\mathbf{n} = \nabla\phi/|\nabla\phi|$. The evolution equation for the signed distance function ϕ is therefore given by

$$\phi_t + F|\nabla\phi| = 0, \tag{66}$$

with some given initial condition $\phi(\mathbf{x}, t = 0)$. This is the level set equation that was originally formulated by Osher and Sethian (1988). The signed distance function ϕ is well behaved when $|\nabla\phi| = 1$ at every grid point in the computational domain. When this occurs, $\phi_t = -F$, so when $F > 0$ the interface moves in the outward normal direction, and when $F < 0$ the interface moves in the inward normal direction.

The signed distance function can become spatially discontinuous in gradient when points exist which are equidistant to at least two points on the front. Thus, numerical methods used for solving Eq. (66) must exhibit reasonable behaviour in the presence of gradient discontinuities; fortunately, a large number of schemes, such as ENO and WENO, have been developed for solving differential equations of this type (e.g. Osher and Shu, 1991; Jiang and Peng, 2000).

When the scalar speed function F is dependent on some vector velocity field $\mathbf{v}(\mathbf{x})$ defined at the grid points, then $F = \mathbf{v} \cdot \mathbf{n}$, so Eq. (66) becomes

$$\phi_t + \mathbf{v}(\mathbf{x}) \cdot \nabla\phi = 0. \tag{67}$$

The strength of the level set method lies in its implicit representation of an interface; topological changes like breaking and merging are handled naturally, and there is no need for resampling of the wavefront as it evolves. However, the interface cannot self-intersect, which at first appears to rule out using the method for multi-valued wavefronts. An approach devised by Osher *et al.* (2002), which overcomes this limitation, is to consider the problem in a reduced phase space, where the bicharacteristic curve is defined as the intersection of two zero level set surfaces (see Fig. 34); these surfaces are perpendicular along their line of intersection. The two surfaces can be separately evolved by solving the following pair of level set equations

$$\left.\begin{aligned} \phi_t + \mathbf{v} \cdot \nabla\phi &= 0 \\ \psi_t + \mathbf{v} \cdot \nabla\psi &= 0 \end{aligned}\right\}, \tag{68}$$

where the velocity vector $\mathbf{v}$ is given by Eq. (29), and ϕ and ψ represent the two level set functions. The bicharacteristic curve is thus implicitly defined by the curve $\phi = \psi = 0$.

An example of a triplicating wavefront that is tracked using the level set approach described above is shown in Fig. 35. The two zero level set surfaces of ϕ and ψ are perpendicular where they intersect, and the line of intersection describes the bicharacteristic curve, which can be projected into normal space to obtain the wavefront. In this case, the level set equations defined by Eq. (68)

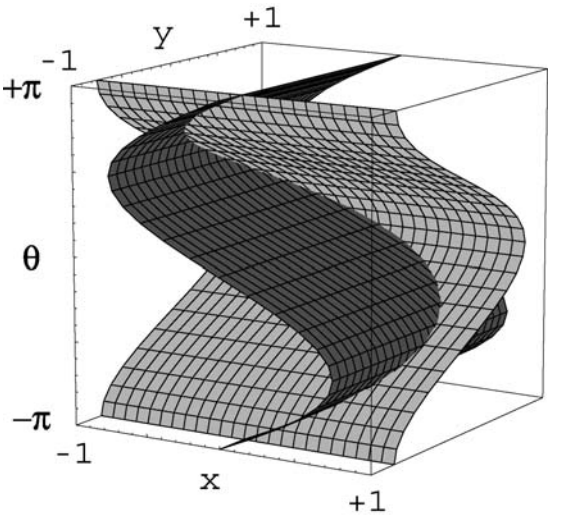


FIG. 34. In 3-D reduced phase space, the bicharacteristic curve can be described by the intersection of two zero level set surfaces. The projection of the bicharacteristic curve into normal space produces a circular wavefront in this example.

are solved using a fifth order WENO scheme in space and a third order TVD Runge–Kutta scheme in time (Hauser *et al.*, 2006). Each signed distance function is updated at each time step for the entire computational domain. When computed on a $51 \times 31 \times 51$ grid using 70 time steps and a time increment of 0.1 s, the CPU time is approximately 45 s on a 3.0 GHz Intel Pentium 4 PC running GNU/Linux; however, reducing the grid spacing by a factor of two ($101 \times 61 \times 101$ points) results in CPU time increasing to approximately 717 s. This increase can be explained by two factors: first, halving the grid spacing in 3-D phase space will increase the number of points by a factor of 8; second, the CFL criterion used (Hauser *et al.*, 2006) requires that a surface does not cross more than one grid cell per time step. Thus, halving the grid spacing means that it is now necessary to use 140 time steps with a time increment of 0.05 s. Combined, these two factors result in a 16-fold increase in CPU time when grid spacing is halved.

The relationship between grid spacing and CPU time is the main limitation of the level set scheme of Osher *et al.* (2002). By comparison, wavefront construction schemes can increase the resolution of a wavefront by a factor of two (doubling the number of rays) with only a factor of two increase in CPU time (Hauser *et al.*, 2006). For strongly curved wavefronts (e.g. Fig. 35), it is often necessary to have high resolution grids, as the zero level set surfaces tend to become more tightly folded as the wavefront evolves; if a surface overturns within several grid cells or less, then information will be lost and the wavefront will be more poorly represented. Although this scheme can be extended to 3-D by track-

