Rapid and Robust Cross-Correlation-Based Seismic Signal Identification Using an Approximate Nearest Neighbor Method

Rigobert Tibi, Christopher Young, Antonio Gonzales, Sanford Ballard, and Andre Encarnacao

Sandia National Laboratories,

P.O. Box 5800,

Albuquerque, NM 87185-0404.

Email: rtibi@sandia.gov

Phone: 505-844-9843

To be submitted to *Bulletin of the Seismological Society of America*

Abstract

The matched filtering technique involving the cross-correlation of a waveform of interest with archived signals from a template library has proven to be a powerful tool for detecting events in regions with repeating seismicity. However, waveform correlation is computationally expensive, and therefore impractical for large template sets unless dedicated distributed computing hardware and software are used. In this study, we introduce an Approximate Nearest Neighbor (ANN) approach that enables the use of very large template libraries for waveform correlation without requiring a complex distributed computing system. Our method begins with a projection into a reduced dimensionality space based on correlation with a randomized subset of the full template archive. Searching for a specified number of nearest neighbors for a query waveform is accomplished by iteratively comparing it against neighbors of its immediate neighbors. We used the approach to search for matches to each of ~2300 analyst-reviewed signal detections reported for May 2010 for the International Monitoring System (IMS) station MKAR. The template library in this case consists of a data set of more than 200,000 analyst-reviewed signal detections for the same station from 2002 to July 2016 (excluding May 2010). Of these signal detections, 73% are teleseismic first *P*, and 17% regional phases (*Pn*, *Pg*, *Sn*, and *Lg*). The analyses performed on a standard desktop computer shows that the proposed ANN approach performs the search of the large template libraries about 25 times faster than the standard full linear search, while achieving recall rates greater than 80%, with the recall rate increasing for higher correlation thresholds.

Introduction

The correlation of incoming waveform data with archived signals from a template library is a simple and powerful means to improve the detection of events in regions with repeating seismicity (e.g. Gibbons & Ringdal, 2006; Schaff, 2010). Depending on the density of the monitoring network and the completeness of the template library, the potential impact of this technique on monitoring for seismicity is significant. For example, it has been estimated that 85% of events reported in the Annual Bulletin of Chinese Earthquakes (ABCE) from 1985 to 2005 could be detected using waveform correlation, and that correlation could find additional events that are beyond the capability of standard detectors (Schaff, 2009). In Northern California, the percentage of events detected by waveform correlation may be as high as 90% (Waldhauser and Schaff, 2008). Correlation can not only lower the detection thresholds beyond that of traditional methods (e.g. Gibbons & Ringdal, 2006; Schaff and Waldhauser, 2010), and hence extend the magnitudes of completeness toward lower values (Slinkard et al., 2016), but it has also been reported that correlation detectors have been able to find aftershocks obscured by the coda of the mainshocks (Peng and Zhao, 2009; Schaff and Waldhauser, 2010). In principle, using waveform correlation to detect events is simpler to implement than traditional multi-step data processing methods where signal detection and event formation are separate processes, hence it might be expected to be a standard approach for monitoring for new events in regions of repeating seismicity. However, it is also computationally expensive, and this has so far limited use of the technique. The computational cost is directly proportional to the number of templates in the library, which varies according to the scale

of the monitoring task. A single focused region such as a mine or a nuclear test site can be easily monitored with a small number of template event waveforms (e.g. Zhang & Wen, 2015), while typical aftershock sequences can be effectively processed with a few hundred templates (Harris & Dodge, 2011; Slinkard et al., 2013). Processing template libraries of these sizes is feasible using a single standard desktop computer, but to use a template library of sufficient size for broad regional or global scale monitoring – i.e. thousands, tens of thousands, or perhaps even hundreds of thousands of templates per station -- dedicated distributed computing hardware and software are needed (Dodge and Walter, 2015; Slinkard et al., 2016). Development of such systems is ongoing and the results are promising, but at least in the near term there is little prospect that these brute-force approaches will result in anything that can be used by the broader community interested in using waveform correlation to monitor seismicity across large regions.

Data mining studies have shown that approximate search algorithms are computationally efficient in finding similar objects in large data sets with only minor loss of accuracy compared to a standard linear search (e.g. Muja and Lowe, 2009). These fast algorithms, which constitute a class of approaches known as Approximate Nearest Neighbor (ANN), involve, among others, hashing, K-dimensional tree (KDT) or a variant of these two methods (Bentley, 1975; Friedman et al., 1976; Gionis et al., 1999; Andoni and Indyk, 2008; Slaney and Casey, 2008; Silpa-Anan and Hartley, 2008). In practice, by providing for comparison only those objects that are likely to be similar to the query object, both hashing and KD tree methods eliminate the computationally very costly task of examining the many

pairs of dissimilar objects. ANN search returns an approximation of the true nearest neighbor set from an archive of data in exchange for large increases in search speed over a standard linear search. For that reason, ANN approaches are best suited for scenarios where the database is expected to have a large number of useful neighbors and the specific subset of those returned is not critical. ANN approaches have been successfully used in different practical scenarios, mainly for image and audio clip search in large databases and for information retrieval in Web text searches (e.g. Kulis and Grauman, 2009; Wang, 2003; Henzinger, 2006)

