ELECTRONIC SEISMOLOGIST

PhasePApy: A Robust Pure Python Package for Automatic Identification of Seismic Phases

by Chen Chen and Austin A. Holland

ABSTRACT

We developed a Python phase identification package: the PhasePApy for earthquake data processing and near-real-time monitoring. The package takes advantage of the growing number of Python libraries including Obspy. All the data formats supported by Obspy can be supported within the PhasePApy. The PhasePApy has two subpackages: the PhasePicker and the Associator, aiming to identify phase arrival onsets and associate them to phase types, respectively. The PhasePicker and the Associator can work jointly or separately. Three autopickers are implemented in the PhasePicker subpackage: the frequency-band picker, the Akaike information criteria function derivative picker, and the kurtosis picker. All three autopickers identify picks with the same processing methods but different characteristic functions. The PhasePicker triggers the pick with a dynamic threshold and can declare a pick with false-pick filtering. Also, the PhasePicker identifies a pick polarity and uncertainty for further seismological analysis, such as focal mechanism determination. Two associators are included in the Associator subpackage: the 1D Associator and 3D Associator, which assign phase types to picks that can best fit potential earthquakes by minimizing root mean square (rms) residuals of the misfits in distance and time, respectively. The Associator processes multiple picks from all channels at a seismic station and aggregates them to increase computational efficiencies. Both associators use travel-time look up tables to determine the best estimation of the earthquake location and evaluate the phase type for picks. The PhasePApy package has been used extensively for local and regional earthquakes and can work for active source experiments as well.

Online Material: Figures illustrating the performance of the AICDpicker and the KTpicker. Some technical details of the PhasePApy are included.

INTRODUCTION

Modern seismic monitoring networks, with broadband and/or strong-motion seismometers, in increasing numbers globally, can easily produce large enough volumes of waveform data so that manual picking of phase arrivals by analysts becomes nearly impossible. The phase identification for a large amount of data is tedious work for a human being. Moreover, to some extent, manual analysis of seismograms and phase picking is subjective and may depend on different analysts who may introduce bias and inconsistencies in phase arrival data (e.g., Leonard, 2000). Therefore, phase identification will always be a critical issue, both for manual phase identification and for automatic algorithmic phase identification. Inaccurate and inconsistent phase arrival times can bias all subsequent analyses, such as earthquake location, travel-time tomography, and focal mechanism analysis. To analyze manual picking errors, Zeiler and Velasco (2009) performed two experiments to define and isolate phase-picking errors. They stated that the signal-to-noise ratio (SNR) is the main source of error for an individual analyst. Leonard (2000) conducted the comparison of 78 teleseismic phases manually picked by four analysts and three automatic picking algorithms to conclude that the average difference between the analysts is greater than the difference between analysts and an automatic picker.

There are many automatic phase arrival pickers based on different algorithms. Each kind of picker has its own advantages and limitations. Effectiveness and computational efficiency are two important measurements of an automatic algorithm. Many factors can affect these two measures of an automatic picker, such as data quality and algorithm simplicity. The SNR can dramatically affect the behavior of automatic phase pickers (e.g., Zeiler and Velasco, 2009). The majority of picking algorithms can be classified into one of the following methods: energy transient methods in time domain or frequency domain (e.g., Withers *et al.*, 1998; Vassallo *et al.*, 2012), autoregressive (AR) methods (e.g., Leonard and Kennett, 1999), high-order statistical methods (e.g., Baillard *et al.*, 2014), neural network methods (e.g., Gentili and Michelini, 2006), and wavelet transform methods (e.g., Bogiatzis and Ishii, 2015). The most popular picking algorithms compare the short-term average (STA) and long-term average (LTA) of the characteristic function (CF) of a signal to determine the phase arrivals; this is a

classic energy transient method (e.g., Allen, 1978; Baer and Kradolfer, 1987; Lomax *et al.*, 2012; Vassallo *et al.*, 2012). These algorithms trigger and declare picks when the ratio of STA to LTA exceeds a predefined or dynamic threshold value. The STA/LTA method is an important automatic picking algorithm for early warning systems, due to its simplicity and effectiveness. Withers *et al.* (1998) discussed several STA/LTA algorithms including classic STA/LTA, delayed STA/LTA, recursive STA/LTA, and Z-detector in time domain and power spectrum density in frequency domain.

Another picking method is based on autoregressive Akaike information criteria (AR-AIC) function (e.g., Akaike, 1974; Kitagawa and Akaike, 1978; Leonard and Kennett, 1999; Sleeman and van Eck, 1999; Zhang *et al.*, 2003). By AR approach, the seismogram window is divided into two different local stationary segments, which contain the noise and signal, respectively. Two segments of different statistical properties are modeled as an AR process. The AR-AIC picker can determine the optimal onset time by looking for the global minimum of the AIC function of a seismogram. For this reason, it is necessary to choose a window that only includes the segments of interest. Therefore, phase arrival identification is required to choose the appropriate window before using the AR-AIC method. The signal predetection processing will increase the computation time and reduce the efficiency.

Higher-order statistics, such as kurtosis, is used to identify seismic phases due to the non-Gaussian distribution attribute of seismic waves (e.g., Saragiotis *et al.*, 2002; Panagiotakis *et al.*, 2008; Baillard *et al.*, 2014; Hibert *et al.*, 2014). Artificial neural network methods are also used to search for phase onset by training the machine to recognize phase arrivals (e.g., Dai and MacBeth, 1995; Wang and Teng, 1997; Gentili and Michelini, 2006). In addition, wavelet transform is another approach for seismic signal analysis (e.g., Anant and Dowla, 1997; Akansu *et al.*, 2010; Bogiatzis and Ishii, 2015). The two main types of wavelet analysis, discrete and continuous wavelet transform, can be used to detect both *P* and *S* phases.

In this article, we introduce a software package to identify phase arrival times and associate the picks to the type of seismic wave. This software package is developed as a Python package, PhasePApy, to leverage the growing number of scientific libraries being written in Python, most notably Obspy (Beyreuther *et al.*, 2010). The choice of making our algorithms compatible with Obspy means that all data formats and access methods supported by Obspy are naturally supported; these include SEED, MiniSEED, SAC, SEGY, and others. The other libraries on which Obspy and our algorithms rely are NumPy (Oliphant, 2007), SciPy (Jones *et al.*, 2001), and Matplotlib (Hunter, 2007). The PhasePApy is separated into two different tasks, which are implemented in two subpackages: the PhasePicker and the Associator, aiming to pick phase arrivals and associate them to earthquakes, respectively. We implement three pickers based on different CFs: the frequency-band picker (FBpicker), the AIC function derivative picker (AICDpicker), and the kurtosis picker (KTpicker). We also implement two associators: the 1D Associator and 3D Associator. Users can choose the appropriate picker and associator that best meet their requirements.

The task of the PhasePicker is simply making picks of phase arrivals from seismograms. Each picker only has a small number of different parameters to input. The processing method FBpicker adapts to sampling rate and instrument gain systematically. The FBpicker algorithm is an algorithm modified from that of Lomax *et al.* (2012), which does not require extensive tuning (Vassallo *et al.*, 2012), because there is no need to specify triggering thresholds and other parameters for each analyzed channel. The AICDpicker and KTpicker determine the phase arrival times based on the derivative of the AIC function and kurtosis, respectively. Many picking algorithms pick and declare phase types by using specified thresholds, but the PhasePicker employs the dynamic threshold. The PhasePicker only detects phase arrivals and identifies the phase onsets but not phase types. Phase polarities and picking uncertainties can be determined as well by the PhasePicker. The picker itself can be set to make picks from global phase arrivals to near-field phase arrivals and reflections.

Because the PhasePicker can pick any number of phase types that have sufficient energy contrast in the filtered waveforms, it is necessary to have a robust approach to automatically associate the picks to seismic phase types. The task of the Associator is to associate the picks that best fit the different phase types associated with a particular earthquake. We developed two earthquake phase associators (fixed depth 1D velocity model associator and 3D velocity model associator) appropriate for local and regional earthquake monitoring, but these methods could be adapted to accommodate any distance. The Associator can provide estimation of event location and origin time for the determined earthquake events. The primary function is not to locate the earthquake, but to properly identify the phases associated with the earthquake.

