Ⓔ

An Algorithm for Detecting Clipped Waveforms and Suggested Correction Procedures

Wenzheng Yang and Yehuda Ben-Zion
University of Southern California

Online material: An algorithm for detecting clipped waveforms and basic performance results.

INTRODUCTION

Waveform clipping due to instrumental limitation is a common problem in seismic data recorded by local and regional networks. With high-amplitude data missing, clipped waveforms, which usually have short epicentral distances and high signal-to-noise ratios, have to be removed from the estimation of magnitudes and other earthquake properties. An efficient automatic algorithm for detecting clipped waveforms can help eliminate artifacts associated with analysis of clipped seismograms. In addition, a reliable algorithm for correcting some clipped waveforms can increase the availability of data closer to the source locations and data for larger magnitude events.

The severity of waveform clipping at a given station is expected to increase with the earthquake size. Since the frequency of earthquakes generally decreases with size, following the Gutenberg-Richter relation, a large number of clipped waveforms are likely to have only a small number of clipped points. In such cases, it may be possible to provide reasonably accurate corrections for the clipped waveforms.

Interpolation methods are often used to reconstruct missing data. By using a spline interpolation method, Karabulut and Bouchon (2007) corrected the peak values of clipped waveforms recorded by strong motion accelerometers during the 1999 *Mw* 7.1 Düzce earthquake in Turkey. Subsets of earthquakes recorded by local or regional networks often have very similar waveforms with high cross-correlation coefficient (CCC) values (*e.g.*, Aster and Scott 1993; Nadeau *et al.* 1994). Another possible way of correcting clipped waveforms may be by scaling up portions of unclipped similar waveforms with sufficiently high CCC values.

Here we analyze statistical features of clipped waveforms in a large data set of 26,080 aftershocks, recorded with sampling frequency of 100 Hz by a ten-station near-fault seismic network in the six months following the 1999 *Mw*7.4, İzmit mainshock (Figure 1). Based on the response of instruments to amplitude saturation, clipped waveforms can generally be classified into two types: Flat-Top (FT), where the values of clipped points are set to be the upper limit value; and Back-to-Zero (BZ), where the values of clipped points are set to zero. In our data set, the clipped waveforms are of the BZ type.

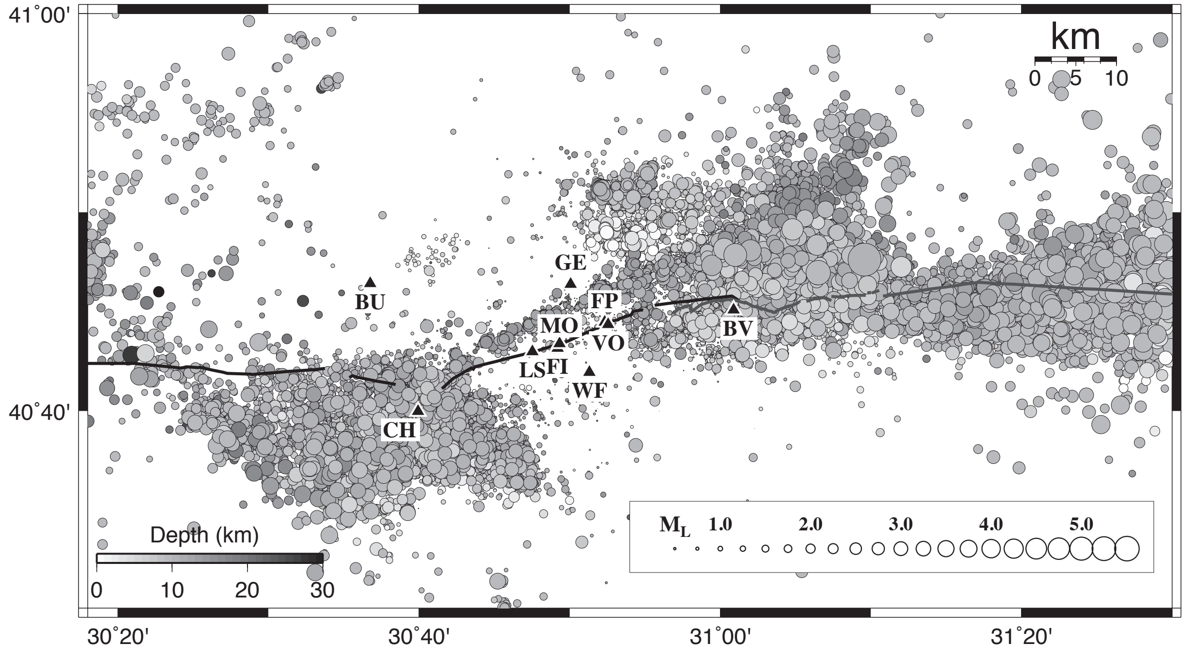
We propose an algorithm for automatic detection of clipped waveforms and discuss the possibility of correcting clipped seismograms. The detection algorithm is based on the observed ranges of amplitudes in the recorded seismograms, expected frequency-size statistics of amplitudes associated with the Gutenberg-Richter distribution, and properties of neighboring points in seismograms. In our data set, about 3.6% of the recorded waveforms are found to be clipped. As expected, the horizontal components have more detected clipped waveforms than the vertical one, and the intensity of waveform clipping increases with proximity to the fault and amplitude of site effects.

We consider waveform corrections based on a linear interpolation with the Kriging method and using unclipped similar waveforms of smaller events. We compare the performance of these two methods using artificially clipped waveforms within a 10-second time window. With the sampling rate and other properties of our data, the results may be summarized as follows. In cases with one–two consecutive clipped points, the interpolation method generally produces corrections with smaller errors. In cases with three, four, or five consecutive clipped points, the similar waveform method performs better if similar waveforms with CCC values larger than, respectively, 0.96, 0.91, and 0.88 are available. In cases with six or more consecutive clipped samples, the similar waveform method generally performs better. However, corrections using similar waveforms with CCC < 0.85 are likely to produce overly large errors for accurate magnitude determinations.

