Bulletin of the Seismological Society of America, Vol. 96, No. 6, pp. 2329–2347, December 2006, doi: 10.1785/0120050172

Rapid Estimation of Earthquake Source Parameters from Pattern Analysis of Waveforms Recorded at a Single Three-Component Broadband Station, Port Vila, Vanuatu

by M. N. Zhizhin, D. Rouland, J. Bonnin, A. D. Gvishiani, and A. Burtsev

Abstract The present study deals with rapid, automatic, estimation of some earthquake parameters (location, focal depth, and magnitude) in a region of rather high seismic activity, in quasi-real time, through the analysis of incoming broadband records. The method can be applied, in particular, in poorly instrumented countries with high seismic-risk potential. It can also be applied when the analysis of a very important flow of data requires rapid, sophisticated, preferably automatic, data processing. The method requires, as a minimum, a three-component broadband seismographic station and a sufficiently populated database, that is, an instrument operating for a time long enough to have accumulated an appropriate data set, used to construct the knowledge base. The more extensive the knowledge base, the better the accuracy of the method.

We proceed in several steps. First, applying the SPARS algorithm to the only vertical component, available waveforms are classified according to the source location taken from National Earthquake Information Center (NEIC) catalog; it results in the sorting out of a subset of waveforms/events which will not be included in the knowledge base. Second, each element of the knowledge base is validated according to the epicentral distance with respect to the reference station (and eventually the azimuth of the corresponding source). Third, new input waveforms are analyzed and compared with one or more elements of the knowledge base to estimate their source location and size. The method can be used to search for doublets (or multiplets); if multiplets are found, their location and focal depth can be determined by using a fuzzy event relocation method.

We have tested the capability of the proposed algorithms, processing (broadband) waveforms collected during four and half years at the GEOSCOPE broadband station PVC, operated by Institut de Recherche pour le Développement, formerly ORSTOM (IRD) at Port Vila, Vanuatu. Among 650 events recorded at this station, 254 ones, meeting a good criterion of quality, have been sorted. The results show that, in a range of distances up to 1000 km, the method is capable of yielding, in a very short time, the location of the input event, the accuracy depending on the local density of known events in the vicinity. We also obtain a reliable estimation of the energy by measuring the maximum surface wave (or *S*-wave) amplitude, related to the classical magnitude MSZ.

Introduction

In countries facing a high level of seismic hazard, more or less dense seismological networks have been installed; and numerous articles have been devoted to describing the capabilities of such networks for determining in real time the source parameters (epicenter location, focal depth, magnitude, and moment tensor). The accuracy of the various parameters depends obviously on the density of the contributing stations (for a general review of the basic detection algorithms, see Cuénot, 2003). On the one hand, in some regions with high seismic hazard, the deployment of such sophisticated networks is not easy, in particular, because of a specific environment (*e.g.*, vast oceanic domain surrounding the region, very rugged relief, etc.), and most often because of economical reasons. An alternative solution consists in using, for a rapid location, data recorded at a single station. This could make up one step toward the design of an efficient early warning system (Nakamura, 1988; Saita and Nakamura, 2003)

Data-processing techniques, using single broadband three-component instruments, have been developed, most of them based on automatic *P*- and *S*- (and multiple-) wave-detection algorithms, leading to an estimation of the source location from azimuth and apparent surface velocity measurements (Magotra *et al.*, 1987; Roberts *et al.*, 1989; Saita and Nakamura, 2003) and/or of seismic moment (Talandier *et al.*, 1987; Reymond *et al.*, 1991). Another, very different approach consists in using pattern recognition methods to extract, from broadband records, the information needed to interpret them with respect to a knowledge base. Bonnin *et al.* (1991) and Zhizhin *et al.* (1994, 1995) have developed a method based on cluster analysis of nonlinearly aligned traces: the Syntactic Pattern Recognition Scheme, hereafter called SPARS. It consists in measuring, in a set of records, the dissimilarities (Levenstein distance) between elements of the set, leading to a nearest-neighbor classification of the parameterized waveforms. The method has been successfully applied to seismological data collected at the GEOSCOPE broadband station in Nouméa (New Caledonia); data processing has permitted clustering almost all earthquakes of a given set among themselves (Zhizhin *et al.*, 1995).

The method leads straightforwardly to the search, in a cluster of events, of the so-called multiplets (Ishida and Kanamori, 1978; Geller and Mueller, 1980). Several such investigations have been conducted largely to increase the resolution of travel-time propagation models, for event relocation purposes and for fine local structure studies (Poupinet *et al.*, 1984; Fréchet *et al.*, 1989; Got *et al.*, 1994; Poupinet *et al.*, 2000). To find multiplets, these authors use a precise time window (generally around the *P* and multiple *P* phases) to compare the records among themselves, whereas comparison of events with the SPARS algorithm implies the analysis of the whole seismogram. This second approach allows for searching multiplets in a very large set with a great variety of events. Gaucher (1998) adapted and implemented SPARS to identify multiplets in a set of 16,000 microseismic events recorded at a geothermal field, allowing for a very precise relocation of events along planar structures presumed to be fracture planes. Battaglia (2001) used the SPARS technique to classify automatically different types of seismic signals recorded at Piton de la Fournaise (Réunion Island) volcano. Both authors then applied the master-event technique to relocate microearthquakes (Gaucher), or seismovolcanic events (Battaglia); both authors needed, for relocation purposes, to pick *P*- and *S*-phase arrival times.

In this article, we improve SPARS and associated algorithms to automatically pick *P*- and *S*-phase arrival times and to locate new events by a fuzzy location technique. For testing the method we used data collected by the three-component broadband station PVC, a contributing GEOSCOPE station, operated at Port Vila (Vanuatu). With these data we generate a knowledge base that consists of waveforms and source parameters of well-located events taken from National Earthquake Information Center (NEIC) catalogs. We present a technique for location of new events based on the similarity of their waveforms with the events in the knowledge base, with additional physical constraints derived from the incoming signals.

Seismicity of New Hebrides Region and Observatory Capabilities

The zone covered (Fig. 1) spreads over the whole part of the New Hebrides arc, from Santa Cruz Islands (10° S) to Matthew and Hunter Islands in the southern part of the Fiji Basin (23° S).

The New Hebrides region, with high-potential seismic and volcanic hazards (Louat and Baldassari, 1989; Eissen *et al.*, 1991; Robin and Monzier, 1994), is characterized by a poor coverage of permanent seismic stations, due partly to the vast oceanic environment, which does not favor a national plan for seismic-hazard assessment. Paucity of seismographic stations in the Southwest Pacific is striking if compared, for example, with the coverage in central and southern Europe, whereas the frequency of occurrence of major events is much higher in the former region than in the latter. Therefore, the application of special data-processing techniques based on single-station observations is of particular interest in the latter region.

The seismicity of New Hebrides, observed with temporary regional networks, has been described in a few articles (e.g., Coudert *et al.*, 1981; Kruger-Knuepfer *et al.*, 1986). Epicentral locations are published regularly in NEIC and International Seismological Center (ISC) catalogs. When comparing the locations obtained by these institutions, large discrepancies are sometimes observed, due to the inclusion of more phase data in the ISC computations. In addition, the focal-depth estimate is poorly constrained because of the lack of contributing regional stations. Figure 1a shows the epicenters corresponding to earthquakes located by NEIC, with good resolution on focal depth, whereas, in Figure 1b, the epicenters without any resolution on focal depth (fixed at the conventional value of 33 km) are plotted; we observe that more than 45% of the events correspond to earthquake focuses poorly determined in depth. This low resolution in focal depth is understandable when considering the very small density of stations operating in this region, within a thousand of kilometers. Rosat (1999) underlines these discrepancies by noticing that, in some cases, the epicenter locations reported by international agencies can differ by more than 100 km from those calculated using local seismological networks. Statistics on the largest differences of epicentral distances calculated for events reported during the year 1994 in both NEIC and ISC catalogs on the one hand, and those reported in the regional Institut de Rechercha pour le Developpement, formerly ORSTOM (IRD) catalog (M. Reg-

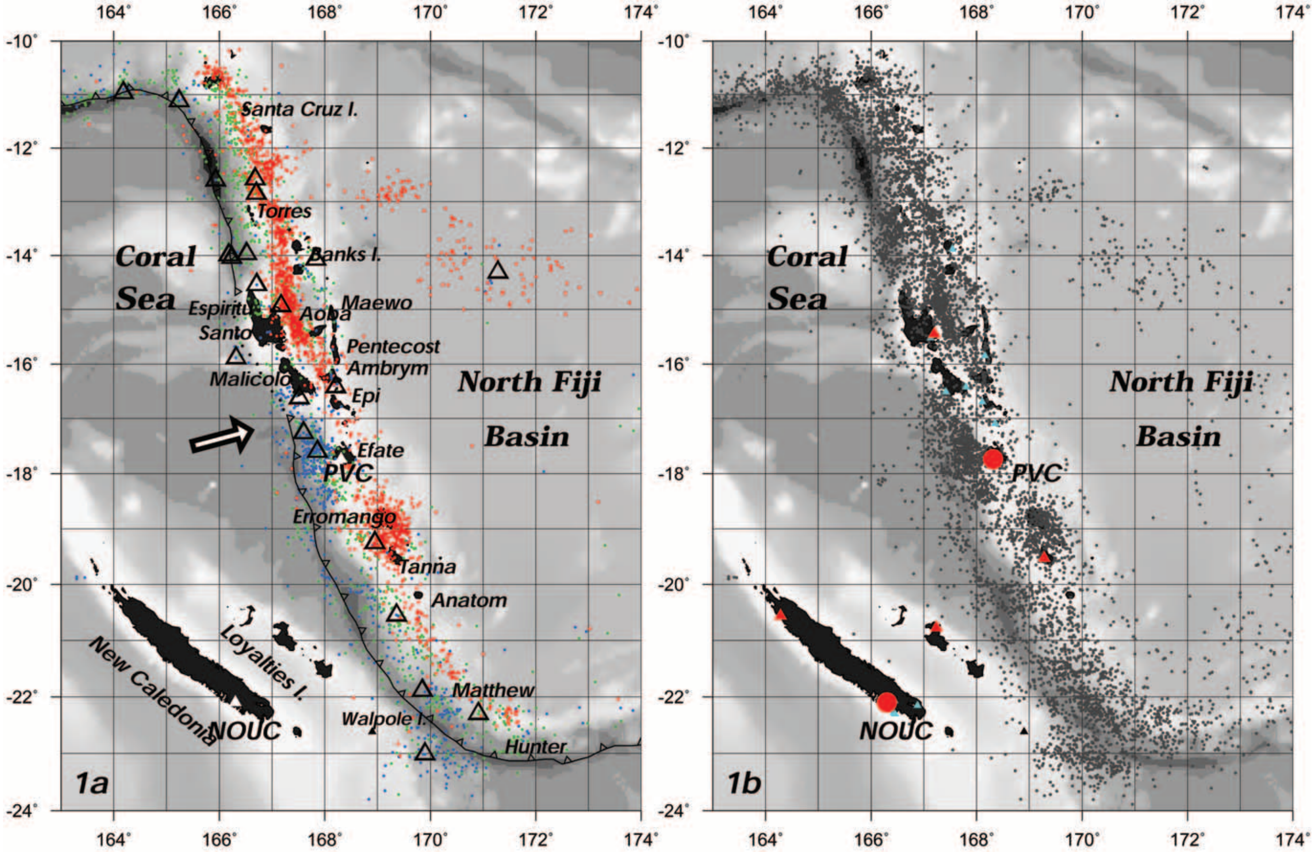

Figure 1. (a) Regional seismicity 1993–1998 (NEIC locations), with green dots for events with focal depth $h \leq 33$ km, blue dots for events with depth $33 < h < 80$ km, and red dots for events with depth $h \geq 80$ km. The strongest events (with $M > 7$) are especially marked with large black triangles. Direction of the relative motion between the Indo-Australian and Pacific plates is indicated by the arrow. The indented solid line indicates the New Hebrides trench and the direction of subduction. (b) NEIC-reported events with unknown depth (black dots). The location of seismic stations operating during this period are shown by red circles for permanent (GEOSCOPE) stations, red triangles for broadband temporary stations, and small blue circles for short-period stations. The locations of the four telemetered short-period stations on Efate Island, in the vicinity of Port Vila, are not reported. Notice the poor coverage of the region by the broadband stations as compared with the strong regional seismic activity.

nier, personal communication) on the other, are summarized in Table 1.