PHASEPICKER ALGORITHMS

Characteristic Function 1: FBpicker

We implement three pickers for pick identification; the first is the FBpicker, which is a modified transient energy method from Lomax *et al.* (2012). Before any processing, the FBpicker removes the mean and trend of the data with a least-squares method. It is important to note that if the user is attempting to make picks for long-period phases, a sufficiently long window of data is required so that this initial processing does not affect the ability to identify phases. The FBpicker applies an octave filter (doubling the central frequency) to the seismograms and generates several frequency bands for each seismogram, which automatically adapt to instrumentation sampling rate. The central frequency of each consecutive band keeps doubling from a user-defined minimum band, which includes low-frequency components of interest, until the high corner of the next band of the last filtering band exceeds the Nyquist frequency (see Ⓔ Fig. S1, available in the electronic supplement to this article). The user can reduce the number of bands by decimating the data, which decreases the Nyquist frequency.

The FBpicker allows users to change the corner order of the filtering, which determines the attenuation for frequencies outside of the bandwidth. To reduce the Gibbs effect from the band-pass filtering, the filtered data of each band are cosine tapered and also user configurable; that is, users can modify the tapering percentage of the data.

Following the method of Lomax *et al.* (2012), the FBpicker calculates the energy E_n of the filtered data for each band

$$E_n[i] = BF_n[i]^2, \tag{1}$$

in which $BF_n[i]$ is the band-pass filtered amplitude of ith sample of the nth band. We implemented root mean square (rms) and standard deviation (st. dev.) mode to calculate the CF for each band in equations (2) and (3), respectively

$$\mathrm{CF}_n^{\mathrm{rms}}[i] = \frac{E_n[i]}{\mathrm{rms}(E_n[i-1-l:i-1])}, \tag{2}$$

$$\mathrm{CF}_n^{\mathrm{st.dev.}}[i] = \frac{E_n[i] - \mathrm{mean}(E_n[i-1-l:i-1])}{\mathrm{st.dev.}(E_n[i-1-l:i-1])}, \tag{3}$$

in which l is the window length in a sample. For rms mode, $\mathrm{CF}_n^{\mathrm{rms}}[i]$ is the ratio of the transient energy of sample i of frequency band n to the rms value of energy in the previous moving window; for st. dev. mode, $\mathrm{CF}_n^{\mathrm{st.dev.}}[i]$ is the ratio of difference between the transient energy of sample i and the mean of its previous window for frequency band n to the standard deviation of the energy in the previous window. The CF quantifies the energy change relative to the energy level in the previous window. Therefore, the high $\mathrm{CF}_n[i]$ indicates the high energy level of a sample compared to that in the previous window. Because of fewer numerical operations, the rms mode can improve the computation time compared to the st. dev. mode without significantly affecting the performance. The reason is that the trends of the data are removed in the data-conditioning phase. The rms mode reduces the computation time by about 30%, when computed on a 6 min segment of data. The FBpicker allows the user to customize the CF calculating mode (rms or st. dev.) depending on their requirements. Finally, the FBpicker summarizes the $\mathrm{CF}_n[i]$ by taking the maximum value over all bands n for each sample i (Lomax *et al.*, 2012) to obtain the CF.

Characteristic Function 2: AICDpicker

The AIC picker has been used in many P-wave picking algorithms for single or multiple component records by searching the global minimum of the AIC function calculated with the AR technique (e.g., Leonard and Kennett, 1999; Sleeman and van Eck, 1999; Zhang *et al.*, 2003). The AIC function of a seismogram of two signal segments can be represented as the function of division point (Sleeman and van Eck, 1999)

$$\mathrm{AIC}(P) = (P-M)\log(\sigma^2_{1,\max}) + (N-M-P)\log(\sigma^2_{2,\max}) + \mathrm{Const.}, \tag{4}$$

in which P is the division point, M is the order of the AR model, N is the total length of the data, and $\sigma^2_{1,\max}$ and $\sigma^2_{2,\max}$ indicate the variance of the seismogram in two segments. Another approach without using the AR method determines the AIC function directly from the seismogram (Maeda, 1985)

$$\mathrm{AIC}(P) = P\log\{\mathrm{var}(x[1,P])\} + (N-P-1)\log\{\mathrm{var}(x[P+1,N])\}, \tag{5}$$

in which division point P ranges over all the samples of the seismogram x. If the seismogram only contains the P wave, the AIC picker can easily determine the phase onset by looking for the global minimum of AIC function. However, if the seismogram also contains the S wave, it is difficult to identify the seismic phases by just searching for the global minimum of the AIC function. We take the absolute value of the first derivative AIC function as the CF for AICDpicker, because the first derivative of the AIC function is sensitive to the change of the AIC function.

Characteristic Function 3: KTpicker

The third picker in the PhasePicker is a kurtosis-based picker: the KTpicker. The kurtosis is a statistical measure to describe the degree of measurements concentration, which can be used as the CF of a seismogram. The kurtosis is high-order statistics of a variable and has been widely used in identifying the seismic waves phase onset (e.g., Panagiotakis *et al.*, 2008; Baillard *et al.*, 2014; Hibert *et al.*, 2014). We apply the kurtosis in a moving window through the time series to characterize the seismic signal. Kurtosis is defined as the standardized fourth moment about the mean of the measurements

$$K = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}, \tag{6}$$

in which E is the expectation operator, X is the measurements, μ is the mean of the measurements, μ_4 is the fourth-moment about the mean, and σ is the standard deviation. White noise is normally (Gaussian) distributed, the kurtosis of which is 3 (DeCarlo, 1997). Values of $K > 3$ result in a higher peak compared to the Gaussian distribution, whereas values < 3 result in lower peak than the Gaussian distribution. Therefore, the kurtosis characterizes the signal based on the shape of the distribution instead of the SNR, and as such, the kurtosis can work well for the signals with low SNR.

Triggering, Declaration, and False Pick Filtering

The PhasePicker identifies, or triggers, the pick by using a floating threshold level, which is determined by multiplying the rms of the CF in a moving window with a user-defined coefficient, instead of a fixed threshold. The floating threshold level may make the pick triggering independent of the noise level of CF

to some extent. This dynamic threshold level is a method often applied in statistical process control called control charts (e.g., Alwan and Roberts, 1988) (e.g., 6-sigma). The threshold is determined by multiplying the rms of the CF to a user-defined coefficient, which is roughly comparable to the number of standard deviations of the CF. The Associator allows the user to change threshold as a multiple of σ based on the requirements of each case. The PhasePicker triggers picks at the time when the CF exceeds the dynamic threshold. We design the PhasePicker with a rollback algorithm to ensure that picks are declared at the same time as much as possible for different threshold levels. The PhasePicker will roll backward the triggering time along the CF until it reaches the first local minimum, at which point a pick is declared. Without the rollback algorithm, there will be a few samples difference between higher and lower threshold levels. Therefore, pick declaration is roughly independent of the triggering threshold level. The SNR of CF (SNRCF) of each pick is determined by the ratio of the first local maximum CF after the pick to the rms of CF in the previous window. This feature can assist users in choosing the appropriate threshold level for the pick-triggering process.

The SNR of a signal on different stations or time can vary over a wide range. Usually, the stations close to an earthquake's epicenter have higher SNR than the distant ones. As would be expected, it is easier to trigger picks for channels with high SNR rather than the ones with low SNR. A lower dynamic threshold level can identify a greater number of smaller amplitude signals (indicated by low CF spikes), but false picks are more likely to be identified. A high threshold-triggering level can trigger high SNR signals and introduce fewer false picks, but may miss weak signals. Therefore, there is a trade-off between the threshold levels for triggering the picks. By tuning the threshold level, one can remove a significant number of false picks. This is different than the method of Lomax *et al.* (2012), where static thresholds are applied and do not change without reconfiguration and as such cannot account for changing noise levels due to things such as wind or cultural noise. From our experience, it is often possible to find a single dynamic threshold coefficient that is appropriate for a broad range of channels within a region.

A high dynamic threshold level can identify picks that correspond to phase with high SNR. However, this approach often discards significant picks for phase arrivals with lower SNR. If users choose a relatively low dynamic threshold level, we implement two optional pick-filtering algorithms that attempts to preserve as many real picks as possible and remove questionable picks. The first utilizes a close-pick cleaning filter, which does not allow a following consecutive pick to occur within a user-defined amount of time. The second filter implements a short-period noise-cleaning filter, which determines whether a pick may be false or not by comparing the σ in the previous and following window of a pick. The energy levels before and after phase onset are different, which can be used to determine whether the detected signal is a false or correct earthquake phase. The short-period noise-cleaning filter removes the pick if the σ of the waveform amplitude in a short preceding window multiplied by a user-defined coefficient is greater than the σ of the waveform amplitude in the window after the pick. Otherwise, the pick is considered a correct phase pick. In the case that another pick is falling in the previous or following window, the window size will automatically shrink to the interval between two picks. Although it is possible that some of the remaining picks are still false, after both false-pick-cleaning filters are applied, a significant number of false picks can be filtered out.

