GEOPHYSICS, VOL. 86, NO. 4 (JULY-AUGUST 2021); P. V299–V315, 22 FIGS., 2 TABLES.
10.1190/GEO2020-0308.1

Automatic seismic phase picking based on unsupervised machine-learning classification and content information analysis

Eduardo Valero Cano[1], Jubran Akram[2], and Daniel B. Peter[1]

ABSTRACT

Accurate identification and picking of P- and S-wave arrivals is important in earthquake and exploration seismology. Often, existing algorithms are lacking in automation, multiphase classification and picking, as well as performance accuracy. We have developed a new fully automated four-step workflow for efficient classification and picking of P- and S-wave arrival times on microseismic data sets. First, time intervals with possible arrivals on waveform recordings are identified using the fuzzy *c*-means clustering algorithm. Second, these intervals are classified as corresponding to P-, S-, or unidentified waves using the polarization attributes of the waveforms contained within. Third, the P-, S-, and unidentified-waves arrival times are picked using the Akaike information criterion picker on the corresponding intervals. Fourth, unidentified waves are classified as P or S based on the arrivals moveouts. The application of the workflow on synthetic and real microseismic data sets indicates that it yields accurate arrival picks for high and low signal-to-noise ratio waveforms.

INTRODUCTION

Hypocenter locations are in most cases estimated either using arrival times (e.g., by linearized inversion grid-search methods; see Buland, 1976; Pavlis, 1986; Moser et al., 1992; Oye and Roth, 2003) or waveform-based approaches (e.g., by time-reverse migration; see Artman et al., 2010; Nakata and Beroza, 2016). In the former approach, accurate arrival picking of P- and S-waves is critical for the accurate estimation of hypocenter locations. These picked arrival times are used in the polarization analysis for receiver orientations and back azimuths, in the velocity model calibration, and ultimately in the direct estimation of hypocenter locations. Therefore, any errors in the arrival-time picks can cause significant uncertainty in the estimated hypocentral parameters.

The arrival picking of P- and S-waves on microseismic data sets is, nonetheless, a challenging endeavor due to the poor signal-to-noise ratio (S/N) of the waveforms and large data volumes (days to weeks of continuous recordings). Previously, numerous automatic arrival picking methods have been proposed (see, e.g., Akram and Eaton [2016] who compare different algorithms such as the short- and long-term average ratio [STA/LTA], Akaike information criterion [AIC], phase arrival identification-kurtosis [PAI-K], and crosscorrelation pickers). Recently, many supervised and unsupervised machine-learning methods have also gained considerable popularity. Gentili and Michelini (2006) pick P- and S-phases using shallow neural networks with four manually defined input features, including variance, absolute values of skewness and kurtosis, and a combination of skewness and kurtosis. Similarly, Maity et al. (2014) use a neural network with two hidden layers and four input features including the variance of the sum of absolute values and other attributes based on the wavelet coefficients and envelope functions. More recently, many applications of deep-learning algorithms for arrival picking have been developed (e.g., Ross et al., 2018; Wang et al., 2019; Zhu and Beroza, 2019).

Neural network approaches typically belong to supervised machine-learning methods (e.g., artificial neural networks) and have a high success rate, in cases in which a good training data set is available. Because the ground truth is known a priori for the training set, direct quantification of the accuracy of the learning algorithm is possible (Ross et al., 2018). Nonetheless, the availability of an adequate training set often serves as a potential bottleneck, affecting the learning process. Research is ongoing on how to construct an

Manuscript received by the Editor 31 May 2020; revised manuscript received 21 February 2021; published ahead of production 29 March 2021; published online 30 June 2021.
[1]King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia. E-mail: eduardo.valerocano@kaust.edu.sa (corresponding author); daniel.peter@kaust.edu.sa
[2]Formerly KAUST; presently at the University of Calgary, Alberta, Canada. E-mail: akramj@ucalgary.ca.

optimal training data set for a given problem. Supervised deep-learning applications thus generally contain ad hoc choices and assumptions on the neural network architecture as well as a sparse coverage of the parameter model space represented by a training data set. In addition, training sets need manual labeling, which can take a considerable amount of user time.

Conversely, unsupervised machine-learning methods can be used in arrival picking without requiring any training data set because they rely on the data itself. For instance, Zhu et al. (2016) and Chen (2020) pick first arrivals on raw, noisy microseismic data using the fuzzy c-means (FCM) clustering algorithm. In the arrival picking context, the FCM method computes a time series, called the membership function, in which abrupt increases indicate wave arrivals. Although FCM is efficient in detecting arrivals, accurate identification of the instance at which the membership function increases is challenging, causing picking inaccuracies. In addition, some data sets (e.g., downhole microseismic monitoring) require picking of P- and S-waves, which the above method cannot determine reliably in its current state. For P- and S-arrival picking, additional modifications to the current workflow, as described in Zhu et al. (2016) and Chen (2020), are therefore necessary.

Here, we present a new fully automated workflow capable of picking P and S arrivals not only on events in which both phases are present but also on events in which only one of them exists (single-phase events). First, we use the FCM as described in Chen (2020) to identify multiple signal intervals (if present) in the analysis window. Second, we classify these intervals either as P-, S-, or an unidentified wave using polarization analysis on the waveforms in the signal intervals. Third, we pick the arrival times of the P-, S-, and unidentified waves using the AIC picker on the waveforms in the corresponding intervals. Fourth, we fit the P and S moveouts with a quadratic function and use them to classify unidentified picks as P or S. Finally, to evaluate the workflow performance, accuracy, and computational cost, we test it on synthetic and real microseismic data, and, for the synthetic data, we conduct a hypocenter location analysis.

FCM CLUSTERING

FCM clustering partitions a set of N points $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_N$ in a F-dimensional Euclidean space into C clusters by minimizing the objective function (Dunn, 1973; Zadeh, 1977; Bezdek, 1981; Bezdek et al., 1984; Zhu et al., 2016; Cano et al., 2019; Chen, 2020):

$$J(\mathbf{U},\mathbf{V}) = \sum_{k=1}^{N}\sum_{i=1}^{C}(u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|^2, \tag{1}$$

where $\mathbf{U}$ is the partition matrix in which elements $u_{ik} \in [0,1]$ indicate the degree of membership of the point $\mathbf{x}_k$ to the cluster i, $\mathbf{V} = \mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_C$ is a set of C points $\mathbf{v}_i$ that represent the centroid of cluster i, $\mathbf{x}_k$ is the kth point of $\mathbf{X}$, $m \in (1,\infty)$ is the controller of cluster fuzziness, and $\|\cdot\|$ is any norm.

One approach to minimize J is to update the set of centroids $\mathbf{V}$ and the partition matrix $\mathbf{U}$ via iterations of

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m}, \quad 1 \le i \le C, \tag{2}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{C}\left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|}\right)^{\frac{2}{m-1}}}, \quad 1 \le k \le N; 1 \le i \le C, \tag{3}$$

where equation 3 has the constraint $\sum_{i=1}^{C} u_{ik} = 1$ for all k.

The similarity metric of the points and the shape of the clusters depend on the choice of norm $\|\cdot\|$. Here, we use the L2 norm, which induces a similarity metric based on the Euclidean distance and clusters of hyperspherical shape.

AIC PICKER

AIC is a model selection technique developed by Akaike (1973), which also can be used for picking the onset of seismic phases on a 1C trace. It assumes that a seismic trace can be divided into locally stationary segments in which each is modeled as an autoregressive process. The onset time of a wave arrival separates two different segments and is associated with the minimum of the AIC values (Sleeman and Eck, 1999; Oye and Roth, 2003; Akram and Eaton, 2016).

Typically, calculation of the AIC function requires the estimation of autoregressive model coefficients, but Maeda (1985) uses the following relation to calculate the AIC function directly from the input trace (waveform):

$$\begin{aligned}\mathrm{AIC}(k) = {} & k\,\log(\mathrm{var}\{x(1,k)\}) \\ & + (N-k-1)\log(\mathrm{var}\{x(k+1,N)\}),\end{aligned} \tag{4}$$

where x is a trace of N samples, k ranges from 1 to N, and $\mathrm{var}\{x\}$ is the variance function. In this study, we use equation 4 to pick the onset of P- and S-wave arrivals. Because AIC picks on the global minimum, it is important that we first identify P and S intervals and then apply the picker to the corresponding intervals (Akram and Eaton, 2016).

AUTOPICKING WORKFLOW

The automatic arrival-time picking workflow (explained in Figure 1) comprises four main stages: (1) signal identification, (2) wave classification, (3) arrival-time picking, and (4) unidentified pick classification. For an adequate performance, our workflow requires a provisionally detected event in which the P- and S-wave arrivals occur only once. This is a regular strategy that usually simplifies arrival picking (Akram and Eaton, 2016). In addition, because we determine S-wave picks as the average of picks on the SV and SH components, our method is limited to isotropic media to avoid inaccuracies due to S-wave splitting.

For the binary clustering problem of signal identification, we define the samples $k = 1, \ldots, N$, of a trace $d(k)$ as a set of points $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_N$, in an F-dimensional Euclidean space. The elements of $\mathbf{x}_k$ represent the value of some feature of $d(k)$ (e.g., mean) at the sample k. The term F represents the number of features and N is the total number of samples in the analysis window. We denote the cluster number by i, where $i = 1$ is the noise cluster and $i = 2$ is the signal cluster, and we denote the centroid of cluster i by $\mathbf{v}_i$. Among many existing features, we use the following three in this study:

DOI:10.1190/geo2020-0308.1

- the mean of the absolute value of the amplitude:

$$M(k) = \frac{1}{N}\sum_{k-w}^{k+w} |d(k)|, \tag{5}$$

- the peak power spectral density:

$$P(k) = \max(|D(k,\omega)|^2), \tag{6}$$

- the STA/LTA ratio:

$$Q(k) = \frac{\mathrm{STA}}{\mathrm{LTA}} = \frac{\frac{1}{\mathrm{SW}}\sum_{j=k}^{k+\mathrm{SW}} |d(j)|}{\frac{1}{\mathrm{LW}}\sum_{j=k-\mathrm{LW}}^{k} |d(j)|}, \tag{7}$$

where, in equation 5, the constant w is half the length of a window around the sample k; in equation 6, $D(k,\omega)$ is the modulus of the discrete short-time Fourier transform of $d(k)$; and in equation 7, SW and LW are the lengths of the short- and long-term windows, respectively. For a correct signal identification, we suggest estimating the dominant period of the arrival of interest (i.e., using time-frequency analysis), T_{dom}, and set $w \approx 0.5 \times T_{\mathrm{dom}}$, $\mathrm{SW} \approx 1.5 \times T_{\mathrm{dom}}$, and $\mathrm{LW} \approx 5 \times \mathrm{SW}$. After computing the features of $d(k)$, we define the points

$$\mathbf{x}_k = [M(k), P(k), Q(k)]^{\mathrm{T}}, \tag{8}$$

and we apply fuzzy clustering (equations 1–3) to obtain $u_2(k) = u_{i=2,k}$, the membership degree of the sample k to the signal cluster. For a 3C record, we carry this process on the components $c = 1, 2, 3$, and, assuming that the signal arrives simultaneously on the three components, we stack their signal-cluster membership degrees $u_{2_c}(k)$ to highlight the wave arrivals:

$$u_s(k) = \frac{1}{3}\sum_{c=1}^{3} u_{2_c}(k), \tag{9}$$

where $u_s(k)$ is the stacked signal-cluster membership degree at sample k. Finally, we apply a threshold β to identify the signal intervals. Any continuous interval in which u_s is greater than β for at least 1.5 times T_{dom} duration is considered to contain a possible wave arrival; therefore, the other remaining intervals are deemed noise and are discarded from further analysis. Here, we set β between 1.0 and 2.0 times the mean value of u_s, depending on the S/N of the data.