Table 1

Statistics on Epicentral Distance Differences, ΔD, Induced by the Different Events' Locations (Epicentral Distance Is Taken with Respect to PVC Station)

ΔD (km)	No. of Events
NEIC/ISC catalogs	
$\Delta D < 10$	200
$10 < \Delta D < 20$	79
$20 < \Delta D < 40$	52
$40 < \Delta D < 80$	21
$80 < \Delta D$	9
Total	361
ISC/IRD catalogs	
$\Delta D < 10$	5
$10 < \Delta D < 20$	8
$20 < \Delta D < 40$	14
$40 < \Delta D < 80$	15
$80 < \Delta D$	9
Total	51

Data Collection and Preprocessing: Building the Knowledge Base

Data were collected at the seismological station PVC, Vanuatu, during 1993–1998. Epicentral distances vary from ~30 to 900 km, the reference station itself being located in the central part of the area under study. The station was equipped, until May 1995, with a Streckeisen broadband BRB instrument (Romanowicz *et al.*, 1991; Morand and Roult, 1996); then it has been upgraded with the very broadband VBB version. For the different instrument characteristics of the station, we refer to Pillet *et al.* (1990), Morand and Roult (1996), and Roult *et al.* (1999). Because of this up-grading of the instruments, it was necessary to select an

Table 2

Statistics on Detection of Events as a Function of Event Geographic Latitude (see Fig. 5)

Latitudes (deg)	Distance (km)	CECM Detector	LP Detector	Not Detected	Subtotal	Region
$-12 < \text{lat} < -10$	$650 < D < 900$	3	8	1	12	Santa Cruz Island
$-14 < \text{lat} < -12$	$450 < D < 650$	11	13	2	26	Banks and Torres Island
$-16 < \text{lat} < -14$	$200 < D < 450$	15	10	6	31	Aoba and Santo Island
$-18 < \text{lat} < -16$	$0 < D < 200$	66	9	2	77	Mallcolo and Efate Island
$-20 < \text{lat} < -18$	$0 < D < 200$	31	8	8	47	Erromango and Tanna Islands
$-22 < \text{lat} < -20$	$200 < D < 550$	17	9	3	29	Walpole Island
$-24 < \text{lat} < -22$	$550 < D < 800$	11	19	2	32	Matthew and Hunter Island
	Subtotal	154	76	24		
	Total				254	

appropriate channel, and to convert data by filtering and deconvolution into a standard homogeneous format, namely the velocity-deconvolved trace with Butterworth bandpass filters. To test, at first, the quality of the records, mainly their signal-to-noise ratio (SNR), we use the vertical 1-Hz continuously sampling channel. Then, in a second step, to construct the final knowledge base, we use the three-component (with 5-Hz sampling rate) channels, velocity deconvolved. We apply a bandpass filter between 1 Hz and 0.01 Hz, with, in addition, a band-reject filter between 0.182 and 0.125 Hz to reduce the microseismic noise. Details concerning the geographical position of the events are reported in Table 2 (columns 1, 2, and 7), and statistics about the focal depth range are given in Table 3. The privileged location of the recording station, in the central part of the studied area, has a main consequence that most of the recorded earthquakes can have their "symmetrical double" (the same epicentral distance with opposite azimuth); this means that the corresponding seismograms can have *a priori* great similarities, disregarding the fault mechanism effect. In the following sections, we call such events "symmetrical." The use of three-component data sets must be therefore of great interest to reduce the number of doubtful results. On the other hand, the focal-depth diversity is relatively large along the arc, the area under study being located at the borders of subducting plates; noticing that the density of collected events decreases rapidly with focal depth, this creates new difficulties again in searching for events with neighbor hypocenters.

Records, sorted for the observation period 25 December 1993 to 30 June 1998), correspond to about 650 earthquakes located by NEIC, with magnitudes greater than 4. Nevertheless, to avoid bias due to the microseismic noise, which is commonly important in such island sites, we disregarded most of the events recorded with magnitude less than 4.5. Moreover, to ensure a good *S*-wave detection, it is necessary to severely control the noise on the horizontal components, because of the presence of a higher noise level on these components. This additional criterion leads to a reduction of the number of waveforms in the knowledge base, in particular, disregarding the lowest-magnitude events. Finally the records meeting a good criterion of quality have been restricted to a set of 254 events, as it is described in the Computational Methods section. Only half of these events are assigned a computed focal depth in NEIC catalogs, the other ones being reported with the 33 km conventionally fixed depth. A set of nine vertical records at the PVC station, showing the large diversity of events occurring from north to south of the New Hebrides arc, is shown on Figure 2.

Table 3

Events' Focal Depth Distribution, Published in NEIC Preliminary Determination of Earthquakes Bulletin

Depth Range	Events Count
$h < 33$	53
$h = 33$ (arbitrary fixed)	154
$33 < h < 80$	12
$80 < h$	35
Total	254

As a result of the data preprocessing we have built a "learning set" or a "knowledge base" of well-identified waveforms, to which the next incoming seismic signal can be compared by using the computational methods presented in the next section. Then we will try to interpolate the similarity between the waveforms for relocation of the new event by analogy with the identification information learned from the knowledge base.

Computational Methods

In this section we combine traditional seismic methods of waveform interpretation based on the wave detectors and physical models of source and velocity structure together with artificial intelligence reasoning by analogy based on the assumption that similar waveforms originate from similar seismic sources. Our three-component *P*- and *S*-wave detector provides estimates for epicentral distance, azimuth, and event magnitude. Source locations of similar waveforms from the compact granule in the knowledge base (earthquake doublet or multiplet) provide fuzzy event relocation with probability densities for the source coordinates, distance, and focal depth. When used independently both approaches can lead to erroneous conclusions due to noise in the data, complexity of both the source and the propagation path, and the

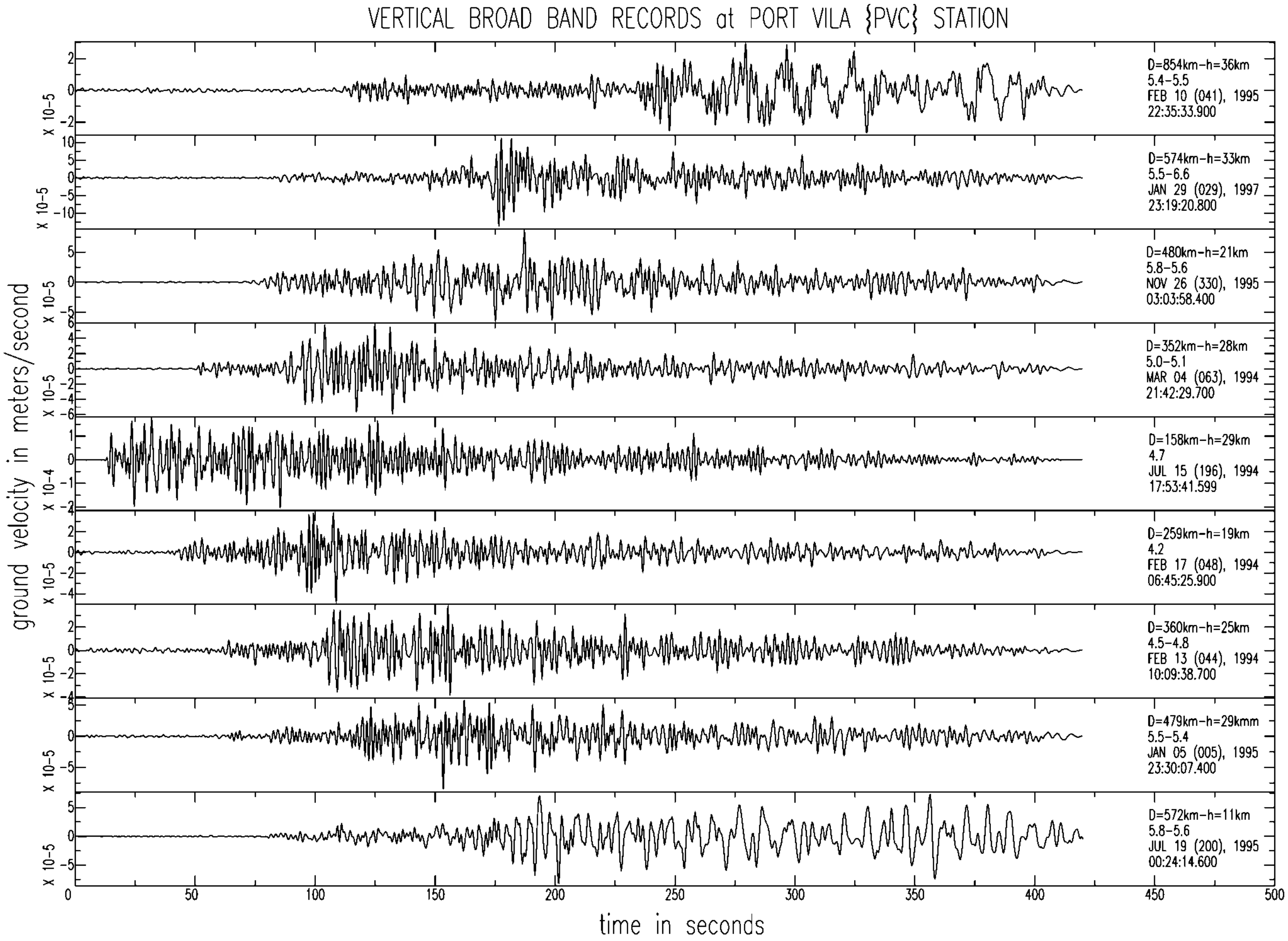

Figure 2. Plot of various vertical seismograms, velocity-deconvolved, bandpass filtered (1–0.01 Hz) with a frequency band-reject filter (0.182–0.125 Hz). The records are plotted from top to bottom according to decreasing latitudes ranging from − 10° S to − 22° S. Epicentral distance and focal depth are indicated on the right of each trace with mention of the body-wave magnitude m_b and, if available, surface-wave magnitude MSZ. Each record (trace) is aligned on its individual event origin time.

limited number of analogies found in the knowledge base. When properly combined, they can constrain the analyst decision space to allow rapid earthquake location and alert, if not as a fully automated procedure, then as a visual decision support component of an expert system.

P- and *S*-Wave Detectors for Regional Seismograms

In our attempt to elaborate a robust and sensitive *P*-wave detector, we have tried several different methods, including the short time average to long time average (STA/LTA) detector, the linear polarization detector, and the multiresolution detector using wavelet-transform (Anant and Dowla, 1997), as well as the so-called LP detector routinely applied for the processing of GEOSCOPE data (Romanowicz *et al.*, 1991). We obtained the best results using a modified version of the component energy comparison method (CECM), proposed in Nagano *et al.* (1989) for automatic detection of *P* arrivals in the location of acoustic emissions. Comparison of capabilities of some detectors is illustrated on Figure 3.

A simple idea behind CECM is that in any *P* wave, the 3D particle motion occurs along one direction, and there is no particle motion in the plane perpendicular to this direction. Thus, a correlation of energy dissipating in the direction of the *P*-wave motion with the energy dissipated in other directions, will be close to zero. On the other hand, in case of random motion in seismic noise, on average, we have the same amount of energy dissipated in all directions. If we assume that the seismic noise is a multivariate Wiener stochastic process, then the energy dissipated in each channel will be proportional to the registration time, $E_x(t) \sim t$ (Wentzell, 1981). Thus, the energy will be correlated between channels. The deterministic component in the stochastic signal (in our case, the *P*-wave arrival) will temporarily destroy the established cross-channel correlation.