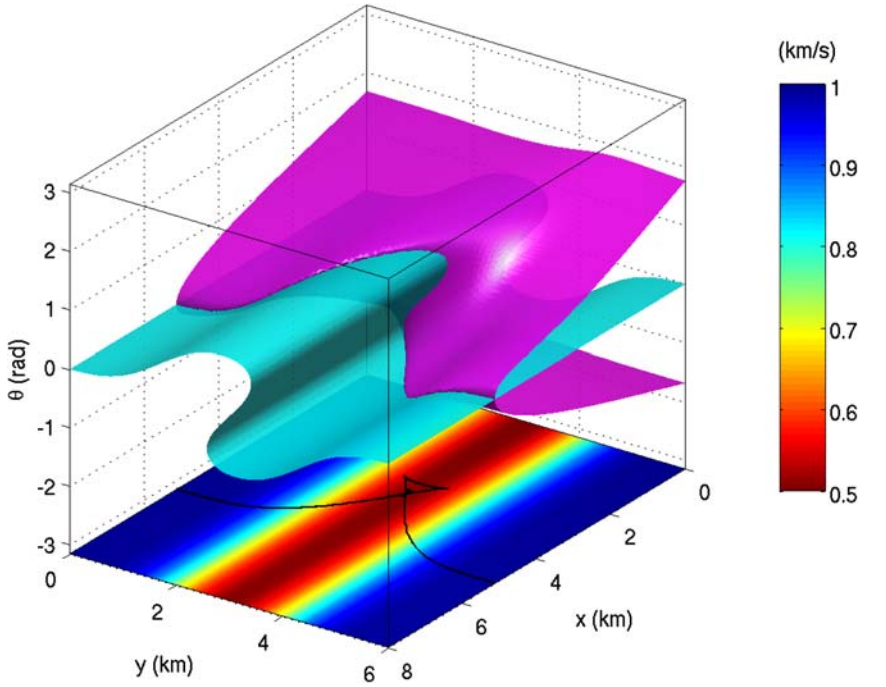


FIG. 35. The evolution of a triplicating wavefront computed using a level set scheme. The zero level set of ϕ is denoted by the magenta surface, and the cyan surface denotes the zero level set of ψ. The bicharacteristic curve is defined by the intersection of these two surfaces, and the projection of the line of intersection into (x, y) space yields the triplicating wavefront. The initial plane wavefront enters the box at $x = 0$.

ing the evolution of three surfaces in five dimensions (Osher *et al.*, 2002), the computational burden would be immense, because increasing grid resolution by a factor of two would increase CPU time by a factor of 64.

Despite the apparent limitations of the level set approach described above, research in this area is ongoing, with particular attention given to improving the efficiency of the scheme. This can be done by using a local level set scheme which only updates the signed distance functions within a narrow band about the bicharacteristic curve (Osher *et al.*, 2002; Osher and Fedkiw, 2003; Qian and Leung, 2006), and adaptive gridding, which provides higher resolution only where it is required. Qian and Leung (2006) devise a local level set scheme which uses a paraxial formulation of the ray equations in which the z coordinate becomes the independent variable. For two dimensional problems, phase space can therefore be defined by (x, θ) rather than (x, z, θ), which simplifies the problem, but imposes limitations on path geometry (in this case, wavefronts can only move in the positive z direction). Using a waveguide model similar to Fig. 35 and

a 120 × 120 grid, Qian and Leung (2006) show that CPU time (on a Pentium IV desktop) is 72 minutes for a global level set method, but only 8 minutes for their local level set scheme, which is a significant improvement. Other papers on level set based methods for wavefront tracking in phase space include those by Leung *et al.* (2004), Qian and Leung (2004) and Cockburn *et al.* (2005).

An alternative approach to computing multi-arrivals with a grid-based scheme has been proposed by Fomel and Sethian (2002). Starting with the Liouville formulation of the ray tracing equations, a system of time-independent PDEs (referred to as "Escape Equations") are formulated which can be solved numerically on a grid in reduced phase space. The solutions correspond to the arrival times at the boundary from every point in the phase space domain. As a result, the computational efficiency of the scheme is $O(M \log M)$, where M is the total number of grid points. Multi-arrival information such as wavefront geometry and two-point traveltimes is extracted with post-processing. In comparison with the level set method described previously, Osher *et al.* (2002) claims that the Fomel and Sethian (2002) scheme is less computationally efficient at each update step (by a factor of $O(M^{1/2})$ in 2-D and $O(M)$ in 3-D), but given that the Escape Equations are boundary value PDEs, and therefore do not require solution at multiple time increments, it is debatable which scheme is ultimately more efficient.

It is interesting to note that a number of new schemes for computing multi-arrivals using a grid based (or "Eulerian") framework in phase space appeared in the literature at the same time. Two of these have been discussed above, but another, referred to as the "segment projection" method (Engquist *et al.*, 2002), which explicitly tracks a wavefront in phase space, will also be mentioned. In three dimensional reduced phase space, the bicharacteristic curve is represented by its projection onto the two-dimensional xy, $x\theta$ and $y\theta$ planes. Each of these curves is subdivided into a contiguous set of single valued line segments. The segments are then evolved by solving the relevant equations of motion, with dynamic creation and elimination of segments used to retain the minimum number of single-valued segments necessary to represent the wavefront. The method bares some resemblance to wavefront construction in phase space, but uses an Eulerian framework. It appears to be robust in the presence of strong velocity contrasts (such as a waveguide) which cause multiple foldings of a wavefront, and Engquist *et al.* (2002) claim that it has a similar level of computational complexity as ray tracing.

A final grid-based technique that is worth considering is Dynamic Surface Extension or DSE (Steinhoff *et al.*, 2000). DSE shares some similarities with the level set scheme of Osher *et al.* (2002), except that the grid lies in normal space and the information carried by each grid point is the coordinate of the closest point on the wavefront rather than the distance. The DSE algorithm proceeds in two steps; in the first, the location of the wavefront is updated throughout the grid using a knowledge of the wavefront velocity; the second step updates the wavefront location information carried by the grid. Self-intersecting wavefronts can be

tracked, but not all types of wavefront behaviour can be captured (Engquist *et al.*, 2002). In addition, the method is not purely Eulerian as interpolation is needed to maintain an even sampling of the wavefront on the grid.

Although a number of grid and ray based schemes for tracking multi-arrival wavefronts has been described above, our coverage has not been exhaustive. For more information, the interested reader is directed to two relatively recent review papers on the subject (Benamou, 2003; Engquist and Runborg, 2003).