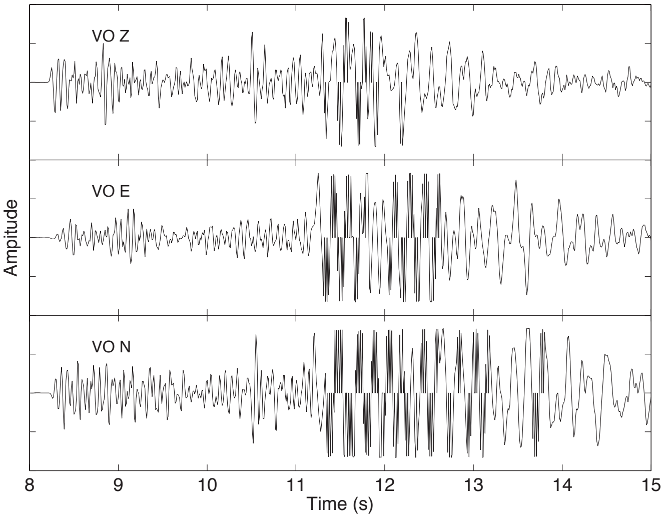
DATA

One week after the occurrence of the 1999 *Mw*7.4 İzmit earthquake, a PASSCAL seismic network with 10 short-period seismic stations was deployed along the Karadere-Düzce branch of the North Anatolian fault (NAF). All stations had REFTEK recorders and three-component L22 velocity sensors with a sampling frequency of 100 Hz (Seeber *et al.* 2000; Ben-Zion *et al.* 2003). This network operated for six months and recorded 26,080 events with magnitudes between about 0 and 7.2 and hypocentral distances less than 100 km (Figure 1).

doi: 10.1785/gssrl.81.1.53



▲ **Figure 1.** Epicentral distribution of ~ 26,000 earthquakes along the Karadere-Düzce branch of the North Anatolian fault. The earthquake symbols indicate depth (gray scale) and magnitude (symbol size). Stations are marked by triangles with adjacent names. The black and gray curves are surface ruptures associated with the İzmit and Düzce mainshocks, respectively.



▲ **Figure 2.** Three-component clipped waveforms recorded by the fault zone station VO. The amplitudes of clipped samples were set to be zero by the instrument.

Waveforms recorded by the various instruments and components were clipped to different degrees. Figure 2 shows an example set of three-component clipped waveforms recorded by station VO located within the İzmit rupture zone and generated by an earthquake with $M_L = 3.3$.

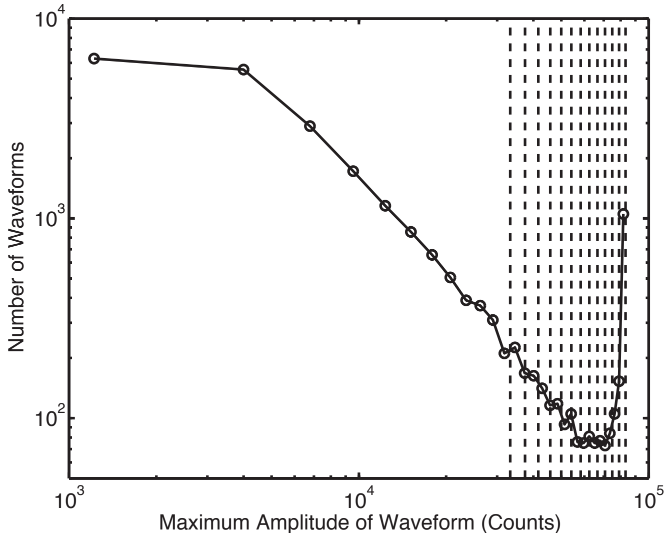
FEATURES OF CLIPPED WAVEFORMS

By visually inspecting clipped waveforms in the data set, we can summarize the features of clipped waveforms as follows: 1) Clipped waveforms contain points with amplitude close to the upper range of the recorded data. 2) Clipped points are in waveform portions with amplitude close to the peak value in the seismogram. 3) Clipped points are located within waveform sections having large amplitude fluctuations. 4) All clipped data points have zero amplitude values. 5) Data on both sides of a clipped point have the same sign or one neighboring point is also zero. We define "the upper range of the recorded data" in feature 1 to be the Observed Range of the relevant instrument.

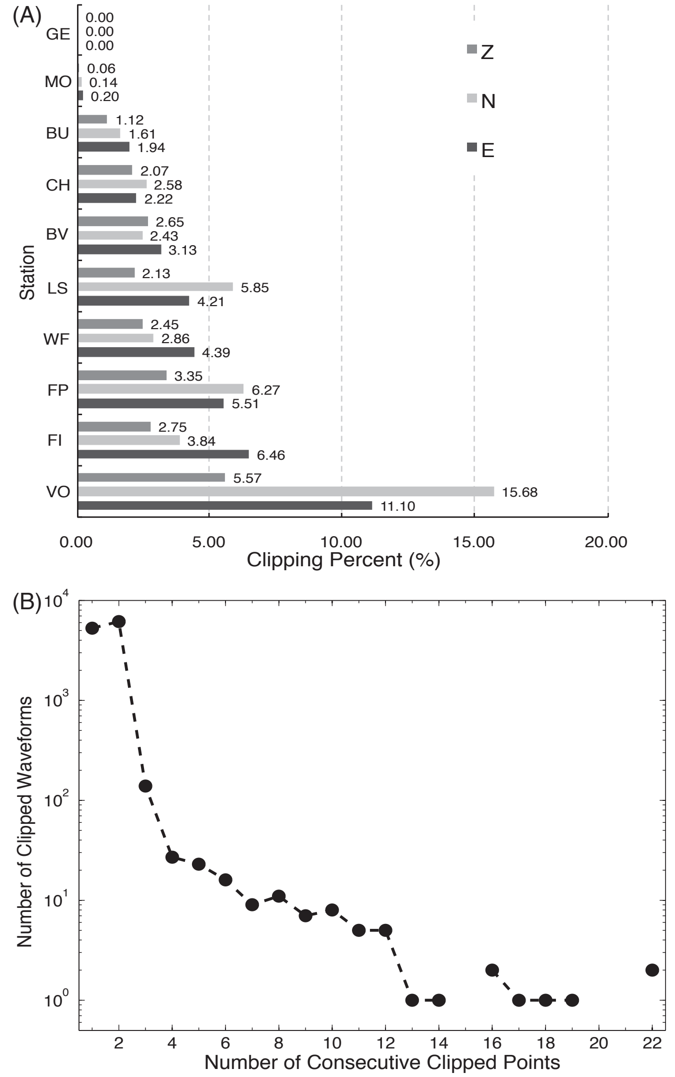
Features 1 and 2 are intuitive, and feature 3 further limits the location of clipped sample points inside the waveform. Features 4 and 5 represent the BZ type of clipped waveforms. If there are two or more consecutive clipped points, at least one side of the clipped point has zero amplitude. It is difficult to judge whether a zero value sample with neighboring samples of different signs is clipped or not. However, the high sampling frequency (100 Hz) in our data set could rule out most such cases.

