Geophysical Journal International

Geophys. J. Int. (2015) **202,** 1164–1179 doi: 10.1093/gji/ggv218
GJI Seismology

A non-parametric method for automatic determination of *P*-wave and *S*-wave arrival times: application to local micro earthquakes

Christopher Rawles* and Clifford Thurber

Department of Geoscience, University of Wisconsin-Madison, 1215 *W. Dayton St., Madison, WI* 53706*, USA. E-mail: clifft@geology.wisc.edu*

Accepted 2015 May 26. Received 2015 May 23; in original form 2014 November 22

SUMMARY

We present a simple, fast, and robust method for automatic detection of *P*- and *S*-wave arrivals using a nearest neighbours-based approach. The nearest neighbour algorithm is one of the most popular time-series classification methods in the data mining community and has been applied to time-series problems in many different domains. Specifically, our method is based on the non-parametric time-series classification method developed by Nikolov. Instead of building a model by estimating parameters from the data, the method uses the data itself to define the model. Potential phase arrivals are identified based on their similarity to a set of reference data consisting of positive and negative sets, where the positive set contains examples of analyst identified *P*- or *S*-wave onsets and the negative set contains examples that do not contain *P* waves or *S* waves. Similarity is defined as the square of the Euclidean distance between vectors representing the scaled absolute values of the amplitudes of the observed signal and a given reference example in time windows of the same length. For both *P* waves and *S* waves, a single pass is done through the bandpassed data, producing a score function defined as the ratio of the sum of similarity to positive examples over the sum of similarity to negative examples for each window. A phase arrival is chosen as the centre position of the window that maximizes the score function. The method is tested on two local earthquake data sets, consisting of 98 known events from the Parkfield region in central California and 32 known events from the Alpine Fault region on the South Island of New Zealand. For *P*-wave picks, using a reference set containing two picks from the Parkfield data set, 98 per cent of Parkfield and 94 per cent of Alpine Fault picks are determined within 0.1 s of the analyst pick. For *S*-wave picks, 94 per cent and 91 per cent of picks are determined within 0.2 s of the analyst picks for the Parkfield and Alpine Fault data set, respectively. For the Parkfield data set, our method picks 3520 *P*-wave picks and 3577 *S*-wave picks out of 4232 station–event pairs. For the Alpine Fault data set, the method picks 282 *P*-wave picks and 311 *S*-wave picks out of a total of 344 station–event pairs. For our testing, we note that the vast majority of station–event pairs have analyst picks, although some analyst picks are excluded based on an accuracy assessment. Finally, our tests suggest that the method is portable, allowing the use of a reference set from one region on data from a different region using relatively few reference picks.

Key words: Time-series analysis; Body waves; Computational seismology.

INTRODUCTION

Correct identification of phase arrival times is essential for seismic event location and identification, source mechanism analysis and seismic tomography. Phase picking is often done manually. Identification of the first *P*-wave arrival is often relatively straightforward, however identification of later arrivals such as the *S* arrival is complicated by noise from the *P*-wave coda, and earlier arriving *S*-to-*P* converted waves can often be misidentified as the *S* arrival.

Beginning with Allen (1978, 1982), there are many published autopickers for *P* waves, but relatively few for *S* waves. Examples of the types of methods used for *S*-wave pickers include higher-order statistics such as kurtosis (Savvaidis *et al.* 2002; Baillard *et al.* 2014), neural networks or trees (Wang & Teng 1997; Zhao & Takano 1999; Gentili & Michelini 2006), autoregressive modelling (Takanami & Kitagawa 1993; Leonard & Kennett 1999; Küperkoch *et al.* 2010), wavelets (Anant & Dowla 1997) and combinations of

*Now at: Pivotal, 875 Howard Street, San Francisco, CA 94103, USA.

Table 1. Summary of *S* autopicker performance from recent studies.

Study	Data set	1σ	Number of picks
Patanè *et al.*			
	Mt Etna Volcano	0.3 s	129
Kuperkoch *et al.*			
	Class 0	0.12 s	23
	Class 0–1	0.62 s	154
	Class 0–2	0.70 s	804
Diehl *et al.*			
	Class 0	0.07 s	73
	Class 1	0.15 s	82
	Class 2	0.31 s	139
Baillard *et al.*			
	Vanuatu	0.23 s	390
	Mid-Atlantic Ridge	0.08 s	1809

techniques (Sleeman & van Eck 1999; Patanè *et al.* 2003; Diehl *et al.* 2009). Many of these studies simultaneously pick *P*-wave arrivals using similar techniques combined with polarization analysis to discern specific phases (Cichowicz 1993; Reading *et al.* 2001; Baillard *et al.* 2014; Ross & Ben-Zion 2014). Only a few of these studies provide readily interpretable information on the accuracy of their picker, comparing the autopicker performance to analyst picks. Table 1 summarizes a number of these performance evaluations for *S*-wave pickers specifically.

Many of the aforementioned pickers provide good results in both *P*-wave and *S*-wave picking, but there is room for improvement in both identifying and making high precision picks, particularly for identifying *S* waves in noisy data. Many of these pickers have a large number of parameters, which must be tuned for each data set, and one goal of our work is to reduce the number of parameters and thus complexity, as we feel complexity is a major barrier for practitioners applying automatic picking methods to real data. We propose a phase picking method using a nearest-neighbour based approach. Our primary motivation for the development of the new method is to reduce or remove human intervention during the phase identification process, by introducing and applying a time-series pattern recognition technique, which has been very successful in the field of data mining, to the problem of seismic phase identification. The nearest neighbour algorithm, particularly one-nearest-neighbour, incorporating the Euclidean distance metric, is among the oldest and most popular methods for time-series pattern recognition problems due to its accuracy and simplicity (Ueno *et al.* 2006; Wei & Keough 2006). The technique has been applied to problems in different domains including finance, medicine, biometrics, chemistry, astronomy, robotics and networking (Keogh & Kasetty 2002).

Most seismic phase detection methods use a parameter-based approach, in which a parameter such as kurtosis is estimated to build a model for determining whether a given window contains a phase arrival. Instead of choosing a parametric model, our model is defined by the data itself. The proposed method is based on the approach developed by Nikolov (2012), which he applied to the early detection of trending Twitter topics. The method is simple and has few model parameters to tune. The algorithm is based on similarity of input signals to one reference set of waveforms containing either *P*-wave or *S*-wave phase onsets and effectively dissimilarity to a second reference set not containing phase onsets. Other non-parametric methods, such as neural networks and neural trees, have been used for phase identification, but such methods are often trained on features or parameters such as AR-model coefficients or STA/LTA values instead of on the raw data (Dai & MacBeth 1995; Wang & Teng 1997; Gentili & Michelini 2006). Our new method is reasonably fast and very accurate.

METHOD

Our phase identification method is adapted from the approach of Nikolov (2012), which takes a nearest neighbour-based approach (Altman 1992). In our method, the first step is the assembly of a reference set consisting of positive and negative examples, where positive examples are windows surrounding analyst *P*-wave or *S*-wave picks from vertical traces (*P* wave) or horizontal traces (*S* wave), and negative examples are windows that do not contain *P*- or *S*-wave picks, instead containing, for example, *P*-wave coda, *S*-wave coda, or just a horizontal line (i.e. all zeros). We generally choose a negative set consisting of windows containing *P*-wave coda. Phase onsets are identified in the test set by taking successive

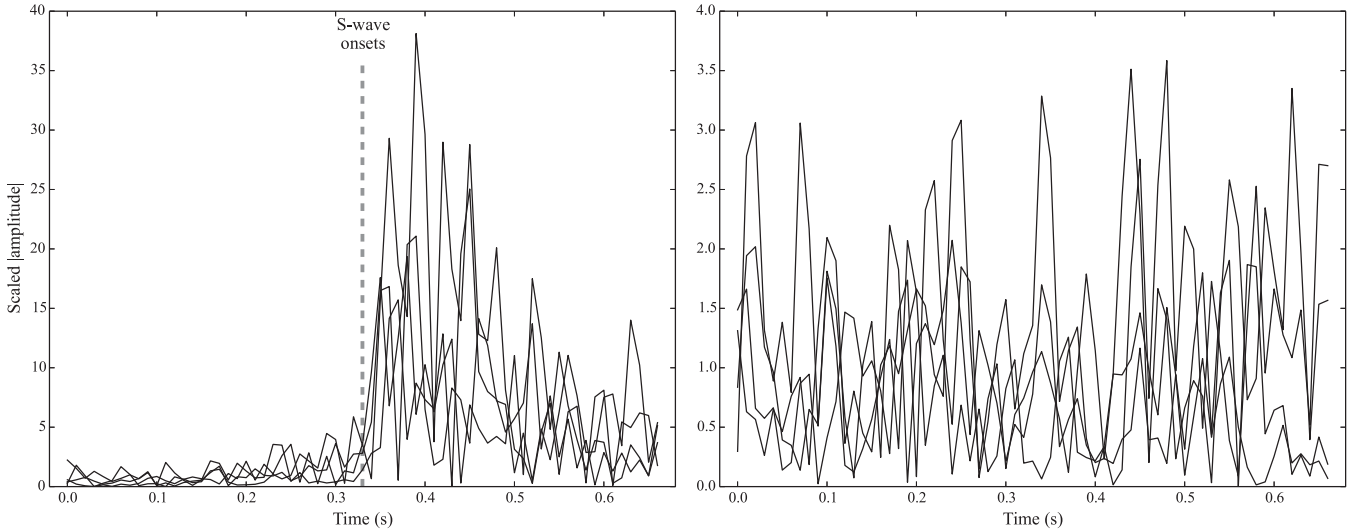

Figure 1. Example of (left-hand panel) a positive reference set using *S*-wave analyst picks from four horizontal component seismograms and example of (right-hand panel) a negative reference set consisting of *S*-wave coda from four horizontal component seismograms. Each window is scaled using the absolute amplitude of the window and divided by the mean of the first half of the window.