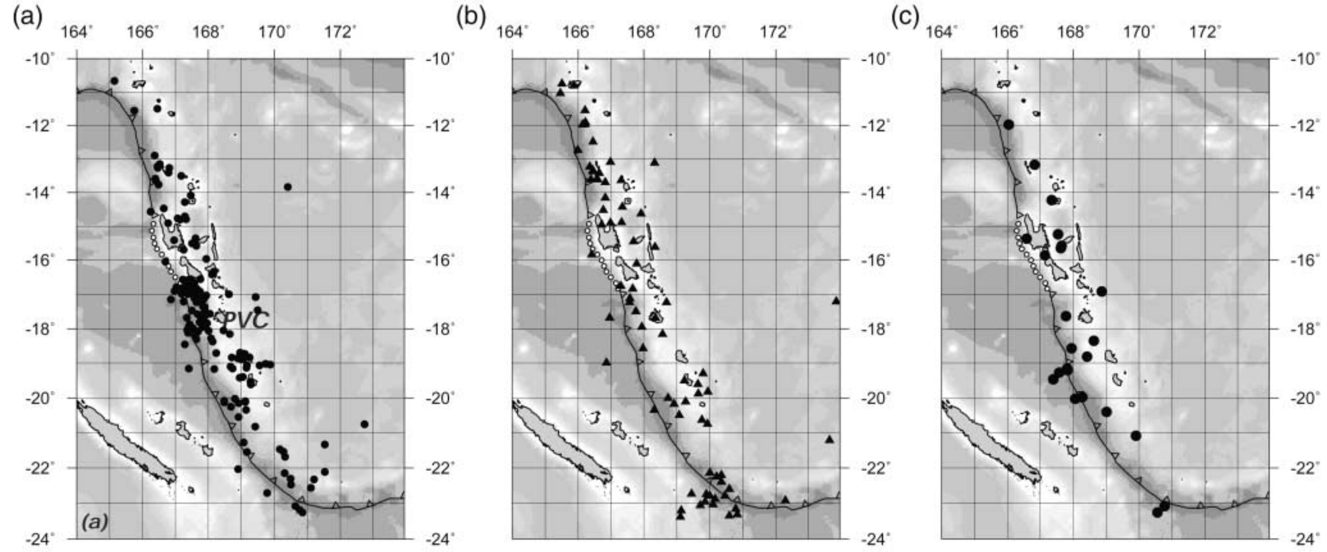

Figure 3. Location of the events used in this study. (a) Events for which both *P* and S waves are detected by the *P-S* detector algorithm. The *P* wave for most of them is also detected by the LP detector. (b) Events detected only by the LP algorithm. (c) Events not detected by either algorithm. The indented solid line indicates the New Hebrides trench and the direction of subduction. The white circles correspond to the lack of this trench between Espiritu Santo and Malicolo Islands.

In our version of the CECM algorithm, we correlate the energy dissipated in a moving time window $T = 25$ sec between the three components of the seismic record. The total energy dissipated in each direction (e.g., along discrete time channel $x(i)$) is defined by

$$E_x(t) = \sum_{i=1}^{t} x^2\ (i). \qquad (1)$$

The choice of the initial time moment $t = 1$ may seem arbitrary: the only requirement is that it is "fixed" long time before t: for example, at each hour boundary. To be less sensitive to the direction of the *P*-wave arrival, we analyze a product of the energy correlation coefficients between *x-z* and *y-z* channels (*z* is the vertical component of the motion; *x* and *y* are in the horizontal plane; *x* is north–south; *y* is west–east):

$$R(t,T) = R_{xz}(t,T)R_{yz}(t,\ T), \qquad (2)$$

where the energy correlation between *x-z* channels is defined by

$$R_{xz}(t,T) = \frac{\sum_{i=t}^{t+T} E_x(i)\,E_z(i)}{\sqrt{\sum_{i=t}^{t+T} E_x^2(i) \sum_{i=t}^{t+T} E_z^2(i)}}\ , \qquad (3)$$

and $R_{yz}(t,\ T)$ is defined by the same formula changing *x* into *y*. The *P*-wave arrival t_0 is preliminarily determined as the time at which the coefficient $R(t,\ T)$ reaches a local minimum, as shown in Figure 4. To reduce the number of false detections, we introduce an experimentally determined threshold for the candidate local minima values $R(t,\ T) \le 0.6$.

Following Nagano *et al.* (1989), we also examine SNR in a time window *T* centered at the preliminary *P*-wave arrival time, $\mathrm{SNR} = 20 \log (P_{\mathrm{signal}}\ (t_0)/P_{\mathrm{noise}}\ (t_0))$, where the average amplitudes of signal and noise are defined as:

$$P_{\mathrm{signal}}(t_0) = \frac{1}{T} \sum_{i=t_0-T/2}^{t_0+T/2} \sqrt{x^2(i) + y^2(i) + z^2(i)}, \qquad (4)$$

$$P_{\mathrm{noise}}(t_0) = \frac{1}{t_0} \sum_{i=1}^{t_0} \sqrt{x^2(i) + y^2(i) + z^2(i)}. \qquad (5)$$

The detected *P* wave is considered to be below noise level if SNR <20 dB. The *P*-wave arrival time t_p is more precisely determined by fitting a parabola to several values of $R(t,\ T)$ before t_0 and finding the time moment when the fitted parabola intersects with time axis. By this we pick the first statistical evidence of the *P* arrival before the well-developed wave train at t_0.

To determine the azimuth from the event source to the station, we use standard polarization analysis of the *P* wave (Frölich and Pullein, 1999) in a fixed-duration time window $T_a = 8$ sec after the arrival time t_p. Other authors (Cichowicz, 1993) have used polarization analysis to detect *P* and *S* arrivals; in the present article, statistics on amplitude distribution between components and wave trains are applied,

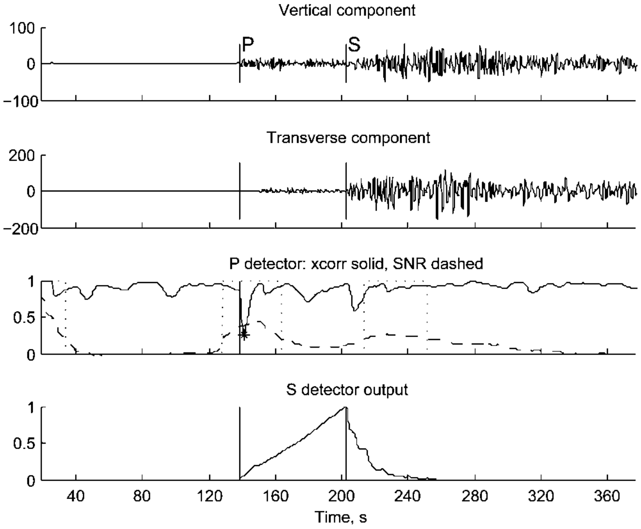

Figure 4. Example of *P* and *S* detections. (top and top middle) Vertical and transverse components with *P*- and *S*-arrival marks and amplitudes in micrometers per second, the transverse component calculated according to NEIC location. (bottom middle) Dimensionless *P*-wave CECM detector output with solid line for energy correlation across the channels; candidate local minima are marked by a star, SNR plotted with a dashed line, and threshold SNR regions marked as dotted segments. (bottom) Dimensionless *S*-wave detector output.

because it proved more robust (see the following text for discussion).

Our *S*-wave detector is based on two assumptions: (1) the probability density functions for distribution of seismic-signal amplitudes $a(t)$ in the Hilbert transform envelopes of the *P*- and *S*-wave trains $p_p\,(a(t))$ and $p_s\,(a(t))$ are normal with significantly different means but similar variances $N(\mu_p, \sigma)$ and $N(\mu_s, \sigma)$; (2) the signals within each of the wave trains can be considered as quasi-stationary (see Nagano *et al.*, 1989). Then arrival time t_s of the *S* wave will divide the signal in the time window between *P*-arrival t_p and maximum degree of polarization $t_{\max}$ (which for regional seismograms occurs within the *S* wave) into two homogeneous segments $t_p \le t_s \le t_{\max}$ with the likelihood

$$L(t_s) = \log \prod_{i=t_p}^{t_s} p_p(a(t_i)) \prod_{i=t_s+1}^{t_{\max}} p_s(a(t_i)). \qquad (6)$$

For normal distribution functions $p_p\,(a(t))$ and $p_s\,(a(t))$, the likelihood formula will be

$$L(t_s) \approx -\frac{t_s - t_p}{2} \log \frac{1}{t_s - t_p} \sum_{i=1}^{t_s} [a(t_i) - \mu_p]^2 - \frac{t_{\max} - t_s}{2} \log \frac{1}{t_{\max} - t_s} \sum_{i=t_s}^{t_{\max}} [a(t_i) - \mu_s]^2 , \qquad (7)$$

where

$$a(t_i) = \sqrt{x(t)^2 + y(t)^2 + z(t)^2}, \qquad \mu_p = \frac{1}{t_s - t_p} \sum_{i=1}^{t_s} a(t_i), \quad \mu_s = \frac{1}{t_{\max} - t_s} \sum_{i=t_s}^{t_{\max}} a(t_i). \qquad (8)$$

The optimal *S*-wave arrival time t_s should make maximum the value of the likelihood function $L(t_s)$ (Fig. 4).

Similarity of Seismic Waveforms

A brief account of the concept of syntactic dissimilarity applied to seismic waveforms is given in the appendix. Waveforms are parameterized by applying a wavelet decomposition, leading to waveform representation in the form of "scalograms."

Analysis of the waveform similarities is based on the assumption that relatively close (in 3D space) seismic sources with similar rupture mechanisms produce similar waveforms at the same recording site. Estimation of the waveform dissimilarity is mathematically equivalent to the embedding of the set of waveforms $w \in \mathbf{W}$ into a metric space $\mathbf{D}$, taking the value of dissimilarity as a metric or distance between waveforms in $\mathbf{D}$. The preceding assumption on the "continuous" mapping from a set of sources $\mathbf{S}$ onto the seismic waveforms $r(\cdot){:}\mathbf{S} \to \mathbf{W}$ with proper choice of observation conditions and waveform dissimilarity measure $d(\cdot,\cdot)$ will manifest itself in the mathematical fact that the multiplets of seismic events will form dense clouds (granules or clusters) in the embedded metric space.

To elaborate a multiplet-detection algorithm, we have to define how to calculate dissimilarity $d(\cdot,\cdot)$ between the seismic waveforms and what is a granule in the resulting metric space $\mathbf{D}$. Obvious candidate for the measure is the maximum cross-correlation (Joswig, 1990):

$$d(w_1, w_2) = \max_t |R(t,T)| = \max_t \left| \frac{\sum_{i=1}^{T} \langle \bar{w}_1(i + t), \bar{w}_2(i) \rangle}{\|w_1\|\,\|w_2\|} \right|, \qquad (9)$$

where by $\langle\cdot,\cdot\rangle$ we denote the inner product of three-component vectors, and by $\|\cdot\|$ we denote their Euclidean norm.

The cross-correlation reaches its maximum when the waveforms are linearly dependent. It is insensitive to the phase shift between waveforms (e.g., different origin time), but it is sensitive to local nonlinear distortions of timescale (e.g., different *S*-*P* delays). In our earlier studies we proposed nonlinear alignment of scalograms by means of dynamic programming technique. A global estimate of dissimilarity (called syntactic distance) between the waveforms takes into account both local delays in arrival times and difference in frequency content and energy envelope of these phases. Syntactic distance has been defined by Levenstein (1965). Similar time-warping technique was developed in the 1970s for speech recognition (Rabiner and Juang, 1993, chapter 4). The first application of time warping to seismic discrimination was reported in Liu and Fu (1982).

Waveform Database Granulation

To analyze the internal structure of a waveform collection and to search for "neighbor" waveforms for fuzzy-event relocation, we use cluster, also known as granulation, analysis based on dissimilarity measure of the entries. Given a set of N waveforms $w_1, \ldots, w_N$ represented by scalograms $S_1, \ldots, S_N$ with known matrix of pairwise syntactic dissimilarities $d_{ij} = D_{\mathrm{Lev}}\,(S_i, S_j)$ (definition of Levenstein distance is given in the appendix), we want to introduce a set of concepts that can regroup the data into a set of granules (clusters, multiplets) and discover unseen (hidden) structure.

A simple definition of database granule around a given waveform w_i will be a set of K-nearest neighbors to w_i, that is, the first K waveforms w_j from the database with the minimal values of dissimilarity d_{ij}, where K is a given constant. In reality, the diversity of waveforms and varying number of events from different granules in the database require a more flexible estimate of the granule size K. In this study, we define an optimal K in the interval $K_{\min} = 1$ to $K_{\max} = 7$ by analyzing how fast the diameter of granule is growing with addition of a new neighbor: slow down of the diameter growth indicates that the optimal granule size is reached. Let us denote by $\Delta_j = d_{i,j+1} - d_{i,j}$, the diameter growth after addition of the neighbor w_j. Addition of waveforms to granule stops when one of the conditions $j = K_{\max}$ or $\Delta_j \geq \Delta_{j+1}$ is reached. We illustrate the granulation algorithm in Figure 5 (top), where the optimal granule size estimate is 2. Granule waveforms are plotted in Figure 5 (bottom), with granule center at the top and neighbor waveforms ordered by increasing syntactic dissimilarity to the center.

Granule-size estimation is a more difficult task than performance measurement in the supervised learning process. A compromise approach was proposed by Gaucher (1998) in his Ph.D. dissertation: he used diameter of the granule source region in space as an additional constraint for the waveform granule growth.

Fuzzy Event Relocation

The fuzzy event relocation method is a nonlinear tool for estimating quantitative seismic-event parameters. It is similar to K-nearest neighbors classifier, but instead of discrete (or nominal) classification (such as explosion, earthquake type of event, etc.) it applies to continuous parameters such as the source depth and epicenter coordinates. Using the values of these parameters for events with the K-nearest neighbor waveforms similar to the unknown event, we are able to construct a nonlinear mapping, namely the kernel-density estimate, which predicts the searched parameters for the unknown event.