Recently, ANN has also found application in seismology with success. For example, Zhang and colleagues have developed a search engine based on multiple randomized K-dimensional (MRKD) trees to quickly estimate earthquake source parameters by searching for similar seismograms from a large database (Zhang et al., 2014). Similarly, Yoon et al. (2015) developed an algorithm for earthquake detection called Fingerprint And Similarity Thresholding (FAST). In that approach, seismic waveforms are first transformed to spectrograms. Spectrograms are then substantially compressed using wavelet transforms to create compact "fingerprints" that are indexed using locality-sensitive hashing (LSH). FAST was successfully applied to the computationally demanding task of correlating continuous data from an aftershock sequence to search for new repeating events.

In this paper, we introduce a new ANN approach that enables the use of very large template libraries for waveform correlation. Our method begins with the creation of lower

dimensional representations of template waveforms based on correlation with a randomized subset of the full template archive. Searching for a specified number of nearest neighbors for a query waveform is accomplished by iteratively comparing it against neighbors of its immediate neighbors. The approach performs a number of small internal queries, each time focusing on a local neighborhood that is refined as new nearest neighbors are identified. For that reason, we called our approach Local Area Focused Search (LAFS). We test the approach on a single standard desktop computer to search for matches to each of ~2300 analyst-reviewed signal detections reported for May 2010 for the International Monitoring System (IMS) station MKAR. The template library in this case consists of a data set of more than 200,000 analyst-reviewed signal detections from the Late Event Bulletin (LEB) produced by the International Data Centre (IDC) for the same station from 2002 to July 2016 (excluding May 2010) (Fig. 1). The IDC is part of the Comprehensive Test Ban Treaty Organization (CTBTO) based in Vienna, Austria. The IDC continuously processes data from the IMS, a globally distributed network of seismic stations, to detect seismic events.

Methods

Proposed ANN Approach

We will use the term "query waveform" hereafter to refer to a waveform that we are trying to identify with our ANN method. Most space-partitioning ANN indexing algorithms like KDTF (K-dimensional tree forest) require a Minkowski distance as a measure of similarity. Minkowski distance $d(X, Y)$ between two points $X(x_1,x_2,\ldots,x_n)$ and $Y(y_1,x_2,\ldots y_n)$ is defined as:

$$d(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{\frac{1}{r}} \quad (1)$$

For $r = 2$, $d(X, Y)$ represents the Euclidian distance. Another common distance measure is Manhattan distance (for $r = 1$). Unfortunately, the Minkowski distance is not appropriate for many types of data. To circumvent this limitation, data can be projected into a higher dimensional space using a non-linear kernel function. This projection results in a new representation of the data objects whose similarity can be assessed via Minkowski distance, therefore, enabling indexing by a space-partitioning ANN method.

A standard method used to accomplish the kernel projection described above is the Kernel Principal Component Analysis (KPCA), which we adopt in this study. The method is discussed in greater details in Wang (2011). From our data set $\boldsymbol{D}$ of m template waveforms, we randomly select a subset of k waveforms. Each of the k waveforms are correlated against all the waveforms in that subset to generate the kernel covariance matrix $\boldsymbol{K}$. Selecting only a limited, representative subset of the template database instead of the entire data set keeps the computational cost associated with covariance computation low. We then perform a Principal Component Analysis (PCA) of the matrix $\boldsymbol{K}$ to reduce dimensionality and create a projection matrix $\boldsymbol{P}$. Using $\boldsymbol{P}$, our template data set $\boldsymbol{D}$ is then projected into $\boldsymbol{D_\phi}$. The projection matrix $\boldsymbol{P}$ is preserved and used in each search to project the query waveform q into the same space as $\boldsymbol{D_\phi}$. The KPCA projection projects the data to a form such that correlation scores are mapped to Euclidean distance, i.e. correlated objects can be found via searching for Euclidean nearest neighbors.

In KDTF indexing for nearest neighbor search, the variance for each dimension of the projected data is evaluated and the variances are ranked. To create a single KD Tree, the dimension with the highest variance is selected as the first division, then the next highest dimension is chosen, and so forth. To generate a set of multiple trees (a "forest"), the dimension to divide on is chosen randomly from the top variance dimensions. This allows us to generate a large number of trees that have different internal structures. By searching these trees in tandem, we are able to return a set of similar items rather than just one, hence we get better recall than by searching any individual tree alone (Silpa-Anan and Hartley, 2008). In a special case of KDTF, which we named KDTF-1, the forest consists of only one tree, which is built with the standard single KD Tree method, i.e. by dividing the projected data successively in half along the dimension with highest variance. KDTF-1 generates a set of potential matches by revisiting the most marginal decision points within KD Tree and making a different choice, i.e. though only one tree is used, it is searched in multiple ways to generate a set of potential matches, similar to what is generated by using a KDTF.