AN ALGORITHM FOR DETECTING CLIPPED WAVEFORMS

Based on the features of clipped waveforms listed above, we propose an algorithm for automatic detection of clipped waveforms with the following steps: 1) Treat waveforms with maximum amplitudes larger than a given threshold (*e.g.*, 60% of the Observed Range) as potentially clipped waveforms. 2) For each potential clipped waveform, assume that points with zero amplitude, between the first and last samples with absolute values larger than 0.5 of the waveform peak amplitude, are tentatively clipped. 3) Require that clipped samples are bounded



▲ **Figure 3.** The distribution of maximum amplitude within 24,135 waveforms recorded by station VO on the vertical component. The vertical dashed lines represent 98%, 95%, 90%, 85%, …, 45%, 40% of the Observed Range.



▲ **Figure 4.** A) Percent values of detected clipped waveforms from the total recorded waveforms by the different stations shown in Figure 1. B). Probability density of clipped waveforms on all stations according to the maximum numbers of consecutive clipped points on each clipped waveform.

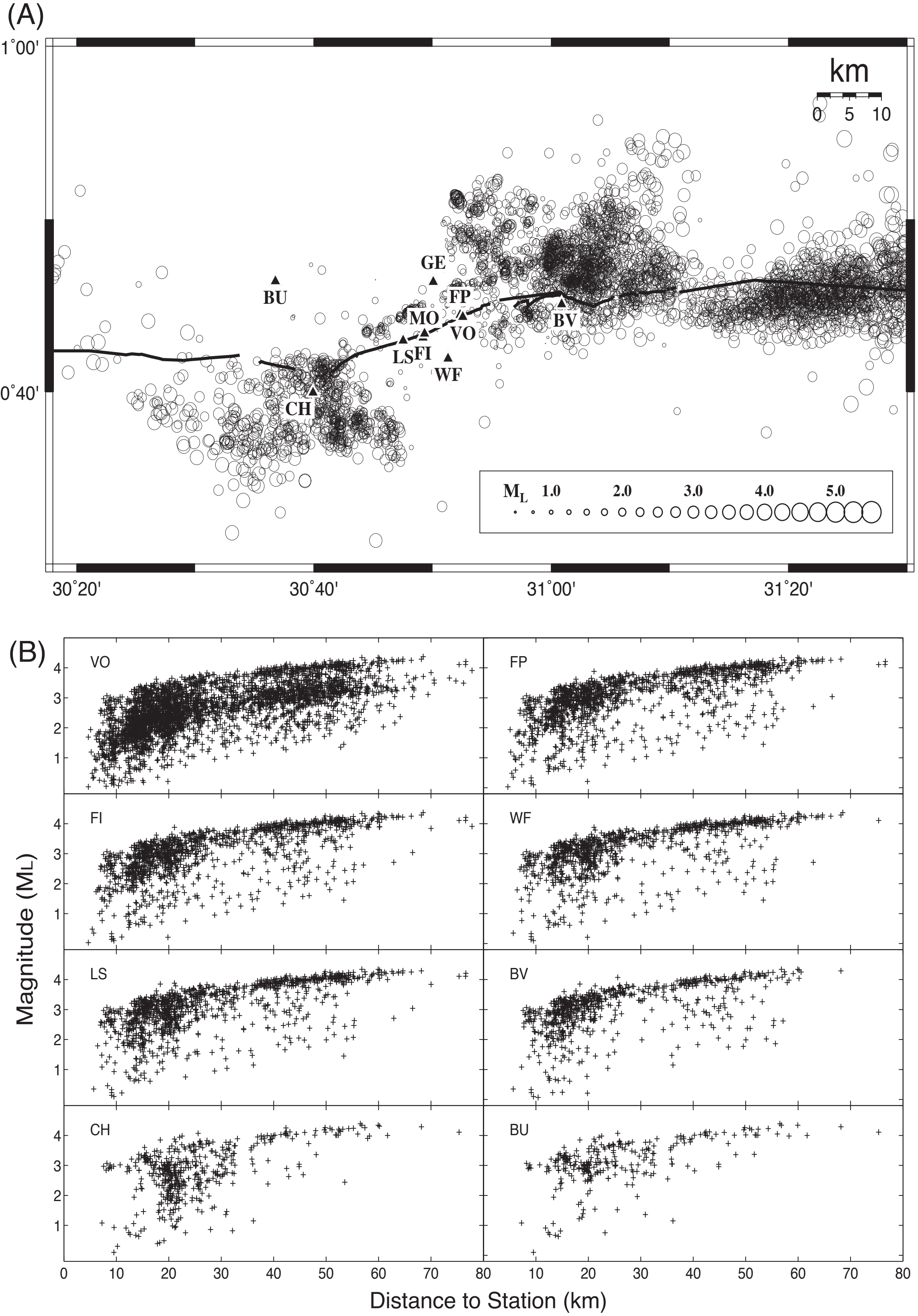
on both sides by points with the same sign or another zero. 4) Require that the maximum or minimum value of ± 10 samples on either side of the candidate clipped points is larger than 0.8 of the waveform peak amplitude.

Figure 3 presents the statistics of the recorded maximum amplitudes on the vertical component of fault zone station VO. In agreement with the Gutenberg-Richter statistics, the number of observed waveforms generally decreases with increasing maximum amplitude. Because of clipping, however, the population increases dramatically beyond a certain amplitude (*e.g.*, 6×10^4 in Figure 3). A reasonable choice for a clipping threshold is somewhat below the end of the power-law range in the frequency-size statistics of the recorded maximum amplitudes.