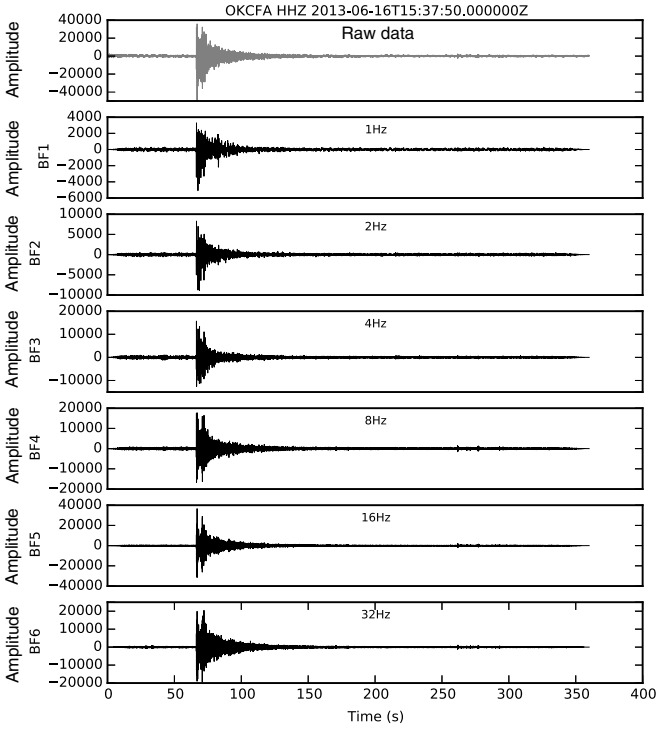
Polarity and Uncertainty

The PhasePicker includes a method to determine the phase polarity or direction of first motion as compression, dilatation, or uncertain. Human beings determine the phase polarity by visually inspecting the difference between the amplitude of the phase and background noise level. If an analyst considers the amplitude large enough, he or she can assign the direction of first motion. If the amplitude is approximately equivalent to the noise level, it is difficult to identify the first arrival and the polarity of the phase. The PhasePicker determines the polarity in a similar way by comparing the phase amplitude to the noise level. When the pick is declared, the PhasePicker searches for a local maximum or minimum value along the waveform after the triggering time. The amplitude of the pick is determined as the difference between the average in the pick epoch and the identified extreme value. If the absolute value of the pick amplitude is greater than σ of the seismic waveform in a previous window times a user-defined coefficient, the polarity will be declared.

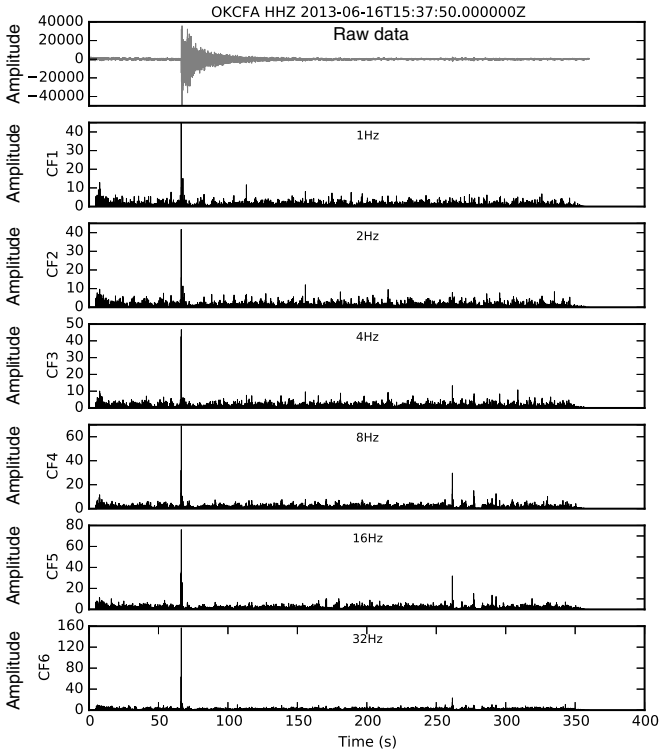
The uncertainty of a pick is not easy to determine, because it is not easy to distinguish how much of the true seismic phase is buried in the noise. The SNR is a significant factor that affects the accuracy of the uncertainty determination. We employ the rms of the CF before the declared picks and a user-defined coefficient to evaluate the picking uncertainty. The uncertainty-triggered picking time $t^{\text{uncert}}_{\text{triggered}}$ is obtained by rolling forward the declared time t_{declared} along the CF until the value of the CF exceeds the rms times the user-defined coefficient. The PhasePicker determines the uncertainty as the time difference between declared pick t_{declared} and uncertainty-triggered pick $t^{\text{uncert}}_{\text{triggered}}$.

RESULTS OF PHASEPICKER

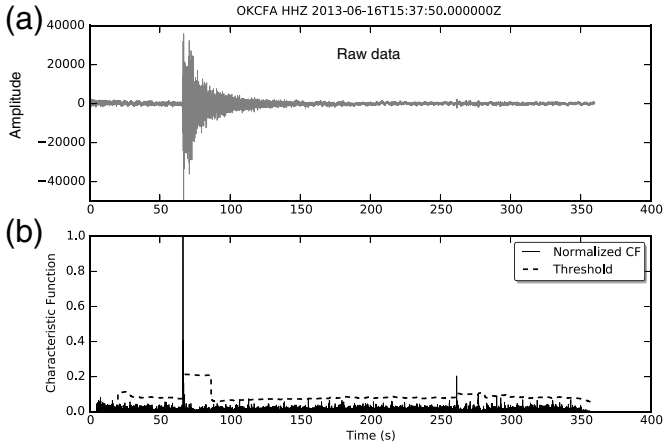
We take a 6 min window of data as an example to show the implementation and performance. The example seismogram is from a magnitude 2.7 earthquake that occurred in central Oklahoma on 16 June 2013, which is recorded on the vertical component of the OKCFA station in the Oklahoma regional network (OK). In this example, we use the FBpicker to demonstrate seismic phase arrival picking performance. The AICDpicker and KTpicker have similar picking performance. The sampling rate of the seismogram is 100 Hz. By using 1 Hz as the central frequency of the starting band, the FBpicker determines six bands with central frequency at 1, 2, 4, 8, 16, and 32 Hz, respectively (Fig. 1). In this example, we use the rms and



▲ **Figure 1.** A detrended example seismogram (top panel) from vertical component of station OKCFA in Oklahoma regional network (OK) is octave-filtered with six determined bands (BF_1–BF_6) from the FBpicker. The sampling rate of the data is 100 Hz. The central frequency of each band is labeled in the middle of each panel. A 10% tapering is applied to all bands. Amplitude is in counts.



▲ **Figure 2.** The detrended seismogram on vertical channel from OKCFA (top panel) and the FBpicker determined the CF_n (CF_1–CF_6) with root mean square (rms) mode. Amplitude is in counts.