Because one or both P- and S-waves can exist in the analysis window, we need some criteria to identify the intervals corresponding to the desired arrivals. We do so by computing the rectilinearity of the waveforms contained in each interval. First, we form an $(N_i \times 3)$ matrix $\mathbf{D}$, where N_i is the number of samples in the analyzed interval. In our workflow, N_i is determined automatically. The limits of each signal interval, and thus N_i, are defined by the intersection points of the threshold β with the peak of u_s corresponding to the interval of interest. Each column of $\mathbf{D}$ contains one of the three waveforms in the interval. After $\mathbf{D}$ is created, we compute the rectilinearity as follows:

$$R = 1 - \frac{\sigma_3^2}{\sigma_1^2}, \tag{10}$$

where R is the rectilinearity and $\sigma_{1,3}^2$ is the first and third eigenvalues of $\mathbf{D}$ (Jurkevics, 1988). Then, we determine the first signal interval with an acceptable high rectilinearity value as the first arrival. If there are no intervals at a later time than the first arrival, we label it as unidentified and classify it as P- or S-wave at the end of the workflow. This is because, in this situation and at this stage, it is complicated to determine whether the first arrival is a P- or S-wave because both waves can exhibit large rectilinearity values. Otherwise, if an interval exists later than the first arrival, we assume the first arrival to be a P-wave. Then, we rotate the waveforms to ray-centered coordinates (p, s1, s2) using the polarization information from the selected P interval and find the interval with the maximum S energy on the s1 and s2 components.

Following the wave-interval classification, we pick the onset of the P-, S-, and unidentified waves in the corresponding intervals. Although a threshold-based arrival-picking methodology from u_s (as given in Chen, 2020) can be adopted, the results it yields can be highly unstable for noisy data sets. Therefore, we apply the AIC algorithm to the P interval on the p components to pick the P-wave arrival time. For the S arrival time, we average the AIC picked times from the S interval on s1 and s2 components. In the case of unidentified waves, we obtain the arrival time using the AIC method on the component with the highest S/N on the corresponding interval. It is worth mentioning that the AIC is not the only algorithm that we can use. Other algorithms, such as PAI-K (Saragiotis et al., 2002) and crosscorrelation-based (VanDecar and Crosson, 1990; Song et al., 2010) pickers, will work equally as well.

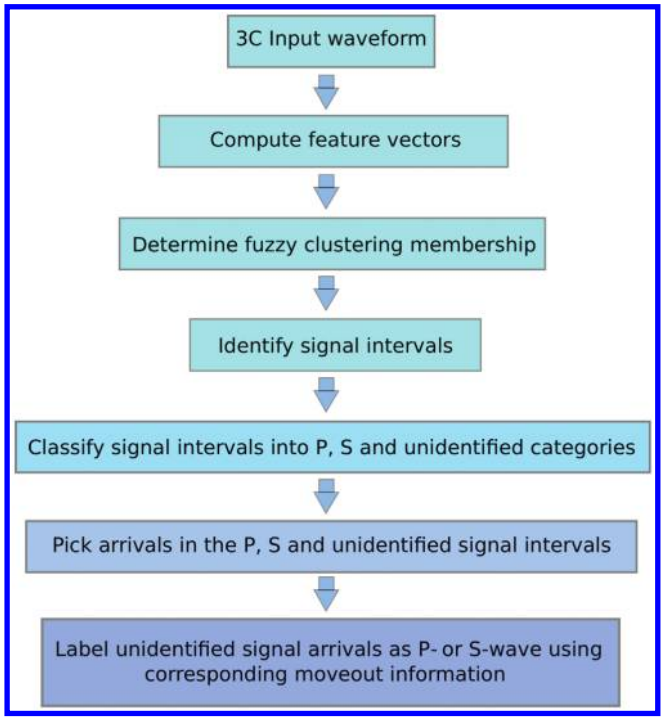

Figure 1. Arrival-picking workflow.

DOI:10.1190/geo2020-0308.1

Figure 2 illustrates the application of the autopicking workflow on 3C waveforms from a single receiver level. The P and S phases on the input waveforms have strong amplitudes on each of the three components (Figure 2a). For each of these components, the mean, power spectral density, and STA/LTA features are computed. Figure 2b shows the features for one of the components. In this case, all features and the signal-cluster memberships (Figure 2c) show a clearly distinguished response for the intervals containing P and S arrivals. These arrivals are easily identified by applying a thresholding criterion (Figure 2d). After the polarization analysis on the identified intervals, a P-wave interval is selected based on the order of occurrence and a high rectilinearity value. The data are then rotated into ray-centered coordinates to maximize the amplitudes of the P- and S-waves on the corresponding components (Figure 2e). The P- and S-wave arrival times are accurately picked using the AIC picker on the p components and the s1 and s2 components, respectively.

Once all of the P and S arrivals on the event are picked, we classify any unidentified picks as P or S, as illustrated in Figure 3. First, we temporarily label unidentified picks as S picks (Figure 3b). Then, we estimate the S-wave moveout by fitting the S picks with a quadratic function using the random sample consensus (RANSAC) method (Fischler and Bolles, 1987). For an explanation of how RANSAC is used to estimate moveout curves, we refer the reader to Zhu et al. (2017). Finally, once the S moveout is estimated (Figure 3c), we compute the time difference between the unidentified picks and the fitted S moveout curve. Any unidentified pick between $\pm T_{\mathrm{dom}}$ from the S-moveout curve is classified as an S pick, and any remaining unidentified picks are classified as P (Figure 3d). We also correct any P picks on S-waves using the same moveout criteria.

For events in which all picks are unidentified (Figure 4a), the previous strategy cannot be used because fitting the P or S moveout is not possible. In this situation, we generate a database of the P and S moveouts, obtained from events in which the P and S picks were available. We then fit the moveout of the unidentified arrivals and compare it with the moveouts database (Figure 4b). We do so by shifting all of the moveouts to a common time and computing the following coefficients:

$$\mathrm{P}_{\mathrm{res}} = \frac{1}{N_p}\sum_{i=1}^{N_p} \|\mathbf{u}_{\mathbf{mov}} - \mathbf{p}_{\mathbf{mov_i}}\|_2, \tag{11}$$

$$\mathrm{S}_{\mathrm{res}} = \frac{1}{N_s}\sum_{i=1}^{N_s} \|\mathbf{u}_{\mathbf{mov}} - \mathbf{s}_{\mathbf{mov_i}}\|_2, \tag{12}$$

where $\mathbf{u}_{\mathbf{mov}}$, $\mathbf{p}_{\mathbf{mov}}$, and $\mathbf{s}_{\mathbf{mov}}$ are the vectors containing the unidentified, P, and S moveout curves, respectively, and N_p and N_s are the number of P and S moveouts in the database, respectively. The unidentified moveout of interest $\mathbf{u}_{\mathbf{mov}}$, and therefore the associated picks, are classified as P or S depending on whether $\mathrm{P}_{\mathrm{res}}$ or $\mathrm{S}_{\mathrm{res}}$ is the smallest value. This approach is useful for identifying the wave type of single-phase events as illustrated in Figure 4c.

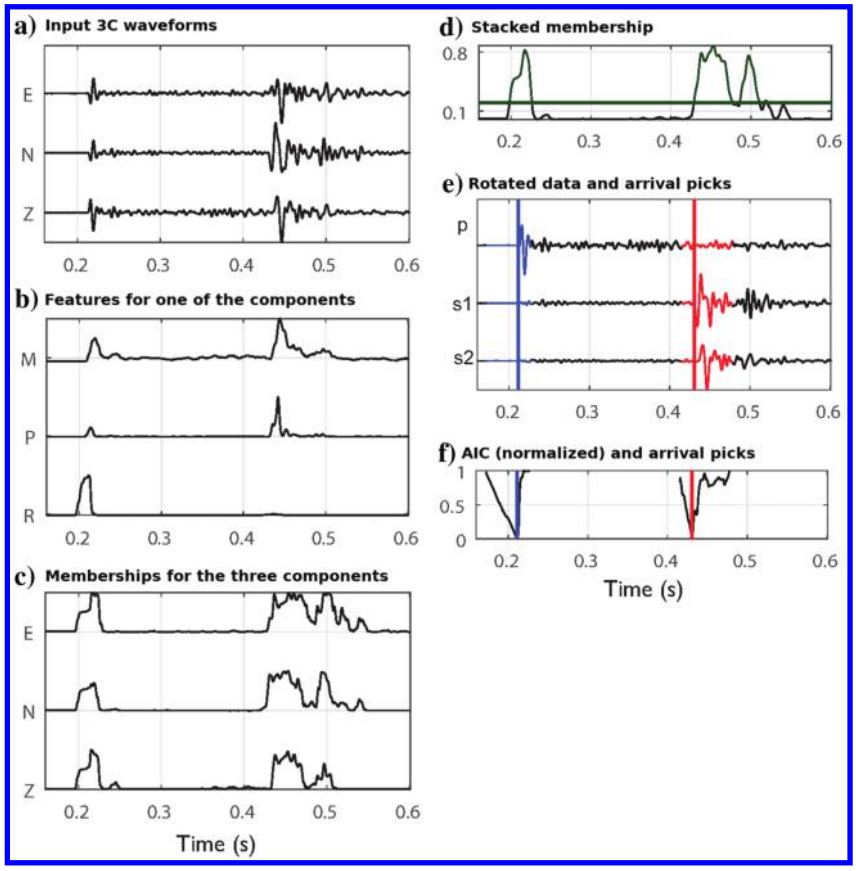

Figure 2. Arrival-time picking using the proposed workflow. (a) Input 3C waveforms, (b) vertical component features, (c) fuzzy memberships for the three components, (d) stacked membership, and (e) rotated data and arrival picks. The green line in (d) indicates the threshold β. In (e), the highlighted blue and red curves show the P and S intervals, respectively, obtained from the FCM clustering. In (e) and (f), the blue and red vertical lines represent the P and S picks, respectively. (f) AIC values for the rotated data. The time associated with the minimum AIC value is the arrival time of a phase arrival.

DOI:10.1190/geo2020-0308.1

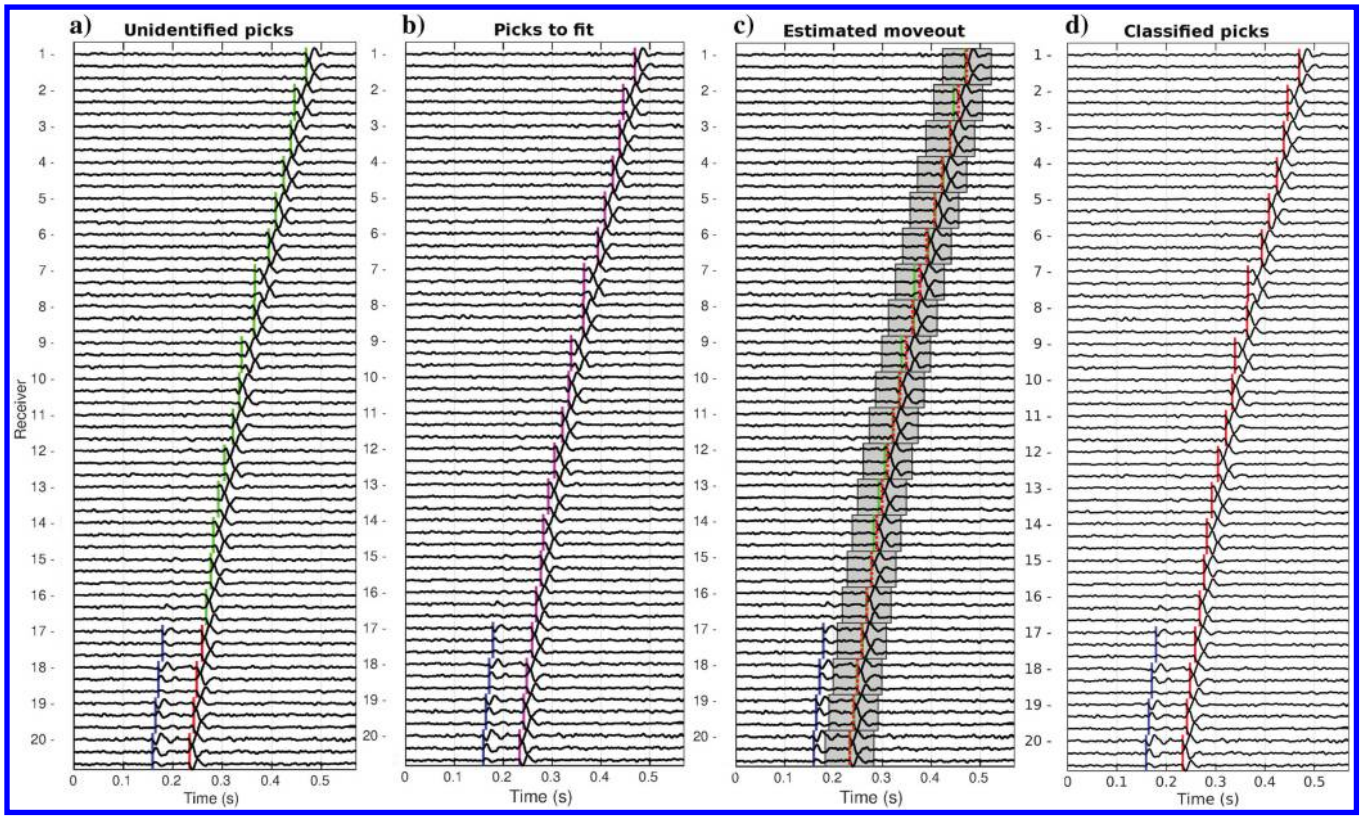

Figure 3. Classification of unidentified picks for events with available P and S picks. (a) Unidentified picks. The P, S, and unidentified picks are indicated by the blue, red, and green lines, respectively. (b) Picks used for S moveout fitting (the magenta lines). (c) Estimated S moveout (the dotted red line). Any pick laying on the dark area will be classified as an S pick. (d) Classified picks. The unidentified picks in (c) are now classified as S picks.