5. Concluding Remarks

In this review paper, we have described a variety of schemes for tracking the kinematic evolution of high frequency seismic waves in heterogeneous 2-D and 3-D structures. Where possible, we have focused on methods that have been used in practical applications; most of these can be characterised as either ray based or grid based. Ray tracing has traditionally been the method of choice in many seismic applications due to its high accuracy and potential for computational efficiency. Common ray tracing schemes include shooting, bending and pseudo bending. Shooting methods formulate the kinematic ray equation as an initial value problem, which allows a complete ray path to be computed given an initial projection vector. The boundary value problem of locating source-receiver trajectories is typically solved using iterative non-linear strategies which exploit information from nearby paths to update the projection parameters and better target the receiver. Bending methods perturb the geometry of an initial path joining source and receiver, until it becomes a true (stationary) path, by iteratively solving a boundary value formulation of the ray equations. Pseudo-bending schemes usually represent a path as a series of points which are perturbed using a simple algorithm based directly on Fermat's principle of stationary time. All three methods are non-robust in the sense that the assumption of local linearity is made even though it is not generally valid.

Grid based methods include those that solve the eikonal equation using finite differences on a grid of points, and shortest path ray tracing, which uses Dijkstra-like algorithms to find traveltimes and paths through a network of nodes with predetermined connectivity. Both techniques are capable of computing traveltimes to all points of a medium in a stable and computationally efficient manner. It was shown that eikonal solvers which use the expanding wavefront as the computational front, such as the fast marching method, share many similarities with SPR. Compared to ray tracing, advantages of these grid based schemes include the ability to compute traveltimes to all points of the medium in one pass, stability, and in many cases, computational efficiency. However, the fact that they only compute the first arrival traveltime field in continuous media is a limitation, and excessive computing time can become an issue if highly accurate traveltimes

are required. In addition, quantities other than traveltime are difficult to compute without first extracting ray paths from the traveltime field and then resorting to ray based techniques.

In Section 4, a variety of ray and grid based schemes for tracking multi-arrival wavefronts were presented. Although this area of research is arguably still in its infancy (particularly with regard to grid-based methods), there has been a proliferation of new schemes appearing in the literature over the past decade. The most developed and practical scheme to emerge so far has been wavefront construction, which uses local ray tracing and interpolation to advance a wavefront surface in discrete time steps. Of the grid-based solvers, segment projection perhaps shows the most promise, although research is still ongoing. The attraction of these techniques is that they are capable of computing all arrivals, which may allow more of an observed seismic wavetrain to be predicted. However, from an application point of view, there are still significant hurdles to be overcome. For instance, it is almost impossible to pick arrivals other than the first from a seismogram unless they represent specific reflection or refraction phases. Thus, the issue of how to best exploit the multi-arrival capability of these schemes is yet to be fully explored.

In practice, the most appropriate scheme for predicting an observational dataset depends on a number of factors including computational efficiency, accuracy, robustness and the type of information that is required. For very large datasets, computational efficiency and robustness are crucial. In cases where there are many sources and few receivers or vice-versa, then eikonal or SPR solvers should be considered. When lateral heterogeneity is not particularly significant and true two-point paths do not deviate significantly from those computed through a laterally averaged model, then shooting or bending methods of ray tracing may be more appropriate. Global phases that penetrate deep into the Earth tend to fall into this category, which perhaps explains why ray based schemes are still the preferred option with this class of dataset. Traditionally, ray tracing has always been preferred in models that include interfaces, but recent developments in grid-based eikonal solvers suggest that they can be just as effective for computing phases that experience multiple refractions and reflections. In cases where multi-arrival information is required, then a method like wavefront construction would be recommended.

In a paper by Leidenfrost *et al.* (1999), six different methods for calculating traveltimes are compared. These include three different varieties of eikonal solver, SPR, wavefront construction and a combined eikonal and Runge–Kutta method. The first of the eikonal solvers is essentially the Vidale (1988) method; the second adapts the same scheme to polar coordinates with an expanding circle rather than square; and the third uses the method of Podvin and Lecomte (1991). The combined eikonal and Runge–Kutta method solves the eikonal equation in polar coordinates by employing first-order finite differences in θ and a fourth-order Runge–Kutta scheme in radius. The SPR method is that of Klimeš and Kvasnička (1994) and the wavefront construction scheme is essentially that of Ettrich

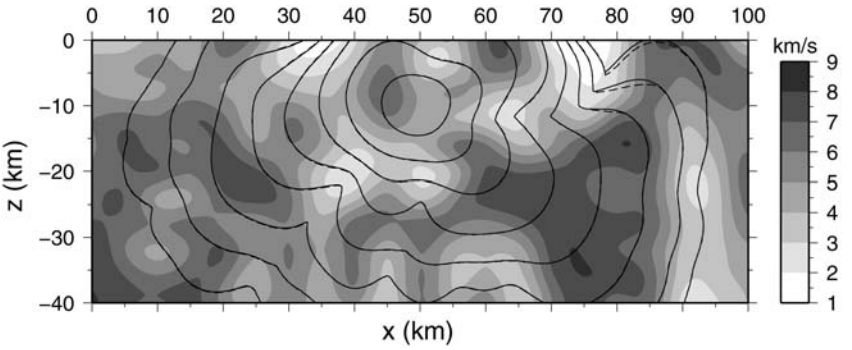


FIG. 36. Comparison of wavefronts computed using FMM (solid lines) and WENO-DNO-PS (dashed lines) for the complex velocity model of Fig. 16. The WENO-DNO-PS scheme clearly overestimates traveltimes in the neighbourhood of $(x, z) = (80, -5)$ km.

and Gajewski (1996). All solvers were benchmarked in 2-D using both a simple model with analytic solutions and the Marmousi model. The tests show that the expanding square eikonal solver offers the best compromise between speed and accuracy for models that are not too complicated, while wavefront construction proved to be the most reliable for the more complex Marmousi model. It is interesting to note that despite computing multi-arrivals, wavefront construction was of the same order of accuracy as the expanding square method and had computation times better than the SPR scheme.