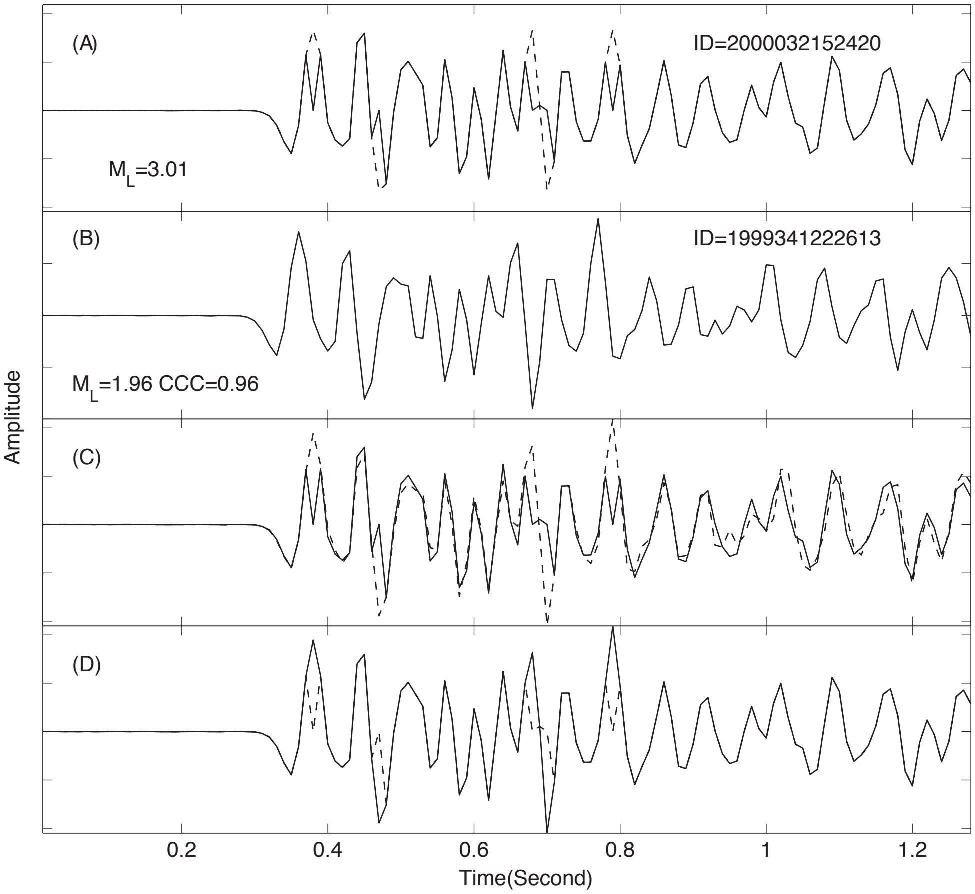
To implement the clipped waveform detection algorithm, we choose a series of thresholds from 98% to 40% of the Observed Range on a given component of a given station (dashed lines in Figure 3). If the algorithm works properly, the numbers of detected clipped waveforms will first increase and then become saturated with the decrease of the assumed threshold. Figure S1 in the electronic supplement demonstrates that such a pattern is obtained for waveforms recorded by the vertical component of station VO. Application of this algorithm to the data recorded by all stations indicates that as the threshold drops below approximately 65%, the numbers of detected clipped waveforms saturate for all stations and instrument components. Table S1 in the electronic supplement lists the main statistical features of the recorded data and detected clipped waveforms. The algorithm and parameters used for automatic detection of clipped waveforms are also provided in the electronic supplement.

Figure 4A summarizes the main results of applying the detection algorithm to our data set. The intensity of clipping generally decreases with distance from the fault, and the only station that apparently did not record any clipped waveforms is station GE. In the other stations, the horizontal components generally have more clipped waveforms than the vertical component, although all components have similar Observed Ranges at each station. One approach to quantitatively classify the severity of clipping is to count the number of consecutive clipped points on a waveform. Figure 4B displays the number of clipped waveforms with different numbers of consecutive clipped points among all detected clipped waveforms. The dominant cases of clipped waveforms are seen to be associated with single and two consecutive clipped points (45% and 53% of all clipped waveforms, respectively). There are 1.2% cases of clipped waveforms with three consecutive clipped samples, and less than 0.8% cases with four or more consecutive clipped samples.

Among the 26,080 aftershocks, 4,177 generated at least one clipped waveform (Figure 5A). The total number of



▲ **Figure 5.** A) Distribution of ~ 4,200 events with detected clipped waveforms. The size of each event scales with magnitude as shown by the legend. B) The magnitude and distance relations of detected clipped waveforms on the vertical components of the various stations (denoted by two letters at the top-left corners). The number of detected clipped waveforms decreases following the order: VO, FP, FI, WF, LS, BV, CH, and BU.



▲ **Figure 6.** Example of correcting clipped waveforms using the similar waveform method. A) Replace the amplitudes of clipped points with the maximum observed value. B) Find a similar unclipped waveform with the highest CCC (here CCC = 0.96). C) Project the similar unclipped waveform (dashed) onto the clipped waveform (solid). D) A comparison of the clipped (dashed) and corrected (solid) seismograms.

detected clipped waveforms is 22,472 among all 617,526 waveforms. Yang *et al.* (2009) found that the observed waveform spectral energy associated with site effects of eight of the stations used decreases following the order: VO, FP, FI, WF, LS, BV, CH, and BU. Figure 5B shows relations between event magnitudes and hypocentral distances producing detected clipped waveforms at each of these eight stations. A comparison of the results with the amplitude of the site spectra at the various stations (Yang *et al.* 2009) illustrates, as expected, that the intensity of waveform clipping is strongly correlated with the amplitude of the local site effects.

THE KRIGING AND SIMILAR WAVEFORM CORRECTION METHODS

In this section we discuss two methods that could be used to correct less severely clipped waveforms. Interpolation is commonly used to replace missing data points (*e.g.*, Kokaram *et al.* 1995; Karabulut and Bouchon 2007). One of the most used interpolation techniques in geo-statistics is the Kriging method. Below we implement the Kriging interpolation using the DACE package (Sacks *et al.* 1989), with a Gaussian correlation model, a zero-order polynomial regression model, and the following parameters: $\theta = 5$, lob = 0.1, and upb = 10. More detailed information on the employed Kriging interpolation is documented in the DACE manual (http://www2.imm.dtu.dk/~hbn/dace).

Various studies demonstrate that in many cases seismicity contains subsets of events with very similar locations and focal mechanisms. These events, referred to as repeating earthquakes, generate very similar waveforms with high CCC values, which in some cases match almost wiggle by wiggle (*e.g.*, Poupinet *et al.* 1984; Nadeau *et al.* 1994; Schaff *et al.* 1998; Peng and Ben-Zion 2006). In cases where sufficiently similar unclipped waveforms exist, scaled-up portions of similar unclipped waveforms may be used to replace missing samples in clipped waveforms.