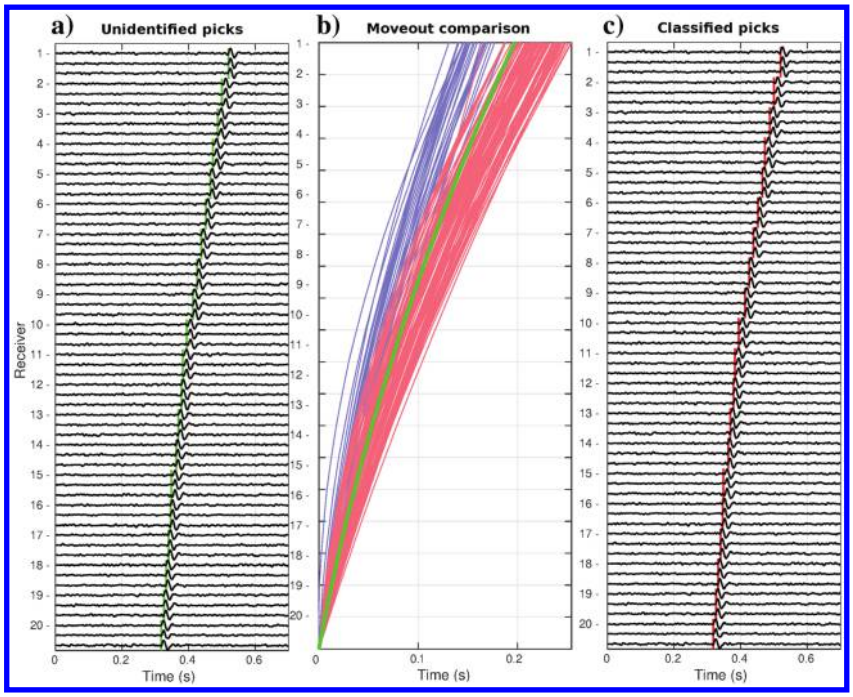

Figure 4. Classification of unidentified picks for events without P and S picks. (a) Unidentified picks and (b) moveout comparison. The P, S, and unidentified moveouts are in blue, red, and green, respectively. (c) Classified picks. The initially unidentified picks are classified as S picks.

RESULTS

To evaluate the performance of the proposed arrival-time picking workflow, we apply it to synthetic and real microseismic data and compare the results with the reference picks. For the synthetic data, we also conduct a hypocenter location analysis.

To compare our workflow with existing methods, we use the STA/LTA and Chen (2020) algorithms on the synthetic and real data sets. Given that for a 3C record, our workflow potentially obtains one P and S pick and the STA/LTA and Chen (2020) methods obtain three first-arrival picks (one per receiver component), we consider the picks of the STA/LTA and Chen (2020) methods with the minimum difference to the reference picks as P picks. For the synthetic data set, we use theoretical arrival times (computed using ray tracing) as reference picks, whereas for the real data set, we use manual picks.

To measure the picking accuracy, we compute the residual between the reference and the automatic picks. We illustrate our results in scatterplots of the arrival-pick residual against the arrival S/N. To compute the arrival S/Ns, we define a window of noise from the start of the record to one dominant period before the first arrival. Then, we set a window centered on the arrival of interest and compute the ratio between the root-mean-square (rms) amplitude of each window. In the scatterplots, residuals of the P and S picks are pictured in blue and red colors, respectively. The residuals of initially unidentified picks are indicated in green, and the residuals of picks skipped by our workflow but determined by the STA/LTA and Chen (2020) methods are indicated in brown. We also plot black lines at −10 and 10 ms to highlight relatively accurate picks. In addition, we show histograms of the residuals and compute their mean μ and standard deviation σ. We compute μ and σ using residuals in the −50 to 50 ms interval to decrease the influence of large picking errors and obtain meaningful indicators of each algorithm's performance.

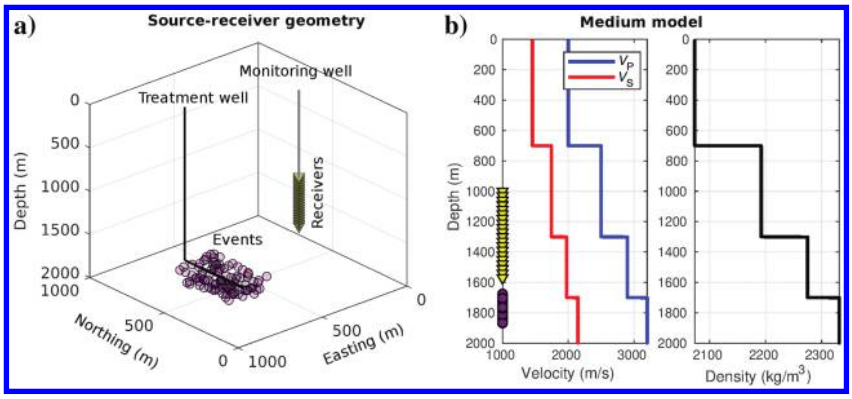

Figure 5. Synthetic experiment setup. (a) Source-receiver geometry and (b) medium model. In (b), the left panel shows the P and S velocity models, and the right panel shows the density model.

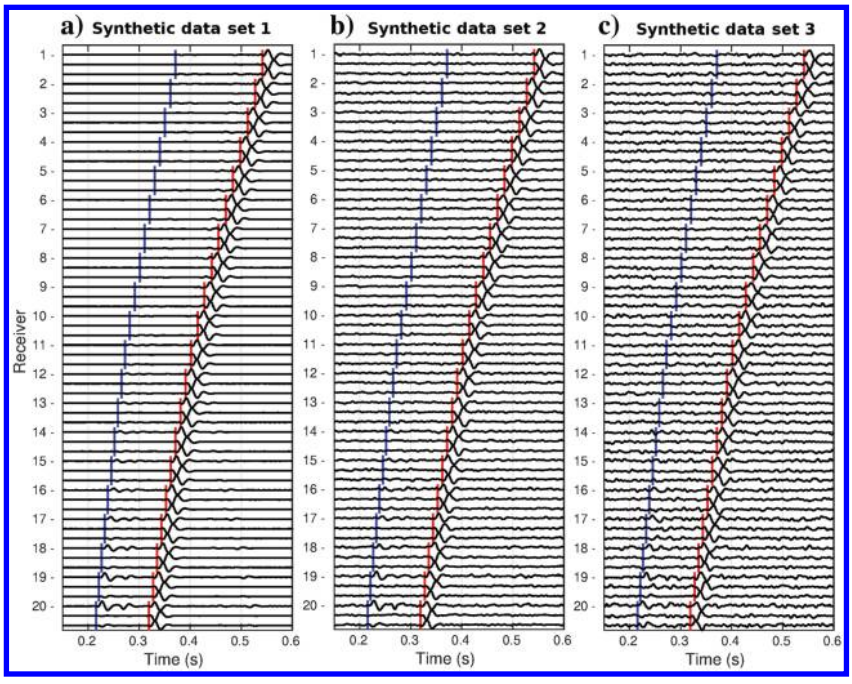

Figure 6. Example of a synthetic event with different noise levels. The illustrated P and S picks are reference picks computed using ray tracing. (a) Synthetic data set one (S/N = 20 dB). (b) Synthetic data set two (S/N = −8 dB). (c) Synthetic data set three (S/N = −13 dB).

Synthetic data

Arrival-time picking

The synthetic data consist of 100 events recorded by a vertical downhole array of 20 3C receivers located in an elastic homogeneous-layered medium (Figure 5). The events are defined by randomly distributed double-couple and tensile sources with moment magnitudes ranging from −3 to 1. A Berlage wavelet (Aldridge, 1990) with a dominant frequency of 30 Hz is used as the source time function, and the wave propagation is simulated using the SPECFEM3D Cartesian package (Komatitsch and Tromp, 1999). We create three synthetic data sets (Figure 6) by adding white Gaussian noise (AWGN) with S/Ns of 20, −8, and −13 dB and filtering the waveforms between 0.1 and 100 Hz with a zero-phase fourth-order Butterworth band-pass filter. As reference picks, we use 2000 P- and 2000 S-arrival times computed with ray tracing (Figure 6).

On synthetic data set one (AWGN S/N of 20 dB), the μ and σ of the P residuals are relatively low among the methods, especially for the proposed and STA/LTA pickers. This data set presents low levels of noise; thus, the three methods have an acceptable outcome, as illustrated in event 33 in Figure 7. Nevertheless, the proposed workflow exhibits the lowest μ and σ values (−0.66 and 2.99 ms), indicating less biased and more reliable picking. Most of the P residuals range between −10 and 10 ms, with some outliers from relatively low S/N arrivals (<20 dB),

related to barely detectable P-waves. For the proposed workflow, three P picks were initially labeled as unidentified (Table 1). Their residuals are close to 0 ms, indicating accurate picking and classification of unidentified arrivals. In addition, in total 221 P arrivals were skipped by our method. This tends to occur when the P arrival is buried by noise, has a poor S/N (receivers 8–11 in Figure 7a), or was picked on an S-wave and corrected by our moveout criteria. On the other hand, the STA/LTA and Chen (2020) methods managed to pick 153 and 42 of these skipped P arrivals, respectively, with residuals between −10 and 10 ms. The number of P arrivals omitted by our workflow and accurately picked by the STA/LTA and Chen (2020) methods is related to the number of arrivals from which our workflow loses information.

Regarding the S residuals on synthetic data set one, most of them range between −10 and 15 ms. Despite the high S/N of the S-wave arrivals (>40 dB), the residuals are relatively large, mainly due to contamination of the S wave with precursory phases, as observed in the waveforms in Figure 7 (receiver 7). This increases the σ value, explaining why σ is 2.09 ms larger for S residuals than for P residuals. Moreover, in total, 221 S picks were determined from unidentified picks (Table 1). These picks correspond to the waveforms in which the P picks were skipped. The associated residuals are in the same range as the rest of the S picks, suggesting a correct classification of unidentified arrivals.

The automatic picking results for synthetic data set two (AWGN S/N of −8 dB) are less favorable than those in data set one. As observed in event 33 in Figure 8 (receivers 1–14), the noise on this data set masks P arrivals that were detectable previously (see Figure 7). In addition, the presence of high-amplitude noise foregoing the first arrivals increases. The previous noise-related effects result in an increase of picks on noise preceding the first arrivals (large positive residuals) and picking of P arrivals on S-waves (large negative residuals). This is especially true for the STA/LTA and Chen (2020) methods, as they yield a higher σ (14.25 and 12.36 ms, respectively) than our workflow (10.49 ms). The proposed workflow yields better results than the other two methods because it skips P arrivals completely covered by noise, as shown in Figure 8a (receivers 1–6). Of the 2000 P arrivals, 1014 were skipped (Table 1). Although this is more than half of the total P arrivals in the data set, most of them have poor S/N (<5 dB) or were not recorded. From the skipped P picks, the STA/LTA and Chen (2020) methods picked 295 and 149 between −10 and 10 ms, respectively.