In a separate study, Kim (2002) carry out a comparison between FMM, GMM and a second-order ENO-DNO-PS scheme using a 2-D smooth velocity model. GMM or the "Group Marching Method" (Kim, 2001), is a variant of FMM which avoids the need to sort narrow band traveltimes; hence computational cost scales as $O(M)$ rather than $O(M \log M)$, where M is the total number of grid points. As described previously, ENO-DNO-PS is an eikonal solver which uses an essentially non-oscillatory finite difference stencil, an expanding square framework, and post sweeping to repair causality breaches introduced by the expanding square. Results using the same grid spacing suggest that ENO-DNO-PS is approximately an order of magnitude more computationally efficient than the other two schemes, yet produces more accurate results. However, it should be pointed out that only first-order FMM and GMM schemes were used, and no source-grid refinement was implemented. Another factor in favour of ENO-DNO-PS was the use of a relatively simple velocity field; in the presence of complex velocity structures, the use of *ad hoc* post-sweeping may not always succeed. Fig. 36 shows a comparison between the WENO-DNO-PS solution of Fig. 16 and a second-order FMM scheme with local grid refinement. Although the solutions are very similar, WENO-DNO-PS significantly overestimates the traveltimes in the region $(x, z) = (80, -5)$. It is possible that a judicious post-sweeping strategy may al-

leviate this problem, but in general, it is not clear that the correct solution can always be guaranteed.

ACKNOWLEDGEMENTS

N. Rawlinson is supported by ARC Discovery Project DP0556282 and J. Hauser is supported by ARC Discovery Project DP0451133. Most figures were created using the freely distributed software packages GMT, xfig, Geomview and PGPLOT.

REFERENCES

Abgrall, R. (1999). Big ray-tracing and eikonal solver on unstructured grids: Application to the computation of a multivalued traveltime field in the Marmousi model. *Geophysics* **64**, 230–239.

Afnimar, Koketsu, K. (2000). Finite difference traveltime calculation for head waves travelling along an irregular interface. *Geophys. J. Int.* **143**, 729–734.

Aki, K., Christoffersson, A., Husebye, E.S. (1977). Determination of the three-dimensional seismic structure of the lithosphere. *J. Geophys. Res.* **82**, 277–296.

Alkhalifah, T., Fomel, S. (2001). Implementing the fast marching eikonal solver: Spherical versus Cartesian coordinates. *Geophys. Prosp.* **49**, 165–178.

Bai, C.-Y. (2005). 3-d multi-step travel time tomography: Imaging the local, deep velocity structure of Rabaul volcano, Papua New Guinea. *Phys. Earth Planet. Inter.* **151**, 259–275.

Bartels, R.H., Beatty, J.C., Barsky, B.A. (1987). An Introduction to Splines for Use in Computer Graphics and Geometric Modelling. Morgan Kaufmann, Los Altos.

Benamou, J.D. (1996). Big ray tracing: Multivalued travel time field computation using viscosity solutions of the Eikonal equation. *J. Comp. Phys.* **128**, 463–474.

Benamou, J.D. (1999). Direct computation of multivalued phase space solutions for Hamilton–Jacobi equations. *Comm. Pure Appl. Math.* **52**, 1443–1475.

Benamou, J.D. (2003). An introduction to Eulerian Geometrical Optics (1992–2002). *SIAM J. Sci. Comput.* **19**, 63–93.

Benz, H.M., Smith, R.B. (1984). Simultaneous inversion for lateral velocity variations and hypocenters in the Yellowstone region using earthquake and refraction data. *J. Geophys. Res.* **89**, 1208–1220.

Bevc, D. (1997). Imaging complex structures with semirecursive Kirchhoff migration. *Geophysics* **62**, 577–588.

Bulant, P. (1996). Two-point ray tracing in 3-D. *Pure Appl. Geophys.* **148**, 421–447.

Bulant, P. (1999). Two-point ray-tracing and controlled initial-value ray-tracing in 3-d heterogeneous block structures. *J. Seismic Explor.* **8**, 57–75.

Buske, S. (1999a). 3-D prestack Kirchhoff migration of the ISO89-3D data set. *Pure Appl. Geophys.* **156**, 157–171.

Buske, S. (1999b). Three-dimensional pre-stack Kirchhoff migration of deep seismic reflection data. *Geophys. J. Int.* **137**, 1243–1260.

Buske, S., Kästner, U. (2004). Efficient and accurate computation of seismic traveltime and amplitudes. *Geophys. Prosp.* **52**, 313–322.

Cao, S., Greenhalgh, S. (1994). Finite-difference solution of the eikonal equation using an efficient, first-arrival, wavefront tracking scheme. *Geophysics* **59**, 632–643.

Cassell, B.R. (1982). A method for calculating synthetic seismograms in laterally varying media. *Geophys. J. Royal Astr. Soc.* **69**, 339–354.

Červený, V. (1987). Ray tracing algorithms in three-dimensional laterally varying layered structures. In: Nolet, G. (Ed.), *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*. D. Reidel, Dordrecht, pp. 99–133.

Červený, V. (2001). Seismic Ray Theory. Cambridge University Press, Cambridge.

Červený, V., Firbas, P. (1984). Numerical modelling and inversion of travel times of seismic body waves in inhomogeneous anisotropic media. *Geophys. J. Royal Astr. Soc.* **76**, 41–51.

Červený, V., Pšenčik, I. (1983). Gaussian beams and paraxial ray approximation in three-dimensional elastic inhomogeneous media. *J. Geophys.* **53**, 1–15.

Červený, V., Klimeš, L., Pšenčik, I. (1984). Paraxial ray approximations in the computation of seismic wavefields in inhomogeneous media. *Geophys. J. Royal Astr. Soc.* **79**, 89–104.

Červený, V., Klimes, L., Psencik, I. (2006). Seismic ray method: Recent developments. *Adv. Geophys.* **48**, 1–126.

Chang, Y.C., Hou, T.Y., Merriman, B., Osher, S. (1996). A level set formulation of Eulerian interface capturing methods for incompressible fluid flows. *J. Comp. Phys.* **124**, 449–464.

Chapman, C. (2004). Fundamentals of Seismic Wave Propagation. Cambridge University Press, Cambridge.

Cheng, N., House, L. (1996). Minimum traveltime calculations in 3-D graph theory. *Geophysics* **61**, 1895–1898.

Chiu, S.K.L., Kanasewich, E.R., Phadke, S. (1986). Three-dimensional determination of structure and velocity by seismic tomography. *Geophysics* **51**, 1559–1571.

Chopp, D.L. (2001). Some improvements of the fast marching method. *SIAM J. Sci. Comput.* **23**, 230–244.