The kernel-density estimate (Silverman, 1986) is a computer-intensive method, which is an alternative to the classical histogram and involves smoothing the data while retaining the overall structure. The nonlinear mapping $\hat{f}(x)$ essentially is a weighted sum of Gaussian kernels $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(x^2/2)$, scaled to have a unit area below the graph:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} w_i \phi\left(\frac{x - x_i}{h}\right). \tag{10}$$

The kernel estimate, when calculated with an appropriate value of h, gives a good estimate of the population-density function without making any assumptions, for example, that it is a normal distribution. The only complication lies in estimating the appropriate values of w_i and h, which control the contributions of individual neighbors and the degree of smoothing. In our study, the weights $w_i = (d_i + 0.1\,\bar{d}) \Big/ \sum_{j=1}^{K}(d_j + 0.1\,\bar{d})$, with $\bar{d} = \frac{1}{K}\sum_{j=1}^{K} d_j$, are proportional to dissimilarities d_i of the waveforms of the K-neighborhood to the unknown event, and their bandwidth $h = \sqrt{1 + \sigma^2}$ is proportional to the searched parameter standard deviation σ within the K-neighbors sample,

$$\sigma = \sqrt{\frac{1}{K-1}\sum_{j=1}^{K}(x_j - \bar{x})^2}.$$

The heuristic formulas for the weight w_i and bandwidth h are designed to assure single-mode shape of the kernel density for typical neighborhood sizes, $K = 2, \ldots, 4$. The method is directly applicable for the source depth estimate. Generalization of the formulas for the 2D case of epicenter coordinates is straightforward.

Our method of fuzzy event relocation relies on the assumption that syntactically similar seismic waveforms are generated by earthquake sources that are close in space (with close epicenters and focal depths) and similar in rupture dynamics (with similar focal mechanisms). Here we do not claim that the spaces of all possible seismic waveforms and earthquake sources are topologically equivalent; we only suppose that the assumption holds true locally, that is, very similar waveforms may be generated by sources on the same fault or in the same source region (it is observed during the earthquake-generation process that focal mechanisms are not necessarily equal, but that they are compatible with an average regional stress tensor; this can be explained by the concept of self-organized criticality; A. Cisternas, personal communication). We also suppose that variation of the earthquake magnitude is reflected by scaling of amplitudes of seismic waveforms and does not affect the local relation between similar waveforms and their sources.

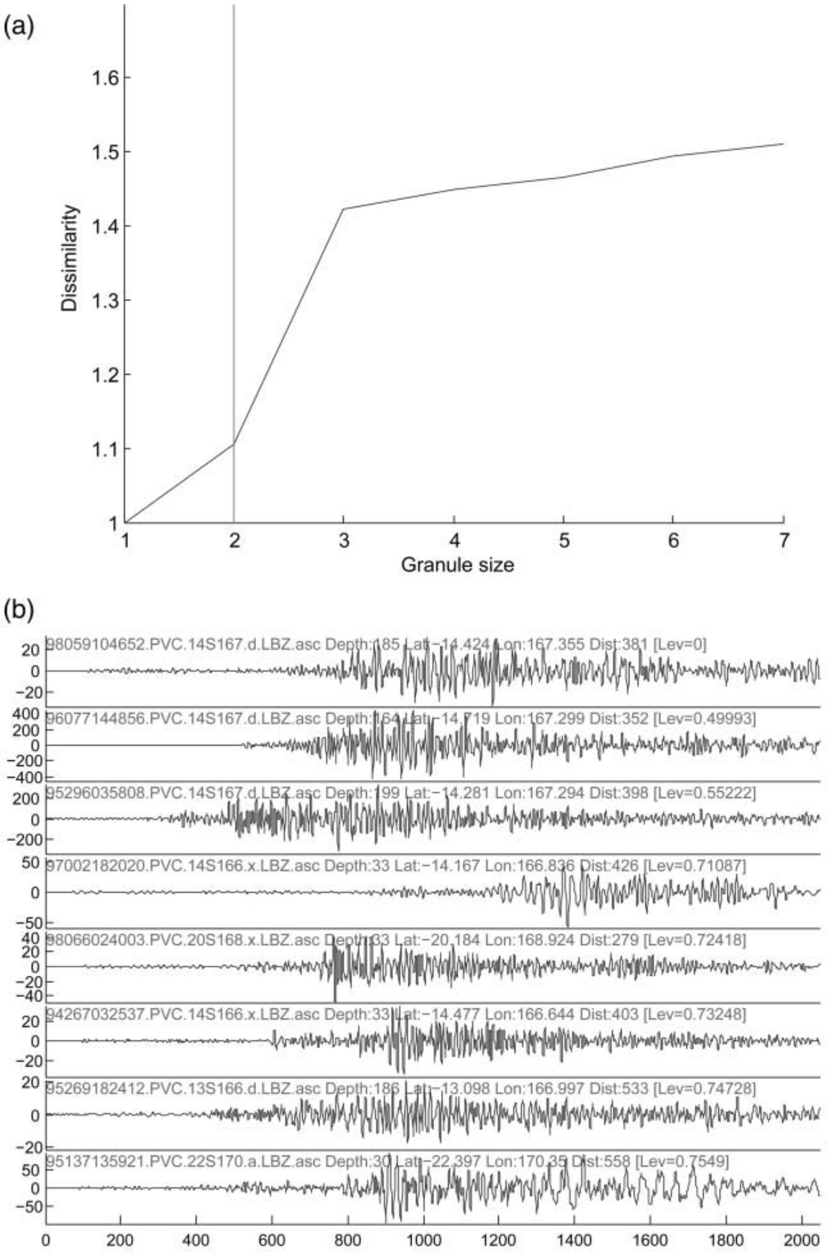

Figure 5. Granulation. (top) Granule diameter growth with the vertical line representing the optimal granule size = 2. (bottom) The upper seismogram (granule center) is compared with the knowledge-base members and the seven nearest neighbors found are shown with increasing dissimilarity from top to bottom.

We illustrate the method for two sets of three-waveforms neighborhoods drawn on Figure 6, with the probabilities for source location and depth. The values of kernel density $\hat{f}(x)$ are calculated on a regular grid, with step of 5 km for depth and 0.25 deg for latitude and longitude estimates. The "arbitrary" depth value of 33 km is ignored in the calculations. We take as the most likely event location the grid cell with the highest values of the kernel density.

Epicentral Distance and Magnitude

The epicentral distance can be estimated from the time difference $\Delta t_{S-P} = t_S - t_P$ of the time arrivals of P- and S-body waves, under the assumption that the crustal structure is known along the path and taking into account the effect of focal depth. Precise determination of the propagation times of body waves can be derived along each path owing to the numerous tomographic models published. In our iterative process, no velocity structure can be defined *a priori*, but we should keep in mind that the main purpose of this work is to locate an event relative to another one using similarity between waveforms; however, a mean value is readily available by referring to regional propagation laws. In the present study, we make use of the hodochrones t_P and t_S derived from regional observation of P and S waves (Dubois, 1971) in the distance range 300 to 800 km:

$$t_p = D/7.878 + 3.18 \quad \text{and} \qquad (11)$$
$$t_s = D/4.506 + 7.67 \, ,$$

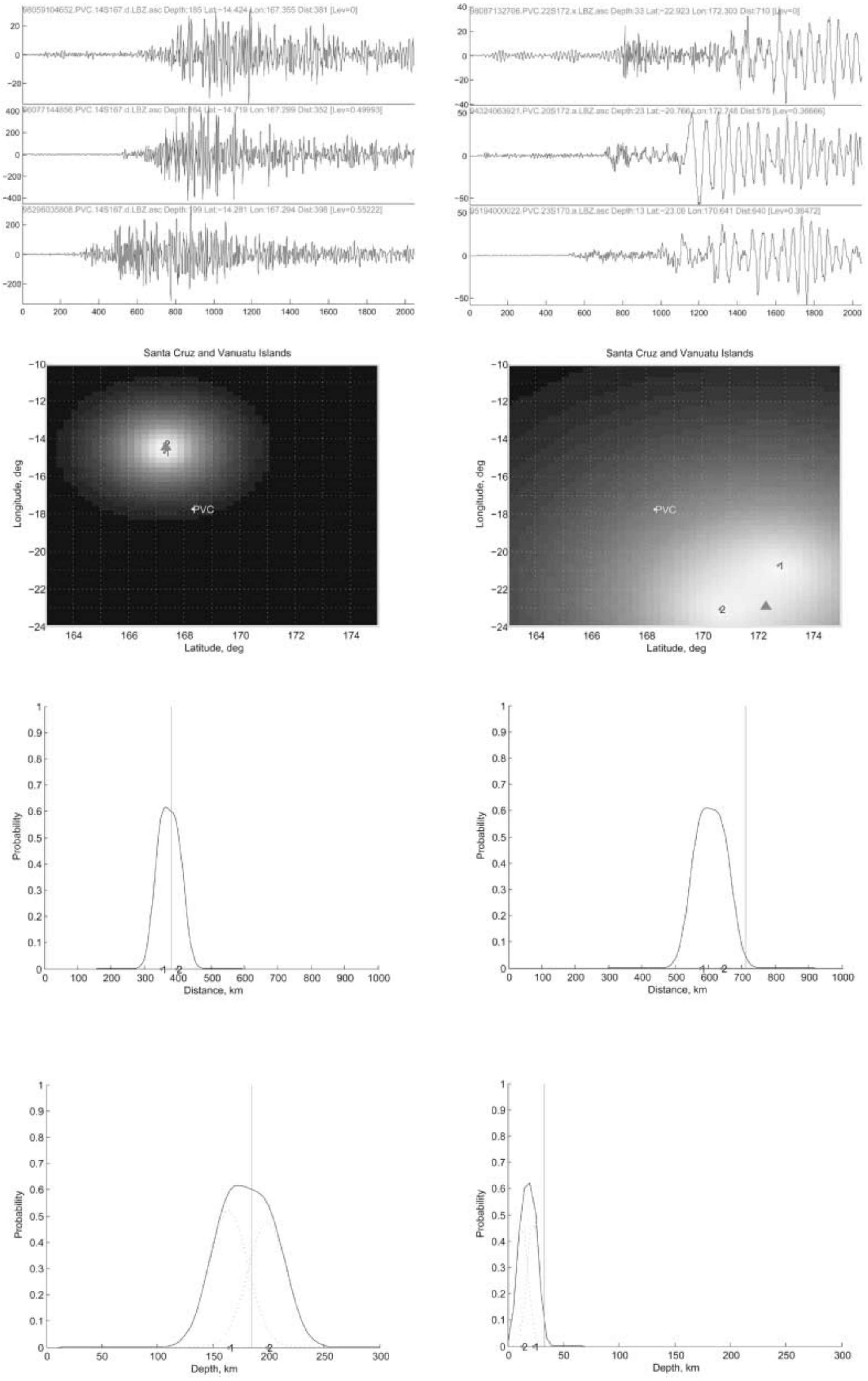

Figure 6. Two examples of fuzzy event relocation. The first example (left half) corresponds to an event that occurred in a region densely covered in the knowledge base; the second example (right half) corresponds to an event that occurred in a region sparsely represented in the knowledge base. (top) The top seismogram is relocated using the locations of its two nearest neighbors. (top middle) Location probability with the brightest zone corresponding to the most likely location of epicenter; longitude/latitude in degrees, dots represent the locations of the neighbor events, numbers represent similarity order, and the triangle shows the catalog location. (bottom middle) Distance probability with the vertical line for the catalog location and the maximum of the black line for the most likely distance estimate from the procedure described in the text. (bottom) Focal depth probability with the same conventions as for the location and individual event kernels shown by the dashed lines.

where D is in kilometers, t_P and t_S are in seconds. A combination of both equations yields the law

$$D = [(t_s - t_p) - 4.5] \times 10.5 = (\Delta t_{s-p} - 4.5) \times 10.5 . \tag{12}$$

Bias due to extrapolation of the law toward short distances (D <100 km) can appear and a more important one can be expected when dealing with deep earthquakes. These possible biases are discussed in the next section.