As observed from Figures 7 and 8, the noise on synthetic data set two has a small influence on the S arrivals. The S residuals have a similar distribution as in the less-noisy data set one, and the increase in the number of large residuals is small. Compared with the P residuals, the S residuals have a lower σ value, suggesting that picking is more reliable in S-waves than in P-waves. The number of S picks determined from unidentified picks also increased considerably as compared to data set one (Table 1). This is expected, as unidentified picks occur on waveforms where the P-wave was not detected. A total of 1014 S picks were determined from unidentified picks, most of which have similar residuals than the rest of the S picks, implying correct classification of unidentified arrivals.

On synthetic data set three (AWGN S/N of −13 dB), the noise covers more P arrivals than in data set two and the amplitude of early noise grows. As a consequence, the number of picks on noise with an amplitude similar to the arrivals and picking of P arrivals on S-waves increases (receivers 1–3 in Figure 9b and 9c). The distributions of P residuals are relatively similar compared with data set two (Figure 10c, 10f, and 10i). The Chen (2020) method has the lowest μ value, which indicates less biased picking than the STA/LTA and proposed methods. However, our workflow exhibits the lowest σ suggesting that it is the most reliable among the three pickers. The number of large P residuals increases on all methods, especially for low-S/N arrivals (<5 dB; Figure 11c, 11f, and

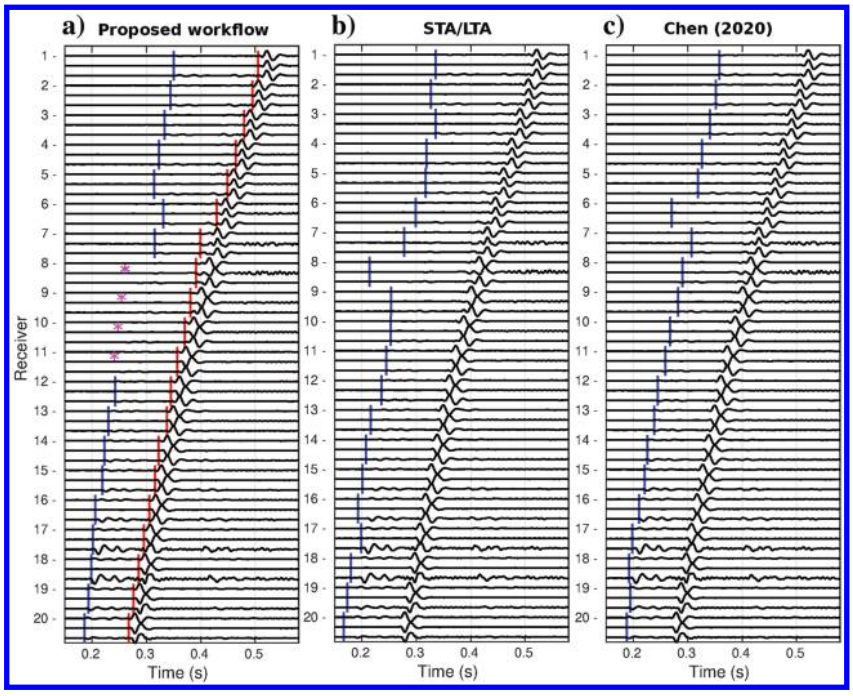

Figure 7. Event 33 of synthetic data set one (average P-wave S/N = 14.9 dB and average S-wave S/N = 43.1 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The blue and red lines indicate the P and S picks. The skipped arrivals are indicated by asterisks.

DOI:10.1190/geo2020-0308.1

11i). For the proposed workflow, some large P residuals occur on relatively high-S/N arrivals (>7 dB). These residuals are related to picking on early noise. The proposed workflow determined 16 P picks from unidentified picks. Some of these reclassified picks have large residuals; however, this is because the unidentified picks occur on early noise and not due to incorrect reclassification. In addition, our workflow skipped 1239 P arrivals (Table 1). As illustrated by the clusters of brown points in Figure 11f and 11i, most of the skipped arrivals have a poor S/N (<5 dB) and were not determined accurately by the STA/LTA and Chen (2020) methods, suggesting that the skipped P arrivals were buried by noise.

Similar to data set two, S arrivals are less affected by noise than P arrivals (Figures 8 and 9). The μ and σ values are considerably lower for S residuals than for P residuals. There is not a high increase of μ and σ compared with previous data sets (Figure 12c). In addition, the number of large S residuals remains almost the same,

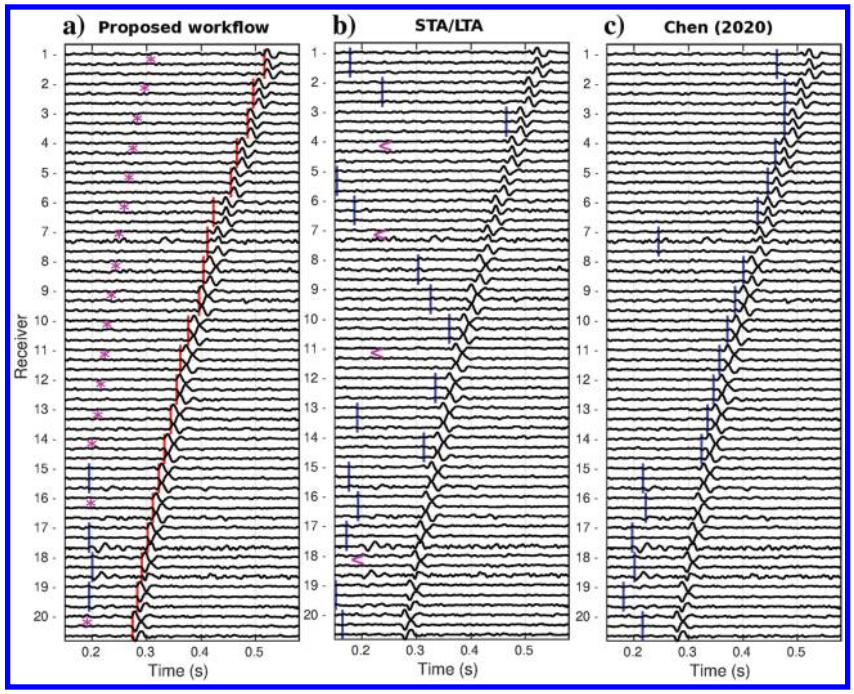

Figure 8. Event 33 of synthetic data set two (average P-wave S/N = −0.3 dB and average S-wave S/N = 17.4 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The blue and red lines indicate the P and S picks, respectively. The $*$ and $<$ symbols indicate skipped arrivals and early picks outside the figure time range, respectively.

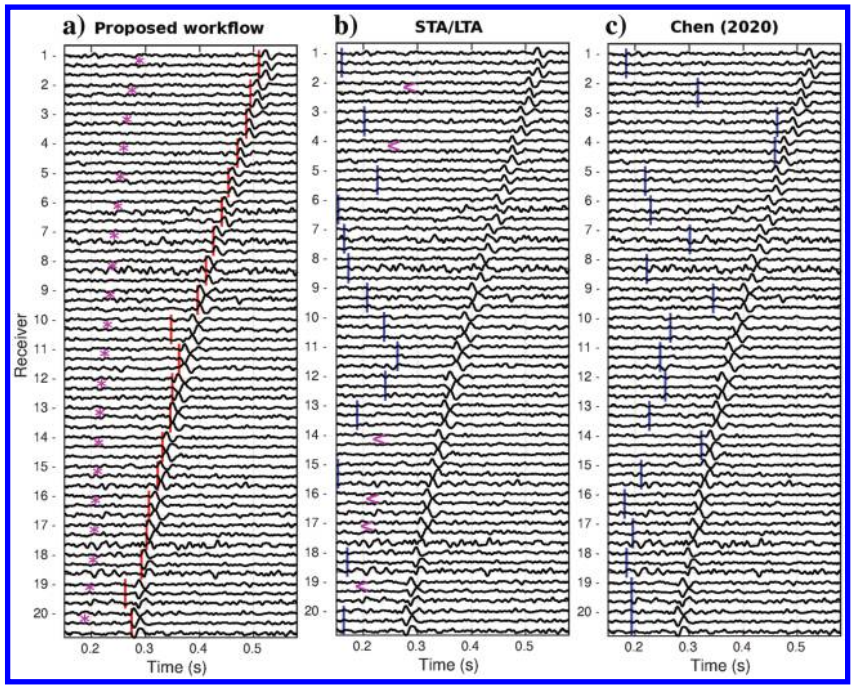

Figure 9. Event 33 of synthetic data set three (average P-wave S/N = 0.2 dB and average S-wave S/N = 12.1 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The symbols are as indicated in Figure 8.

with the exception of an outlier at approximately 140 ms related to the incorrect detection of P- and S-waves (Figure 13e). As hinted at by the similar residuals between the 1237 S picks determined from unidentified arrivals and the remaining S picks (Figure 13c), the unidentified pick classification was successful.

Hypocenter location

To determine if our workflow results are useful for accurate event location, we locate the 100 events of the synthetic data using the true velocity model and the picks' and events' back azimuths estimated by our workflow. Then, we compute the location error using the true hypocenters as a reference. For each event in the following analysis, we define the location rms error as the rms of residuals on the north, east, and depth coordinates. We also define the arrival-picking rms error as the rms of the P- and S-pick residuals. In addition, we compute the difference between the true event back azimuth and the back azimuth estimated by our workflow on each 3C record. We define the back-azimuth rms error as the rms of the back-azimuth residuals.

In Figure 14, the dots show the location errors on the north, east, and depth coordinates of the located hypocenters. We can observe that, on the three data sets, the largest location uncertainty occurs on the north coordinate (the transverse direction), followed by the depth and by the east coordinate (the radial direction). For synthetic data set one, most of the events (84) have a location rms error of less than 30 m, a low value considering the array geometry. On data set two, five single-phase (S-wave) events were not located. A total of