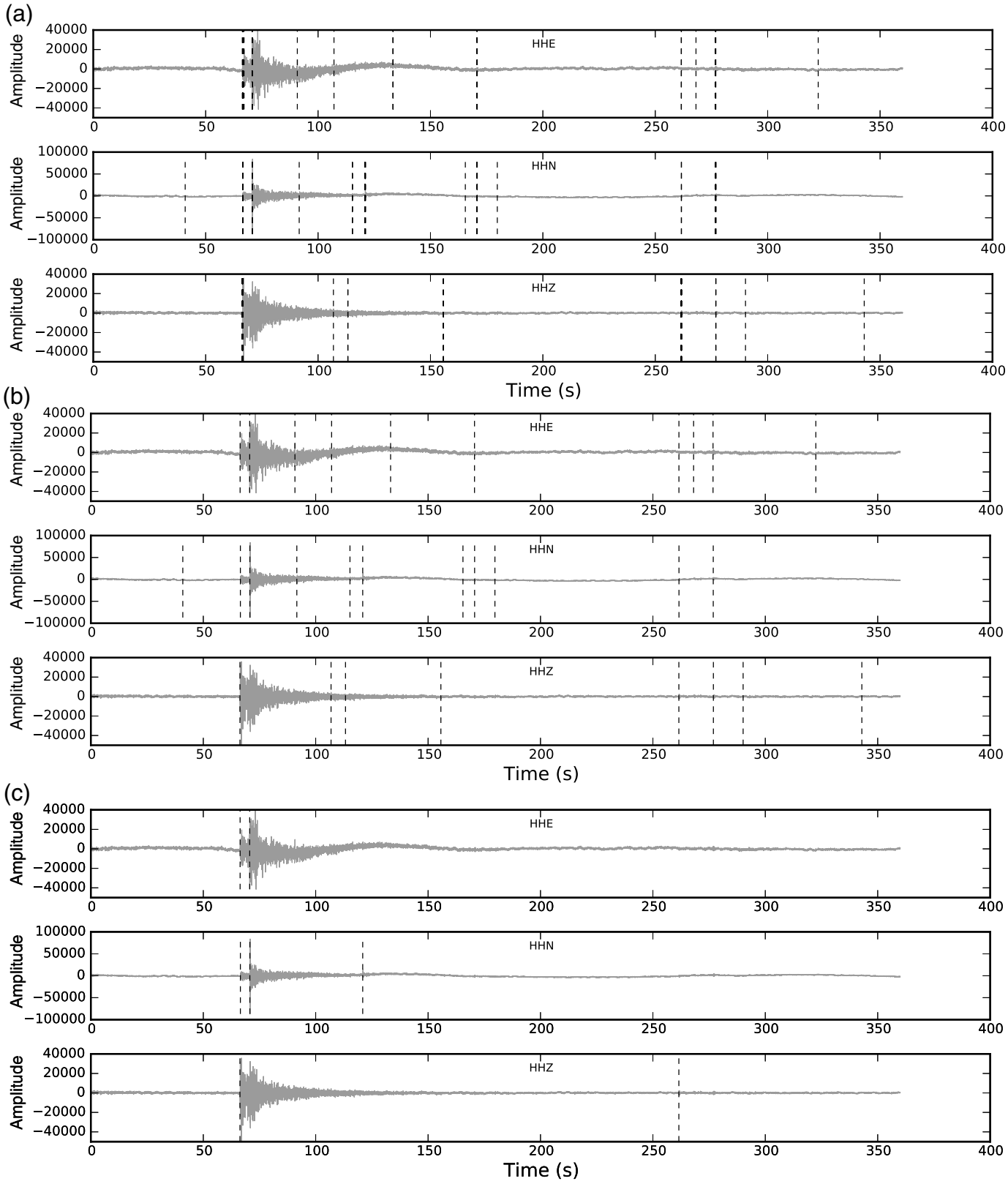


▲ **Figure 3.** (a) The detrended seismogram on vertical channel from OKCFA. (b) The FBpicker summarizes and normalizes the characteristic function (CF) (solid line), and the dynamic threshold level of 6 is the dashed line. Amplitude is in counts.

a 5-s moving window to calculate the CF for each band (Fig. 2). The strongest signal frequency component of this example is in the highest frequency band (32 Hz), which suggests that the seismogram corresponds to a local earthquake with higher-frequency components.

Figure 3 demonstrates the behavior of the CF, which is normalized by the maximum absolute value, and the dynamic threshold with 20 s of the moving window and 6σ of dynamic threshold. We assign a value of zero for the very beginning windows (5 and 20 s, respectively) of the normalized CF and dynamic threshold. In this example, the threshold of 6σ can identify not only the real phase arrivals, but also introduce some false triggering picks (Fig. 4a). By tuning up the dynamic threshold level, such as 8 or even higher, most false picks can be removed. To demonstrate the removal of false picks from each component by applying the two cleaning filters, we use the window length of 0.78 s for the close-pick-cleaning filter to remove close picks, but considerable false picks still remain (Fig. 4b). Then, we use the window length of 2 s and the coefficient of 2 as the short-period noise-cleaning filter setup. After applying the short-period noise-cleaning filter, most of the false picks are filtered out (Fig. 4c). The last picks on component N and Z in Figure 4c are false picks that have not been removed. They could possibly be from some microearthquakes or high-frequency cultural noise.

The FBpicker determines the polarity for declared picks (see Ⓔ Fig. S2). We also show the normalized CF of the AICDpicker and KTpicker in Ⓔ Figures S3 and S4, but we only use picks from the FBpicker in the following processing



▲ **Figure 4.** The FBpicker false pick filtering on east (E), north (N), and vertical (Z) components. (a) All declared picks without applying false picks filters, (b) only applying the close-pick cleaning filter, (c) applying both close-pick cleaning filter and short-period noise-cleaning filter. Amplitude is in counts.

demonstration. In these test data from vertical component of OKCFA, the picks from different automated picks are almost identical, with only a few samples difference. In our experience, the SNRCF of the declared picks is particularly useful information for choosing the appropriate coefficient for the dynamic threshold level, which can significantly affect the picking performance. By evaluating the SNRCF values of picks, the user will have a clue as to what threshold level can trigger the picks for actual phase arrivals and avoid most false picks. The Associator can remove the remaining false picks by associating correct picks to the best-fitting earthquakes.

ASSOCIATOR ALGORITHM

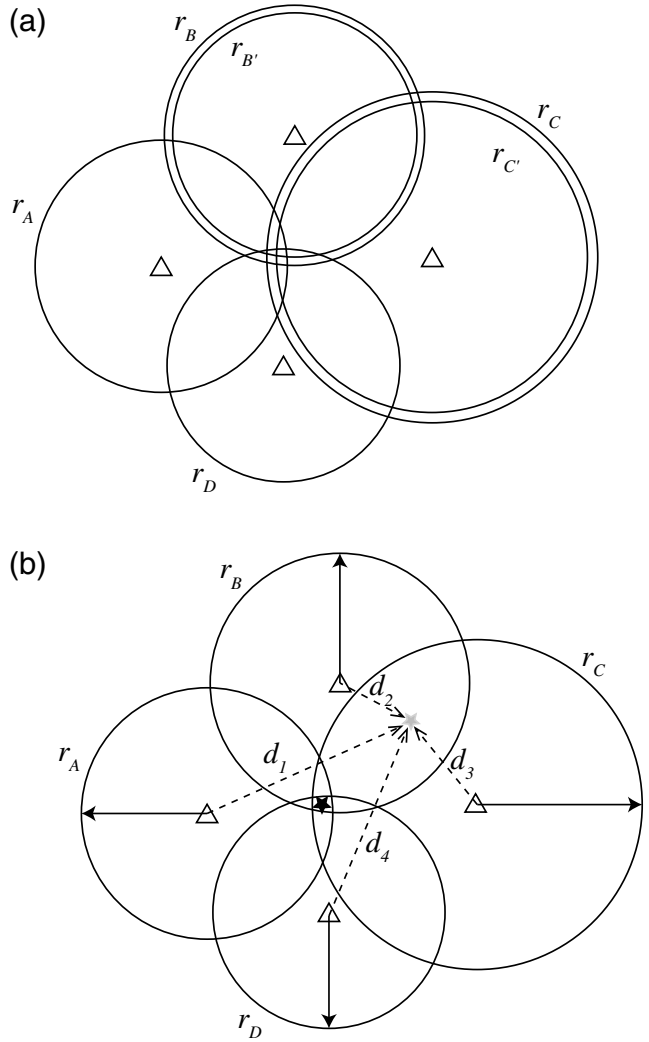
The Associator uses identified picks and associates the picks with earthquakes using the following steps: pick aggregation, event candidate creation, origin time analysis, and phase association. There are two associators in the subpackage: the 1D Associator and the 3D Associator. Both associators are using travel-time look-up tables for *P* and *S* waves to determine earthquake location and to associate phases in this case for locally-to-regionally recorded earthquakes. The picks to associate can come from the PhasePicker, other picking algorithms, manual picks, or other sources.

Step 1. The Associator aggregates picks from different components on the same station if the seismometer has multiple components. Before aggregating picks, the Associator first stacks picks from different components on the same station (see Ⓔ Fig. S5). Ideally, *P* and *S* waves are only identified on vertical and horizontal components, respectively. However, in some cases, *P* and *S* waves are picked on the other components, respectively. Pick stacking without differentiating components can make use of as many identified picks as possible for the further derivative analysis, such as tomography studies. The additional picks may allow for future advancements of phase association or further tuning of the picking algorithms. Automatic picks may occur on multiple channels at a seismic station and may have slightly different times. It is necessary to group the picks from different components of the same station. If the picks from different components have a time separation less than a user-defined pick link length, the picks are linked to create what we call modified picks. The Associator uses the minimum *S*–*P* time of the travel-time table multiplying a user-defined coefficient to determine the link length. The picks can be aggregated with either mean or median statistics mode, which takes the average value and median pick of the linked picks, respectively. Therefore, the aggregated pick is referred to as the modified pick. All the raw picks are associated with the modified picks for further processing. Aggregating picks can reduce the number of picks necessary for the phase association and improves the computing efficiency.