The correction of clipped points with the similar waveforms method is done with the following steps:

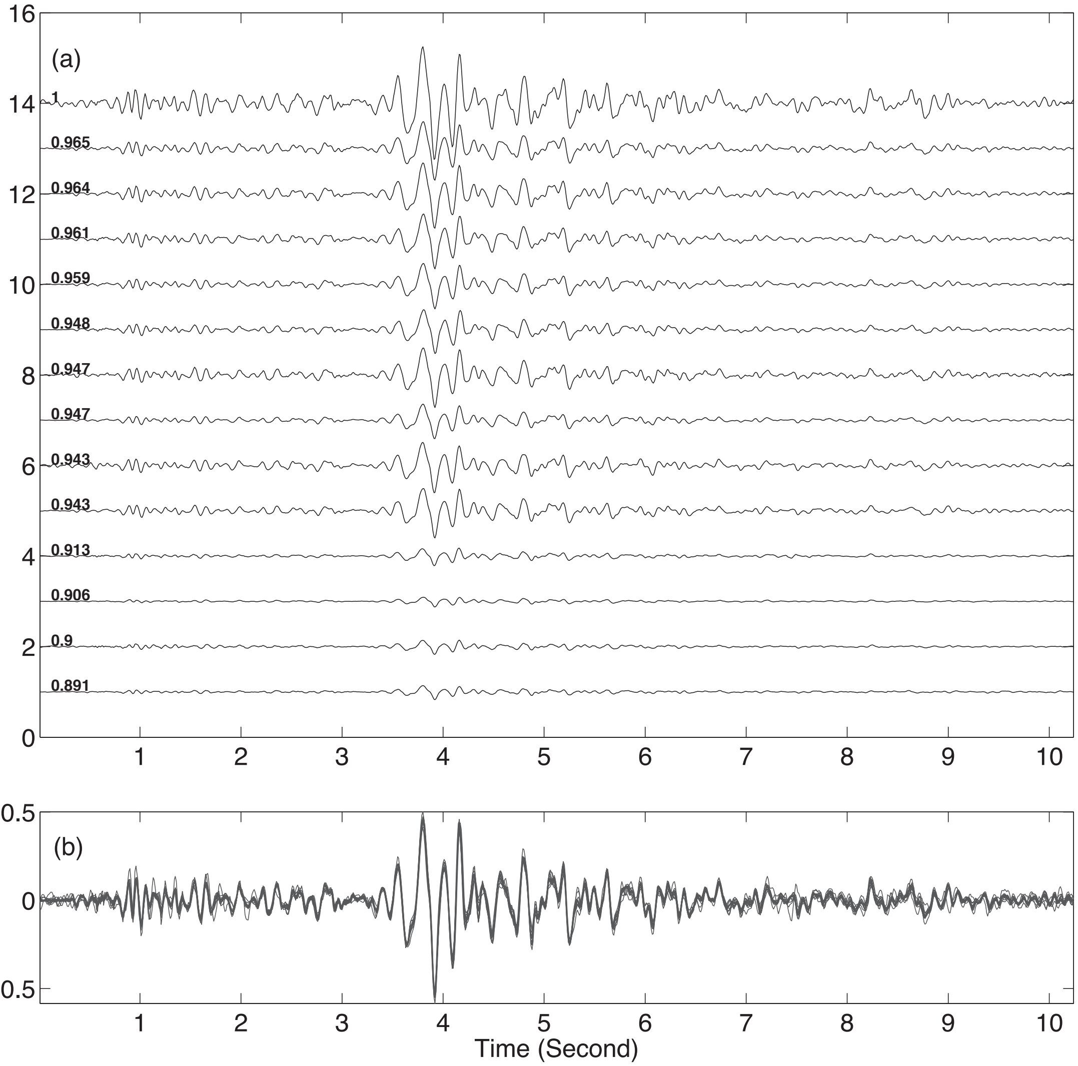
1. For every clipped point of the BZ type in each clipped waveform, we first replace its amplitude with the maximum observed value for that instrument component and use the same sign as that of one of its neighboring unclipped points with larger absolute amplitude (Figure 6A). We then search for similar unclipped waveforms in the data set and use in the next step the unclipped waveform with the highest CCC value (Figure 6B).

2. The similar unclipped waveform is aligned with the clipped waveform and scaled-up to match the amplitudes of the data points with the clipped waveform (Figure 6C). The employed scaling-up factor is the median value of the amplitude ratios between all the unclipped points of the clipped seismogram and the corresponding points of the unclipped similar waveform. The values of clipped points are then replaced with the scaled-up versions of the corresponding points on the unclipped similar waveform (Figure 6D).