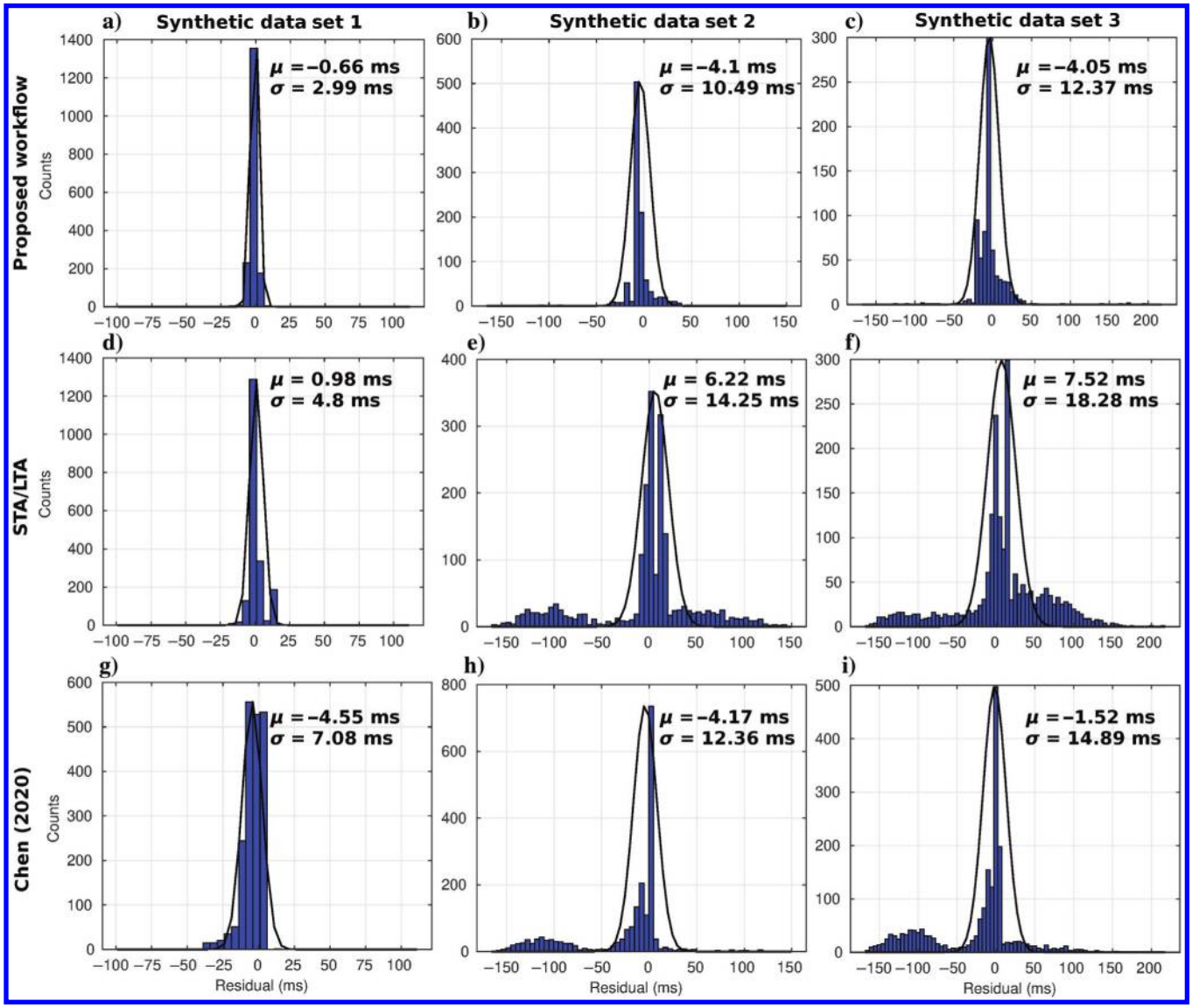

Figure 10. Histograms of P-pick residuals obtained by the proposed workflow and the STA/LTA and Chen (2020) methods on the synthetic data sets. The black line indicates the residuals' probability distribution in the −50 to 50 ms interval. The mean and standard deviation values are indicated by μ and σ, respectively.

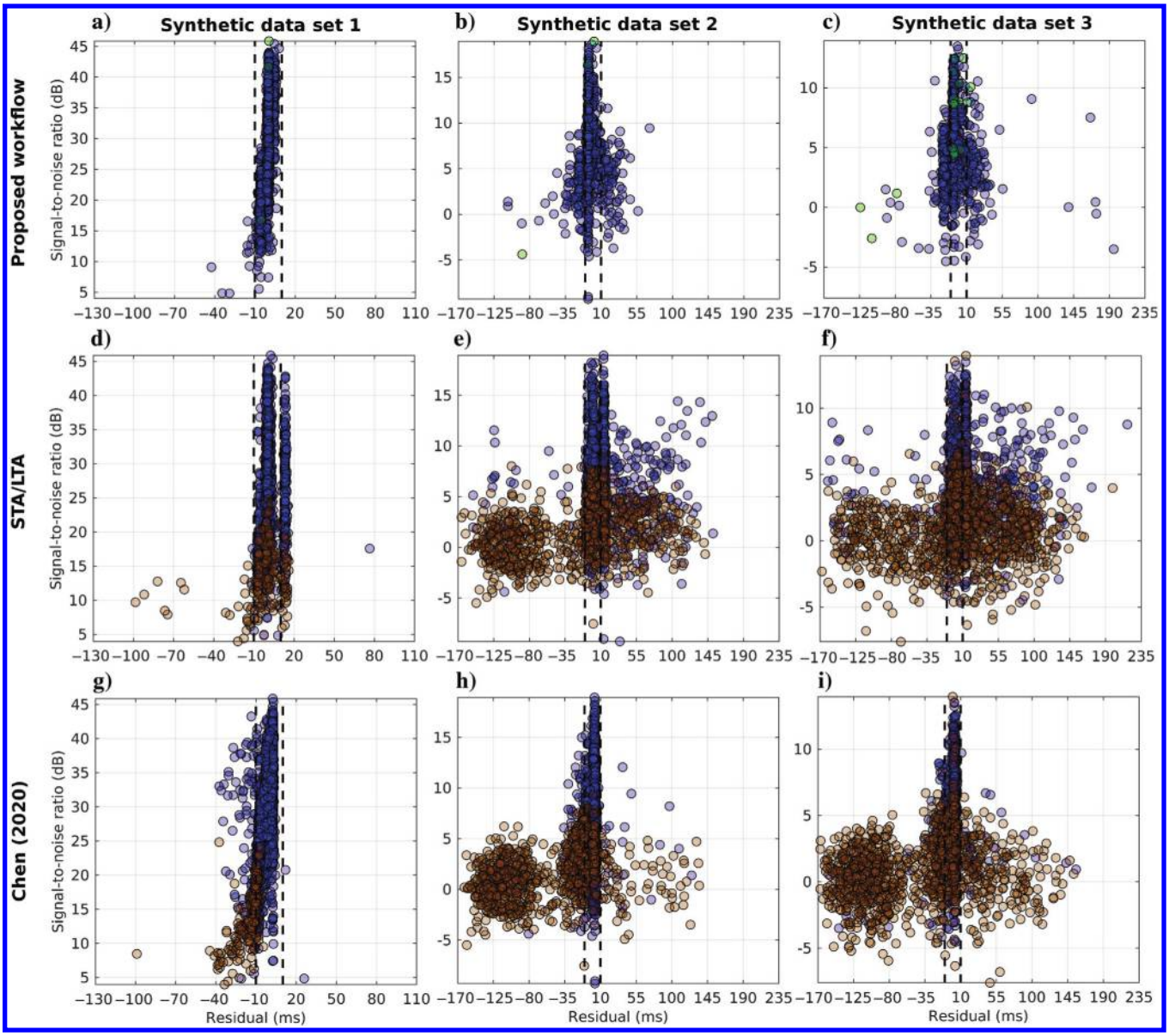

Figure 11. P-pick residuals obtained by the proposed workflow and the STA/LTA and Chen (2020) methods on the synthetic data sets. The residuals of the P picks initially labeled as unidentified are in green, and the residuals of picks skipped by the proposed workflow are in brown.

Table 1. Number of picked, initially unidentified, and skipped arrivals by the proposed workflow on the synthetic data sets.

	P arrivals			S arrivals		
Data set	Picked	Initially unidentified	Skipped	Picked	Initially unidentified or P	Skipped
1	1779	3	221	1997	221	3
2	986	3	1014	1997	1014	3
3	761	16	1239	1982	1237	18

52 events have a location rms error of less than 45 m, which is a realistic value considering the level of noise on this data set. The location errors increase on data set three. The location rms error is less than 58 m for 50 of the 95 located events.

Figure 15 shows scatterplots of the location rms error against arrival-picking and back-azimuth rms error. The dots' color indicates the number of receivers used during the event localization (receivers with available P and S picks). As expected, the location rms error is positively correlated with arrival-picking and back-azimuth rms error in all data sets. That is, the location is poor for events with inaccurate arrival picks and estimated back azimuths. In addition, we can observe that the location rms error tends to be lower for events in which more than 10 receivers were used. Nonetheless, this is not general because some events with low arrival-picking and back-azimuth rms errors were located relatively accurately using fewer than five receivers.

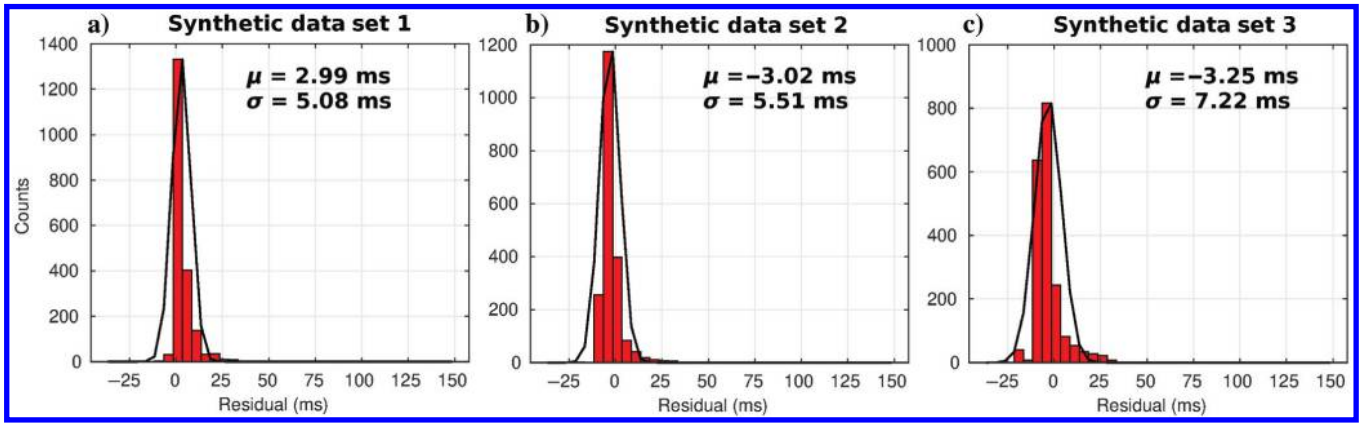

Figure 12. Histograms of the S-pick residuals obtained by the proposed workflow on synthetic data sets (a) one, (b) two, and (c) three. The symbols are as indicated in Figure 10.

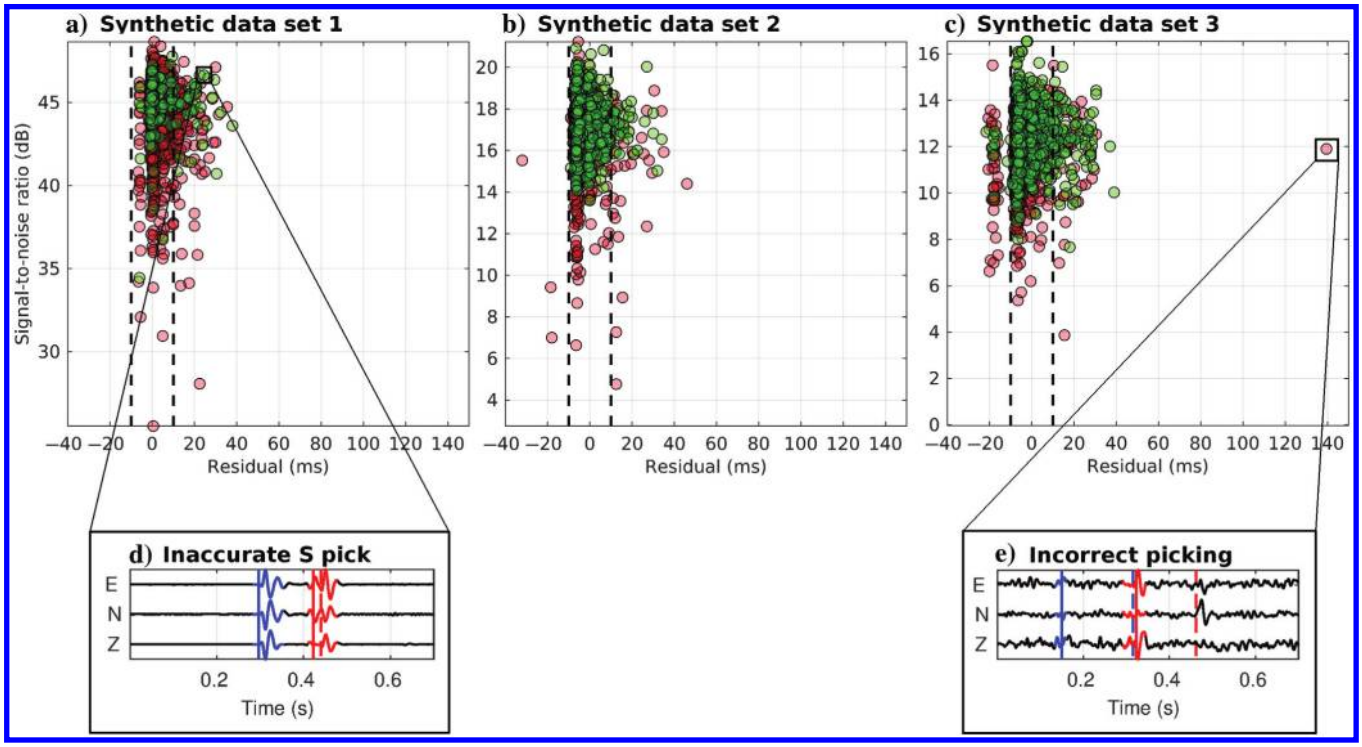

Figure 13. (a–c) The S-pick residuals obtained by the proposed workflow on the synthetic data sets. The residuals of S picks initially labeled as unidentified are shown in green. (d) Inaccurate S pick due to precursory phase contamination. (e) Incorrect P and S picking. The P interval is erroneously set as noise with high rectilinearity, resulting in an incorrect determination of the S-wave. (d and e) The blue and red colors indicate P- and S-waves, respectively. The reference picks are indicated by the dashed lines, and the automatic picks are indicated by the continuous lines. The highlighted curves indicate the windows where picking was conducted using the AIC picker.