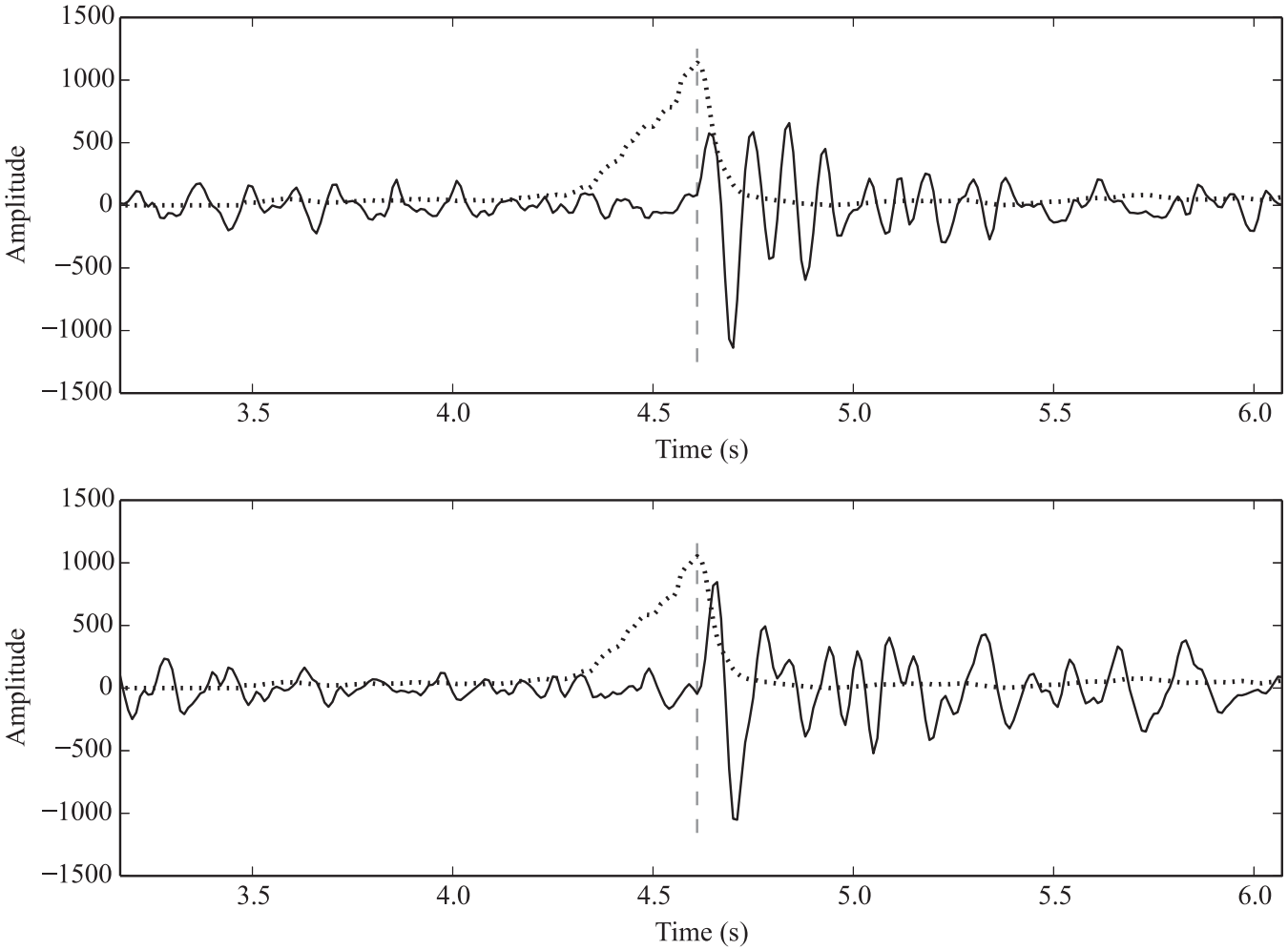

Figure 2. Example of an R_{pair} curve for an automatic *S*-wave pick (dotted line) and the associated horizontal traces. The automatic *S*-wave pick is shown (vertical dashed line). Our method windows through the input trace and computes a ratio of the similarity for the given window to the negative and positive references sets $\mathcal{R}_+$ and $\mathcal{R}_-$ (Fig. 1). An *R* curve representing the score of an *S*-wave arrival at each time along the input trace is thus quantified. The *S* pick is chosen by multiplying the *R* curves of the two horizontal components to form R_{pair} and choosing the centre of the window that maximizes R_{pair}.

windows of a given trace and computing a similarity metric to the set of positive and negative examples. For *P*-wave identification, this process is applied to the vertical component, producing a *P*-wave score function. For *S*-wave identification, this process is applied to both horizontal traces and an *S*-wave score function is computed for the two traces at each time step, and these two scores are combined together to produce a final *S*-wave score. For both phases, the arrival is chosen at the centre of the window corresponding to the maximum score.

Each trace is processed independently and the following analysis is applied to the vertical, east–west and north–south components. A moving-window procedure is performed, using a user-defined window length. A distance measure $d(\boldsymbol{r}, \boldsymbol{s})$ is computed at each time window $\boldsymbol{s}$ to each reference trace $\boldsymbol{r}$ defined as the square of the Euclidean distance between the reference signal and observed data:

$$\mathrm{d}(\boldsymbol{r}, \boldsymbol{s}) = \sum_{i=1} (r_i - s_i)^2.$$

A small distance between two waveforms indicates the two waveforms are similar. The reference set, $\mathcal{R}$, contains both positive and negative examples. Fig. 1 shows an example of *S*-wave positive and negative sets $\mathcal{R}_+$ and $\mathcal{R}_-$. The votes are aggregated for the positive and negative comparisons, and a score $R(\boldsymbol{s})$ is calculated for each window by taking the ratio of the negative to positive distances:

$$R(\boldsymbol{s}) = \frac{\sum_{r \in \mathcal{R}_-} \mathrm{d}(\boldsymbol{r}, \boldsymbol{s})}{\sum_{r \in \mathcal{R}_+} \mathrm{d}(\boldsymbol{r}, \boldsymbol{s})}$$

For *S*-wave identification, two scores R_{H1} and R_{H2} are computed for the two horizontal traces, and a single score R_{pair} is computed by combining the two horizontal *R* values using multiplication:

$$R_{\text{pair}}(\boldsymbol{s}) = R_{H1}(\boldsymbol{s}) \times R_{H2}(\boldsymbol{s}).$$

Fig. 2 shows an example pair of horizontal traces and their associated R_{pair} curve. Next, both the *P*-wave and *S*-wave phase arrival windows are determined by maximizing the vertical component score R_Z and the merged horizontal component score R_{pair}, respectively:

$$\boldsymbol{s}_{P\,\text{wave}} = \arg\max_{\boldsymbol{s}} [R_Z(\boldsymbol{s})]$$

$$\boldsymbol{s}_{S\,\text{wave}} = \arg\max_{\boldsymbol{s}} [R_{\text{pair}}(\boldsymbol{s})]$$

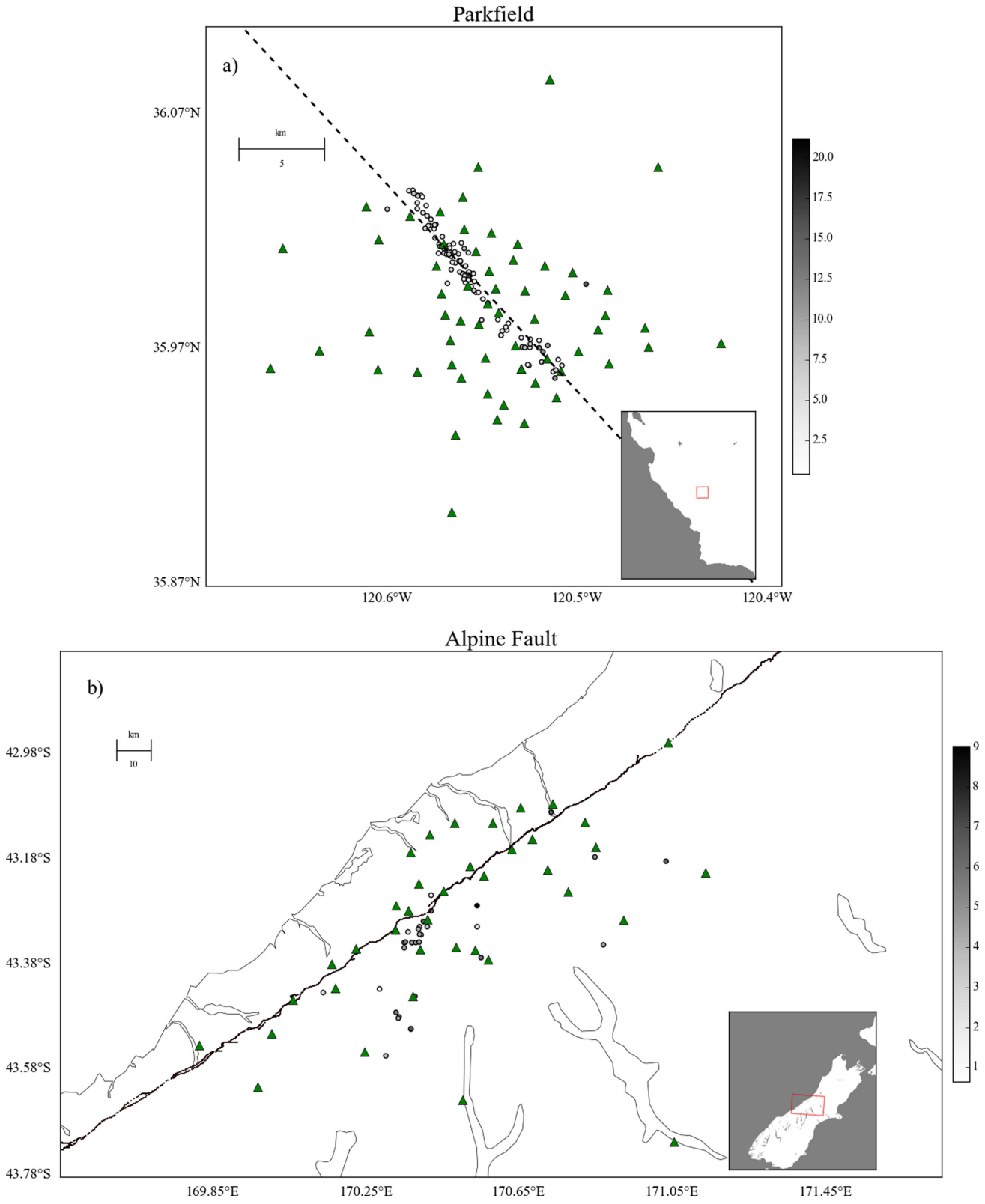

Figure 3. Maps of the two data sets used for analysis in this study from (a) California and (b) New Zealand, and approximate locations of the San Andreas Fault and Alpine Fault traces. The greyscale bar indicates the depth of the earthquakes (circles) in kilometres, and the triangles are the stations.

The phase pick is selected as the time of the centre of the $\boldsymbol{s}_{P\text{-wave}}$ and $\boldsymbol{s}_{S\text{-wave}}$ windows. The score values at these given windows must exceed a user-supplied threshold parameter. As the method picks the phase arrival window using the maximum score value from the entire trace, only one *P*-wave and *S*-wave phase is allowed, thus, excluding the possibility of picking multiple events per window. This could likely be modified to pick additional local maxima that exceed the user-supplied threshold, but more development and

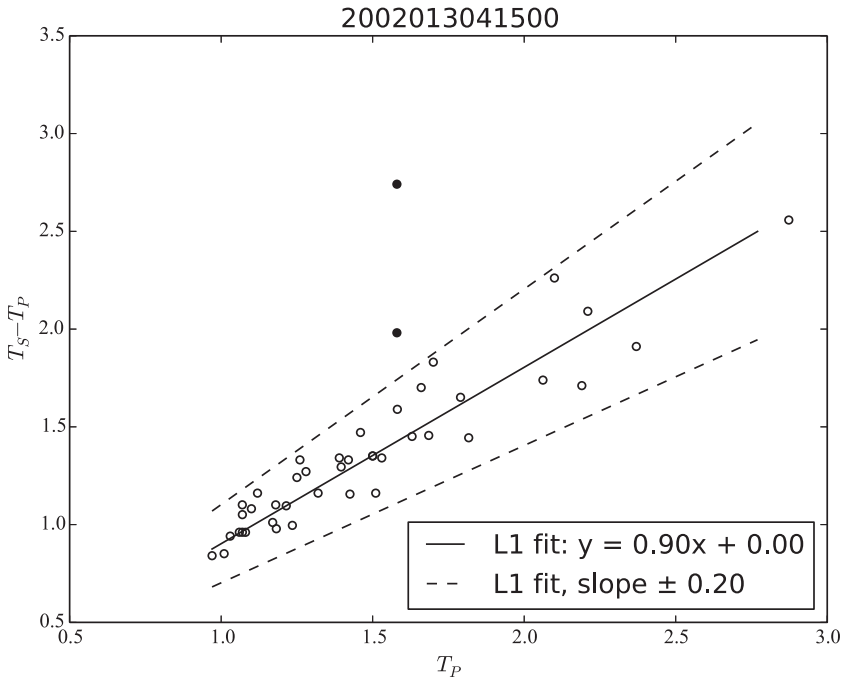

Figure 4. Wadati diagram and L1-norm fit for Parkfield event 2002013041500. Analyst pick outliers (solid black) are identified as erroneous picks and removed from the analysis if points fall outside of lines with slope ±0.2 (dashed lines) from that of the L1-norm fit (solid line). Similarly, after automatic *P* and *S* picks are identified, a new Wadati diagram is constructed for each event using automatic *S* picks and automatic *P* picks. Outliers are marked as incorrect picks and are therefore not kept by the automatic picker.