Magnitude for local and regional events can be obtained using a variety of formulas, depending on the instruments used and on the seismic waves under investigation. Differences in the results, in particular, due to the periods of the sampled amplitudes, lead to a lot of discussions that are out of the scope of the present paper. Because we are dealing with broadband instruments, we decided to use the Praha formula:

$$M_s = \log(A/T)_{\max} + 1.66 \log(D) + 3.3 + C_s, \tag{13}$$

where A, T, and D are, respectively, the displacement in micrometers, the period in seconds, and the epicentral distance in degrees; and C_S is a calibration constant for station correction (Vanek *et al.*, 1962). This magnitude, according to the authors, can be computed in a large-frequency domain, for a very large range of distances; in addition, it converges to MSZ values published by NEIC for epicentral distances greater than 10°. This magnitude M_S can be easily determined directly from the velocity-deconvolved velocity record, $(A/T)_{\max} = V_{\max}/2\pi$; it can be used for oceanic path as well as continental, at distances lower than 1000 km, without restricting the period value (Rouland *et al.*, 1986). The procedure for calculating the magnitude at each station consists in picking automatically the maximum of the signal from the velocity-deconvolved record, whatever the apparent period is, in the temporal window defined by the extreme group velocities 4.0 and 3.0 km$\cdot$sec^{-1} (Rouland *et al.*, 1992, 2003). Notice once again that this calculated value is equivalent to the common surface magnitude value MSZ, issued by NEIC, for distances greater than 1000 km. This procedure does not apply to deep events and can yield values differing strongly from m_b.

Results and Discussion

Syntactic Recognition of Vertical Components Only

When trying to compare events from the New Hebrides Archipelago using SPARS algorithm to assign them to different subregions, we have to take into account the fact that the seismic activity spreads over a long, narrow north–south area. In our study we have divided the area into seven latitudinal sections each with a 2degree span. The selected 254 events have then been classified, applying SPARS, according to the K-nearest neighbor criterion: for each latitude section, the nearest neighbors have been compared and the records displayed for a visual inspection allowing validation of the quality of the performance of the clustering procedure. At this stage, eight events with a bad SNR, or with additional disturbing signal (e.g., two events superimposed), have been disregarded and 246 events have been retained in the knowledge base.

The results of the classification according to the shortest Levenstein distance criterion are the following. For 146 records, another record is found in the knowledge base that corresponds to an epicentral distance differing from the initial one by less than 10%. For 48 other events, the epicentral distance between the initial event and the associated cluster differs by less than 20%. And for 18 more events, the epicentral distance between the event and the cluster differs by less than 30%. In the first category (less than 10%), 43 events have an epicentral distance that differs by 2% only from the corresponding nearest event found in the knowledge base (for some of them the difference in distance is less than a few kilometers). Figures 7 and 8 illustrate selected cases of very good fit: doublets (Fig. 7) and multiplets (Fig. 8). To summarize, among the 246 events processed, 212 events (86%) had, at least, one neighbor according to the K-nearest classifier, with which the corresponding relative epicentral distances are less than 30%. These observations justify the use of the syntactic waveform similarity for fuzzy event relocation.

On the other hand, the 34 events for which the "nearest" waveform events found are too far away in space from the one under consideration (failure of the fuzzy relocation procedure) must be further examined. Among them, a significant portion of "false" associations corresponds to deep events; this is explained by the fact that the density of events included in the data set drops dramatically when focal depth increases and this is why the classification fails in many of these cases. Other cases correspond either to records with a low SNR value, or multiple-events records.

We should also notice that, in several cases which we claimed as successfully paired, the "nearest" waveform neighbor is in fact located in space "symmetrically" to the initial event with respect to the reference seismic station (PVC). Indeed, the paths of seismic waves for these "symmetrical" events are geometrically similar and, therefore, no radical differences in the propagation times, nor in the waveforms, are expected: the Levenstein distance between the records is short. The case of symmetrical events is just a particular one owing to the tectonic structure (curvilinearity) of the region. To explain the similarity of the waveforms, we notice that the structures encountered along the ray paths are comparable. Indeed, for example, records from earthquakes occurring either near Santa Cruz Islands (to the north of the Archipelago) or on Hunter Fracture Zone (to the south) are similar because the structures traveled both correspond to the uppermost part of the subducted Pacific plate, at the same distance from the plate border. This type of symmetrical relocation errors may occur in other areas with linear tectonic structure, especially at the plate boundaries, but it can be solved by combined use of the fuzzy event relocation and the polarization analysis of the detected *P*-wave arrivals.

Other cases of observed faulty clustering, not due either to bad SNR or to equivalent epicentral distances with very different azimuths, are less easy to get successful results from, and it becomes hazardous to try to specify the location of an incoming event, in particular, in using the fuzzy relocation method. Therefore, at this stage, the method needs improvement with a more advanced signal processing, in particular, using three-component records and developing an appropriate *P*-*S* detector is desirable.

Phase Detection in Three-Component Records

To ensure having a secure knowledge base for further applications, we consider that each element of this base must show well-identified seismic phases, principally the body waves, disregarding the remote events located at the border of the area displaying fully developed surface waves. The *P*-*S* detector algorithm was applied to all three-component records collected in the primary knowledge base (254 events); 154 of them provide reliable *P* and *S* detections (Table 2). The distances calculated through regionalized formula (see previous section) are in good agreement with those published in ISC and/or NEIC catalogs and confirm our choice of Dubois's regional *P* and *S* travel-time tables. In

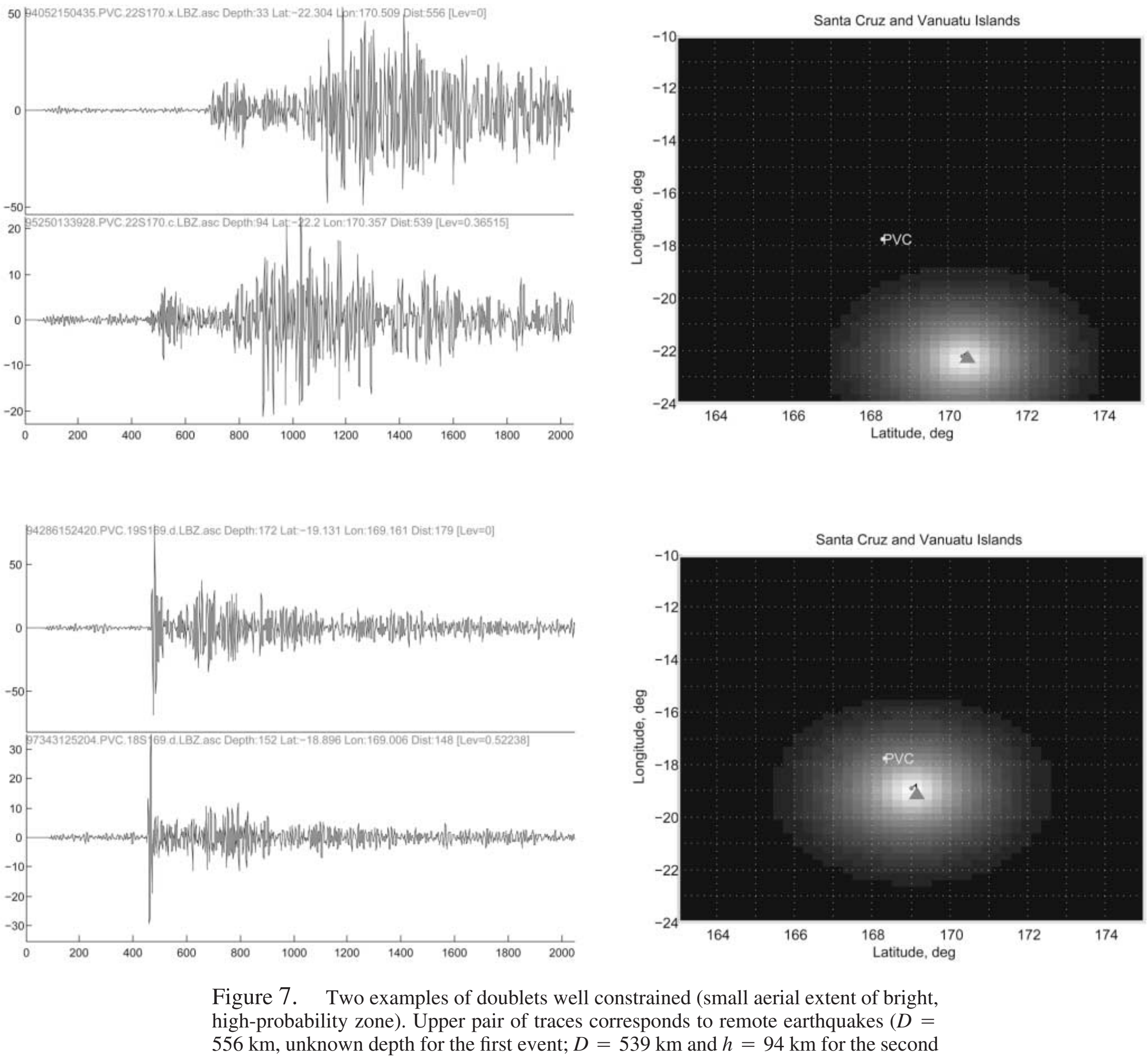

Figure 7. Two examples of doublets well constrained (small aerial extent of bright, high-probability zone). Upper pair of traces corresponds to remote earthquakes (D = 556 km, unknown depth for the first event; D = 539 km and h = 94 km for the second event). Lower pair of traces corresponds to deep earthquakes (D = 179 km, h = 172 km, for the first event; D = 148 km and h = 152 km for the second event). In each case, the upper trace is from the event under test and the lower trace corresponds to the closest neighbor in the SPARS knowledge base.

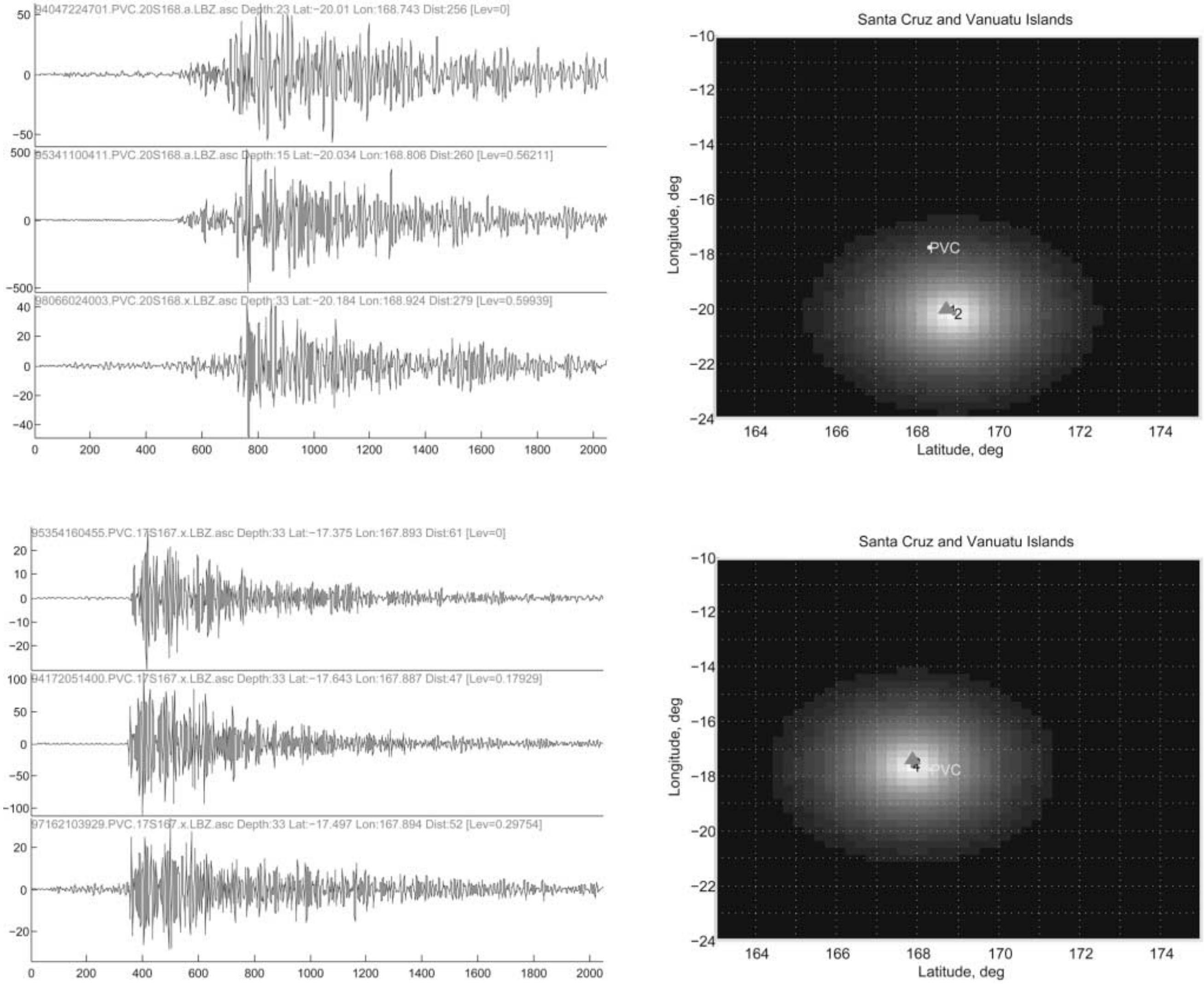

Figure 8. Examples of multiplets well constrained in this study (small aerial extent of bright, high-probability zone). The upper three traces correspond to regional earthquakes ($D = 256$ km, $h = 23$ km for the first event; $D = 260$ km and $h = 15$ km for the second event; $D = 279$ km and depth is unknown for the third event). The lower three traces correspond to local earthquakes of unknown depth at distances $D = 61$ km, $D = 47$ km, and $D = 52$ km. In each case, the upper trace is from the event under test and the lower traces correspond to the nearest neighbors in a granule found in the SPARS knowledge base.

the case of nearby earthquakes (epicentral distance less than 150 km), the difference of a few tens of kilometers between observed values (this study) and published ones (catalogs) are compatible with the errors generally accepted in routine location and the uncertainty in the procedures of focal depth estimation. Comparison between these measured and published epicentral distances is illustrated in Figure 9a. A bias is observed at large distances (greater than 350 km); it can be explained because we use only one formula to link the epicentral distance to the observed *S-P* time difference. Another reason to observe a quite large difference between the two values is that we do not take into account the contribution of depth in the regionalized formula: the *P-S* detector algorithm does not yield depth information (see, for example, event marked “1” on Fig. 9a). In addition, differences observed at shortest distances are most probably coming from uncertainties in the catalog’s locations.