DOI:10.1190/geo2020-0308.1

Real data

The real microseismic data set was acquired during a hydraulic fracturing operation by an array of 20 3C receivers placed in a vertical monitoring well. The recorded waveforms were sampled at a 0.5 ms interval. Using time-frequency analysis, we estimate that the dominant frequency of the arrivals is 100 Hz. To denoise the data, we filter the waveforms using a zero-phase fourth-order Butterworth band-pass filter. Due to receiver limitations, we set the lower cutoff frequency to 10 Hz. We set upper cutoff frequencies between 200 and 300 Hz. For this study, we use only a data set containing 40 previously detected events. To compare the automatic picks, we carry out manual picking on waveforms in which the P and S arrivals were recorded. Of the potential 800 P arrivals and 800 S arrivals, we retrieve 694 P and 714 S picks (106 P and 86 S missing reference picks).

Figures 16, 17, and 18 show examples of arrival picking from all three methods on real waveform data with different S/N. In Figures 16 and 17, all three methods yield relatively precise arrival picking results. As in the synthetic data, the arrival-picking omission on waveforms with P arrivals buried in noise (receivers 16 and 19 in Figure 18a) reduces the number of incorrect picks by our workflow. Figure 19 shows the P residuals' distribution, which suggests that most of the residuals are in the range of –10 to 10 ms. The erroneous STA/LTA picks on presignal noise (Figure 18b, receivers 2–4) result in the largest μ and σ values of the P residuals' distribution among all methods. In this data set, 95 P arrivals were skipped by our method (Table 2), from which 61 correspond to

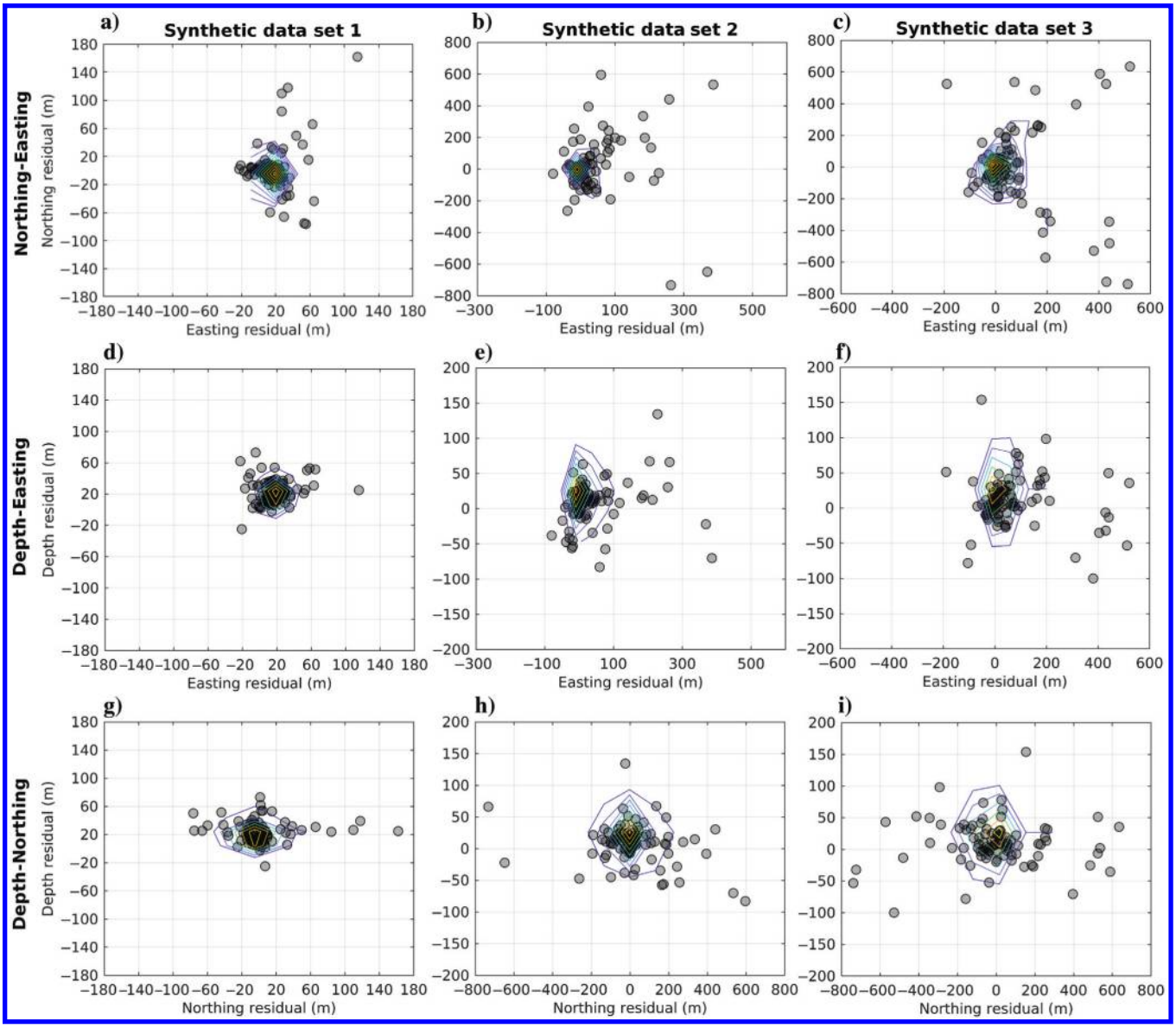

Figure 14. Hypocenter location errors on the north, east, and depth coordinates. Each dot represents the hypocenter localization error of one event.

missing reference picks. Of the remaining 34 arrivals, the STA/LTA and Chen (2020) methods picked 10 and 22 with a residual between −10 and 10 ms, and the rest correspond to outliers (Figure 20). In addition, a total of five P picks were correctly determined from unidentified picks.

Compared with P residuals, S residuals have a slightly higher μ and σ (Figure 21). A great part of the residuals occurs between −10 and 10 ms, indicating relatively accurate picking as shown in Figures 16a, 17a, and 18a. In addition, high- (>30 dB) and low- (<20 dB) S/N outliers are present. The high-S/N outliers in

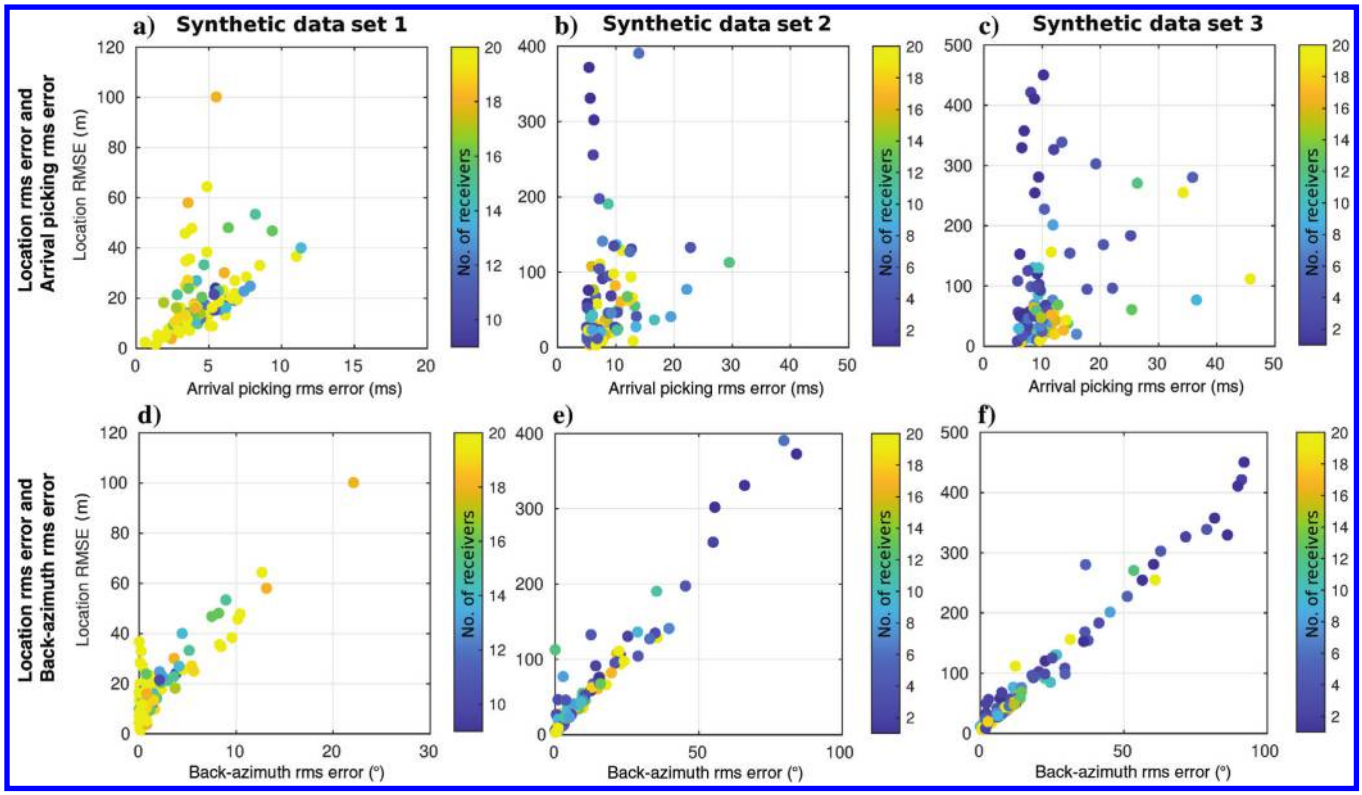

Figure 15. Hypocenter location rms error (a–c) against the arrival-picking rms error and (d–f) against the back-azimuth rms error.

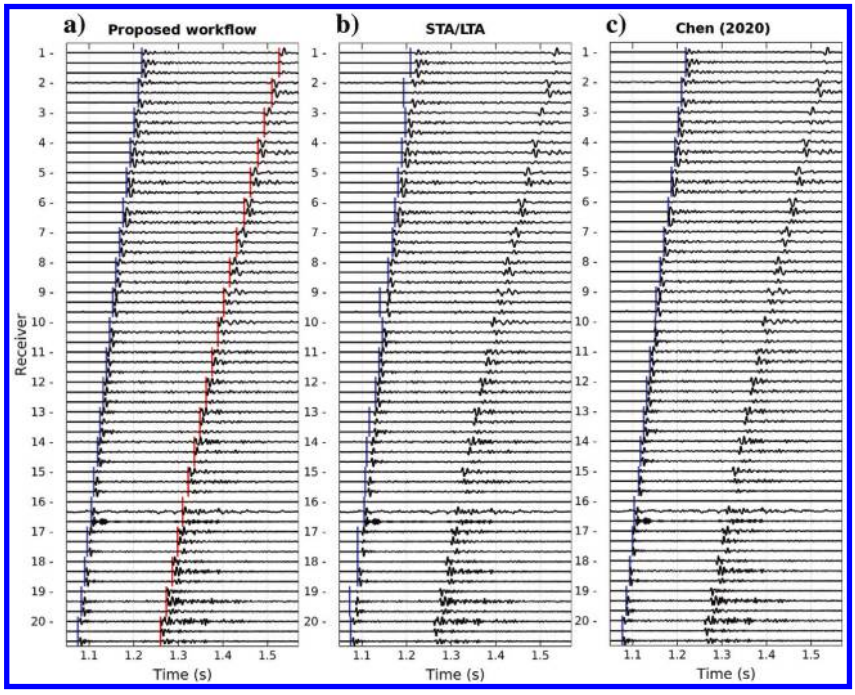

Figure 16. Event 8 of the real data set (average P-wave S/N = 40.8 dB and average S-wave S/N = 41 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The symbols are as indicated in Figure 8.