Step 2. The Associator creates event candidates for every modified pick pair from the same station. The candidate pick pair separation (*S*–*P* interval in time) can determine *S*–*P* interval in distance (a measure of distance from the earthquake epicenter to a seismometer, *S*–*P* interval in distance refers to the epicentral distance for all the following cases in the 1D Associator) by looking up the travel-time table. With the determined *S*–*P* interval in distance, the Associator calculates the origin time for the candidate by backprojecting the *P*-wave travel time for the determined *S*–*P* distance. Each earthquake only has a single origin time; there must be some event candidates from different stations having similar origin times within some amount of uncertainty. Well-recorded or large earthquakes will naturally have a larger number of event candidates with origin times close to the actual earthquake origin time. The Associator is able to only process candidates with the *S*–*P* interval less than a certain user-configurable cutoff distance to improve processing efficiency and force the regional restriction we chose.

Step 3. The Associator analyzes the origin time clusters after all the event candidates are created. With a user-defined moving window through time, the Associator counts the number of unassociated event candidates with origin times falling within the moving window range. The origin time cluster analysis may roughly estimate when an earthquake occurred if the number of origin times in the cluster is greater than a predefined value. The minimum predefined value is 3, because triangulation for earthquake location requires at least three stations.

Step 4. The Associator determines the location for event candidate cluster by the count from high-to-low and associates seismic phases to the unassociated picks. The Associator selects the first target cluster from all the candidate clusters that has the greatest number of event candidates. Once the event candidates in the first target cluster are used to associate to an earthquake, they will not be used in the cluster analysis for the second target cluster. Then, the Associator reruns the cluster analysis to determine the second target cluster among all the remaining clusters, which has the greatest number of event candidates. An accurate earthquake location enhances pick association, although locating the earthquake is not the final objective of the Associator. The objective of the earthquake location algorithm in the Associator is to determine which pick pairs can fit the earthquake location best, and allow for picks not associated into pairs to possibly be identified as a phase arrival. The basis of the location algorithm in the Associator is to find a location with minimized rms of misfits between *S*–*P* interval (in distance or epicentral distance for 1D Associator and in time for 3D Associator) and potential selected event candidates. Because any automatic phase picking algorithm may introduce false picks, a false pick close to a correct pick can introduce a false candidate with its origin time close to the one of the real events. In our package, one can reduce the false picks by tuning the PhasePicker, but meanwhile it is possible to miss the weak real seismic phase picks. If applying low-dynamic threshold level, one may introduce false picks, which cause the false candidates. Because of the false candidates, two or even more event candidates from the same station may occur in an origin time cluster. The event candidates with close origin times can be viewed as concentric circles in a 2D perspective (Fig. 5a). The Associator divides the cluster with



▲ **Figure 5.** (a) Six event candidates in the origin time cluster, which are A, B, B′, C, C′, and D with the determined radius: r_A, r_B, $r_{B'}$, r_C, $r_{C'}$, and r_D. Event candidate B and B′ are from the common station. Event candidate C and C′ are from another common station. There are four combinations to create the subcluster with event candidates only from different station: $S_1 = [A, B, C, D]$, $S_2 = [A, B', C, D]$, $S_3 = [A, B, C', D]$, and $S_4 = [A, B', C', D]$. Triangles represent the stations from which the event candidates come. (b) After excluding concentric circles, location determined by minimizing the rms of the misfits between the *S*–*P* interval in distances r_i and the distances d_i from stations to trial epicenter (gray star). The final determined location is around the circles crossing area (black star).

concentric circles into several subclusters with event candidates from unique stations (no event candidates from the common station), calculates the rms residuals for all the subsets, and determines the subset that minimizes rms residuals as the best solution. In this way, the best-fitting location can be determined, and false candidate(s) can be identified. The Associator includes an outlier evaluation function to determine whether an event candidate is an outlier by evaluating the residual. If the residual of an event candidate is greater than a user-defined parameter, the Associator will remove the candidate from the cluster as an outlier and repeat the location determination process until no outlier appears. Finally, in order to control the processing quality, we design a user-configurable rms threshold and set the minimum number of observations to declare an earthquake. An earthquake is declared only if the rms of misfits is less than a threshold, and the number of observations is greater than the minimum observation number.

Step 5. After all the earthquakes are identified, and all *P*- and *S*-phase arrival pairs are associated, it is still possible that individual modified picks have not been associated with any earthquakes. Modified picks that do not have a *P*- and *S*-phase arrival pair may also represent phase arrivals associated with the associated earthquakes. These unassociated modified picks are then evaluated to determine whether the picks may be a single-phase arrival for an earthquake. The travel time of an individual single phase is calculated for phase arrivals in the travel-time look-up table based on the distance between the estimate of earthquake location and the seismometer. It is then possible to determine if the modified pick corresponds within uncertainty to the origin time of an identified earthquake. If an individual modified pick can be associated to an earthquake, then all its aggregated raw picks are associated to the earthquake with the appropriate phase.

Both the 1D Associator and 3D Associator take advantage of the fact that the PhasePicker does well at picking both *P*- and *S*-phase arrivals for local and regional earthquakes. For the 1D Associator, the location algorithm determines the hypocenter of an earthquake by searching the best estimation that minimizes the rms residuals between the *S*–*P* intervals in distance and the distance from the trial epicenter to stations

$$R = \sqrt{\frac{\sum_{i=1}^{N}(r_i - d_i)^2}{N}}, \tag{7}$$

in which N is the number of stations. A location determination example with four event candidates is shown in Figure 5b. With an initial trial epicenter, the 1D Associator iterates the location search until the best estimation of hypocenter is determined. The final location will be close to the intersection of the circles (black star in Fig. 5b). The location uncertainty of the 1D Associator is determined as the rms residuals of distance

$$U_{\mathrm{loc}} = \sqrt{\frac{\sum_{i=1}^{N}(r_i - D_i)^2}{N}}, \tag{8}$$

in which D_i is the distance from each station i to the final determined earthquake location.

For the 3D Associator, the hypocenter is determined by using a grid-search algorithm, minimizing rms residuals between the predicted travel times t_i and the *S*–*P* intervals in time t_i^{S-P}

$$R = \sqrt{\frac{\sum_{i=1}^{N}(t_i^{S-P} - t_i)^2}{N}}, \quad (9)$$

in which N is the number of stations. The 3D Associator uses a pyramid grid-search algorithm. The pyramid search is a search scheme breaking the searching grids into several sub-blocks (at most eight sub-blocks in one search iteration) and finding the one with the least rms of travel-time misfits. The identified sub-block that minimizes the rms is then broken up and grid searched for the node that minimizes the rms residuals until the sub-block cannot be split (less than or equal to three grids in each dimension). Finally, we employ a fine search to the node from the last iteration of pyramid search in order to identify the node with the global minimum rms residuals. The fine search compares the rms residuals of all the surrounding nodes to the central node. If the node of minimum rms residuals is in surrounding nodes, the fine search process is recentered with the new node and continues comparing the rms residuals of the central node to the surrounding nodes until the node of minimum rms residuals is in the center. The location uncertainty in the case of the 3D Associator is determined as the rms residuals of travel time

$$U_{\mathrm{loc}} = \sqrt{\frac{\sum_{i=1}^{N}(t_i^{S-P} - T_i)^2}{N}}, \quad (10)$$

in which T_i is the predicted travel time from each station i to the final determined earthquake location. The grid size used for the 3D velocity model can affect the uncertainty, because dense girds will improve the resolution of location search but dramatically increase search times required to associate phases to an earthquake.

The location determination algorithms for the 1D Associator and the 3D Associator are based on minimizing rms misfits of S–P interval in distance (hypocentral distance) and S–P time, respectively. The hypocentral distance relates the S–P time with the known velocity model. Therefore, the 1D Associator and the 3D Associator determine the earthquake location with the same method, but by minimizing different rms misfits.