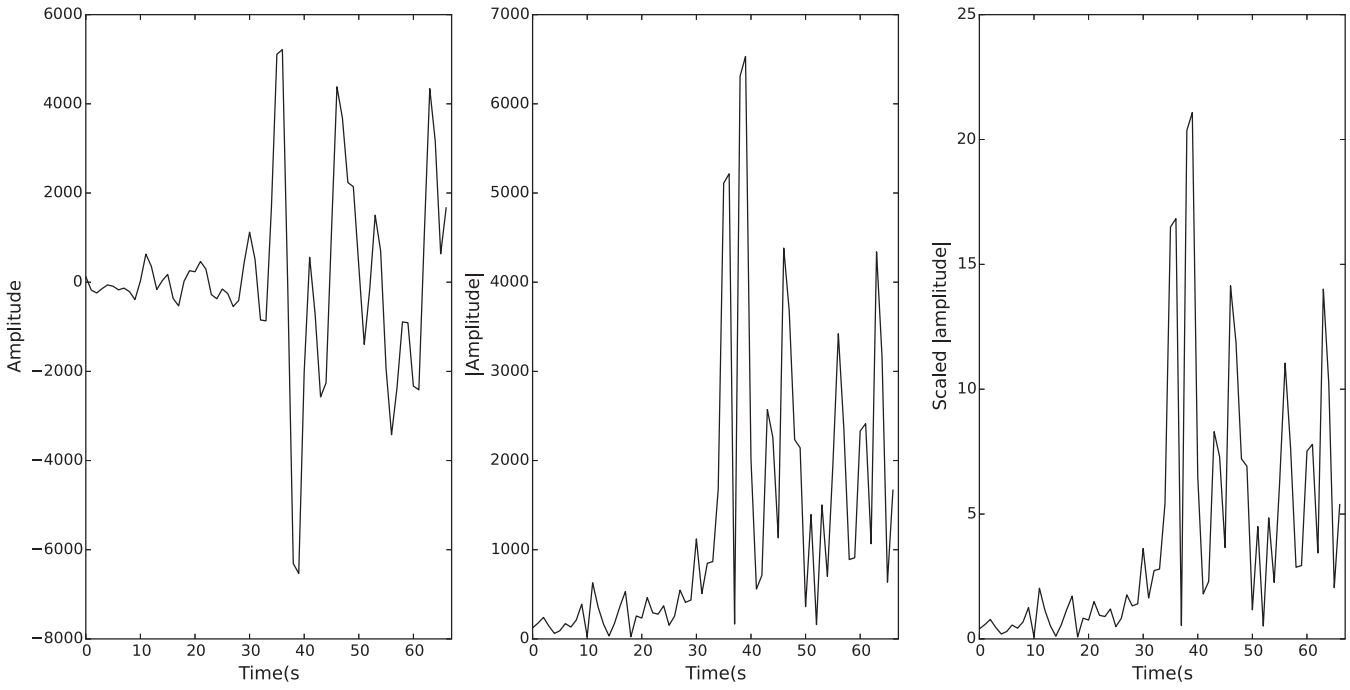

Figure 5. Rescaling scheme demonstrated on a positive *S*-wave example. Left-hand panel: unscaled example of an *S*-wave arrival at 0.33 s. Middle panel: absolute value of the amplitude. Right-hand panel: absolute value of the amplitude divided by the mean of the first half of the signal (0.00–0.33 s).

testing would need to be done to realize that capability. We also use a quality rating consisting of three classes for *P* waves and two classes for *S* waves defined using the score values at the maximum windows, which is discussed in further detail in the implementation section. Classes are defined using the values of the score functions R and R_{pair} at the *P* and *S* pick, respectively. For *P* waves, Class 1 picks have score values between 0.05 and 1.0, and Class 0 picks have score values greater than 1.0. For *S* waves, Class 2 picks have score values between 1.2×10^{-4} and 0.001, Class 1 picks have score values between 0.001 and 0.05, and Class 0 picks have score values greater than 0.05. We find that using two classes for *P*-wave picks and three classes for *S*-wave picks is sufficient for our needs, but the class definitions and number of classes can easily be modified for future studies.

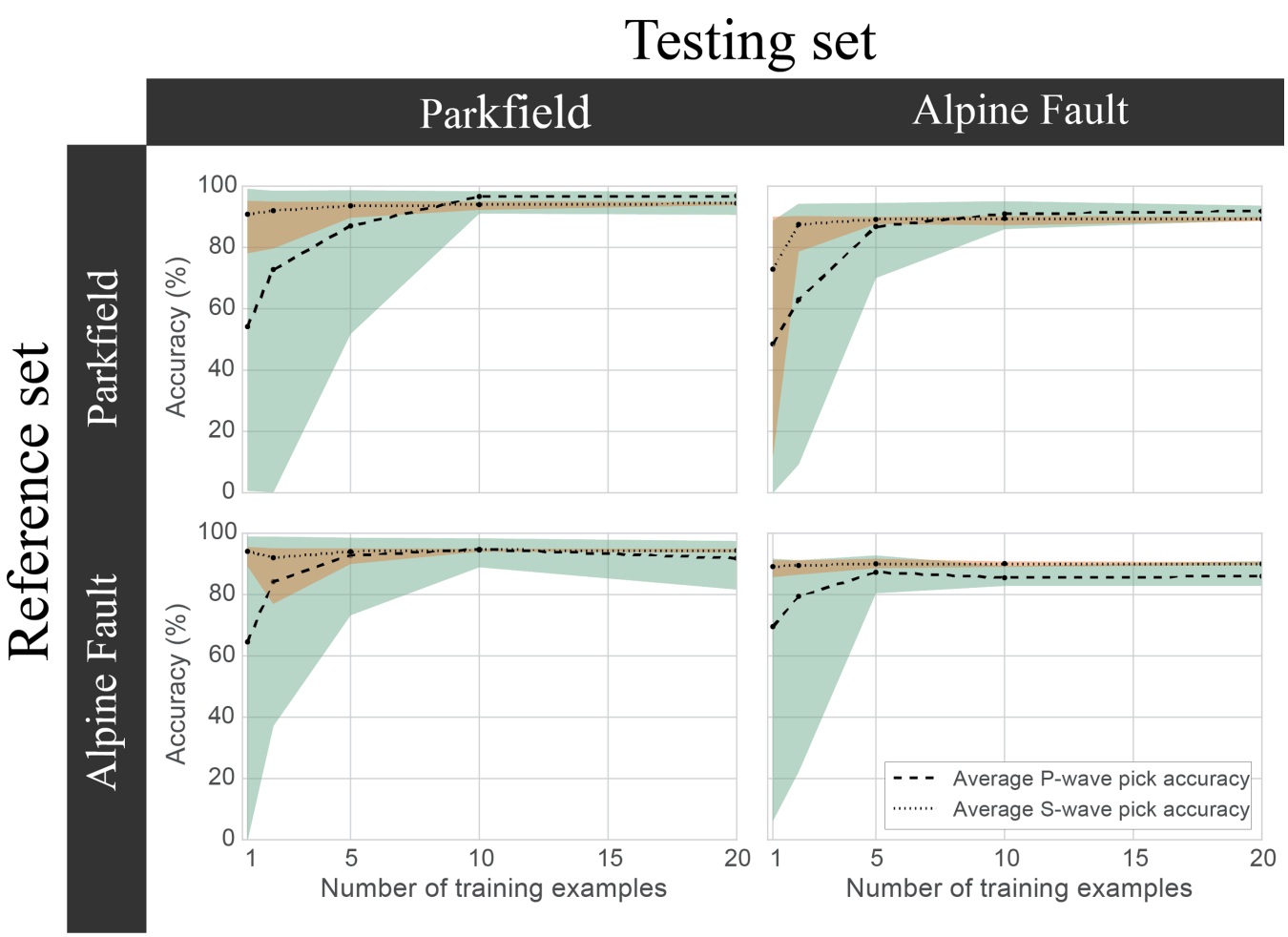

Figure 6. Learning curves for Parkfield and the Alpine Fault showing accuracy as a function of number of reference examples. The dotted lines correspond to the average accuracy of the automatic picker for *P* waves, where *P*-wave picking accuracy is defined as percentage of picks with residuals less than or equal to 0.1 s. The dashed lines correspond to the average accuracy of the automatic picker for *S* waves, where *S*-wave picking accuracy is defined as percentage of picks with residuals less than or equal to 0.2 s. The shaded area surrounding each line corresponds to the interval between the minimum and maximum accuracy observed using 10 trials consisting of random reference data. As expected, using more reference examples stabilizes the accuracy. However, declining returns in stability are observed after approximately 10 reference examples.

Table 2. Parameter values.

Parameters	Both data sets
Moving window length	0.67 s (67 samples)
Vertical component band-pass filter	3–30 Hz
Horizontal components band-pass filter	2–30 Hz
P-wave *R* threshold	0.05
S-wave R_{pair} threshold	1.2e-4
Size of *P*-wave reference set	2 picks
Size of *S*-wave reference set	2 picks
P-wave score threshold for classes: C_1, C_0	0.05, 1
S-wave score threshold for classes: C_2, C_1, C_0	1.2×10^{-4}, 0.001, 0.05
Minimum, maximum T_S–T_P	0.6 s, 10.0 s

DATA

We test the accuracy of our autopicker using two different data sets containing reviewed *P*-wave and *S*-wave picks from two different analysts, one from the Parkfield region in central California and the other from the Alpine Fault region on the South Island of New Zealand (Fig. 3). These events were initially identified using a multifrequency STA/LTA detector on the vertical component. A 30-second window was cut around the detected events, and a human analyst picked both *P* waves and *S* waves from these events. One *S*-wave pick was identified per station. The data in our data sets contain traces from known events that a human analyst was able to pick either *P* waves or *S* waves on, and these picks were carefully reviewed as part of our analysis. As a result, in our testing scheme, there are few chances for a false positive pick to be made. Thus in this paper we are mostly measuring the error in terms of the residual between human picks and our automatic picks.