Cockburn, B., Qian, J., Reitich, F., Wang, J. (2005). An accurate spectral/discontinuous finite-element formulation of a phase-space-based level set approach to geometrical optics. *J. Comp. Phys.* **208**, 175–195.

Curtis, A., Snieder, R. (1997). Reconditioning inverse problems using the genetic algorithm and revised parameterization. *Geophysics* **62**, 1524–1532.

Day, A.J., Peirce, C., Sinha, M.C. (2001). Three-dimensional crustal structure and magma chamber geometry at the intermediate-spreading, back-arc Valu Fa Ridge, Lau Basin – results of a wide-angle seismic tomographic inversion. *Geophys. J. Int.* **146**, 31–52.

Dębski, W., Ando, M. (2004). Spectral ray tracer: A class of accurate two-point ray tracers. *Acta Geophysica Polonica* **52**, 1–14.

de Kool, M., Rawlinson, N., Sambridge, M. (2006). A practical grid based method for tracking multiple refraction and reflection phases in 3d heterogeneous media. *Geophys. J. Int.* **167**, 253–270.

Dijkstra, E.W. (1959). A note on two problems in connection with graphs. *Numer. Math.* **1**, 269–271.

Dyn, N., Levin, D., Gregory, J.A. (1990). A butterfly subdivision scheme for surface interpolation with tension control. *ACM Transaction on Graphics* **9**, 160–169.

Dziewonski, A.M., Hager, B.H., O'Connell, R.J. (1977). Large-scale heterogeneities in the lower mantle. *J. Geophys. Res.* **82**, 239–255.

Eberhart-Phillips, D. (1986). Three-dimensional velocity structure in northern California coast ranges from inversion of local earthquake arrival times. *Bull. Seism. Soc. Am.* **76**, 1025–1052.

Eberhart-Phillips, D. (1990). Three-dimensional *P* and *S* velocity structure in the Coalinga Region, California. *J. Geophys. Res.* **95**, 15,343–15,363.

Eberhart-Phillips, D., Reyners, M. (1997). Continental subduction and three-dimensional crustal structure: The northern South Island, New Zealand. *J. Geophys. Res.* **102**, 11,848–11,861.

Engquist, B., Runborg, O. (2003). Computational high frequency wave propagation. *Acta Numerica* **12**, 181–266.

Engquist, B., Runborg, O., Tornberg, A.-K. (2002). High-frequency wave propagation by the segment projection method. *J. Comp. Phys.* **178**, 373–390.

Ettrich, N., Gajewski, D. (1996). Wave front construction in smooth media for prestack depth migration. *Pageoph* **148**, 481–502.

Farra, V., Madariaga, R. (1987). Seismic waveform modelling in heterogeneous media by ray perturbation theory. *J. Geophys. Res.* **92**, 2697–2712.

Farra, V., Madariaga, R. (1988). Non-linear reflection tomography. *Geophys. J.* **95**, 135–147.

Fischer, R., Lees, J.M. (1993). Shortest path ray tracing with sparse graphs. *Geophysics* **58**, 987–996.

Fomel, S., Sethian, J.A. (2002). Fast-phase space computation of multiple arrivals. *Proc. Nat. Acad. Sci.* **99**, 7329–7334.

Graeber, F.M., Asch, G. (1999). Three-dimensional models of P wave velocity and P-to-S velocity ratio in the southern central Andes by simultaneous inversion of local earthquake data. *J. Geophys. Res.* **104**, 20,237–20,256.

Gray, S.H., May, W.P. (1994). Kirchhoff migration using eikonal equation traveltimes. *Geophysics* **59**, 810–817.

Hammer, P.T.C., Dorman, L.M., Hildebrand, J.A., Cornuelle, B.D. (1994). Jasper Seamount structure: Seafloor seismic refraction tomography. *J. Geophys. Res.* **99**, 6731–6752.

Hauser, J., Sambridge, M., Rawlinson, N. (2006). Phase space methods for multi arrival wavefronts. *Explor. Geophys.* **37**, 331–339.

Hildebrand, J.A., Dorman, L.M., Hammer, P.T.C., Schreiner, A.E., Cornuelle, B.D. (1989). Seismic tomography of Jasper Seamount. *Geophys. Res. Lett.* **16**, 1355–1358.

Hole, J.A. (1992). Nonlinear high-resolution three-dimensional travel-time tomography. *J. Geophys. Res.* **97**, 6553–6562.

Hole, J.A., Zelt, B.C. (1995). 3-D finite-difference reflection travel times. *Geophys. J. Int.* **121**, 427–434.

Iyer, H., Hirahara, K. (1993). Seismic Tomography: Theory and Practice. Chapman & Hall, London.

Jeffreys, H., Swirles, B. (1966). Methods of Mathematical Physics. Cambridge University Press, Cambridge.

Jiang, G.S., Peng, D.P. (2000). Weighted ENO schemes for Hamilton–Jacobi equations. *SIAM J. Sci. Comput.* **21**, 2126–2143.

Jiang, G.S., Shu, C. (1996). Efficient implementation of weighted ENO schemes. *J. Comp. Phys.* **126**, 202–228.

Julian, B.R., Gubbins, D. (1977). Three-dimensional seismic ray tracing. *J. Geophys.* **43**, 95–113.

Kim, S. (2001). An $O(N)$ level set method for eikonal equations. *SIAM J. Sci. Comput.* **22**, 2178–2193.

Kim, S. (2002). 3-D eikonal solvers: First-arrival traveltimes. *Geophysics* **67**, 1225–1231.

Kim, S., Cook, R. (1999). 3D traveltime computation using second-order ENO scheme. *Geophysics* **64**, 1867–1876.

Kimmel, R., Sethian, J.A. (1998). Fast marching methods on triangulated domains. *Proc. Nat. Acad. Sci.* **95**, 8341–8435.

Kirkpatrick, S., Gelatt Jr, C.D., Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Klimeš, L. (1989). Gaussian packets in the computation of seismic wavefields. *Geophys. J. Int.* **99**, 421–433.

Klimeš, L., Kvasnička, M. (1994). 3-D network ray tracing. *Geophys. J. Int.* **116**, 726–738.