DOI:10.1190/geo2020-0308.1

Figure 22 are related to complex waveforms where high-amplitude phases exist after the S-wave, resulting in P picking on the S-wave and S picking on a late phase (Figure 22b). Regarding the classification of unidentified arrivals, a total of 85 S picks were obtained from unidentified picks, all with residuals similar to the rest of the S picks. This indicates a successful classification of the unidentified arrivals.

DISCUSSION

Overall, the proposed workflow obtains lower mean and standard deviation values of P residuals than the STA/LTA and Chen (2020) methods on the synthetic and real data sets. This suggests that the P-arrival picking carried out by our workflow is less biased and more stable than that of the other two methods. The main observed

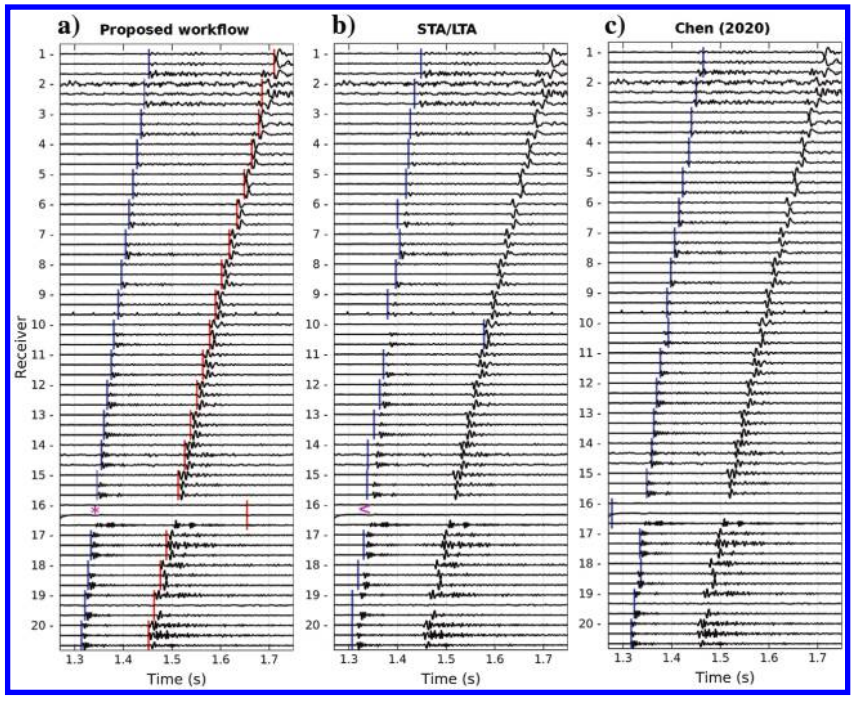

Figure 17. Event 13 of the real data set (average P-wave S/N = 23.2 dB and average S-wave S/N = 30.2 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The symbols are as indicated in Figure 8.

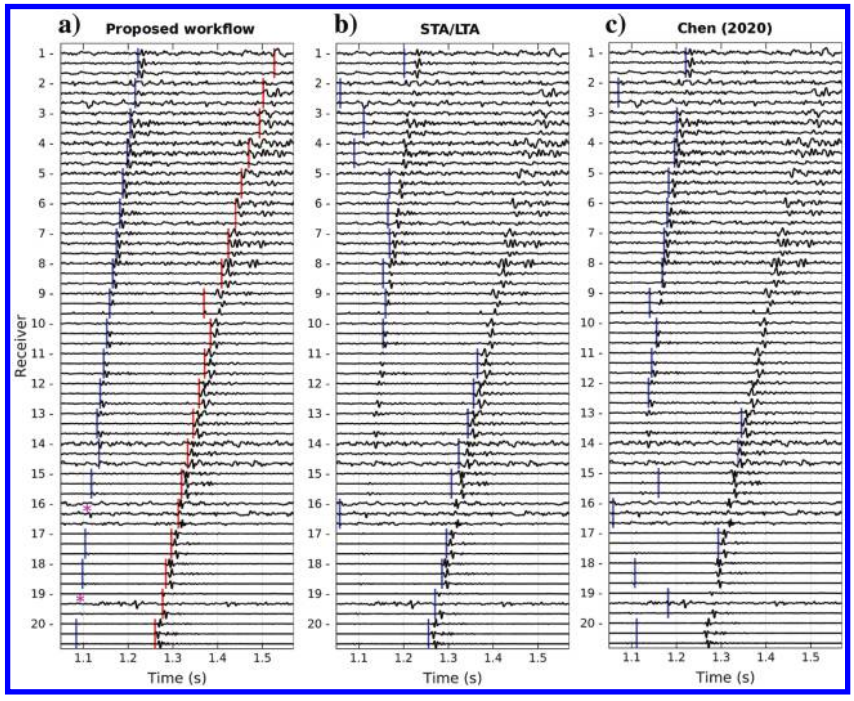

Figure 18. Event 28 of the real data set (average P-wave S/N = 13.8 dB and average S-wave S/N = 19 dB). (a) Proposed workflow picks, (b) STA/LTA picks, and (c) Chen (2020) picks. The symbols are as indicated in Figure 8.

DOI:10.1190/geo2020-0308.1

problems of the STA/LTA and Chen (2020) methods are picking on presignal noise and P picking on S-waves. High-amplitude noise tends to generate high STA/LTA values, and depending on the used trace features, this also occurs for FCM. This results in picks on presignal noise because these methods set the arrival time on the earliest jump of STA/LTA and membership values. On the other hand, in cases in which P-waves are masked by noise, the earliest increase in STA/LTA and membership values is usually related to S-waves, resulting in P picks on S-waves. Because our workflow selects P arrivals based on the rectilinearity of waveforms contained in signal intervals determined from FCM membership values, high-amplitude noise is avoided most of the time. In addition, the capacity of our workflow to detect waveforms containing one "unidentified" arrival and to classify it as a P or S aids in the picking omission of not-recorded phases and allows picking and phase identification on single-phase events.

Based on the mean and standard deviation values of S residuals on synthetic data set one and the real data set, P picking is slightly more accurate and reliable than S picking. The S residuals present slightly higher mean and standard deviation values than those of

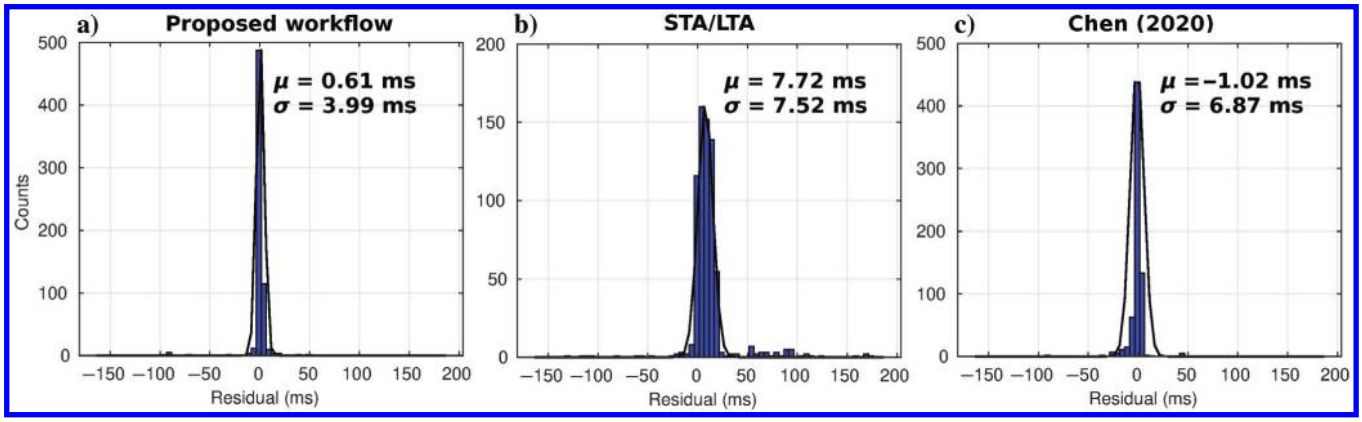

Figure 19. Histograms of the P-pick residuals obtained by (a) the proposed workflow, (b) the STA/LTA method, and (c) the Chen (2020) method on the real data set. The symbols are as indicated in Figure 10.

Table 2. Number of picked, initially unidentified, and skipped arrivals by the proposed workflow on the real data set.

P arrivals			S arrivals		
Picked	Initially unidentified	Skipped	Picked	Initially unidentified or P	Skipped
720	5	95	780	85	15

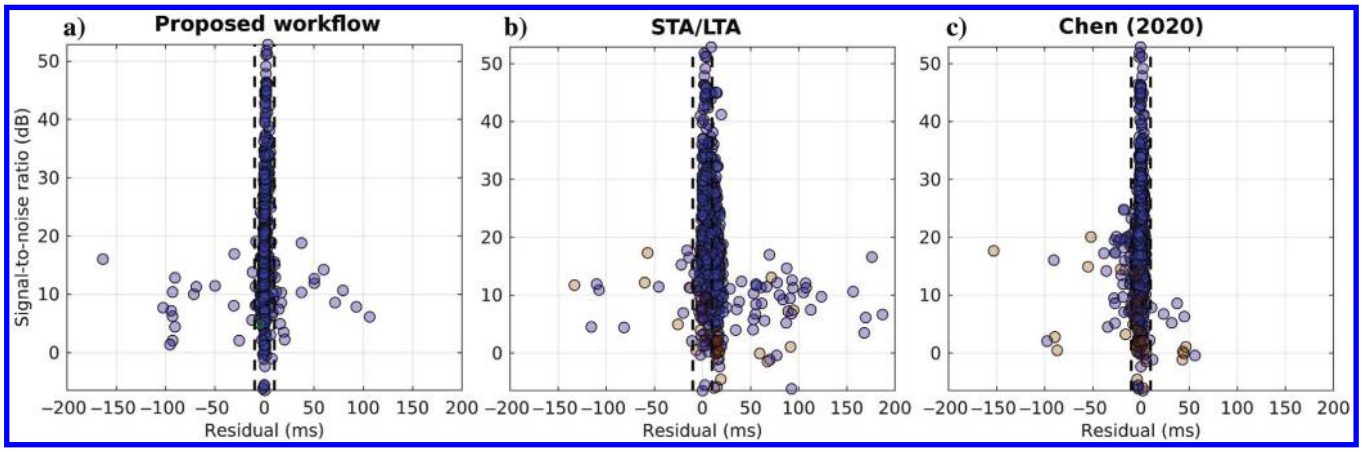

Figure 20. P-pick residuals obtained by (a) the proposed workflow, (b) the STA/LTA method, and (c) the Chen (2020) method on the real data set. The color code is as indicated in Figure 11.

DOI:10.1190/geo2020-0308.1

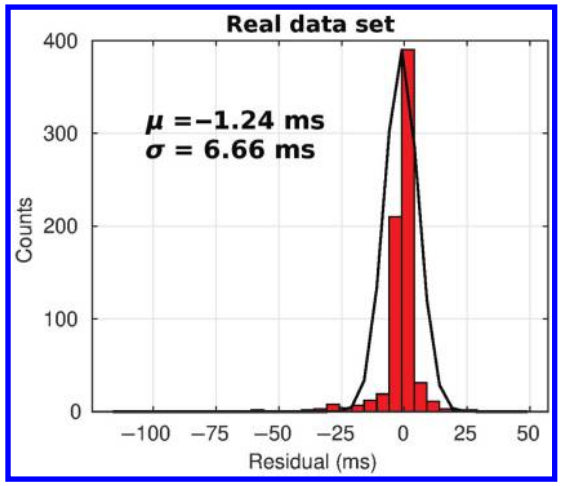

Figure 21. Histograms of S-pick residuals obtained by the proposed workflow on the real data set. The symbols are as indicated in Figure 10.