The preceding discussion is applicable when comparing magnitudes estimated in this study and those published in catalogs. Comparison between m_b values and the ones obtained in this study is illustrated in Figure 9b. Miscalculations in evaluating the epicentral distance have a direct impact on magnitude estimation. We must also be careful in comparing m_b, which is most often understood as a mean value, with magnitude calculated using records of individual stations. Moreover, the extension at short distances of the

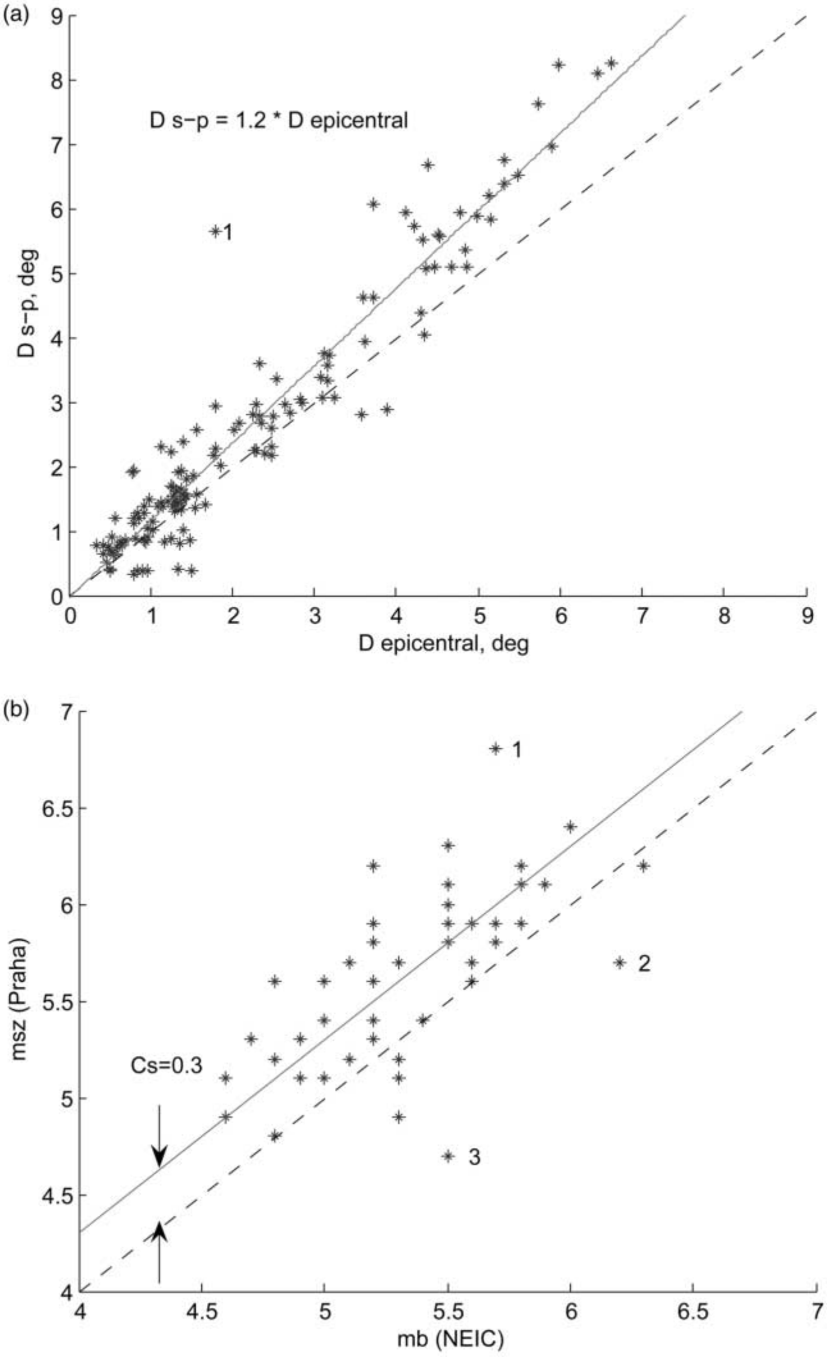

Figure 9. (a) Distances determined in this study by means of the *P-S* detector algorithm are plotted versus the epicentral distance computed from NEIC catalog. $D_{s\text{-}p}$ is the distance derived from the picking of *P* and *S* arrivals. The solid line is the best-fit regression line; its slope (1.2) comes mainly from the difference between the adopted values of T_P/T_S hodochrones and the actual ones, but also results partly from a focal depth effect (see, for example, point labeled "I," which corresponds to an event 178 km deep). (b) Magnitudes calculated according to the Praha formula (this study) are plotted versus published values from NEIC catalog. The solid line is the regression line constrained to a slope of 1. C_S corresponds to the so-called "station correction" in Praha formula (see text). Values in excess correspond either to a distance not well constrained by the *P-S* detector algorithm (point labeled "1"), or to a strong focal depth effect (points labeled "2" and "3").

use of Praha formula (13) is questionable; nevertheless, the observed differences are in the same order as those published by different institutions (see, *e.g.*, numerous examples in Lebreton, [1997]).

In the preceding section, we suggest constructing a reliable and robust knowledge base by analyzing *P* and *S* waves: this requires essential information contained in each set of three-component records. This step in data processing allows elimination of, in particular, noisy seismograms, but also "complex" ones for which we encounter difficulties analyzing automatically the body waves. By running SPARS algorithm on this new "clean" database, we considerably increase the possibility of finding events that are statistically close to each other, commonly called doublet (or multiplet, if more than two items) if their relative locations correspond to *a priori* given criteria, as summarized in Table 4. In this study, the search for such events includes similarity of all the waves characterizing the whole record, instead of precisely adjusting part of the waveforms. Our approach could be a first step in dealing with a very large amount of data. The initial purpose of the present study, however, was to classify events according to their geographical location.

Table 4
Events Included in the Database Showing Good *P* and *S* Waves, and Located According to Similarity of the Waveforms

Latitude Range (deg)	No. of Events	With Fuzzy Location*	With Single Neighbor[†]	With a Symmetrical Neighbor[‡]	*P-S* only[§]	Satisfactory Relocation (%)[¶]
$-12 < \text{lat} < -10$	3	0	2	1	0	67
$-14 < \text{lat} < -12$	11	3	4	3	1	64
$-16 < \text{lat} < -14$	15	3	3	4	5	40
$-18 < \text{lat} < -16$	66	35	17	6	8	79
$-20 < \text{lat} < -18$	31	4	8	6	13	38
$-22 < \text{lat} < -20$	17	4	6	6	1	59
$-24 < \text{lat} < -22$	11	4	4	3	0	73
Total	154	53	44	29	33	63

*"With fuzzy location" means that the event is located from a well-defined cluster considered as a multiplet.

[†] "With a single neighbor" means that the event to locate forms a doublet with another event.

[‡] "With a symmetrical neighbor" means that the event to locate is associated by SPARS with an event at the same distance but in the opposite direction relative to PVC station.

[§] "*P-S* only" means that no waveform similar to the one exhibited by the event under examination has been found in the knowledge base; nevertheless *P* and *S* are well detected, corresponding to "good" epicentral distance with respect to PVC.

[¶] The last column summarizes the percentage of events properly relocated (With Fuzzy Location + With Single Neighbor); the latitude of PVC is about 16° S: notice that in this latitude section there are many satisfactory relocations (79%).

We have tried to apply polarization analysis; however, because of the slab structure, the *P* arrivals are complex due to the combination of direct and multiply reflected *P* waves in a very short time after the *P* onset, not always compatible with the accuracy of *P* detection. In addition, a given uncertainty on *P* detection results in a relatively small error on epicentral distance, whereas the location of the source could induce large errors on azimuth which would have a strong impact on polarization. The location of the reference source (from the knowledge base) at a given epicentral distance does not provide the right azimuth, used for validation of the estimated polarization, to be checked against the output of the polarization analysis.

Conclusions

We have successively investigated two complementary applications of artificial intelligence techniques associated with well-established physical models of earthquakes, among many other possible, to the automatic recognition of the incoming seismic signal. First, we have applied the techniques of pattern recognition to seismic waveforms. Indeed, this technique requires a learning approach, which must be performed under the supervision of an expert, and ends up with a 'learning set' of well-identified waveforms, to which the next incoming signal can be compared; similarity between the new signal and the waveforms of the learning set can then be measured.

A second application has been investigated: how picking *P*- and *S*-phase arrival times can improve and complement the fuzzy relocation of seismic events based on similarity of their waveforms. Once *P*- and *S*-wave arrivals are picked, similarity of waveforms can be more certainly interpreted in terms of common source zones. Indeed, as more waveforms are analyzed, the knowledge base is enlarged and improved, and it allows better and better controlled clustering of the earthquake sources zones.

Computing time is not a limiting factor for the usefulness of the process: the most time-consuming step is the computation of Hilbert and wavelet transforms on a few hundreds of samples; the techniques investigated could be used for automatically offering the seismologist on duty in an observatory a very first, just-in-time, estimate of the source parameters of an incoming waveform: very few minutes (1–2?) are sufficient to have a long-enough waveform (i.e., up to the incoming of Rayleigh waves).

Implementation of this approach for surveillance of local earthquake activity (within a distance of a few hundreds of kilometers) allows a single recording station to yield a rapid insight, within a few tens of seconds, onto potentially damaging events. In a certain sense, this technique is complementary to, but almost more rapid, than the usual detection networks. Recent developments of signal analysis linked with a very good knowledge of the local structure allow to process only the beginning of the *P* wave leading in a few seconds to a very fast alarm, that is, an early warning system (Wu and Kanamori, 2005). In addition, our technique could be used to implement information for rapid assessment of the damage potential of earthquakes.

The efficiency of the proposed signal-processing algorithms applied to local seismic-network observations and combined with the know-how of seismologists, may also help to better locate and understand the seismicity of an active and tectonically complex seismic region. On the other hand, the routine application of the algorithms developed in this article may be attractive to determine event parameters in large volume data flows provided by risk-monitoring seismological networks, as well as in monitoring volcanic activity (Rouland *et al.*, in preparation).

Acknowledgments

This work was supported in the framework of a C.N.R.S "Programme d'Intérêt Commun" project involving scientists from the Institute of Physics of the Earth, Moscow, and the Institute of Physics of the Earth, Strasbourg. We thank all persons who contributed to it, in particular, R. Decourt, C. Baldassari, F. Bondoux, and L. Mollard, who participated actively in field operations and in data processing in Nouméa, New Caledonia, and Port-Vila, Vanuatu, as well as the GEOSCOPE organization. We thank D. Boulanger for his precious aid in up-grading the computing facilities by implementing software. We are very grateful to A. Cisternas and J. Fréchet for carefully reading various early versions of the manuscript and suggesting a number of improvements. We are also very grateful to the W. H. K. Lee and an anonymous reviewer for their valuable, detailed, and constructive comments.