Koketsu, K., Sekine, S. (1998). Pseudo-bending method for three-dimensional seismic ray tracing in a spherical earth with discontinuities. *Geophys. J. Int.* **132**, 339–346.

Kreyszig, E. (1993). Advanced Engineering Mathematics. John Wiley & Sons, Inc., New York.

Lambaré, G., Lucio, P.S., Hanyga, A. (1996). Two-dimensional multivalued traveltime and amplitude maps by uniform sampling of a ray field. *Geophys. J. Int.* **125**, 584–598.

Langan, R.T., Lerche, I., Cutler, R.T. (1985). Tracing of rays through heterogeneous media: An accurate and efficient procedure. *Geophysics* **50**, 1456–1465.

Leidenfrost, A., Ettrich, N., Gajewski, D., Kosloff, D. (1999). Comparison of six different methods for calculating traveltimes. *Geophys. Prosp.* **47**, 269–297.

Leung, S., Osher, S.J., Qian, J. (2004). A level set method for three-dimensional paraxial geometrical optics with multiple point sources. *Commun. Math. Sci.* **2**, 643–672.

Li, X.G., Ulrych, T.J. (1993). Traveltime computation in discrete heterogeneous layered media. *J. Seismic Explor.* **2**, 305–318.

Liu, X., Osher, S., Chan, T. (1994). Weighted essentially non-oscillatory schemes. *J. Comp. Phys.* **115**, 200–212.

Lucio, P.S., Lambaré, G., Hanyga, A. (1996). 3D multivalued travel time and amplitude maps. *Pageoph* **148**, 449–479.

McCaughey, M., Singh, S.C. (1997). Simultaneous velocity and interface tomography of normal-incidence and wide-aperture seismic traveltime data. *Geophys. J. Int.* **131**, 87–99.

Michelini, A. (1995). An adaptive-grid formalism for traveltime tomography. *Geophys. J. Int.* **121**, 489–510.

Moser, T.J. (1991). Shortest path calculation of seismic rays. *Geophysics* **56**, 59–67.

Mulder, W., Osher, S., Sethian, J.A. (1992). Computing interface motion in compressible gas dynamics. *J. Comp. Phys.* **100**, 209–228.

Nakanishi, I. (1985). Three-dimensional structure beneath the Hokkaido–Tohoku region as derived from a tomographic inversion of *P*-arrival times. *J. Phys. Earth* **33**, 241–256.

Nakanishi, I., Yamaguchi, K. (1986). A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. *J. Phys. Earth* **34**, 195–201.

Neele, F., VanDecar, J., Snieder, R. (1993). The use of *P* wave amplitude data in a joint inversion with travel times for upper mantle velocity structure. *J. Geophys. Res.* **98**, 12,033–12,054.

Nolet, G. (1987). Waveform tomography. In: Nolet, G. (Ed.), *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*. D. Reidel, Dordrecht, pp. 301–322.

Osher, S., Cheng, L.-T., Kang, M., Shim, H., Tsai, Y.-H. (2002). Geometric optics in a phase-space-based level set and Eulerian framework. *J. Comp. Phys.* **179**, 622–648.

Osher, S., Fedkiw, R. (2003). Level Set Methods and Dynamic Implicit Surfaces, 2nd ed. Springer-Verlag.

Osher, S., Sethian, J. (1988). Fronts propagating with curvature dependent speed: Algorithms based on Hamilton–Jacobi formulations. *J. Comp. Phys.* **79**, 12–49.

Osher, S., Shu, C.W. (1991). High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations. *SIAM J. Numer. Anal.* **28**, 907–922.

Pereyra, V. (1996). Modelling, ray tracing, and block nonlinear travel-time inversion in 3d. *Pure Appl. Geophys.* **148**, 345–386.

Pereyra, V., Lee, W.H.K., Keller, H.B. (1980). Solving two-point seismic-ray tracing problems in a heterogeneous medium. *Bull. Seism. Soc. Am.* **70**, 79–99.

Podvin, P., Lecomte, I. (1991). Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys. J. Int.* **105**, 271–284.

Popovici, A.M., Sethian, J.A. (2002). 3-D imaging using higher order fast marching traveltimes. *Geophysics* **67**, 604–609.

Prothero, W.A., Taylor, W.J., Eickemeyer, J.A. (1988). A fast, two-point, three-dimensional raytracing algorithm using a simple step search method. *Bull. Seism. Soc. Am.* **78**, 1190–1198.

Pulliam, J., Snieder, R. (1996). Fast, efficient calculation of rays and travel times with ray perturbation theory. *J. Acoust. Soc. Am.* **99**, 383–391.

Qian, J., Leung, S. (2004). A level set based eulerian method for paraxial multivalued traveltimes. *J. Comp. Phys.* **197**, 711–736.

Qian, J., Leung, S. (2006). A local level set method for paraxial geometrical optics. *SIAM J. Sci. Comput.* **28**, 206–223.

Qian, J., Symes, W.W. (2002). An adaptive finite-difference method for traveltimes and amplitudes. *Geophysics* **67**, 167–176.

Qin, F., Luo, Y., Olsen, K.B., Cai, W., Schuster, G.T. (1992). Finite-difference solution of the eikonal equation along expanding wavefronts. *Geophysics* **57**, 478–487.

Rawlinson, N., Houseman, G.A. (1998). Inversion for interface structure using teleseismic traveltime residuals. *Geophys. J. Int.* **133**, 756–772.

Rawlinson, N., Sambridge, M. (2003a). Irregular interface parameterization in 3-D wide-angle seismic traveltime tomography. *Geophys. J. Int.* **155**, 79–92.

Rawlinson, N., Sambridge, M. (2003b). Seismic traveltime tomography of the crust and lithosphere. *Advances in Geophysics* **46**, 81–198.

Rawlinson, N., Sambridge, M. (2004a). Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics* **69**, 1338–1350.

Rawlinson, N., Sambridge, M. (2004b). Wavefront evolution in strongly heterogeneous layered media using the Fast Marching Method. *Geophys. J. Int.* **156**, 631–647.