The Parkfield data set consists of 98 events recorded by over 50 stations of the PASO array along the San Andreas Fault operated by UW-Madison and Rensselaer Polytechnic Institute in 2001–2002. The events are local earthquakes with average *S*–*P* times of 1.5 s, epicentral distances less than 40 km and depths ranging from 0.5 to 21 km. The magnitudes range from 0.18 to 2.83. In total, there are 4452 station–event pairs containing 4416 *P*-wave picks and 4435 *S*-wave picks. These events were selected from our database of 238 events occurring in 2002 because they had origin and magnitude information in the Northern California Earthquake Data Center (NCDEC) database.

The Alpine Fault data set consists of 355 station–event pairs containing 352 analyst *P*-wave and 353 analyst *S*-wave picks from 32 events recorded by over 30 different broad-band and

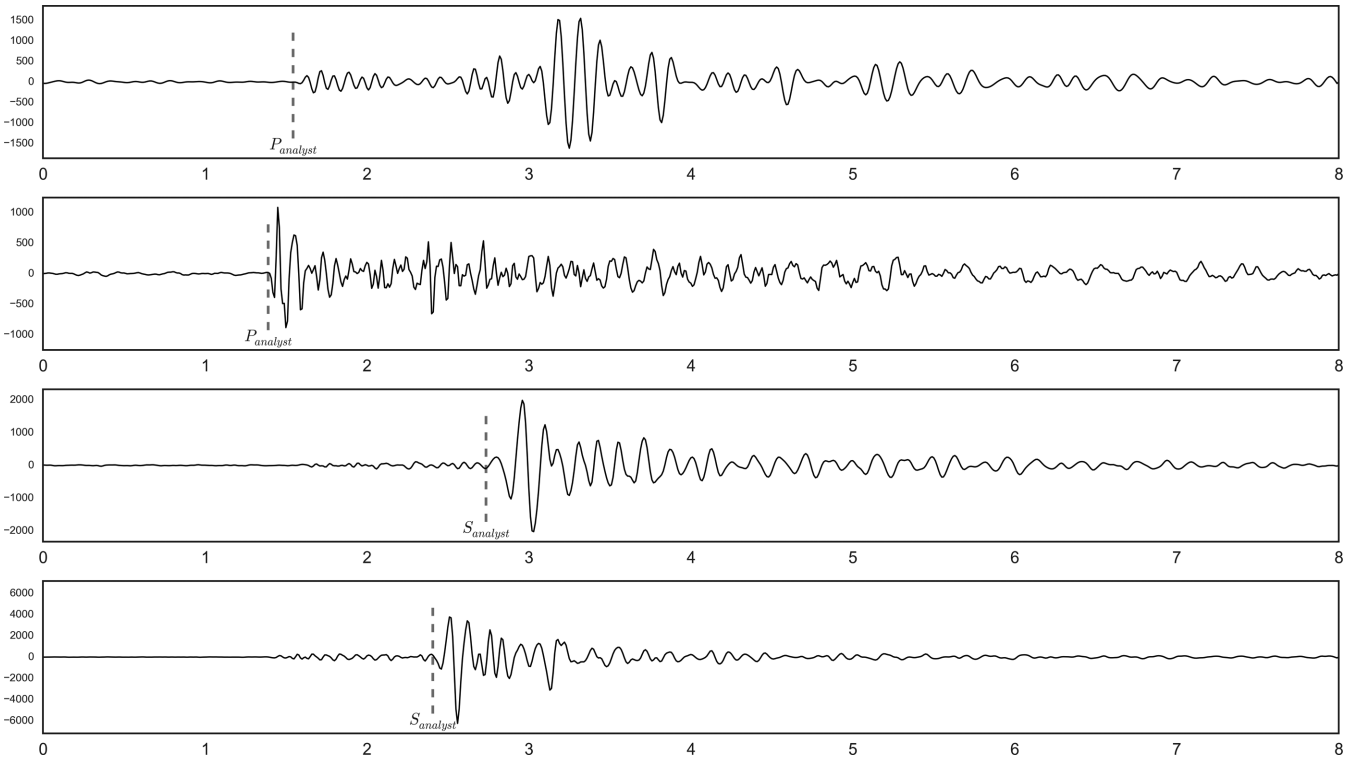

Figure 7. The (top two traces) *P*-wave and (bottom two traces) *S*-wave reference waveforms used in the 'designed' reference set and associated analyst picks. The *P*-wave reference traces are on the vertical channel and the *S*-wave reference traces are on the horizontal channels. All four traces are from different stations and events and were empirically chosen to maximize accuracy and robustness. The positive reference set windows are created by cutting 0.33 s before and after the analyst pick.

short-period stations along the Alpine Fault, New Zealand, in 2012–2014 from the WIZARD array operated by the University of Wisconsin-Madison, Rensselaer Polytechnic Institute and Victoria University of Wellington, and from the SAMBA array operated by Victoria University of Wellington. The events are also local earthquakes with average *S*–*P* times of 3.3 s, epicentral distances less than 100 km and depths ranging from 0.6 to 9 km. The magnitudes range from 1.6 to 3.6.

Both data sets are sampled at 100 Hz. Some erroneous analyst picks are identified and removed using a Wadati diagram, excluding outliers from a least absolute deviation (L1) fit (Fig. 4). After this process there are 4200 *P*-wave and 4219 *S*-wave analyst picks in the Parkfield data set and 343 *P*-wave and 344 *S*-wave analyst picks in the Alpine Fault data set.

IMPLEMENTATION

For our initial assessment, we test two approaches: using reference sets from a randomly selected set of seismograms and using a 'designed' reference set. The autopicker has few parameters to tune, and Table 2 shows the parameter values used for both data sets. We choose a moving-window length of 0.67 s. The vertical traces are bandpass filtered between 3 and 30 Hz, and the horizontal traces are bandpass filtered between 2 and 30 Hz. *P*-wave and *S*-wave positive reference examples are created using the portion of the waveform 0.33 s before and after the analyst *P*-wave pick and *S*-wave pick, whereas negative reference waveforms are created by cutting a 0.67 s window in the *P*-wave coda. In our experience, we find the method is relatively invariant to using a negative reference set containing *P*-wave coda, *S*-wave coda, or a horizontal line. The reference examples are then scaled by taking the absolute value of the amplitude and dividing by the mean amplitude of the first half of the 0.67 s window of the positive examples, that is, in the positive examples, the waveform portion prior to the phase arrival (Fig. 5).

First, a potential *P*-wave arrival is selected using the procedure as described in the Methods section. A maximum of one *P*-wave pick is made per trace. If the maximum of the score function of the entire trace exceeds our user-defined *P*-wave threshold value of 0.05, a *P*-wave pick is made. A potential *S*-wave pick is identified by maximizing the score function and searching for a value exceeding the *S*-wave threshold in a window starting 0.6 s after the *P*-wave arrival and ending 10.0 s after the *P*-wave arrival. We choose the lower bound to ensure adequate separation between *P*-wave and *S*-wave picks and the upper bound because we do not expect to see extremely large separation between picks for local events. Also, if no *P*-wave pick is chosen, the entire trace is searched for a potential *S*-wave arrival. If no point of the *S*-wave score function exceeds the threshold in this user-defined window, the method will not pick an *S*-wave. Also, it is noted if the automatic picker makes an incorrect *P*-wave pick before or at the time of true *S*-wave arrival, the method may not correctly pick the *S*-wave arrival.

Subsequently, probable erroneous autopicks are identified using a Wadati diagram, plotting automatic *S* pick minus automatic *P* pick versus automatic *P* pick and removing outliers from an L1 fit. This is similar to the process applied in the Data section with the manual picks using a known origin time, however, in the case of automatically picking a new event, the origin time is determined as the *P* time when the *S*–*P* time equals 0. Next, we exclude picks

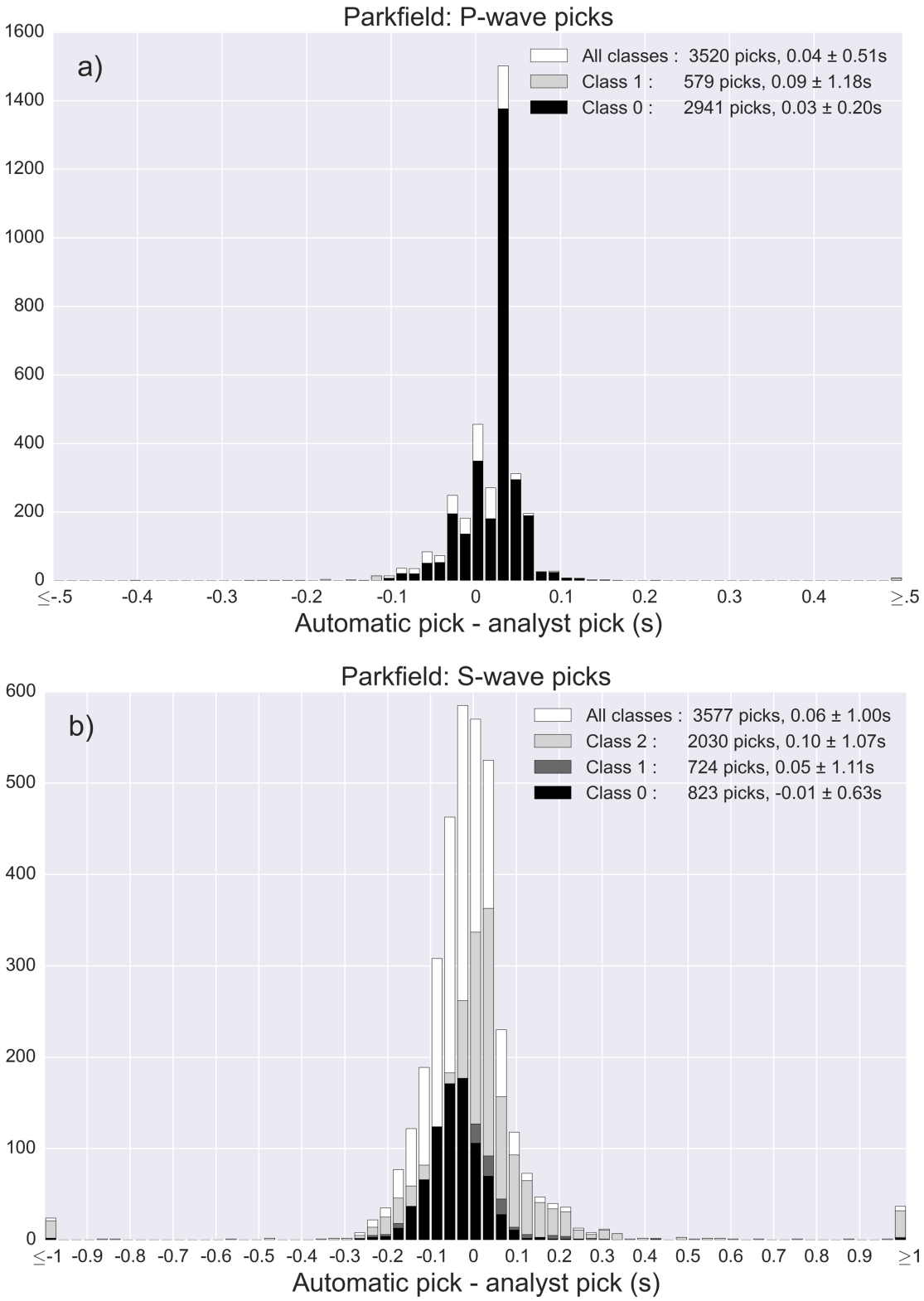

Figure 8. Autopick error histograms for the Parkfield and Alpine Fault data sets and corresponding mean error and standard error for each class. (a) Parkfield *P* waves, (b) Parkfield *S* waves, (c) Alpine Fault *P* waves and (d) Alpine Fault *S* waves.

outside of lines with slopes ±0.20 of the L1 fit. The goal of this process is to exclude egregious outliers.