References

Anant, K., and F. Dowla (1997). Wavelet-transform methods for phase identification in three-component seismograms, *Bull. Seism. Soc. Am.* **87,** no. 6, 1598–1612.

Battaglia, J. (2001). Quantification sismique des phénomènes magmatiques sur le Piton de la Fournaise entre 1991 et 2000, *Ph.D. Thesis*, Université Paris 7-Denis Diderot, Paris, 322 pp.

Bonnin, J., A. Gvishiani, M. Zhizhin, B. Mohammadioun, and J. Sallantin (1991). Strong motion data classification using syntactic pattern recognition scheme, in *Proc. Fourth Int. Conf. Seismic Zonation*, Stanford, Vol. II, 549–555.

Cichowicz, A. (1993). An automatic S-phase picker, *Bull. Seism. Soc. Am.* **83,** 180–189.

Coudert, E., B. L. Isacks, M. Barazangi, R. Louat, R. Cardwell, A. Chen, J. Dubois, G. Latham, and B. Pontoise (1981). Spatial distribution and mechanisms of earthquakes in the southern New Hebrides arc from a temporary land and ocean bottom seismic network and from world-wide observations, *J. Geophys. Res.* **86,** 5905–5925.

Cuénot, O. (2003). Les algorithmes de detection automatique d'ondes sismiques, Mémoire probatoire, Centre National des Arts et Métiers, Lyon, 9 pp., http://ocuenot.online.fr/ProbatoireCNAM_Olivier CUENOT.pdf (last accessed October 2006).

Dubois, J. (1971). Propagation of P waves and Rayleigh waves in Melanesia: structural implications, *J. Geophys. Res.* **76,** no. 29, 7217–7240.

Eissen, J. P., C. Blot, and R. Louat (1991). Chronologie de l'activité volcanique historique de l'arc insulaire des Nouvelles-Hébrides de 1595 à 1991, Rapports scientifiques et techniques, Sciences de la Terre, Centre ORSTOM de Nouméa, Nouvelle-Calédonie, 69 pp.

Fréchet, J., L. Martel, L. Nikolla, and G. Poupinet (1989). Application of the cross-spectral moving-window technique (CSMWT) to the seismic monitoring of forced fluid migration in a rock mass, *Int. J. Rock. Mech. Min. Sci. Geomech. Abstr.* **26,** 221–233.

Frölich, C., and J. Pulliam (1999). Single-station location of seismic events: a review and a plea for more research, *Phys. Earth Planet. Interiors* **113,** 277–291.

Gaucher, E. (1998). Comportement hydromécanique d'un massif fracturé: apport de la microsismicité induite. Application au site géothermique de Soultz-sous-Forêts, *Ph.D. Thesis*, Université Paris 7-IPGP, Paris, 245 pp.

Geller, R. J., and C. S. Mueller (1980). Four similar earthquakes in Central California, *Geophys. Res. Lett.* **7,** 821–824.

Got, J.-L., J. Fréchet, and F. W. Klein (1994). Deep fault plane geometry inferred from multiplet relative location beneath the south flank of Kilauea, *J. Geophys. Res.* **99,** 15,375–15,386.

Ishida, M., and H. Kanamori (1978). The foreshock activity of the 1971 San Fernando earthquake, California, *Bull. Seism. Soc. Am.* **68,** 1265–1279.

Joswig, M. (1990). Pattern recognition for earthquake detection, *Bull. Seism. Soc. Am.* **80,** 170–186.

Kruger-Knuepfer, J.-L., J.-L. Chatelain, M. W. Hamburger, B. L. Isacks, M. Barazangi, G. Hade, R. Prévot, and J. Kelleher (1986). Evaluation of seismic risk in the Tonga-Fiji-Vanuatu region of the Southwest Pacific, A country report, Republic of Vanuatu, Report submitted to Office of U.S. Foreign Disaster Assistance, 118 pp.

Lebreton, S. (1997). Comparaison des magnitudes locales diffusées par différents instituts européens, *Diplôme d'Ingénieur*, École et Observatoire des Sciences de la Terre, Université Louis Pasteur, Strasbourg, 132 pp.

Levenstein, V. I. (1965). Binary codes with use of deletions, insertions and substitutions of symbols, *Dokl. Akad. Nauk. SSSR*, 132 pp.

Liu, H. H., and K. S. Fu (1982). A syntactic approach to seismic pattern recognition, *IEEE Trans. Pattern Anal. Mach. Intell. PAMI* **4,** 136–140.

Louat, R., and C. Baldassari (1989). Chronologie des séismes et des tsunamis ressentis dans la région Vanuatu-Nouvelle Calédonie, Rapports scientifiques et techniques, Sciences de la Terre, 1, ORSTOM, Centre de Nouméa, 48 pp.

Magotra, N., N. Ahmed, and E. Chael (1987). Seismic event detection and source location using single station (three-component) data, *Bull. Seism. Soc. Am.* **77,** 958–971.

Morand, M., and G. Roult (1996). Géoscope Station Book, *Int. Rep. Institut de Physique du Globe, Paris*, 166 pp.

Nagano, K., H. Niitsuma, and N. Chubachi (1989). Automatic algorithm for triaxial hodogram source location in downhole acoustic emission measurement, *Geophysics* **54,** no. 4, 508–513.

Nakamura, Y. (1988). On the urgent earthquake detection and alarm system (UrEDAS), in *Proceedings of the 9th World Conference on Earthquake Engineering*, Tokyo-Kyoto, Japan.

Pillet, R., J.-M. Cantin, and D. Rouland (1990). Acquisition numérique pour sismomètre large bande, *Géodynamique* **2,** 14–20.

Poupinet, G., W. L. Ellsworth, and J. Fréchet (1984). Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault, California. *J. Geophys. Res.* **89,** 5713–5719.

Poupinet, G., A. Souriau, and O. Coutant (2000). The existence of an inner core super-rotation questioned by teleseismic doublets, *Phys. Earth Planet. Interiors* **118,** 77–88.

Rabiner, L., and B.-H. Juang (1993). *Fundamentals of Speech Recognition*, Prentice Hall, New York, 496 pp.

Reymond, D., O. Hyvernaud, and J. Talandier (1991). Automatic detection, location and quantification of earthquakes. Application to tsunami. *Pageoph* **135,** 361–382.

Roberts, R. G., A. Christofferson, and F. Cassidy (1989). Real-time detection, phase identification and source location estimation using single station three-component seismic data, *Geophys. J.* **97,** 471–480.

Robin, C., and M. Monzier (1994). Risque volcanique au Vanuatu, Rapport ORSTOM et Département de Géologie, des Mines et des Ressources en Eau du Gouvernement de Vanuatu, 23 pp.

Romanowicz, B., J.-F. Karczevsky, M. Cara, P. Bernard, J. Borsenberger, J.-M. Cantin, B. Dole, D. Fouassier, J.-C. Koenig, M. Morand, R. Pillet, and D. Rouland (1991). The Géoscope program: present status and perspective, *Bull. Seism. Soc. Am.* **81,** 243–264.

Rosat, S. (1999) Étude des capacités de location du réseau Cavascope, Rapport de stage Sciences de la Terre, École de Physique du Globe de Strasbourg, Univ. Louis Pasteur-Centre IRD, Nouméa (Nouvelle Calédonie), 30 pp.

Rouland, D., C. Condis, C. Parmentier, and A. Souriau (1992). Previously undetected earthquakes in the southern hemisphere using long-period GEOSCOPE data, *Bull. Seism. Soc. Am.* **82,** 2448–2463.

Rouland, D., C. Condis, and G. Roult (2003). Overlooked earthquakes on and around the Antarctica plate; identification and location of 1999 shallow depth events, *Tectonophysics* **376,** 1–17.

Rouland, D., J.-J. Lévêque, and M. Cara (1986). Magnitude determination from the Strasbourg broad-band stations, *Terra Cognita* **6,** 445.

Roult, G., J.-P. Montagner, E. Stutzmann, S. Barbier, and G. Guivineux (1999). The GEOSCOPE program: its data center, *Phys. Earth Planet. Interiors* **113,** 25–43.

Saita, J., and Y. Nakamura (2003). UrEDAS: the early warning system for mitigation of disasters caused by earthquakes and tsunamis, in *Early Warning Systems for Natural Disaster Reduction*, J. Zschau and A. N. Kuppers (Editors), Springer-Verlag, Berlin, 834 pp.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, Great Britain, 175 pp.

Talandier, J., D. Reymond, and E. A. Okal (1987). M_m: use of variable mantle magnitude for the rapid one-station estimation of teleseismic moments, *Geophys. Res. Lett.* **14,** 840–843.

Thurber, C. H., H. Given, and J. Berger (1989). Regional seismic event location with a sparse network: application to eastern Kazakhstan, USSR, *J. Geophys. Res.* **94,** 17,767–17,780.

Vanek, J., V. Zatopek, V. Karnik, N. V. Kondorskaya, Y. V. Rizmitchenko, E. F. Savarensky, S. L. Solovyov, and N. V. Shebalin (1962). Standardization of magnitudes scales, *Izv. Akad. Nauk SSSR Ser. Geofiz.* **2,** 153–158.

Wentzell, A. D. (1981). *A course in the Theory of Stochastic Processes*, McGraw-Hill International, New York, 304 pp.

Wu, Y., and H. Kanamori (2005). Rapid assesment of damage potential of earthquakes in Taiwan from the beginning of P waves, *Bull. Seism. Soc. Am.* **95,** 1181–1185.

Zhizhin, M., A. Gvishiani, J. Bonnin, R. Madariaga, B. Mohammadioun, and D. Rouland (1995). Syntactic pattern recognition scheme (SPARS) applied to seismological waveforms analysis, *Cahiers Centre Européen Géodyn. Séism.* **9,** 17–26.

Zhizhin, M., A. Gvishiani, D. Rouland, J. Bonnin, and B. Mohammadioun (1994). Identification of a geological region for earthquakes using syntactic pattern recognition, *Nat. Hazards* **10,** 139–147.

Appendix

Syntactic Dissimilarity of Seismic Waveforms

Our definition of the single-component waveform dissimilarity is derived from the speech recognition field. First, we parameterize the waveforms by using time-frequency representation. We use continuous time wavelet transform especially developed for broadband signal analysis. Second, we compare the time-frequency diagrams and use the "best match" residual as the waveform dissimilarity measure. The obvious candidate for the measure is maximum cross-correlation (Joswig, 1990). The cross-correlation reaches its maximum when the waveforms are linearly dependent. It is insensitive to the phase shift between waveforms (e.g., different origin time), but it is sensitive to local nonlinear distortions of timescale (e.g., different *P*-*S* delays). In this study we propose nonlinear alignment of scalograms by means of dynamic programming technique. A global estimate of dissimilarity (called syntactic distance) between the waveforms takes into account both local delays in arrivals times and difference in frequency content and energy envelope of these phases. Syntactic distance was defined by B. Levenstein in 1965 (Levenstein, 1965). Similar time-warping technique was developed in the 70 sec for speech recognition (Rabiner, 1993, chapter 4). The first application of time warping for seismic discrimination was reported by K. Fu in 1982 (Liu and Fu, 1982).

Continuous wavelet transform (CWT) is defined as a L_2 projection of signal $x(t)$ onto a family of analyzing functions:

$$\begin{aligned} W_h x(a,b) &= \langle x, h_{a,b} \rangle \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\bar{h}\left(\frac{t-b}{a}\right) dt \quad (a \neq 0), \end{aligned} \tag{A1}$$

where the family of functions

$$h_{a,b}(t) = \frac{1}{\sqrt{|a|}}\, h\left(\frac{x-b}{a}\right)$$

is generated from a "mother" wavelet $h(t)$ by means of translation in time by translation parameter $b \in \mathbf{R}$ (to select a part of signal to be analyzed) and dilation taking $|a| > 1$ or compression taking $|a| < 1$ by scale parameter $a \in \mathbf{R} \setminus \{0\}$ (to focus on a given range of oscillations in the selected part). The term $\frac{1}{\sqrt{a}}$ keeps the energy of the scaled wavelet equal to the energy of the original mother wavelet.

This timescale expression has an equivalent time-frequency expression, obtained by using the formal identification $f = \frac{f_0}{a}$, where f_0 is the central frequency of the mother wavelet at a scale $a = 1$. When the scale factor a is changed, the duration Δt and the bandwidth Δf of the wavelet are both changed, but its shape remains the same; the CWT can be seen as a filter bank analysis composed of geometrically distributed bandpass filters with constant relative bandwidth $Q = \Delta f/f$. In this article we utilize Gaussian weighted tone (or Morlet) wavelet $h(t) = \exp\{-t^2/\Delta t_0^2\} * \exp\{2\pi i f_0 t\}$ with experimentally selected time resolution Δt_0 and f_0 set to the Nyquist frequency of the discrete time signal.