Rawlinson, N., Houseman, G.A., Collins, C.D.N. (2001a). Inversion of seismic refraction and wide-angle reflection traveltimes for 3-D layered crustal structure. *Geophys. J. Int.* **145**, 381–401.

Rawlinson, N., Houseman, G.A., Collins, C.D.N., Drummond, B.J. (2001b). New evidence of Tasmania's tectonic history from a novel seismic experiment. *Geophys. Res. Lett.* **28**, 3337–3340.

Rawlinson, N., Kennett, B.L.N., Heintz, M. (2006a). Insights into the structure of the upper mantle beneath the Murray basin from 3-d teleseismic tomography. *Australian J. Earth Sci.* **53**, 595–604.

Rawlinson, N., Reading, A.M., Kennett, B.L.N. (2006b). Lithospheric structure of Tasmania from a novel form of teleseismic tomography. *J. Geophys. Res.* **111**, doi:10.1029/2005JB003803.

Riahi, M.A., Juhlin, C. (1994). 3-D interpretation of reflected arrival times by finite-difference techniques. *Geophysics* **59**, 844–849.

Riahi, M.A., Lund, C.E., Pederson, L.B. (1997). Three-dimensional image of the Moho undulations beneath the Gulf of Bothnia using wide-angle seismic data. *Geophys. J. Int.* **129**, 461–471.

Rietdijk, R.H. (1999). Notes on big ray tracing. *J. Comp. Phys.* **148**, 149–168.

Sadeghi, H., Suzuki, S., Takenaka, H. (1999). A two-point, three-dimensional seismic ray tracing using genetic algorithms. *Phys. Earth Planet. Inter.* **113**, 355–365.

Saltzer, R.L., Humphreys, E.D. (1997). Upper mantle *P* wave velocity structure of the eastern Snake River Plain and its relationship to geodynamic models of the region. *J. Geophys. Res.* **102**, 11,829–11,841.

Sambridge, M.S. (1990). Non-linear arrival time inversion: Constraining velocity anomalies by seeking smooth models in 3-D. *Geophys. J. Int.* **102**, 653–677.

Sambridge, M., Faletic, R. (2003). Adaptive whole Earth tomography. *Geochem. Geophys. Geosyst.* **4**, doi:10.1029/2001GC000213.

Sambridge, M., Gudmundsson, O. (1998). Tomographic systems of equations with irregular cells. *J. Geophys. Res.* **103**, 773–781.

Sambridge, M.S., Kennett, B.L.N. (1990). Boundary value ray tracing in a heterogeneous medium: A simple and versatile algorithm. *Geophys. J. Int.* **101**, 157–168.

Sambridge, M., Rawlinson, N. (2005). Seismic tomography with irregular meshes. In: Levander, A., Nolet, G. (Eds.), In: *Seismic Earth: Array Analysis of Broadband Seismograms*, vol. 157. American Geophysical Union, pp. 49–65.

Sambridge, M.S., Braun, J., McQueen, H. (1995). Geophysical parameterization and interpolation of irregular data using natural neighbours. *Geophys. J. Int.* **122**, 837–857.

Scott, J.S., Masters, T.G., Vernon, F.L. (1994). 3-D velocity structure of the San Jacinto fault zone near Anza, California-I. *P* waves. *Geophys. J. Int.* **119**, 611–626.

Sethian, J.A. (1996). A fast marching level set method for monotonically advancing fronts. *Proc. Nat. Acad. Sci.* **93**, 1591–1595.

Sethian, J.A. (1999). Level Set Methods and Fast Marching Methods. Cambridge University Press, Cambridge.

Sethian, J.A. (2001). Evolution, implementation, and application of level set and fast marching methods for advancing fronts. *J. Comp. Phys.* **169**, 503–555.

Sethian, J.A., Popovici, A.M. (1999). 3-D traveltime computation using the fast marching method. *Geophysics* **64**, 516–523.

Shapiro, N.M., Campillo, M., Stehly, L., Ritzwoller, M.H. (2005). High-resolution surface wave tomography from ambient seismic noise. *Science* **307**, 1615–1618.

Shu, C.-W., Osher, S. (1988). Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comp. Phys.* **77**, 439–471.

Shu, C.-W., Osher, S. (1989). Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii. *J. Comp. Phys.* **83**, 32–78.

Slawinski, M.A., Slawinski, R.A., Brown, R.J., Parkin, J.M. (2000). A generalized form of Snell's law in anisotropic media. *Geophysics* **65**, 632–637.

Smith, W.H.F., Wessel, P. (1990). Gridding with continuous curvature splines in tension. *Geophysics* **55**, 293–305.

Snieder, R., Sambridge, M. (1992). Ray perturbation theory for traveltimes and ray paths in 3-d heterogeneous media. *Geophys. J. Int.* **109**, 294–322.

Snieder, R., Spencer, C. (1993). A unified approach to ray bending, ray perturbation and paraxial ray theories. *Geophys. J. Int.* **115**, 456–470.

Spakman, W., Bijwaard, H. (2001). Optimization of cell parameterizations for tomographic inverse problems. *Pure Appl. Geophys.* **158**, 1401–1423.

Steck, L.K., Thurber, C.H., Fehler, M., Lutter, W.J., Roberts, P.M., Baldridge, W.S., Stafford, D.G., Sessions, R. (1998). Crust and upper mantle *P* wave velocity structure beneath Valles caldera, New Mexico: Results from the Jemez teleseismic tomography experiment. *J. Geophys. Res.* **103**, 24,301–24,320.

Steinhoff, J., Fan, M., Wang, L. (2000). A new Eulerian method for the computation of propagating short acoustic and electromagnetic pulses. *J. Comp. Phys.* **157**, 683–706.

Sussman, S., Smereka, P., Osher, S.J. (1994). A level set approach for computing solutions to incompressible two-phase flow. *J. Comp. Phys.* **114**, 146–159.

Symes, W.W., Qian, J. (2003). A slowness matching Eulerian method for multivalued solutions of eikonal equations. *SIAM J. Sci. Comput.* **19**, 501–526.

Thomson, C.J., Gubbins, D. (1982). Three-dimensional lithospheric modelling at NORSAR: linearity of the method and amplitude variations from the anomalies. *Geophys. J. Royal Astr. Soc.* **71**, 1–36.