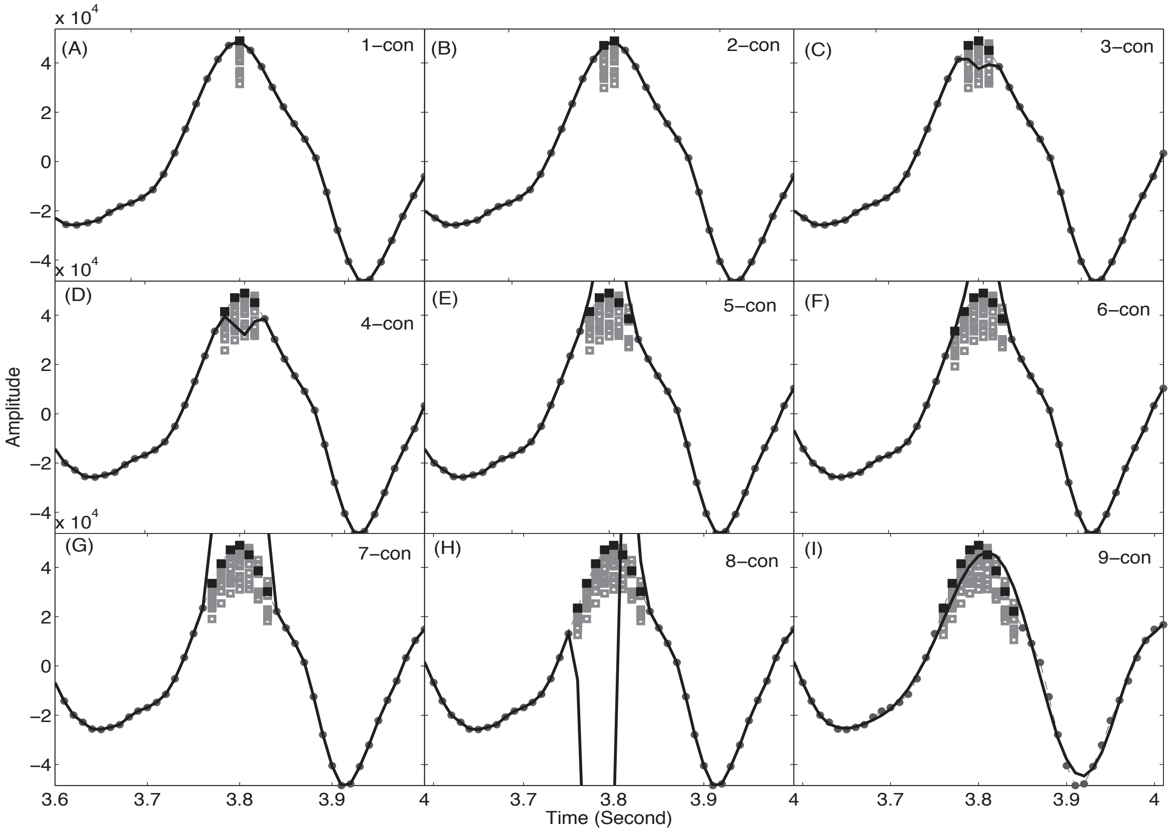
A COMPARISON OF TWO CORRECTION METHODS

We compare the performance of the two correction methods using artificially clipped waveforms associated with cases of one–nine consecutive clipped samples (Figure 4B), which represent 99.9% of all clipped waveforms detected in our data set. To prepare artificially clipped waveforms, we use unclipped waveforms and clip samples with the largest amplitude and in some cases also one–eight neighboring points. We use earthquake clusters with high CCC values identified by Peng and Ben-Zion (2005, 2006). The systematic tests employ 10-second waveforms with all the main seismic phases, recorded on the horizontal component that usually contains the largest amplitudes. Similar waveforms of one cluster recorded by the fault zone station VO on the EW component are displayed in Figure 7. Among all unclipped similar waveforms, we select the waveform with the largest amplitude to be the target waveform. On the selected target waveform, we artificially clip a number of points with amplitude larger than a given threshold.

Figure 8 shows a comparison of the two methods using example artificially clipped waveforms with one–nine consecutive clipped points. With more consecutive points clipped, additional information associated with the waveform structure is lost, the interpolated data becomes further removed from the original amplitudes of the artificially clipped samples, and the corrected interpolated results become less accurate. With the similar waveforms method, the CCC values between a clipped



▲ **Figure 7.** A) A selected large amplitude waveform (top trace) and its similar waveforms arranged following the CCC values above traces. B) A stacking of the normalized similar waveforms in A.



▲ **Figure 8.** Corrections with the two methods for cases of one–nine consecutive clipped points. The solid circles are original unclipped samples. The dark squares are artificially clipped samples. The curves show the correction with the Kriging method for each case. The light squares are corrections with the similar waveform method.

waveform and its similar waveforms are nearly the same for cases with a different number of consecutive clipped points.

To compare cases with a single clipped sample, we consider the sample with the largest amplitude on the selected unclipped waveform to be a clipped sample (solid dark square in Figure 8A). We calculate CCC values between the artificially clipped waveform and each of its similar waveforms in the 10-second time window excluding the clipped sample and then correct the amplitude of the clipped sample from similar waveforms with different CCC values (gray squares in Figures 8A–I). We also apply the Kriging method to interpolate for the amplitude of the clipped sample using data with 17 samples at each side adjacent to the clipped part (curve in Figures 8A–I). To compare cases with two consecutive clipped samples, we consider the sample with largest amplitude and one of its neighboring points that has larger amplitude to be clipped samples (solid dark squares in Figure 8B). For three and more consecutive clipped samples, we consider the sample with the largest amplitude and its neighboring samples as clipped samples (solid dark squares in Figures 8C–I). In all cases the amplitudes of the artificially clipped samples are larger than the amplitudes of unclipped samples along the waveform.

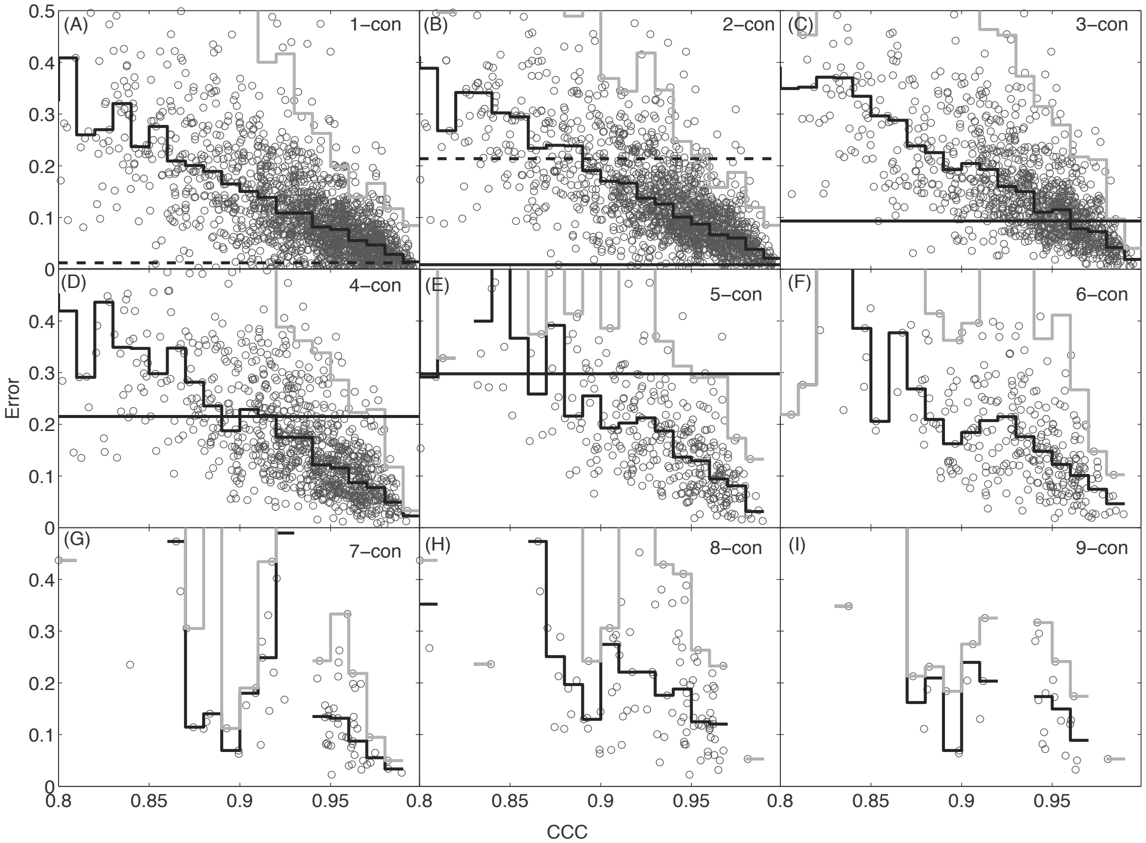
We define the error associated with the corrections as:

$$err = \max(\left| \log_{10} |A_{\mathrm{org}}| - \log_{10} |A_{\mathrm{cor}}| \right|) \ , \tag{1}$$

where A_{org} is the original amplitude of the selected clipped point and A_{cor} is the corrected amplitude. For cases with two or more consecutive clipped points, the error in Equation 1 is defined as the largest correction error, since such error is associated with the peak position along the consecutive clipped samples. The employed log scale in Equation 1 produces a correspondence between the correction errors and associated uncertainties on the local Richter magnitude scale.

We apply both methods to correct a group of 1,600 artificially clipped waveforms with one–nine consecutive clipped points and calculate the corresponding correction errors using Equation 1. We compare both methods using the median values of the correction errors in each case. The median values of the errors associated with the similar waveform method are calculated in every 0.01 CCC bin (stair-curves in Figure 9). For each method, we also calculate the 97.5% confidence level for reference.

For cases of one–two consecutive clipped samples, the Kriging method always performs better than the similar waveform method. For three clipped samples, the similar waveform method performs better with CCC values larger than 0.96. For four consecutive clipped samples, the similar waveform method performs better with CCC values larger than 0.91. For five consecutive clipped samples, the similar waveform method performs better with CCC values larger than 0.88. This also holds for cases of six or more consecutive clipped



▲ **Figure 9.** A comparison between errors of the two correction methods for cases with different numbers of consecutive clipped points (name at the top-left corner). In each panel, the dots are the correction errors of 1,600 waveforms using their similar waveforms. The black and gray stair-curves are the median values and 97.5% confidence levels of the correction errors for the similar waveform method in every 0.01 CCC value interval. The horizontal solid and dashed lines are, respectively, the median value and 97.5% confidence level of the correction errors for the Kriging method. For cases of six–nine consecutive clipped points, the median correction errors with the Kriging method are larger than 0.5.

samples, although the amount of available data for such tests is considerably smaller. Table 1 and Figure 10 summarize the comparison results extracted from Figure 9. The dark line in Figure 10 separates regions in parameter-space, spanned by the number of consecutive clipped points and CCC values of similar unclipped waveforms, where the different methods produce better correction results. The errors associated with corrections are denoted by a gray scale and the unhatched regions produce errors less than 0.3, which would lead to reasonably small errors in magnitude derivation.

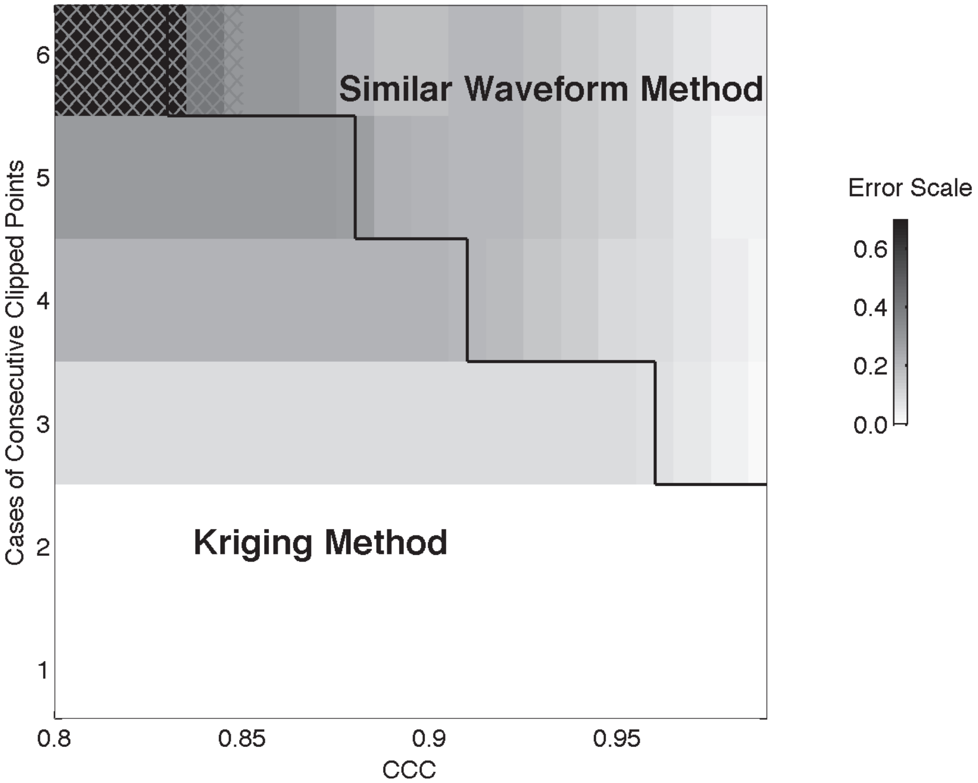
TABLE 1
A Statistical Summary of Correction Errors for Methods Comparison

	Kriging Method		Similar Method (with smaller median value)
Cases	Median	97.5% Confidence Level	
Single	0.0007	0.012	NA
2-Consecutive	0.009	0.2	NA
3-Consecutive	0.09	1.0	CCC > 0.96
4-Consecutive	0.21	1.6	CCC > 0.91
5-Consecutive	0.29	1.7	CCC > 0.88
6-Consecutive	0.7	2.2	CCC > 0.83

DISCUSSION AND CONCLUSIONS

We designed and implemented an automatic detection algorithm to identify clipped waveforms in a near-fault data set generated by 26,080 earthquakes, with amplitudes of clipped points set to zero instrumentally. Applications of the algorithm to other types of instrumental clipping (*e.g.*, Flat-Top) will require some adjustments, but the general statistical features of clipped waveforms should remain the same. As expected, the algorithm indicates that clipping on the horizontal components is more prevalent than on the vertical one, and that the intensity of clipping increases with increasing proximity to the fault and larger local site effects.