Finally, the method assigns class ratings using the value of the score function at the automatic pick. There are three classes for *P* waves and two for *S* waves based on examples from the literature.

RESULTS

The accuracy of the proposed method is tested as a function of reference set quantity and geographic location. Autopicks are compared to analyst picks on both data sets using reference waveform sets

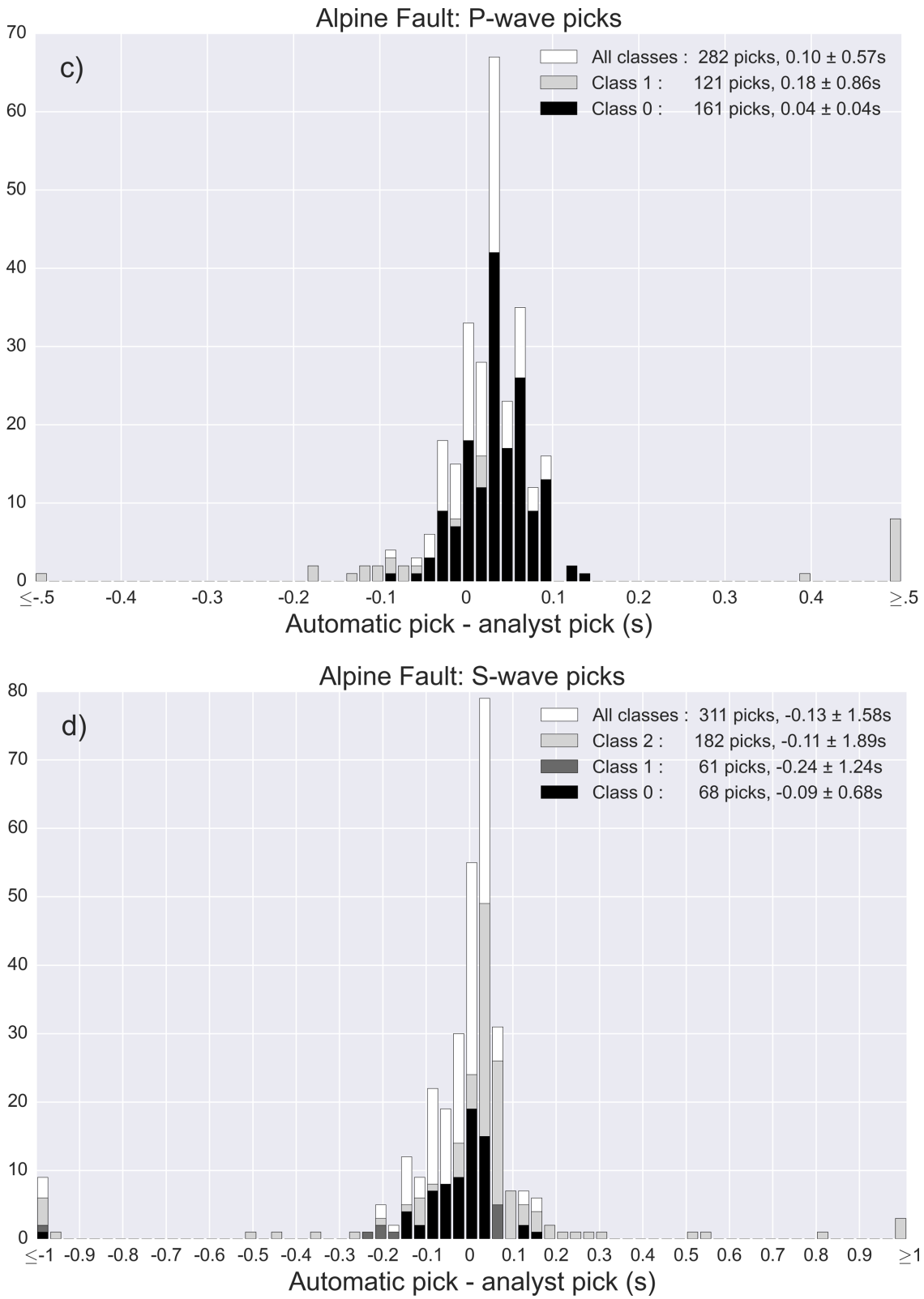

Figure 8 – *Continued*

from both data sets varying in size from 1 to 20 picks. Fig. 6 shows 'learning curves' for the percentage of *P* and *S* autopicks classified within 0.1 and 0.2 s, respectively, of the analyst pick with increasing number of reference waveforms, using 10 trials with randomly chosen reference *P*-wave and *S*-wave picks. For testing the *S*-wave picking accuracy, we use the same 2 *P*-wave picks but vary the *S* reference picks for every trial. The learning curves flatten out relatively quickly, which indicates the method 'learns' rather quickly, requiring a small number of reference traces to achieve adequate results. For both of these data sets, the method becomes slightly more

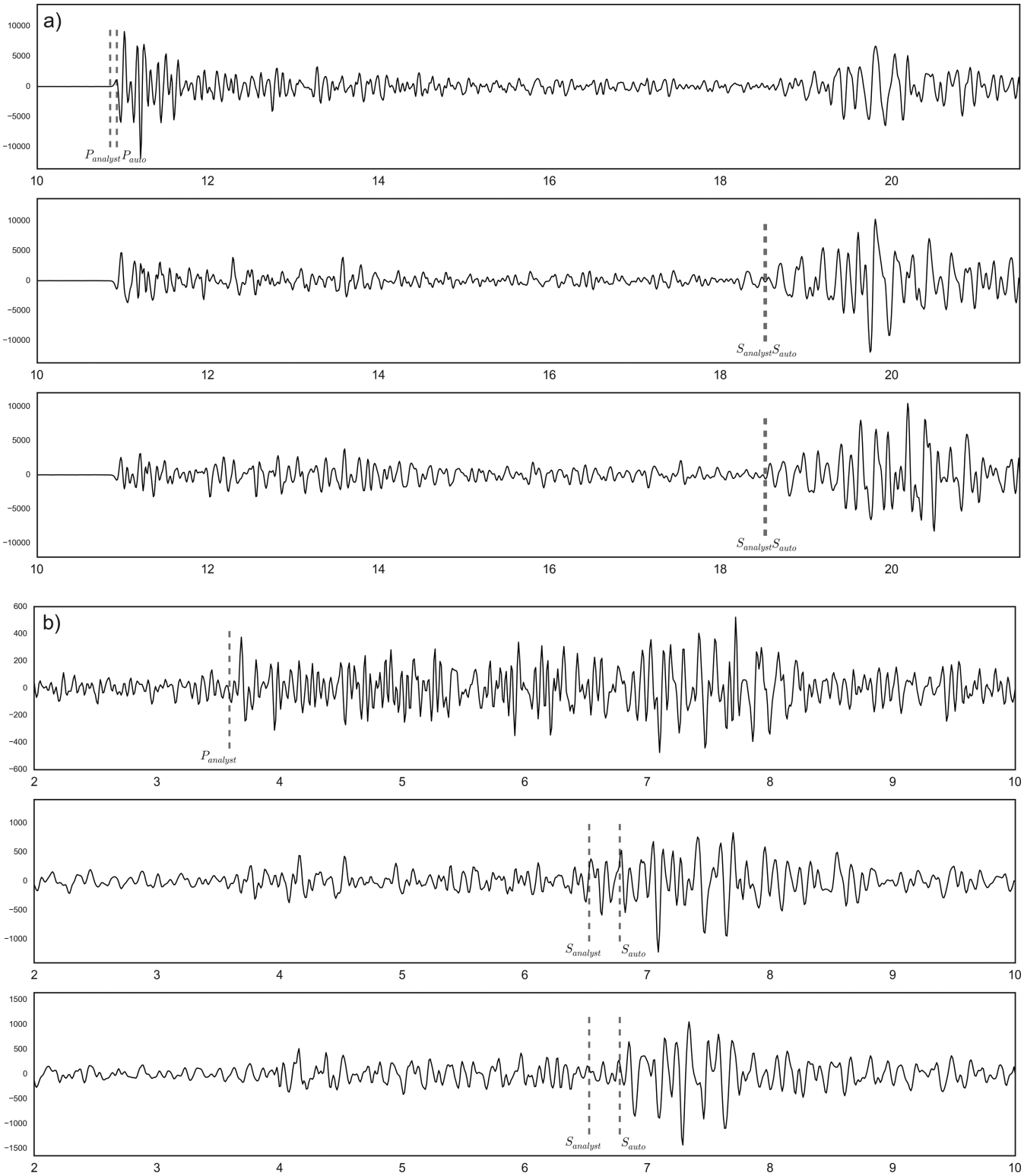

Figure 9. Waveforms and associated automatic and analyst picks for two different station–event combinations. (a) The autopick and analyst pick agree on both the *P*- and *S*-wave picks. (b) The automatic picker is unable to pick a *P*-wave arrival, however it is able to pick an *S*-wave arrival. The *S*-wave pick residual is greater than 0.2 s, which may be due to either anisotropy or the analyst having picked a *P*-to-*S* converted phase arrival.

stable if more reference picks are used, as indicated by the smaller error bars. For all trials using at least 10 reference examples, over 80 per cent of *P*-wave picks fall within 0.1 s of the analyst pick for both data sets. In addition, for all trials using at least five reference examples, over 90 per cent of *S*-wave picks fall within 0.2 s of the analyst pick for both data sets. We note that 0.1 and 0.2 s are typical estimates of analyst pick uncertainty for local earthquake *P* and *S* waves, respectively. We also note that there is less stability in the learning curves for *P*-wave picks, particularly at reference sizes less than 10. In addition, we note that the accuracy is relatively