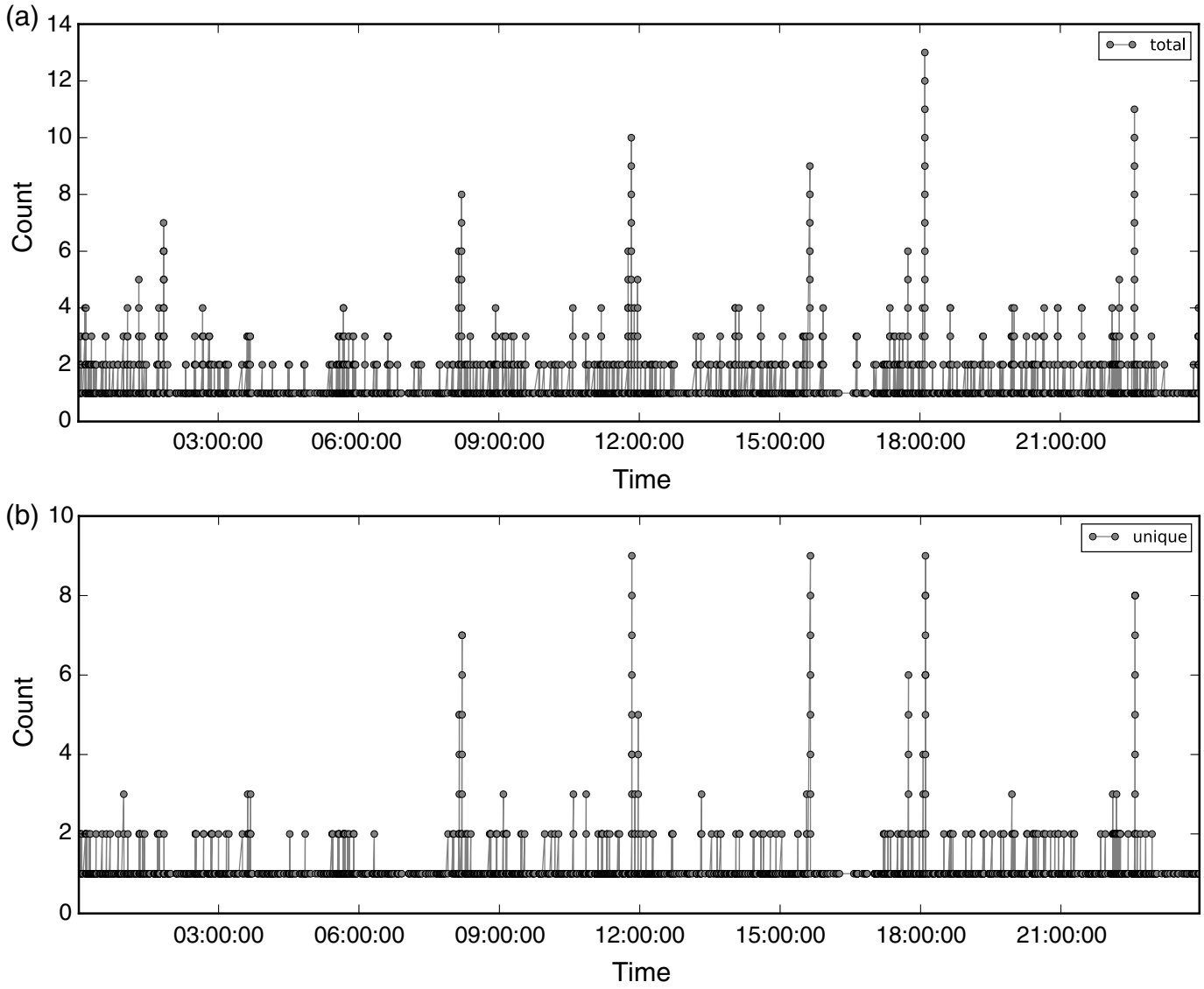
RESULTS OF ASSOCIATOR

We choose minimum S–P time (0.78 s) to aggregate the picks for the 1-day data test. The median mode is chosen in the picks aggregation for our example, so that outliers will not bias our results. There are 6023 picks identified, and 4832 modified picks remain after pick aggregation, which indicates that 1191 (~20%) picks are aggregated. Based on the dimensions of our study area, we use 350 km as the cutoff distance to create an event candidate. A window length of 7 s is selected to process 1 day of data on 16 June 2013. A statistical result of the origin times falling in the moving window is shown in Figure 6a, including candidates from the common station (concentric circles in Fig. 5a). Figure 6b demonstrates the cluster analysis with concentric circles excluded, and clusters with counts greater than three probably indicate earthquakes. One of the events, which occurred at 15:38:50 on 16 June 2013, is located as an example shown in Figure 7. The star and its surrounding circle (very small: ~0.03°) indicate the epicenter and the location uncertainty of the earthquake. The example in Figure 7a indicates that the earthquake is a well-recorded event that occurred in central Oklahoma. The 1D Associator determined S–P intervals could match the modeled travel-time curve well in the subplot to the left. The cross-section plot (Fig. 7b) shows the waveforms and associated phases from all the stations in Figure 7a. The short bars are the associated picks, and the gray dots indicate where the bars cross the waveforms, which match the modeled P and S curves quite well. The 1D Associator assigns picks matching the upper and lower curve as P and S phase, respectively.

The same earthquake is processed with the 3D Associator shown in Figure 8. We use 30, 40, and 4 grids in x, y, and z directions, respectively. Grid spacing is 0.1° for both latitude and longitude, and 3 km for depth. In the initial pyramid searching stage, grids in each direction are split into two segments, thereby having eight sub-blocks for the first search. The 3D Associator determines the deepest central node of the southeastern sub-block as the best solution for the first search (hexagon in Fig. 8a and the node depth in Fig. 8b). The pyramid search continues breaking up the identified sub-block and searching the central node with the minimum rms residuals for each iteration (pentagon, square, and inverted triangle in Fig. 8). The depths of the determined node in the pyramid search are indicated in Figure 8b to the right with corresponding colors. When the pyramid search finished, the fine search keeps comparing the rms residuals of all surrounding nodes to the central node for several iterations to find the node with global minimum rms residuals (circles in Fig. 8a and 8b, respectively).

CONCLUSIONS

We designed a Python automatic phase identification package: PhasePApy. The PhasePApy consists of two subpackages: the PhasePicker and the Associator, which can identify picks and determine phase types, respectively. The PhasePicker is designed to automatically detect the seismic-wave phase arrival times. Three different pickers: the FBpicker, AICDpicker, and KTpicker, are implemented in PhasePicker. The CFs for these three pickers are based on transient energy ratio, derivative of AIC function, and kurtosis, respectively. We employ the dynamic threshold method based on statistical process control (e.g., Alwan and Roberts, 1988) in the PhasePicker for pick triggering to automatically adapt to instrumentation differences. The threshold level is determined based on both the noise level of the CF and a level control coefficient, which simplifies picker tuning over the FilterPicker (Lomax *et al.*, 2012; Vassallo *et al.*, 2012). To ensure that the floating triggering threshold does not affect the declared picks, we implement a rollback algorithm to roll backward the triggering point along the CF to search for the local minimum to declare a pick. The