To increase the range of available data it may be useful to correct some clipped waveforms. Toward this end, we compare two possible correction methods associated with the Kriging interpolation and using scaled-up versions of similar unclipped



▲ **Figure 10.** A comparison of the performance of the two correction methods. The black line separates regions where the different methods produce smaller errors (gray scale). The hatched region in the top-left corner has correction errors larger than 0.3.

waveforms. The availability of the latter generally increases with increasing geometrical simplicity of the fault (*e.g.*, Ben-Zion and Sammis 2003). As examples, along the Parkfield section of the San Andreas fault more than 50% of the ongoing seismicity belongs to clusters of repeating events that have CCC values larger than 0.9 (*e.g.*, Nadeau *et al.* 1994). In contrast, along the San Jacinto fault in southern California only about 2% of the seismicity belongs to clusters of repeating events with CCC > 0.9 (Aster and Scott 1993). For the seismicity along the Karadere-Düzce branches of the North Anatolian fault used in this study, about 18% of the seismicity belongs to clusters of repeating events with CCC > 0.9 (Peng and Ben-Zion 2005, 2006).

Typical median value of magnitude uncertainty in a high-quality seismic catalog (*e.g.*, the one based on the Northern California Seismic Network) is about 0.3 (Werner and Sornette 2008). Since the correction error associated with Equation 1 is based on the log amplitude, errors less than about 0.3 should produce errors in magnitude estimates that are within the typical range. Our synthetic tests based on data with 100-Hz sampling frequency and one–nine consecutive clipped samples can be summarized as follows. For clipped waveforms with one–two clipped samples, the Kriging method performs better. For clipped waveforms with three–five consecutive clipped samples, the Kriging method produces corrections with smaller errors unless unclipped similar waveforms with high CCC values (0.88–0.96) are available. For clipped waveforms with six or more consecutive clipped samples, the similar waveform method performs better, but unclipped waveforms with CCC > 0.85 are needed to produce errors smaller than 0.3. The CCC values separating cases where the different methods perform better, and the calculated errors, will change somewhat for data with different frequency content and sampling rates.

ACKNOWLEDGMENTS

We thank Richard Aster and Zhigang Peng for useful discussions. The manuscript benefited from useful comments by an anonymous reviewer and *SRL* Editor Luciana Astiz. The work was supported by the National Science Foundation (grant EAR-0823695).

REFERENCES

Aster, R. C., and J. Scott (1993). Comprehensive characterization of waveform similarity in microearthquake data sets. *Bulletin of the Seismological Society of America* **83**, 1,307–1,314.

Ben-Zion, Y., J. G. Armbruster, N. Ozer, A. J. Micheal, S. Baris, M. Aktar, Z. Peng, D. Okaya, and L. Seeber (2003). A shallow fault-zone structure illuminated by trapped waves in the Karadere-Düzce branch of the North Anatolian fault, western Turkey. *Geophysical Journal International* **152**, 699–717.

Ben-Zion, Y., and C. G. Sammis (2003). Characterization of fault zones. *Pure and Applied Geophysics* **160**, 677–715.

Karabulut, H., and M. Bouchon (2007). Spatial variability and nonlinearity of strong ground motion near a fault. *Geophysical Journal International* **170**, 262–274.

Kokaram, A. C., R. D. Morris, W. J. Fitzgerald, and P. J. W. Rayner (1995). Interpolation of missing data in image sequences. *IEEE Transactions on Image Processing* **4**, 1,509–1,519.

Nadeau, R. M., M. Antolik, P. Johnson, W. Foxall, and T. V. McEvilly (1994). Seismological studies at Parkfield III: Microearthquake clusters in the study of fault-zone dynamics. *Bulletin of the Seismological Society of America* **84**, 247–263.

Peng, Z., and Y. Ben-Zion (2005). Spatio-temporal variations of crustal anisotropy from similar events in aftershocks of the 1999 **M** 7.4 İzmit and **M** 7.1 Düzce, Turkey, earthquake sequences. *Geophysical Journal International* **160**, 1,027–1,043.

Peng, Z., and Y. Ben-Zion (2006). Temporal changes of shallow seismic velocity around the Karadere-Düzce branch of the North Anatolian fault and strong ground motion. *Pure and Applied Geophysics* **163**, 567–599.

Poupinet, G., W. L. Ellsworth, and J. Frechet (1984). Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. *Journal of Geophysical Research* **89**, 5,719–5,731.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical Science* **4,** 409–423.

Schaff, D. P., G. C. Beroza, and B. E. Shaw (1998). Postseismic response of repeating aftershocks. *Geophysical Research Letters* **25**, 4,549–4,552.

Seeber, L., J. G. Armbruster, N. Ozer, M. Aktar, S. Baris, D. Okaya, Y. Ben-Zion, and N. Field (2000). The 1999 earthquake sequence along the North Anatolian transform at the juncture between the two main ruptures. In *The 1999 İzmit and Düzce Earthquakes: Preliminary Results,* ed. A. Barka, O. Kozaci, S. Akyuz, and S. Altunel, 209–223. Istanbul: Istanbul Technical University.

Yang, W., Z. Peng, and Y. Ben-Zion (2009). Variations of strain drops in aftershocks of the 1999 İzmit and Düzce earthquakes along the Karadere-Düzce branch of the North Anatolian fault. *Geophysical Journal International* **177**, 235–246.

Werner, M. J., and D. Sornette (2008). Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments. *Journal of Geophysical Research* **113**, B08302; doi:10.1029/2007JB005427.

Department of Earth Sciences
University of Southern California
Los Angeles, California 90089-0740 U.S.A.
wenzheny@usc.edu
(W. Y.)
benzion@usc.edu
(Y. B.-Z.)