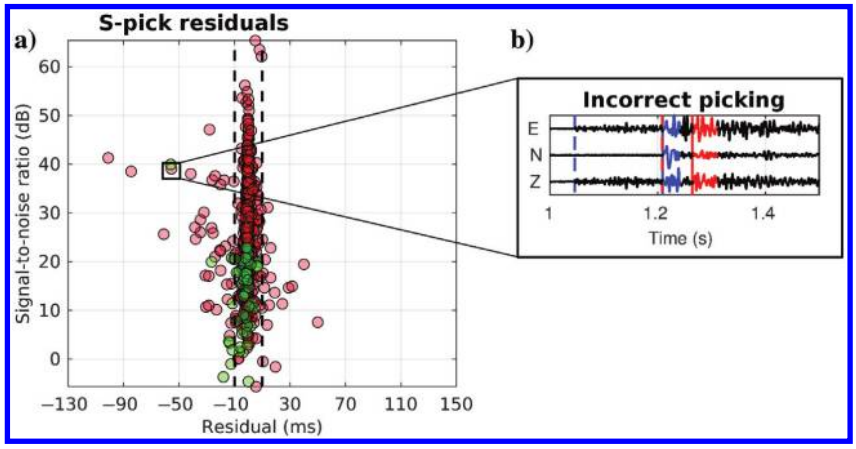

Figure 22. (a) S-pick residuals obtained by the proposed workflow on the real data set. (b) Incorrect picking on the complex waveform. The colors are as indicated in Figure 13.

P residuals on these two data sets. However, when the noise level increases and buries P-waves, S picking by our workflow is more robust. As illustrated in Figures 7–9, S-waves are less affected by noise increments, facilitating arrival picking.

The proposed workflow is not exempt from drawbacks. Because the first arrival is set as the earliest signal interval with high rectilinearity, noise with high rectilinearity values may be set as the first arrival. We also need to consider that the rectilinearity estimation of P-waves can be affected by noise. Another factor that decreases the workflow performance is the waveform complexity. Contamination of S-waves by preceding phases can reduce picking accuracy, and the presence of phases other than direct P- and S-waves may result in incorrect arrival windowing. In addition, on all data sets, our workflow skipped weak P arrivals that were picked by the STA/LTA and Chen (2020) methods with acceptable precision. The previous occurred especially on the synthetic data, where most of these weak arrivals were recorded only on one component. Because we average the three components' membership values to determine signal intervals, low-S/N arrivals recorded on one component may be missed by our workflow. Despite the STA/LTA and Chen (2020) methods picking these arrivals with decent accuracy, it is important to remember that these two methods obtain three picks per 3C record, and here we only consider the pick with minimum residual. In practice, one pick per receiver must be determined, which may reduce the picking accuracy of these methods when low-S/N arrivals are recorded only in one component.

Despite the previous drawbacks, the picking conducted by our workflow is accurate enough to obtain acceptable hypocenter locations. For events with high-S/N waveforms (>20 dB) such as those in synthetic data set one, the arrival picks yielded by our method result in relatively accurate locations for 84 of 100 events (location rms error <30 m). As the noise level increases, the computed arrival times and event back azimuths are less accurate, increasing the hypocenter location error. In synthetic data set two, 52 events were located with an rms error of less than 45 m, and in data set three, the location rms error of 50 events was below 58 m. These are acceptable values considering the low S/N of the waveforms on these data sets (−5 to 15 dB for data set two and −5 to 10 dB for data set three).

Regarding the speed and computational costs of the presented workflow, we ran the algorithm on MATLAB using a single core of the Intel Core i7-9750H CPU processor at 2.6 GHz clock speed. For one 3C record of 0.8 s duration at 0.5 ms sampling rate, the algorithm picks the P and S arrivals in approximately 0.1 s. For existing large data sets, the proposed algorithm could be further parallelized to analyze each 3C trace independently, for speed up.

CONCLUSION

We present a new AIC assisted FCM clustering-based autopicking workflow for efficient identification and picking of P- and S-wave arrival times. The workflow is capable of skipping picking of phases (direct P- and S-waves) not recorded on the waveforms, reducing the number of inaccurate picks. Our workflow also allows picking and phase identification of single-phase events by estimating and comparing the P and S moveouts of analyzed events. This workflow is fully automatic, meaning that almost all of the parameters (e.g., the window duration for trace feature computation and the time difference threshold to classify unidentified picks based on arrival moveouts) are associated with a user-specified estimate of the signal's dominant period. The computational costs of this workflow are very low compared with supervised machine-learning approaches.

As with other arrival-time pickers, this workflow has limitations. First, this workflow only works on previously detected events because the main workflow component involves an FCM clustering to partition signals from noise. For noise-only traces, FCM generates two clusters containing noise with different behavior, making signal detection erratic. Second, the workflow accuracy may decrease in the presence of phases different than the direct P- and S-waves. Third, high-rectilinearity noise may result in incorrect identification of P-waves. Fourth, our workflow may omit picking on arrivals with a minuscule amplitude recorded only on one component. By analyzing the hypocenter locations with different noise levels, we

determine that the picking accuracy is compromised for waveforms with S/Ns <20 dB. Despite these drawbacks, tests on synthetic and real data show that our method is more robust than existing methods. Furthermore, there is room to improve the proposed workflow by using more sophisticated trace features or different picking methods on the detected P and S intervals.

ACKNOWLEDGMENTS

The research reported in this publication was supported by funding from the King Abdullah University of Science and Technology. The authors would like to thank L. Eisner, editor, and the anonymous reviewers for their helpful comments that greatly improved this work. The real data are granted by an anonymous provider.

DATA AND MATERIALS AVAILABILITY

Data associated with this research are available and can be accessed via the following URL: http://doi.org/10.5281/zenodo.4553858.

REFERENCES

Akaike, H., 1973, Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and F. Csaki, eds., Second International Symposium on Information Theory: Budapest Akademiai Kiado, 267–281.

Akram, J., and D. Eaton, 2016, A review and appraisal of arrival-time picking methods for downhole microseismic data: Geophysics, **81**, no. 2, KS67–KS87, doi: 10.1190/geo2014-0500.1.

Aldridge, D. F., 1990, The Berlage wavelet: Geophysics, **55**, 1508–1511, doi: 10.1190/1.1442799.

Artman, B., I. Podladtchikov, and B. Witten, 2010, Source location using time-reverse imaging: Geophysical Prospecting, **58**, 861–873, doi: 10.1111/j.1365-2478.2010.00911.x.

Bezdek, J. C., 1981, Pattern recognition with fuzzy objective function algorithms: Plenum Press.

Bezdek, J. C., R. Ehrlich, and W. Full, 1984, FCM: The fuzzy c-means clustering algorithm: Computers & Geosciences, **10**, 191–203, doi: 10.1016/0098-3004(84)90020-7.

Buland, R., 1976, The mechanics of locating earthquakes: Bulletin of the Seismological Society of America, **66**, 173–187.

Cano, E. V., J. Akram, and D. Peter, 2019, A fuzzy c-means assisted AIC workflow for arrival picking on downhole microseismic data: 89th Annual International Meeting, SEG, Expanded Abstracts, 3056–3060, doi: 10.1190/segam2019-3215089.1.

Chen, Y., 2020, Automatic microseismic event picking via unsupervised machine learning: Geophysical Journal International, **222**, 1750–1764, doi: 10.1093/gji/ggaa186.

Dunn, J. C., 1973, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters: Journal of Cybernetics, **3**, 32–57, doi: 10.1080/01969727308546046.

Fischler, M. A., and R. C. Bolles, 1987, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *in* M. A. Fischler and O. Firschein, eds., Readings in computer vision: Morgan Kaufmann, 726–740.

Gentili, S., and A. Michelini, 2006, Automatic picking of P and S phases using a neural tree: Journal of Seismology, **10**, 39–63, doi: 10.1007/s10950-006-2296-6.

Jurkevics, A., 1988, Polarization analysis of three-component array data: Bulletin of the Seismological Society of America, **78**, 1725–1743.

Komatitsch, D., and J. Tromp, 1999, Introduction to the spectral element method for three dimensional seismic wave propagation: Geophysical Journal International, **139**, 806–822, doi: 10.1046/j.1365-246x.1999.00967.x.

Maeda, N., 1985, A method for reading and checking phase time in autoprocessing system of seismic wave data: Zisin, **38**, 365–379, doi: 10.4294/zisin1948.38.3_365.

Maity, D., F. Aminzadeh, and M. Karrenbach, 2014, Novel hybrid artificial neural network based autopicking workflow for passive seismic data: Geophysical Prospecting, **62**, 834–847, doi: 10.1111/1365-2478.12125.

Moser, T. J., T. Van Eck, and G. Nolet, 1992, Hypocenter determination in strongly heterogeneous earth models using the shortest path method: Journal of Geophysical Research, **97**, 6563–6572, doi: 10.1029/91JB03176.

Nakata, N., and G. C. Beroza, 2016, Reverse time migration for microseismic sources using the geometric mean as an imaging condition: Geophysics, **81**, no. 2, KS51–KS60, doi: 10.1190/geo2015-0278.1.

Oye, V., and M. Roth, 2003, Automated seismic event location for hydrocarbon reservoirs: Computers & Geosciences, **29**, 851–863, doi: 10.1016/S0098-3004(03)00088-8.

Pavlis, G. L., 1986, Appraising earthquake hypocenter location errors: A complete, practical approach for single-event locations: Bulletin of the Seismological Society of America, **76**, 1699–1717.

Ross, Z. E., M. A. Meier, and E. Hauksson, 2018, P wave arrival picking and first-motion polarity determination with deep learning: Journal of Geophysical Research, Solid Earth, **123**, 5120–5129, doi: 10.1029/2017JB015251.

Saragiotis, C. D., L. J. Hadjileontiadis, and S. M. Panas, 2002, PAI-S/K: A robust automatic seismic P phase arrival identification scheme: IEEE Transactions on Geoscience and Remote Sensing, **40**, 1395–1404, doi: 10.1109/TGRS.2002.800438.

Sleeman, R., and T. V. Eck, 1999, Robust automatic P-phase picking: An online implementation in the analysis of broadband seismogram recordings: Physics of the Earth and Planetary Interiors, **113**, 265–275, doi: 10.1016/S0031-9201(99)00007-2.

Song, F., H. S. Kuleli, M. N. Toksöz, E. Ay, and H. Zhang, 2010, An improved method for hydrofracture-induced microseismic event detection and phase picking: Geophysics, **75**, no. 6, A47–A52, doi: 10.1190/1.3484716.

VanDecar, J. C., and R. S. Crosson, 1990, Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares: Bulletin of the Seismological Society of America, **80**, 150–169.

Wang, J., Z. Xiao, C. Liu, D. Zhao, and Z. Yao, 2019, Deep learning for picking seismic arrival times: Journal of Geophysical Research, Solid Earth, **124**, 6612–6624, doi: 10.1029/2019JB017536.

Zadeh, L. A., 1977, Fuzzy sets and their application to pattern classification and clustering analysis, *in* J. Van Ryzin, ed., Classification and clustering: Proceedings of an advanced seminar conducted by the mathematics research center: Elsevier, 251–299.

Zhu, D., Y. Li, and C. Zhang, 2016, Automatic time picking for microseismic data based on a fuzzy C-means clustering algorithm: IEEE Geoscience and Remote Sensing Letters, **13**, 1900–1904, doi: 10.1109/LGRS.2016.2616510.

Zhu, L., E. Liu, J. McClellan, Y. Zhao, W. Li, Z. Li, and Z. Peng, 2017, Estimation of passive microseismic event location using random sampling-based curve fitting: 87th Annual International Meeting, SEG, Expanded Abstracts, 2791–2796, doi: 10.1190/segam2017-17730445.1.

Zhu, W., and G. C. Beroza, 2019, Phasenet: A deep-neural-network-based seismic arrival time picking method: Geophysical Journal International, **216**, 1831–1841, doi: 10.1093/gji/ggy529.

Biographies and photographs of the authors are not available