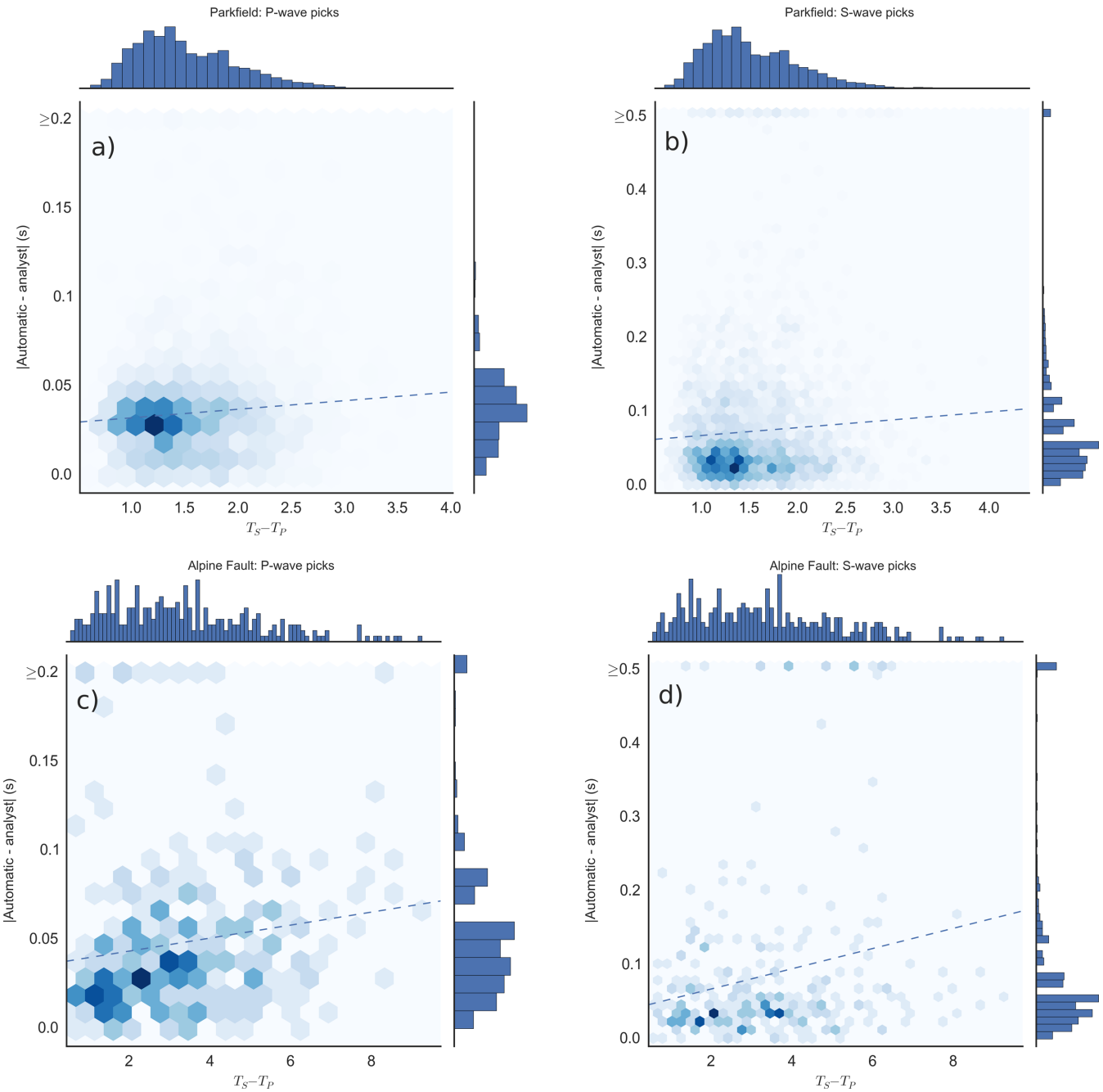

Figure 10. Scatter plots of deviation between automatic and analyst pick as a function of *S*–*P* time and associated best-fitting lines for (a) Parkfield *P*-wave picks, (b) Parkfield *S*-wave picks, (c) Alpine Fault *P*-wave picks and (d) Alpine Fault *S*-wave picks. Cumulative histograms are shown at the top and side edges. Dashed line shows least squares fit.

insensitive to the geographic location of the reference set—results for both data sets are similar using reference waveforms from either data set, although on average using the Parkfield reference data yields better accuracy for picking *P* waves on both data sets. This indicates that our method may be transportable for other local event studies.

For further analysis, we present results for both data sets using reference picks from 'designed' reference sets. The designed reference sets consists of a *P*-wave reference set containing two reference picks and an *S*-wave reference set consisting of two reference picks, both of which are from the Parkfield data set (Fig. 7). Fig. 8 shows the histograms of residuals (autopick minus analyst pick) for both data sets. We adhere to the above autopick convention of error defined as residuals greater than 0.1 and 0.2 s for *P* and *S* waves, respectively. For *P*-wave picks, using the two *P*-wave reference picks from the Parkfield data set, 3520 picks are identified with 98 per cent accuracy and 282 Alpine Fault picks are identified with 94 per cent accuracy. For *S*-wave picks, using *S*-wave reference picks from the Parkfield data set, 94 per cent of the 3577 Parkfield picks and 91 per cent of the 311 Alpine Fault picks are obtained successfully. We attribute the lower success rate for the Alpine Fault data set to the larger epicentral distance ranges for the events. Fig. 9 shows examples of automatic *P*-wave and *S*-wave picks, compared to analyst picks.

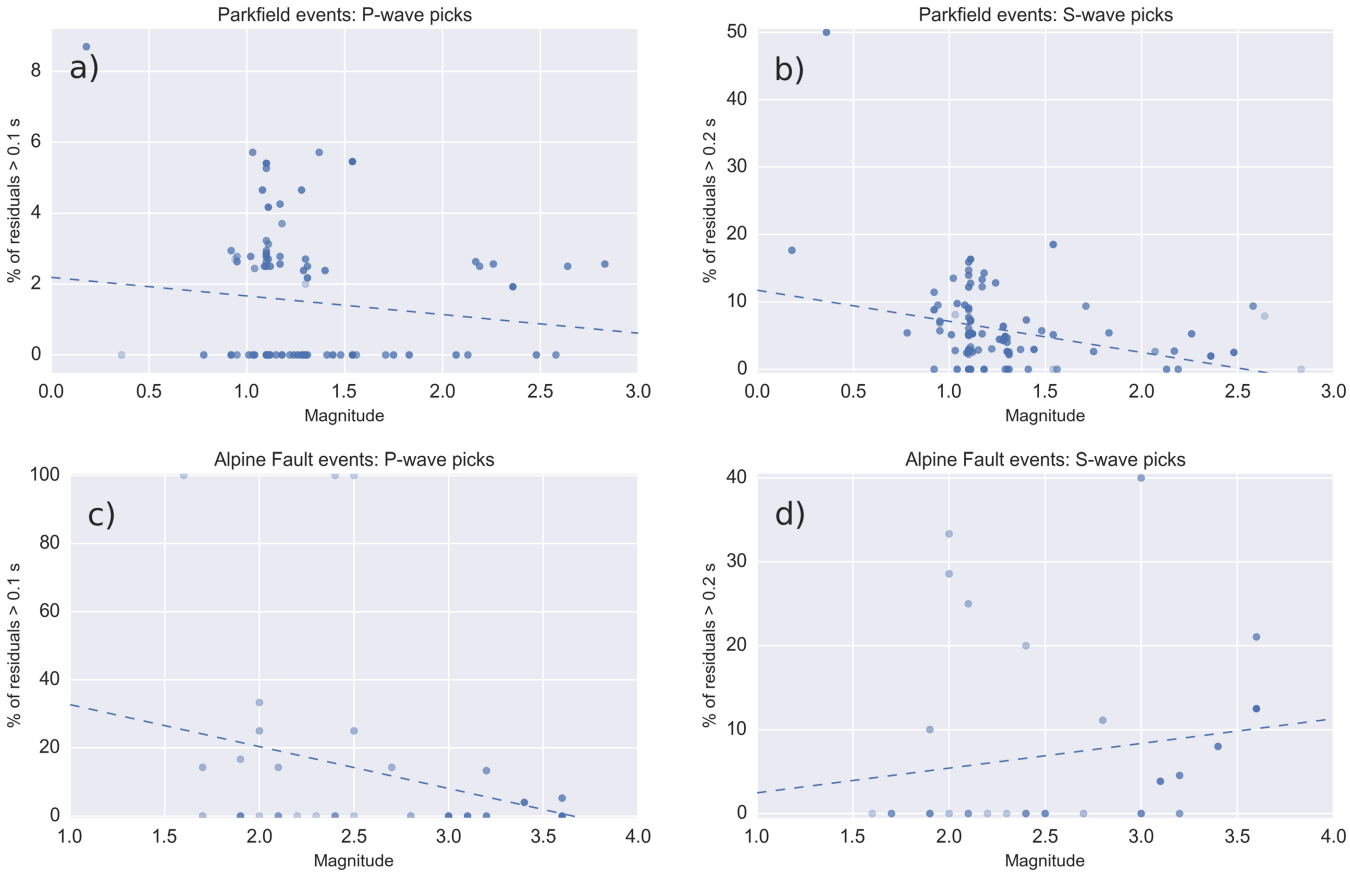

Figure 11. Deviation between automatic and analyst pick as a function of magnitude and associated best-fit lines for (a) Parkfield *P*-wave picks, (b) Parkfield *S*-wave picks, (c) Alpine Fault *P*-wave picks and (d) Alpine Fault *S*-wave picks. Dashed line shows least squares fit.

DISCUSSION

As with other phase detection methods, it is important to assess the generality of the method. Specifically, we evaluated the method's accuracy with respect to *S*–*P* time, magnitude, and depth.

Evaluating the method's accuracy as a function of *S*–*P* time helps to determine the method's ability to apply to different distance ranges. Fig. 10 displays a joint density plot of pick error versus *S*–*P* times. In the Parkfield data set, most events have an *S*–*P* time less than 2.5 s. For *P*-wave picks, there appears to be little correlation between error and *S*–*P* times with a correlation coefficient ρ equal to 0.09. For *S*-wave picks, there is also little correlation between error and *S*–*P* times ($\rho = 0.06$). In the Alpine Fault data set, for automatic *P*-wave picks, pick error is positively but weakly correlated with an increasing *S*–*P* time ($\rho = 0.16$). This suggests accuracy may slightly decrease with distance for *P*-wave picks. Finally, for *S*-wave picks on the Alpine Fault data set, pick error is slightly more positively correlated with *S*-*P* times ($\rho = 0.21$). These results are consistent with the above attribution of a lower Alpine Fault success rate to having events spanning a greater distance range.

In the following analysis of error versus magnitude and error versus depth, for a given event, we continue to define *P*-wave error as the percentage of picks with residuals greater than 0.1 s and *S*-wave error as the percentage of picks with residuals greater than 0.2 s. Fig. 11 shows plots of error versus magnitude. In regards to magnitude versus error rate, for the Parkfield data set, the *P*-wave pick error is weakly negatively correlated with magnitude ($\rho = -0.12$). For *S*-wave picks, pick error is somewhat more strongly negatively correlated with magnitude ($\rho = -0.31$). For Alpine Fault events, *P*-wave pick error is negatively correlated with magnitude ($\rho = -0.24$) and *S*-wave pick error is weakly positively correlated with magnitude ($\rho = 0.15$). In general, one would expect small magnitude events to result in larger error, so a negative correlation is expected. The weak positive correlation for the Alpine Fault events may simply be due to the statistics of small numbers, as we have very few events with magnitude >3.5, and without those events the correlation would become weakly negative. Fig. 12 shows pick error versus depth. For Parkfield *P*-wave picks, pick error is very weakly positively correlated with depth ($\rho = 0.04$), and for *S*-waves picks, pick error is weakly negatively correlated with depth ($\rho = -0.10$). For Alpine Fault *P*-wave picks, pick error is negatively correlated with depth ($\rho = -0.09$), and for *S*-wave picks, pick error is positively correlated with depth ($\rho = 0.44$). The stronger *S*-wave correlation with depth for the Alpine Fault may be a reflection of anisotropy. The other depth correlations are too weak to be considered significant. A greater sampling range of magnitudes and depths would further increase confidence in these results.