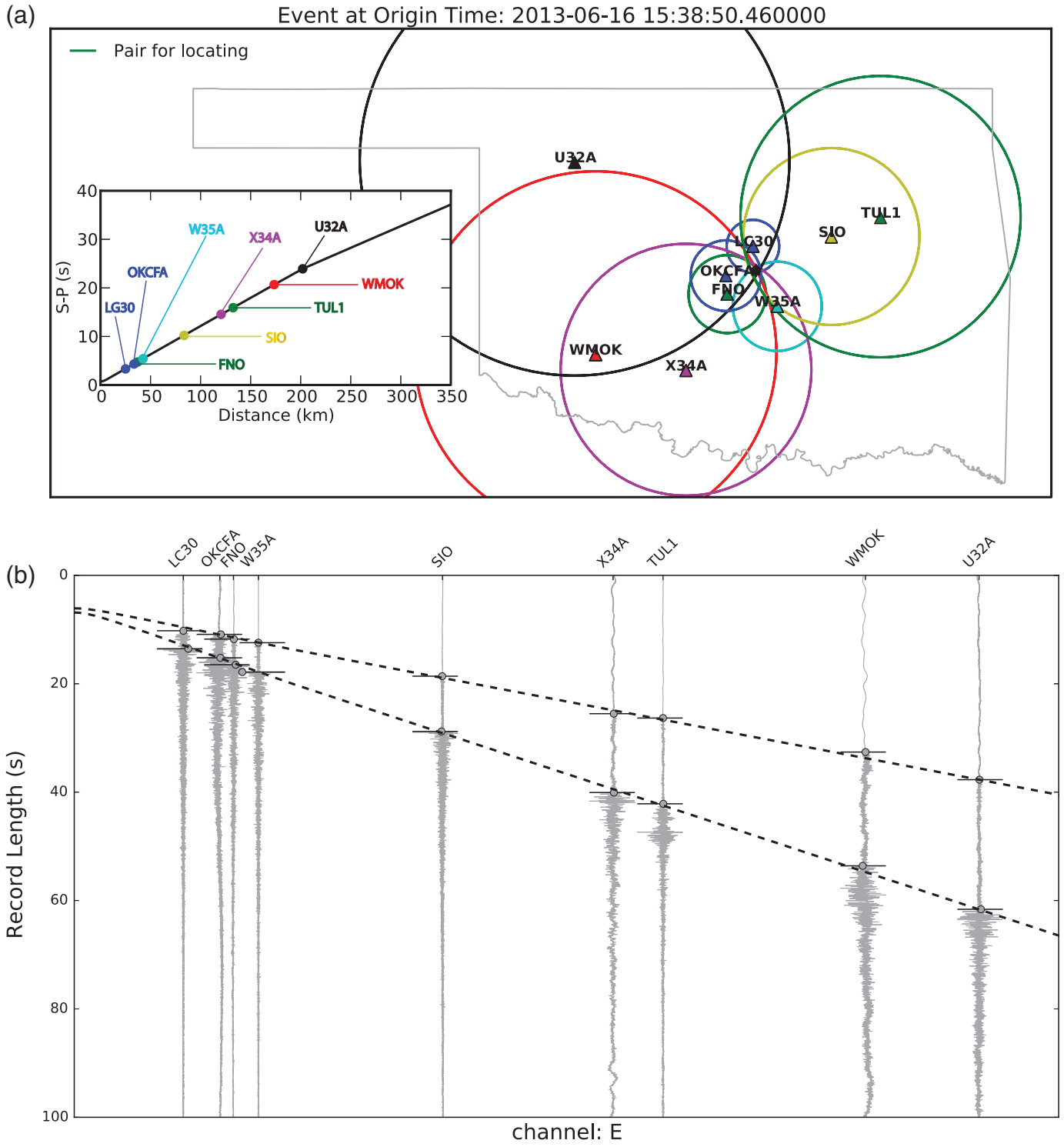


▲ **Figure 6.** Cluster analysis of 1-day data on 16 June 2013. (a) Clusters including concentric circles and (b) clusters excluding concentric circles.

PhasePicker has functions to determine the pick's polarity and uncertainty, which are designed for other processing purposes, such as focal mechanism analysis. The Associator identifies the phase types by searching for the picks that can best fit the determined earthquakes. Both 1D Associator and 3D Associator aggregate the picks from different components on the stations to modify picks in order to improve computation efficiency and make use of as many picks as possible. By creating and back projecting the event candidates, the Associator conducts cluster analysis to search for the potential origin time cluster, in which event candidates have close origin times. The Associator evaluates rms residuals to determine which picks can associate to the earthquakes.

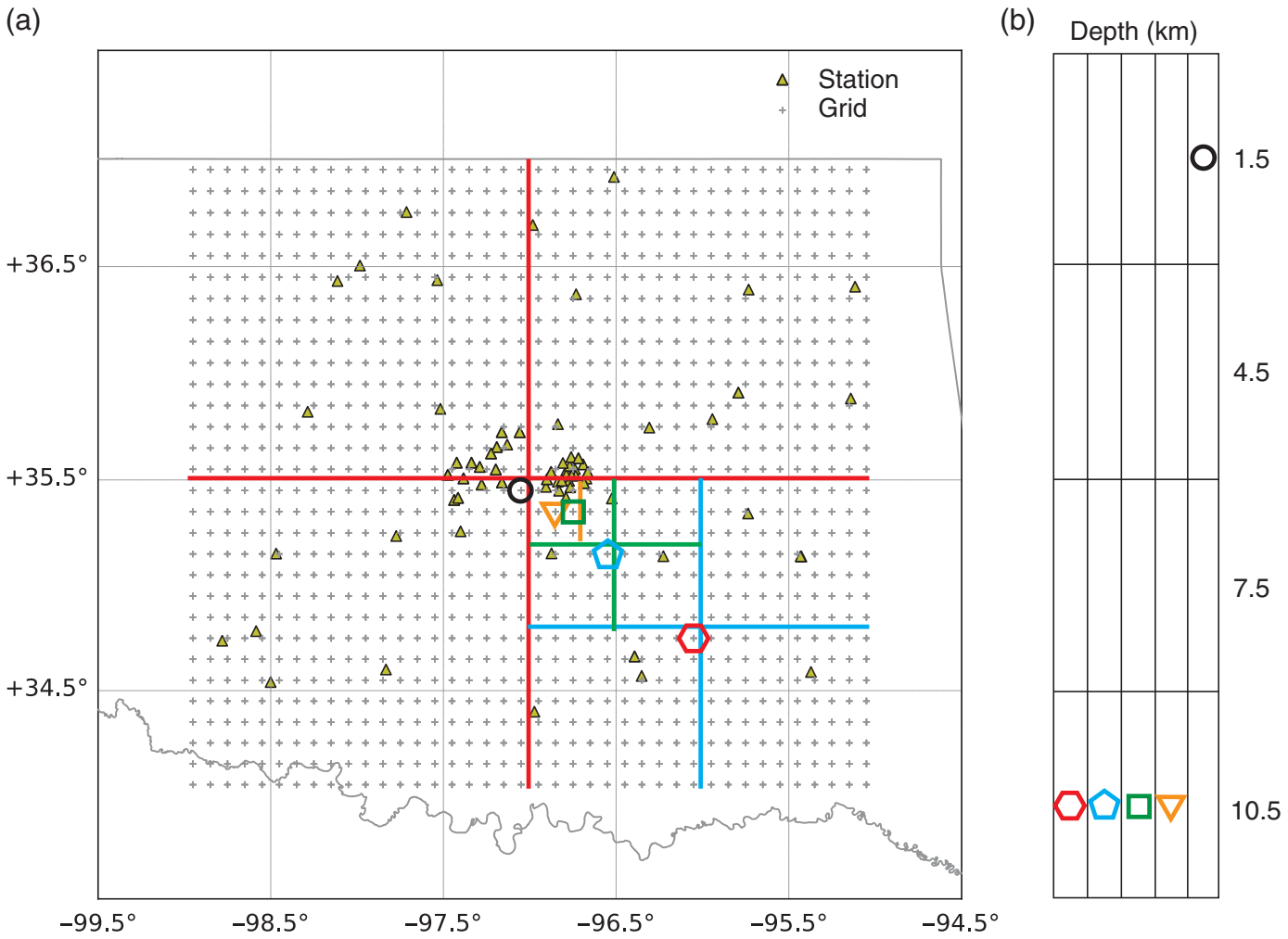
There are only a few parameters to tune in the PhasePApy package. All parameters of the picker and associator are intuitive and easy to set. Many phase-picking algorithms can only identify *P*-wave phase arrival times. One advantage of our seismic phase identification system is that it has the capability to identify both *P*- and *S*-phase arrival times. Moreover, other phase types with sufficient SNR can be recognized by adapting this method to accommodate any distance. In addition, the PhasePApy can process and integrate the new incoming data with existing data, which can improve the earthquake location, origin time, magnitude, number of observed first motion of *P* waves, and so can help further seismological analysis, such as tomography and focal mechanism studies. Depending on users' requirements, the PhasePicker and the Associator can work separately or jointly.

Currently, PhasePApy is not designed to run in real time, but it can be utilized in a near-real-time environment such as that employed for autolocations and event notification at the Oklahoma Geological Survey (OGS). An older version of the



▲ **Figure 7.** Seismograms from an earthquake that occurred at 15:38:50 on 16 June 2013. (a) Map view of the event location determination. The triangles are the stations. Circles are *S*–*P* interval in distance of event candidates. The star in the crossing area indicates the epicenter. The subplot shows the modeled *S*–*P* time curve and the 1D Associator determined *S*–*P* times for event candidates. (b) Cross-section plot of the FBpicker and 1D Associator performance on channel E. Two dashed lines are phases travel-time curves (upper, *P*; lower, *S*). Short bars indicate the associated picks from the stations and gray dots where the bars cross waveforms.