Thurber, C.H. (1983). Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California. *J. Geophys. Res.* **88**, 8226–8236.

Thurber, C.H., Ellsworth, W.L. (1980). Rapid solution of ray tracing problems in heterogeneous media. *Bull. Seism. Soc. Am.* **70**, 1137–1148.

Toomey, D.R., Solomon, S.C., Purdy, G.M. (1994). Tomographic imaging of the shallow crustal structure of the East Pacific Rise at 9°30′N. *J. Geophys. Res.* **99**, 24,135–24,157.

Um, J., Thurber, C. (1987). A fast algorithm for two-point seismic ray tracing. *Bull. Seism. Soc. Am.* **77**, 972–986.

van Trier, J., Symes, W.W. (1991). Upwind finite-difference calculation of traveltimes. *Geophysics* **56**, 812–821.

VanDecar, J.C., James, D.E., Assumpção, M. (1995). Seismic evidence for a fossil mantle plume beneath South America and implications for plate driving forces. *Nature* **378**, 25–31.

Velis, D.R., Ulrych, T.J. (1996). Simulated annealing two-point ray tracing. *Geophys. Res. Lett.* **23**, 201–204.

Velis, D.R., Ulrych, T.J. (2001). Simulated annealing ray tracing in complex three-dimensional media. *Geophys. J. Int.* **145**, 447–459.

Vesnaver, A., Böhm, G., Madrussani, G., Rossi, G., Granser, H. (2000). Depth imaging and velocity calibration by 3D adaptive tomography. *First Break* **18**, 303–312.

Vidale, J.E. (1988). Finite-difference calculations of traveltimes. *Bull. Seism. Soc. Am.* **78**, 2062–2076.

Vidale, J.E. (1990). Finite-difference calculations of traveltimes in three dimensions. *Geophysics* **55**, 521–526.

Vinje, V., Åstebøl, K., Iversen, E., Gjøystdal, H. (1999). 3-D ray modelling by wavefront construction in open models. *Geophys. Prosp.* **64**, 1912–1919.

Vinje, V., Iversen, E., Åstebøl, K., Gjøystdal, H. (1996). Estimation of multivalued arrivals in 3D models using wavefront construction – Part I. *Geophys. Prosp.* **44**, 819–842.

Vinje, V., Iversen, E., Gjøystdal, H. (1993). Traveltime and amplitude estimation using wavefront construction. *Geophysics* **58**, 1157–1166.

Virieux, J., Farra, F. (1991). Ray tracing in 3-d complex isotropic media: An analysis of the problem. *Geophysics* **56**, 2057–2069.

Wang, Y., Houseman, G.A. (1994). Inversion of reflection seismic amplitude data for interface geometry. *Geophys. J. Int.* **117**, 92–110.

Wang, Y., Pratt, R.G. (1997). Sensitivities of seismic traveltimes and amplitudes in reflection tomography. *Geophys. J. Int.* **131**, 618–642.

White, D.J. (1989). Two-dimensional seismic refraction tomography. *Geophys. J.* **97**, 223–245.

Wiggins, S.M., Dorman, L.M., Cornuelle, B.D., Hildebrand, J.A. (1996). Hess deep rift valley structure from seismic tomography. *J. Geophys. Res.* **101**, 22,335–22,353.

Williamson, P.R. (1990). Tomographic inversion in reflection seismology. *Geophys. J. Int.* **100**, 255–274.

Xu, S., Lambaré, G. (2004). Fast migration/inversion with multivalued rayfields: Part 1 – Method, validation test, and application in 2D to Marmousi. *Geophysics* **69**, 1311–1319.

Xu, S., Lambaré, G., Calandra, H. (2004). Fast migration/inversion with multivalued rayfields: Part 1 – Applications to the 3d seg/eage salt model. *Geophysics* **69**, 1320–1328.

Yao, H., van der Hilst, R.D., de Hoop, M.V. (2006). Surface-wave array tomography in SE Tibet from ambient seismic noise and two-station analysis – I. Phase velocity maps. *Geophys. J. Int.* **166**, 732–744.

Zelt, B.C., Ellis, R.M., Zelt, C.A., Hyndman, R.D., Lowe, C., Spence, G.D., Fisher, M.A. (2001). Three-dimensional crustal velocity structure beneath the Strait of Georgia, British Columbia. *Geophys. J. Int.* **144**, 695–712.

Zelt, C.A. (1996). Seismic velocity structure of the central Chilean margin near the subducting Juan Fernández ridge: effective inversion of traveltime data across complex, laterally varying structure. In: Abstracts. 7th Int. Symp., Deep Seismic Profiling of the Continents. Asilomar, CA.

Zelt, C.A. (1999). Modelling strategies and model assessment for wide-angle seismic traveltime data. *Geophys. J. Int.* **139**, 183–204.

Zelt, C.A., Barton, P.J. (1998). Three-dimensional seismic refraction tomography: A comparison of two methods applied to data from the Faeroe Basin. *J. Geophys. Res.* **103**, 7187–7210.

Zelt, C.A., Smith, R.B. (1992). Seismic traveltime inversion for 2-D crustal velocity structure. *Geophys. J. Int.* **108**, 16–34.

Zhang, J., Rector III, J.W., Hoversten, G. (2005). Eikonal solver in the celerity domain. *Geophys. J. Int.* **162**, 1–8.

Zhang, J., Toksöz, M.N. (1998). Nonlinear refraction traveltime tomography. *Geophysics* **63**, 1726–1737.

Zhao, A., Zhongjie, Z., Teng, J. (2004). Minimum travel time tree algorithm for seismic ray tracing: improvement in efficiency. *J. Geophys. Eng.* **1**, 245–251.

Zhao, D., Hasegawa, A., Horiuchi, S. (1992). Tomographic imaging of *P* and *S* wave velocity structure beneath Northeastern Japan. *J. Geophys. Res.* **97**, 19,909–19,928.

Zhou, B., Greenhalgh, S.A. (2005). 'Shortest path' ray tracing for most general 2d/3d anisotropic media. *J. Geophys. Eng.* **2**, 54–63.

This page intentionally left blank

SUBJECT INDEX

This page intentionally left blank