Scalogram Definition

We define the scalogram of signal $x(t)$ as the squared modulus of the CWT $S_h x(a,b) = |W_h x(a,b)|^2$. It is an energy distribution of the signal in the timescale plane.

In practice, the effective frequency band of digital seismic waveforms is limited. We may compress the discrete version of scalogram by changing the timestep from the original sampling interval to the experimentally defined translation step (time frame) Δt. We also restrict the number of central frequencies (and number of scales) to K logarithmically equally spaced values $f_{\min} \leq f_m \leq f_{\max}$, $a_m = f_{\max}/f_m$, $1 \leq m \leq K$. That leads to a discrete scalogram:

$$\begin{aligned} &DS_h(m,n) = \left| \left\langle x, \frac{1}{\sqrt{a_m}}\, h\left(\frac{t - n\Delta t}{a_m}\right) \right\rangle \right|^2, \\ &m,n \in \mathbf{Z}, \quad 0 \leq n < T/\Delta t, \quad 0 \leq m < K. \end{aligned} \tag{A2}$$

For a given n, the values $DS_h(\cdot,n)$ represent an "instantaneous energy spectrum" at the nth frame.

We use logarithmic normalization to reduce variations of absolute amplitudes in discrete scalograms:

$$\mathrm{NDS}_h(m,n) = 10 \log\left(\frac{DS_h(m,n)}{R}\right), \qquad \text{(A3)}$$

where the maximum $R = \max\limits_{m,n}(DS_h(m,n))$ is taken over all the frames and scales. Finally, we apply a simple threshold denoising to the normalized scalograms (typical value of the threshold $D = 60$ dB):

$$\mathrm{DNDS}_h(m,n) = \begin{cases} D + \mathrm{NDS}_h(m,n), & \text{if } \mathrm{NDS}_h(m,n) \geq D \\ 0, & \text{if } \mathrm{NDS}_h(m,n) < D \end{cases} \qquad \text{(A4)}$$

Steps (A2) to (A4) give us a sequence of frames each parameterized by the instantaneous spectrum $S_n = \mathrm{DNDS}_h(\cdot,n)$, and we use this structural pattern next to estimate waveform dissimilarities. An example of the structural pattern (A4) is shown in Figure A1.

Suppose that two seismic records are represented by the discrete scalograms $\mathbf{S} = S_1,\ldots,S_{M_S}$ and $\mathbf{T} = T_1,\ldots,T_{M_y}$; we call **S** a source pattern and **T** a target one. We determine a structural dissimilarity (Levenstein distance) between the records using a nonlinear alignment of their patterns, thus the durations M_S and M_T need not to be very different. The purpose of the alignment is to find a monotonic transform of timescales of the two records, which synchronizes the onsets of similar structural phenomena (such as *P*- and *S*-phase arrivals) by global minimization of accumulated sum of local spectral distortions $d(S_i, T_j)$. The result of alignment is a pair of time-warping functions $i = k(l)$, $j = m(l)$, $l = 1, \ldots, L$, satisfying the constraints of:

- endpoints

$$\begin{aligned} &k(1) = 1, \quad m(1) = 1 \\ &k(L) = M_\mathrm{S}, \quad m(L) = M_\mathrm{T} \\ &L \leq M_\mathrm{S} + M_\mathrm{T} \end{aligned} \qquad \text{(A5)}$$

- monotonicity and local continuity

$$\begin{aligned} &k(l + 1) - k(l) = 0 \text{ or } 1, \\ &m(l+1) - m(l) = 0 \text{ or } 1 \end{aligned} \qquad \text{(A6)}$$

which minimizes the accumulated distortion between the patterns

$$D_{Lev}(\mathbf{S},\mathbf{T}) = \min_{k(\cdot),m(\cdot)}\left[\sum_l d_{loc}(F_{k(l)}, G_{m(l)})\right]. \qquad \text{(A7)}$$

The sum is over the path $k(\cdot),m(\cdot)$ in the i,j-plane satisfying (A5) to (A6) (Fig. A2).

To define local spectral distortion values, consider three elementary editing operations: insertion of a frame into a pattern, deletion of a frame from a pattern, and match (or substitution) of two frames in two patterns. Then the mappings $k(\cdot),m(\cdot)$ may be interpreted as a composition of these editing operations applied to the source pattern to obtain a target one, and the accumulated distortion (A7) may be seen as a sum of weights of the editing operations involved in the composition.

To visualize the alignment results, we introduce an empty frame (gap), denoted by "null":

- If $k(l) = k(l + 1)$, then we insert the null frame into the target pattern (= delete the frame $S_{k(l)}$ from the source pattern)
- If $m(l) = m(l + 1)$, then we insert the null frame into the source pattern (= insert the frame $T_{m(l)}$ into the source pattern)
- Otherwise $k(l + 1) - k(l) = m(l + 1) - m(l) = 1$, match of instantaneous spectra (= substitute the frame $S_{k(l)}$ by $T_{m(l)}$ in the source pattern).

Interpolation of the local spectral distortion values for the null frame (weights of insertions and deletions) brings nonlinearity to the alignment in contrast to the widely used linear cross-correlation (Joswig, 1990). To make the gap null indistinguishable from a zero energy frame (with no signal), we use:

$$\begin{aligned} \text{deletion:} \quad & d_\mathrm{loc}(S_i, \text{ null}) = \|S_i\|^2 \\ \text{insertion:} \quad & d_\mathrm{loc}(\text{null}, T_j) = \|T_j\|^2 \\ \text{substitution:} \quad & d_\mathrm{loc}(S_i, T_j) = \|F_i - G_j\|^2 \end{aligned} \qquad \text{(A8)}$$

where $\|\cdot\|$ stands for Euclidean vector norm.

Extended patterns $\mathbf{S}^* = S_1^*, \ldots, S_L^*$ and $\mathbf{T}^* = T_1^*, \ldots, T_L^*$ including the null frames help to "visualize" the result of the nonlinear alignment. For example, the path in Figure A2 can be interpreted as:

$$\begin{array}{rcllllll} \mathbf{S^*} & = & S_1, & S_2, & \text{null}, & S_3, & S_4, & S_5 \\ & & \text{match} & \text{match} & \text{insert} & \text{match} & \text{delete} & \text{match} \\ \mathbf{T^*} & = & T_1, & T_2, & T_3, & T_4, & \text{null}, & T_6 \\ \\ k(l) & = & 1, & 2, & 2, & 3, & 4, & 5 \\ \text{aligned time } l & = & 1, & 2, & 3, & 4, & 5, & 6 \\ m(l) & = & 1, & 2, & 3, & 4, & 5, & 5 \end{array}$$

We search for the optimal path $k(\cdot)$, $m(\cdot)$ and accumulated distortion (A6) using a dynamic programming algorithm with the complexity $O(N^2)$, where N stands for maximum number of frames in a pattern. Because of the endpoint constraints, we can rewrite (A7) in terms of M_S and M_T as $D_\mathrm{Lev}(\mathbf{S},\mathbf{T}) = D(M_\mathrm{S},M_\mathrm{T})$. Since the local spectral distortions do not depend on the positions of frames in the pattern, the

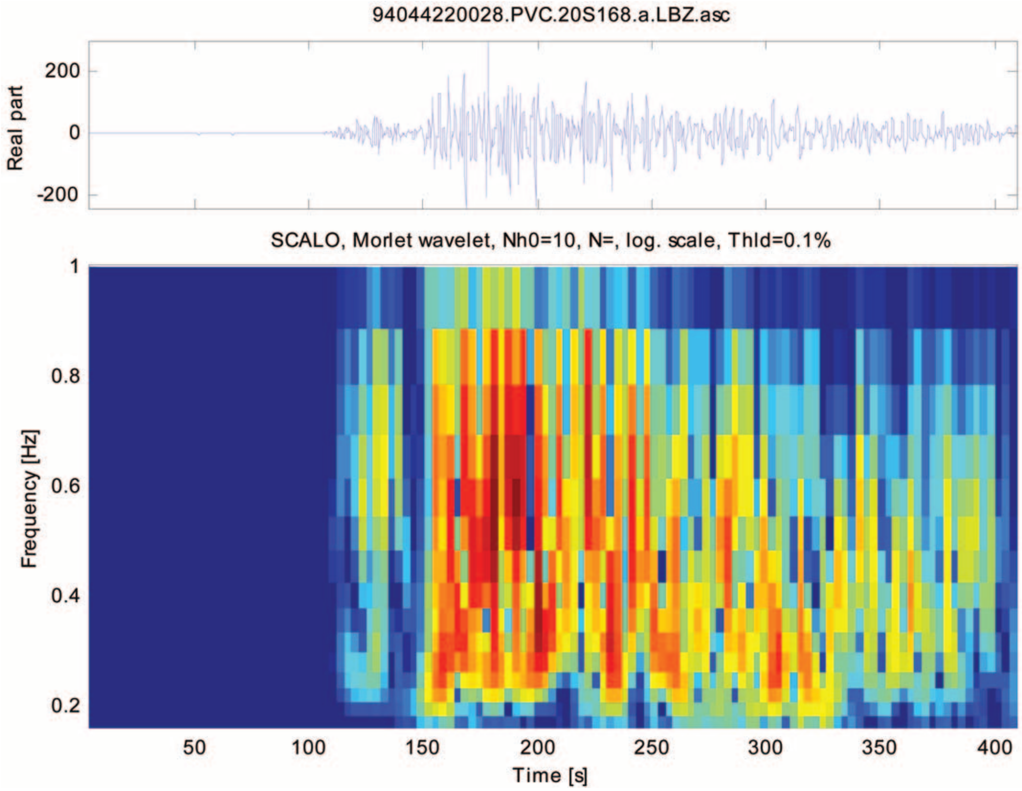

Figure A1. Example of the discrete scalogram. (top) Vertical component of a seismic record. (bottom) Time-frequency surface of normalized wavelet transform coefficients (scalogram, see text).

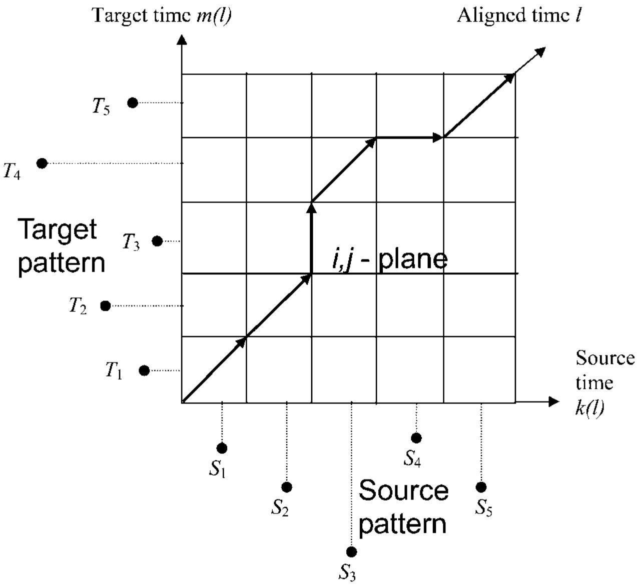

Figure A2. Optimal path for nonlinear alignment of the two patterns $\mathbf{S}^* = S_1^*, \ldots, S_L^*$ and $\mathbf{T}^* = T_1^*, \ldots, T_L^*$. Vertical arrows are used for insertions, diagonal arrows are for substitutions, and horizontal arrows are for deletions.

minimal partial accumulated distance along a path $k(\cdot)$, $m(\cdot)$ connecting (1,1) and (i,j) is

$$D(i,j) = \min\begin{cases} D(i-1,j) + d_{\text{loc}}(S_i,\text{null}) \\ D(i-1,j-1) + d_{\text{loc}}(S_i,T_j) \\ D(i,j-1) + d_{\text{loc}}(\text{null},T_j). \end{cases} \quad \text{(A9)}$$

Starting from $D(1,1) = d_{\text{loc}}(S_1,T_1)$ and recursively filling the $M_S \times M_T$ matrix $D(i,j)$ using (A7), the algorithm stops at the sought value $D_{\text{Lev}}(\mathbf{S}, \mathbf{T}) = D(M_S,M_T)$.

Institute of Physics of the Earth
Russian Academy of Sciences
119991 Moscow, Russia
(M.N.Z., A.D.G., A.B.)

Institute of Physics of the Earth
Louis Pasteur University
67084 Strasbourg, France
(D.R., J.B.)

Manuscript received 11 August 2005.