We note that there is a small percentage of significant outliers (here defined as error great than 0.5 s for *P* and 1.0 s for *S*) in the histograms in Fig. 8. For the Parkfield *P* outliers, half are due to noise from cows, one is due to two events in the time window, and the others are *S* arrivals erroneously picked as *P*. For the Parkfield *S* outliers, 39 per cent are later arriving phases, 37 per cent are *P* arrivals erroneously picked as S, 15 per cent are due to cows, 7 per cent are identified as actual analyst errors, and one case is due to two events in the time window. For the Alpine Fault, the *P* outliers are all cases for which the autopicker missed the *P* arrival and picked the *S* arrival instead. For the *S* outliers, most of them are *P* arrivals

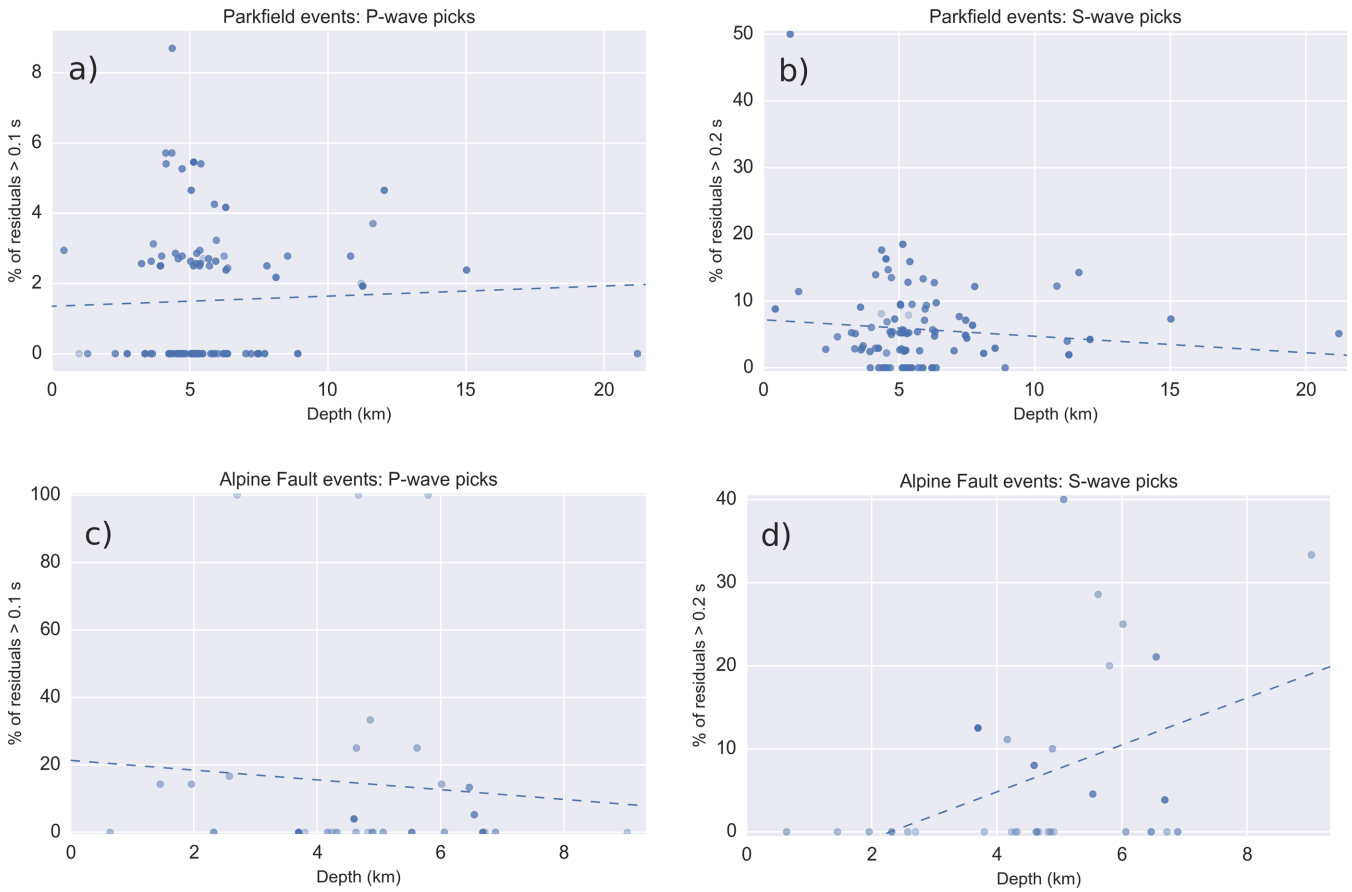

Figure 12. Scatter plot of deviation between automatic and analyst pick as a function of depth and associated best-fitting lines for (a) Parkfield *P*-wave picks, (b) Parkfield *S*-wave picks, (c) Alpine Fault *P*-wave picks and (d) Alpine Fault *S*-wave picks. Dashed line shows least squares fit.

erroneously picked as *S*, a few are cases in which the *S* arrival had been picked as *P* and the *S* was picked on a later phase, and in one case there are two events in the time window. Additional screening criteria would readily identify most of these kinds of errors.

We note that in the presence of significant anisotropy and resultant shear wave splitting, there is likely value to picking *S*-wave arrivals individually on the fast and slow horizontal components. Our method could be modified to make individual *S*-wave picks on rotated components, rather than one pick using the R_{pair} score function, although there may be a trade-off in stability by using one trace at a time rather than two.

So far, we have only tested our method on local earthquakes, thus we cannot generalize our results to regional and teleseismic events. We speculate that in order to generalize to data sets covering greater epicentral distances, modification of the method's parameters may be required (i.e. window size, bandpass filter, threshold value, etc.) as observed in other publications on phase detection.

For completeness, we also tested the method on some examples that the human analysts were unable to pick. These traces contain signals that have (1) high noise levels with ambiguous *P*-wave arrivals, (2) extremely high noise levels, where an earthquake signal is barely recognizable and phase picks could not be accurately picked by an analyst and (3) instrument malfunctions and noise spikes (Fig. 13). There are 21 examples where the analyst was unable to pick a *P* wave, but was able to pick an *S* wave. For these examples, we find 10 of the *S*-wave picks have residuals less than 0.2 s, six have residuals greater than 0.2 s, and five for which no automatic pick was made. There were 12 examples where the analyst was unable to detect phase arrivals for either *P* waves or *S* waves. In this case, the automatic picker did not pick any *P* waves, but incorrectly picked *S* waves on the 12 examples. This problem of detecting *S*-wave false positives could be remedied by raising the *S*-wave threshold parameter. However, doing so would decrease the number of total *S*-wave picks the method will make. Alternatively, as these traces are particularly noisy, using a signal-to-noise ratio analysis could potentially help discriminate between false positives and true positives. Finally, there were two examples containing noise spikes—the auto picker incorrectly locked on the noise spikes for both *P*-wave and *S*-wave picks in these cases. Incorporating a despiking analysis in the preprocessing stage may help decrease the rate of false positives. In practice, for a given event, if noise spikes or malfunctions occur on a small number of stations, the Wadati diagram quality control we utilize may remove such stations, as they will likely appear as outliers on the Wadati diagram.

In Fig. 8, it is observed that the method is slightly biased to picking *P* waves several samples after the analyst pick for both data sets. This indicates that, using the two designed reference picks, the method tends to pick the *P*-wave arrival slightly late. In addition, the method is slightly biased to picking *S* waves several samples after the analyst for the Alpine Fault. This may be due to differences in the data sets, converted phase arrivals, and/or bias due to different analysts having picked the two data sets. Furthermore, in general, the method picks with a higher degree of accuracy on the Parkfield data set. This could be due to a variety of factors including, but not limited to, different geologic settings, small *S*–*P* times and higher signal-to-noise ratios.

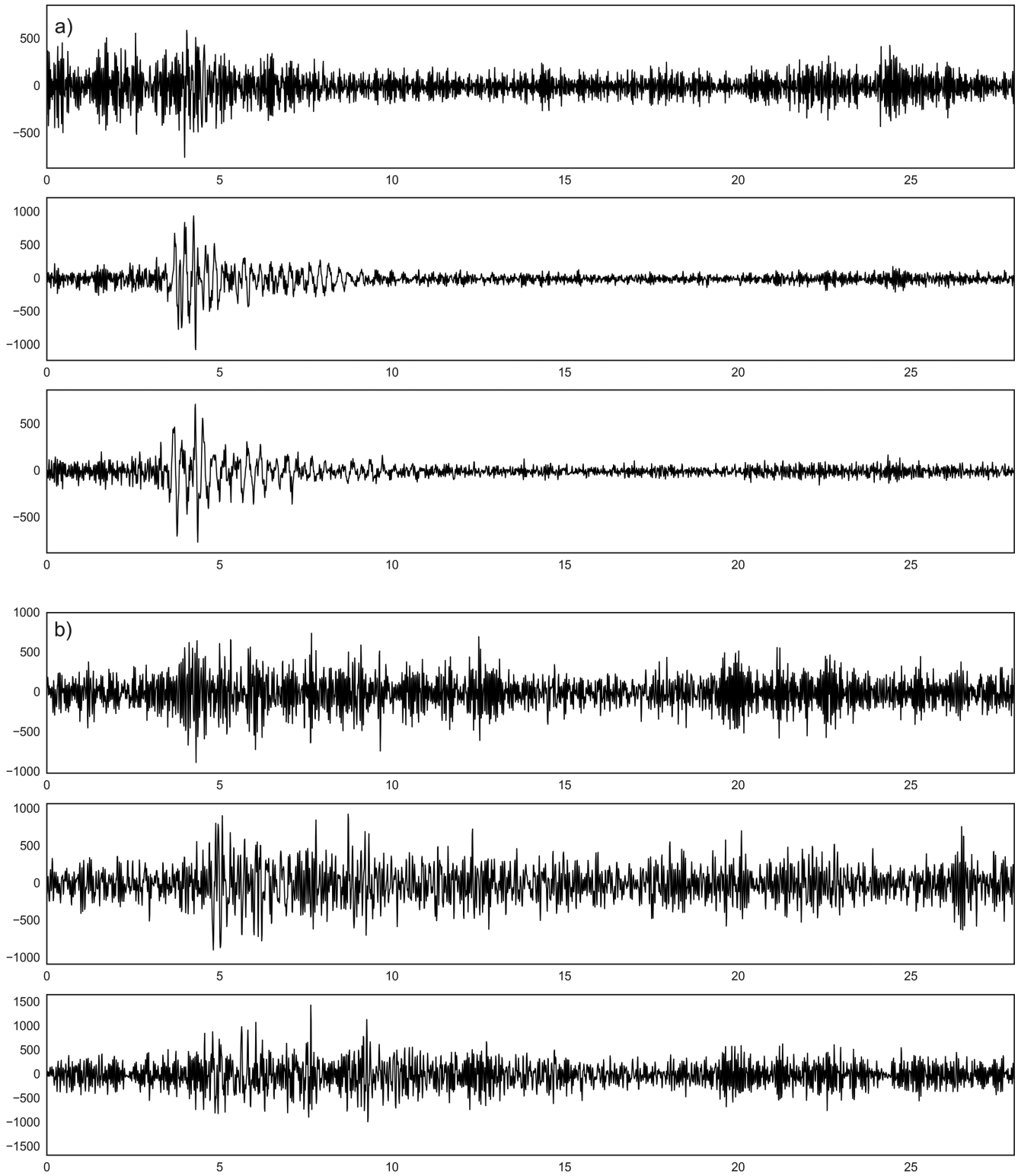

Figure 13. Examples for which the human analysts were unable to pick either *P* waves or both *P* wave and *S* waves. Examples containing (a) high noise levels with ambiguous *P*-wave arrivals, (b) extremely high noise levels, where an earthquake signal is barely recognizable and phase picks were impossible to be accurately made by an analyst and (c) instrument malfunctions and noise spikes (cows).