PhasePApy package is being used to monitor the earthquakes in near-real time for OGS. The OGS, currently, automatically identifies and locates more than 1000 earthquakes each month in Oklahoma and surrounding regions. We compared the events identified by the PhasePApy to those manually identified, and the method presented here can automatically identify



▲ **Figure 8.** Grid search for the hypocenter of the same earthquake in Figure 7a. (a) Map view of the searching process. Gray crosses, searching grids; triangles, seismic stations; thicker straight lines, sub-block boundaries; polygons and circles, the determined grid with least rms residuals in each search. (b) Depths of the grids. The polygons and circles are consistent with the ones in (a).

nearly all earthquakes of magnitude 2.0 or greater by using just a subset of the available stations. The PhasePApy can be applied beyond earthquake phase association and has been seen to perform well for active source experiments. The PhasePApy can clearly identify phase onsets and associate those phases to the active source. In addition, the inclusion of the 3D Associator makes the use of the PhasePApy appropriate for microseismic monitoring of well stimulations, as well as other applications. The PhasePApy should work well for processing microseismic data independent of the surface or down-hole geometry.

DATA AND RESOURCES

The data from the Oklahoma Geological Survey OK network are archived. The PhasePApy package is available on GitHub at https://github.com/austinholland/PhasePApy (last accessed June 2016).

ACKNOWLEDGEMENTS

Funding for this project is provided by Research Partnership to Secure Energy for America (RPSEA) through the "Ultra-Deepwater and Unconventional Natural Gas and Other Petroleum Resources" program authorized by the U.S. Energy Policy Act of 2005. RPSEA (www.rpsea.org) is a nonprofit corporation whose mission is to provide a stewardship role in ensuring the focused research, development and deployment of safe and environmentally responsible technology that can effectively deliver hydrocarbons from domestic resources to the citizens of the United States. RPSEA, operating as a consortium of premier U.S. energy research universities, industry, and independent research organizations, manages the program under a contract with the U.S. Department of Energy's National Energy Technology Laboratory. We would like to particularly thank the Oklahoma Geological Survey for providing the sustained financial support on this work and G. Randy Keller for

providing helps and reviews to improve this work. The authors are grateful to anonymous reviewers, the editors, and reviewers from U.S. Geological Survey: Ole Keven, Michelle Guy, Daniel McNamara, and Jill McCarthy for their valuable thoughts and suggestions.

REFERENCES

Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes, *Ann. Inst. Stat. Math.* **26,** no. 1, 363–387.

Akansu, A. N., W. A. Serdijn, and I. W. Selesnick (2010). Emerging applications of wavelets: A review, *Phys. Comm.* **3,** no. 1, 1–18, doi: 10.1016/j.phycom.2009.07.001.

Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces, *Bull. Seismol. Soc. Am.* **68,** no. 5, 1521–1532.

Alwan, L. C., and H. V. Roberts (1988). Time-series modeling for statistical process control, *J. Bus. Econ. Stat.* **6,** no. 1, 87–95, doi: 10.2307/1391421.

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

Baer, M., and U. Kradolfer (1987). An automatic phase picker for local and teleseismic events, *Bull. Seismol. Soc. Am.* **77,** no. 4, 1437–1445.

Baillard, C., W. C. Crawford, V. Ballu, C. Hibert, and A. Mangeney (2014). An automatic kurtosis-based *P*- and *S*-phase picker designed for local seismic networks, *Bull. Seismol. Soc. Am.* **104,** no. 1, 394–409, doi: 10.1785/0120120347.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010). ObsPy: A Python toolbox for seismology, *Seismol. Res. Lett.* **81,** no. 3, 530–533, doi: 10.1785/gssrl.81.3.530.

Bogiatzis, P., and M. Ishii (2015). Continuous wavelet decomposition algorithm for automatic detection of compressional- and shear-wave arrival times, *Bull. Seismol. Soc. Am.* **105,** no. 3, 1628–1641, doi: 10.1785/0120140267.

Dai, H., and C. MacBeth (1995). Automatic picking of seismic arrivals in local earthquake data using an artificial neural network, *Geophys. J. Int.* **120,** no. 3, 758–774, doi: 10.1111/j.1365-246X.1995.tb01851.x.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis, *Psychol. Meth.* **2,** no. 3, 292–307, doi: 10.1037/1082-989X.2.3.292.

Gentili, S., and A. Michelini (2006). Automatic picking of *P* and *S* phases using a neural tree, *J. Seismol.* **10,** no. 1, 39–63, doi: 10.1007/s10950-006-2296-6.

Hibert, C., A. Mangeney, G. Grandjean, C. Baillard, D. Rivet, N. M. Shapiro, C. Satriano, A. Maggi, P. Boisser, V. Ferrazzini, and W. Crawford (2014). Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano, *J. Geophys. Res.* **119,** no. 5, 1082–1105, doi: 10.1002/2013JF002970.

Hunter, J. D. (2007). Matplotlib: A 2D graphics evironment, *Comput. Sci. Eng.* **9,** no. 3, 90–95, doi: 10.1109/MCSE.2007.55.

Jones, E., E. Oliphant, and P. Peterson (2001). *SciPy: Open source scientific tools for Python*, available at http://www.scipy.org (last accessed January 2016).

Kitagawa, G., and H. Akaike (1978). A procedure for the modeling of non-stationary time series, *Ann. Inst. Stat. Math.* **30,** no. 1, 351–363.

Leonard, M. (2000). Comparison of manual and automatic onset time picking, *Bull. Seismol. Soc. Am.* **90,** no. 6, 1384–1390, doi: 10.1785/0120000026.

Leonard, M., and B. T. N. Kennett (1999). Multi-component autoregressive techniques for the analysis of seismograms, *Phys. Earth Planet. In.* **113,** no. 1, 247–263, doi: 10.1016/S0031-9201(99)00054-0.

Lomax, A., C. Satriano, and M. Vassallo (2012). Automatic picker developments and optimization: FilterPicker: A robust, broadband picker for real-time seismic monitoring and earthquake early warning, *Seismol. Res. Lett.* **83,** no. 3: 531–540, doi: 10.1785/gssrl.83.3.531.

Maeda, N. (1985). A method for reading and checking phase times in autoprocessing system of seismic wave data, *Zisin* **38,** no. 2, 365–379.

Oliphant, T. E. (2007). Python for scientific computing, *IEEE Comput. Sci. Eng.* **9,** 10–20.

Panagiotakis, C., E. Kokinou, and F. Vallianatos (2008). Automatic *P*-phase picking based on local-maxima distribution, *IEEE Trans. Geosci. Remote Sens.* **46,** no. 8, 2280–2287, doi: 10.1109/TGRS.2008.917272.

Saragiotis, C. D., L. J. Hadjileontiadis, and S. M. Panas (2002). PAI-S/K: A robust automatic seismic *P* phase arrival identification scheme, *IEEE Trans. Geosci. Remote Sens.* **40,** no. 6, 1395–1404, doi: 10.1109/TGRS.2002.800438.

Sleeman, R., and T. van Eck (1999). Robust automatic *P*-phase picking: An on-line implementation in the analysis of broadband seismogram recordings, *Phys. Earth Planet. In.* **113,** no. 1, 265–275, doi: 10.1016/S0031-9201(99)00007-2.

Vassallo, M., C. Santriano, and A. Lomax (2012). Automatic picker developments and optimization: A strategy for improving the performances of automatic phase pickers, *Seismol. Res. Lett.* **83,** no. 3, 541–554, doi: 10.1785/gssrl.83.3.541.

Wang, J., and T. Teng (1997). Identification and picking of *S* phase using an artificial neural network, *Bull. Seismol. Soc. Am.* **87,** no. 5, 1140–1149.

Withers, M., R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo (1998). A comparison of select trigger algorithms for automated global seismic phase and event detection, *Bull. Seismol. Soc. Am.* **88,** no. 1, 95–106.

Zeiler, C., and A. A. Velasco (2009). Seismogram picking error from analyst review (SPEAR): Single-analyst and institution analysis, *Bull. Seismol. Soc. Am.* **99,** no. 5, 2759–2770, doi: 10.1785/0120080131.

Zhang, H., C. Thurber, and C. Rowe (2003). Automatic *P*-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings, *Bull. Seismol. Soc. Am.* **93,** no. 5, 1904–1912, doi: 10.1785/0120020241.

Chen Chen[1]
ConocoPhillips School of Geology and Geophysics
University of Oklahoma
Norman, Oklahoma 73019 U.S.A.
c.chen@ou.edu

Austin A. Holland[1]
Albuquerque Seismological Laboratory
U.S. Geological Survey
PO Box 82010
Albuquerque, New Mexico 87198 U.S.A.

Published Online 31 August 2016

[1] Also at Oklahoma Geological Survey, University of Oklahoma, Norman, Oklahoma U.S.A.