As the algorithm is currently implemented, it is not suitable to apply to continuous data for simultaneous event detection and phase picking. However, if combined with event detection software, the method could potentially be incorporated in an automated, continuous data processing workflow, particularly if suitable detection criteria are used, such as requiring a minimum number of *P*- and *S*-wave phase picks to be identified for a potential event.

Finally, it has been shown in machine learning and statistical applications that ensemble methods (techniques that incorporate multiple methods) outperform single-model techniques when there is

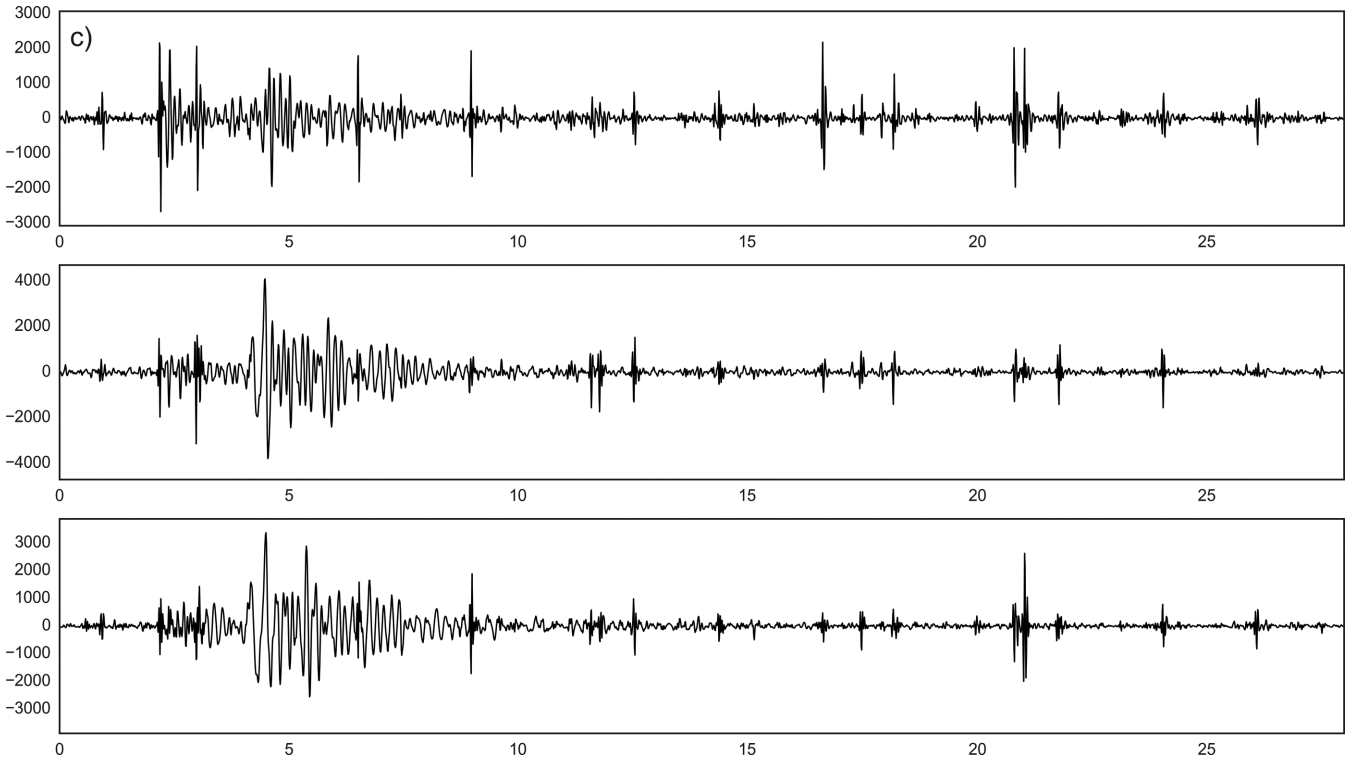

Figure 13 – *Continued*

diversity among the methods (Kuncheva & Whitaker 2003). Furthermore, this phenomenon has been observed in phase picking. Specifically, combining STA/LTA, polarization, and AR-AIC proves more effective than a single-algorithm approach utilizing these methods (Diehl *et al.* 2009). As such, we advocate combining our approach with other current event and phase detection methods to further improve picking accuracy.

CONCLUSION

We have developed a simple, fast, and accurate method for automatically identifying both *P*-wave and *S*-wave onsets by applying a nearest neighbour-based technique to bandpassed seismograms. The method identifies potential phase onsets by comparing the similarity of a given window to a set of positive examples containing known phase onsets and to a set of negative examples not containing phase onsets. For both *P* waves and *S* waves, a score function is calculated using the ratio of similarity to the positive examples divided by the similarity to the negative examples, all of which have been processed by taking their absolute value and rescaling them. The method processes each trace independently and does just one pass through the band-passed data with a fixed window size.

Two data sets containing local earthquakes, one from the Parkfield region in central California and the other from the Alpine Fault region on the South Island of New Zealand, are analysed using various combinations of reference waveforms from both data sets. The results are quite similar for *P* waves and *S* waves for both data sets regardless of the reference set used and with little sensitivity to the number of reference waveforms. Furthermore, we also test both data sets using identical algorithm parameters and two 'designed' reference picks from the Parkfield data set. In this scheme, using an error rate defined as having an absolute residual between the analyst pick and automatic pick greater than 0.1 s, the method picks *P* waves on the Parkfield data set and on the Alpine Fault data set with an error of 2 and 6 per cent (98 and 94 per cent correct), respectively. Furthermore, using an error rate defined as having an absolute residual between the analyst pick and automatic pick greater than 0.2 s, the method picks *S* waves on the Parkfield data set and on the Alpine Fault data set with an error of 6 and 9 per cent (94 and 91 per cent correct), respectively. We assess the accuracy of the method as a function of *S*–*P* time, magnitude, and depth, finding weak correlations except between accuracy and *S*–*P* time (Alpine Fault *S*), accuracy and magnitude (Parkfield S, Alpine Fault *P*), and between accuracy and depth (Alpine Fault *S*).

The method has few parameters to tune, and as the results above indicate, for our data sets, the method shows little dependence to the reference set used and achieves a good pick accuracy with just two reference analyst picks. The success of the method using identical algorithm parameters on both data sets and reference sets from the same or different data sets suggests that, at least in some cases, the method can work as an out-of-the-box autopicker using similar parameters and a provided reference set.

ACKNOWLEDGEMENTS

This material is based upon research supported by the National Science Foundation grant EAR-1114228. We are grateful to two anonymous reviewers for their constructive comments. We thank the Wisconsin Alumni Research Foundation for their assistance in submitting a U.S. patent application for this method (P140387US01).

REFERENCES

Allen, R., 1978. Automatic earthquake recognition and timing from single traces, *Bull. seism. Soc. Am.,* **68,** 1521–1532.

Allen, R., 1982. Automatic phase pickers: their present use and future prospects, *Bull. seism. Soc. Am.,* **72,** S225–S242.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor non-parametric regression, *Am. Stat.,* **46,** 175–185.

Anant, K.S. & Dowla, F.U., 1997. Wavelet transform methods for phase identification in three-component seismograms, *Bull. seism. Soc. Am.,* **87,** 1598–1612.

Baillard, C., Crawford, W.C., Ballu, V., Hibert, C. & Mangeney, A., 2014. An automatic kurtosis-based P- and S-phase picker designed for local seismic networks, *Bull. seism. Soc. Am.,* **104,** 394–409.

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

Dai, H. & MacBeth, C., 1995. Automatic picking of seismic arrivals in local earthquake data using an artificial neural network, *J. geophys. Res.,* **120,** 758–774.

Diehl, T., Kissling, E., Husen, S. & Aldersons, A., 2009. Consistent phase picking for regional tomography models: application to the greater Alpine region, *Geophys. J. Int.,* **176,** 542–554.

Gentili, S. & Michelini, A., 2006. Automatic picking of P and S phases using a neural tree, *J. Seismol.,* **10,** 39–63.

Keogh, E. & Kasetty, S., 2002. On the need for time series data mining benchmarks: a survey and empirical demonstration, in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 102–111.

Kuncheva, L. & Whitaker, C., 2003. Measures of diversity in classifier ensembles, *Mach. Learn.,* **51,** 181–207.

Küperkoch, L., Meier, T., Lee, J. & Friederich, W., 2010. Automated determination of P-phase arrival times at regional and local distances using higher order statistics, *Geophys. J. Int.,* **181,** 1159–1170.

Leonard, M. & Kennett, B.L.N., 1999. Multi-component autoregressive techniques for the analysis of seismograms, *Phys. Earth. planet. Int.,* **113,** 247–263.

Nikolov, S., 2012. Trend or no trend: a novel nonparametric method for classifying time series, *M.S. thesis,* Massachusetts Institute of Technology, Cambridge, MA, USA.

Patanè, D., Ferrari, F., Giampiccolo, E. & Gresta, S., 2003. A PC-based computer package for automatic detection and location of earthquakes: application to a seismic network in eastern Sicily (Italy), methods and applications of signal processing in seismic network operations, in *Lecture Notes in Earth Sciences,* Vol. 98, pp. 89–129.

Reading, A.M., Mao, W. & Gubbins, D., 2001. Polarization ffiltering for automatic picking of seismic data and improved converted phase detection, *Geophys. J. Int.,* **147,** 227–234.

Ross, Z.E. & Ben-Zion, Y., 2014. Automatic picking of direct P, S seismic phases and fault zone head waves, *Geophys. J. Int.,* **199,** 368–381.

Savvaidis, A., Papazachos, C., Soupios, P., Galanis, O., Grammalidis, N., Saragiotis, Ch., Hadjileontiadis, L. & Panas, S., 2002. Implementation of additional seismological software for the determination of earthquake parameters based on MatSeis and an automatic phase-detector algorithm, *Seismol. Res. Lett.,* **73**(1), 57–69.

Sleeman, R. & van Eck, T., 1999. Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings, *Phys. Earth. planet. Int.,* **113,** 265–275.

Takanami, T. & Kitagawa, G., 1993. Multivariate time-series model to estimate the arrival times of *S*-waves, *Comp. Geosci.,* **19,** 295–301.

Ueno, K., Xi, X., Keogh, E. & Lee, D.J., 2006. Anytime classification using the nearest neighbor algorithm with applications to stream mining, in *Proceedings of the Sixth International Conference on Data Mining*, pp. 623–632.

Wang, J. & Teng, T., 1997. Identification and picking of S-phase using an artificial neural network, *Bull. seism. Soc. Am.,* **87,** 1140–1149.

Wei, L. & Keogh, E., 2006. Semi-supervised time series classification, in *Proc. Knowledge Discovery in Databases II*, pp. 748–753.

Zhao, Y. & Takano, K., 1999. An artificial neural network approach for broadband seismic phase picking, *Bull. seism. Soc. Am.,* **89**(3), 670–680.