Our ANN approach, which we called Local Area Focused Search (LAFS) is an extension of KDTF. LAFS was inspired by the K-Nearest Neighbor Graph (K-NNG) method of Dong et al. (2011) and the basic principle is that a neighbor of a neighbor is also likely to be a neighbor. In LAFS, searching for nearest neighbors is accomplished by iteratively comparing the query waveform against neighbors of its immediate neighbors using a KDTF in a number of small queries, i.e. neighbors that are returned for the query object are then

used in turn as query objects themselves to return more potential neighbors. The search is optimized by focusing on a local neighborhood that is refined as new nearest neighbors are identified. The main idea behind the proposed approach is that the potential nearest neighbors that are identified during the search process provide better information on locality than backtracking mechanisms offered by standard ANN indexing methods. In LAFS the similarities between a query object and its potential nearest neighbors are assessed by comparing the non-projected data, i.e. we calculate the correlation coefficients of the potential nearest neighbors with the query object to determine similarity. This makes it more likely that the optimal local area around the query object will be searched. The iterative search process ends if either the specified number of requested nearest neighbors is reached or no additional potential matches are found. LAFS is discussed in greater detail in Gonzales and Blazier (2016).

With suitably chosen values for the number of trees, and requested matches, the ANN set returned should include the majority of the true nearest neighbors. However, the number of low-quality matches can be very high. Hence a pruning step is added wherein the true similarity between the query waveform and the returned ANN set is evaluated and all returned items with a similarity below a specified static threshold are screened. Note that LAFS calculates correlation coefficients as part of its neighborhood expansion described above, so these do not need to be recalculated in the pruning step. As shown below, by using search and prune steps and with well-chosen parameters, LAFS can be very fast compared to a full linear search and have a high recall.

Detection Screening Algorithm

As discussed below in the section "Signal Identification Screening", even for high correlation values, not all the matches returned after search and pruning turn out to be waveforms from the same source area as the query waveform. For that reason, we add a screening step applied after pruning but before any further analyses of the signal detections. Our detection screening algorithm is a hybrid approach based both on a cluster algorithm and high correlation threshold. In the method, a detection is accepted only if, depending on the number of matches returned, one of the following is true:

- The detection has four or more matches above the specified threshold and the events associated with at least three of the top four matches belong to a cluster of radius less than 2.5°.
- The detection has two or three matches above the specified threshold and the associated events belong to a cluster of radius less than 2.5°.
- The detection has only one match and the corresponding correlation coefficient value is larger than 0.88. The idea here is that if the waveform similarity is very good, we don't require multiple matches.

To estimate the adequate maximum cluster radius and the minimum correlation threshold for the screening algorithm, we performed a full linear search of the template archive for each query waveform in our test data set (see section "Data Set and Data Processing"), and generated a plot of distances between each query event and the matching template events against the correlation coefficient values (Fig. 2). The chosen maximum cluster radius of 2.5° corresponds to the thickness of the apparent tail defined by the highly correlating data

219 points at shorter distances. This tail represents high-correlating templates whose events
220 occur in the same region as the corresponding query event (allowing for source
221 mislocations), i.e. the tail represents mostly correct matches. The value of 0.88 for the
222 correlation threshold for single matches corresponds roughly to the minimum correlation
223 coefficient that allows only a very limited number of false positives in the set of potential
224 matches. False positive is defined as any potential match whose corresponding event
225 occurs at a distance beyond 2.5°. The occurrence of false positives, some with correlations
226 as high as 0.90, shows why the detection screening based on spatial clustering is a
227 necessary step. These false positives may result from the 19-second template window
228 capturing a signal from a different event than the one the template waveform is associated
229 with (Ballard et al., 2017).
230
231 We understand that, in particular, the chosen maximum cluster radius may be
232 questionable. Therefore, we experimented with other values for that parameter. Radii
233 smaller than 2.5° appeared to be too restrictive, as it resulted in rejection of a number of
234 true detections; and we found no argument to justify increase beyond 2.5°. Using a radius
235 of 2.5°, we found that on average, depending on the values of the correlation threshold,
236 about 80 to 85% of detections that passed the screening process were validated by
237 comparison with the known analyst-reviewed event associations, arguing for the
238 effectiveness of the screening algorithm.

239 ***ANN Performance Score***

240 One common metric for evaluating the accuracy of an ANN method is "recall". The recall of
241 a query object is defined as the number of its true nearest neighbors returned by the ANN
242 algorithm divided by the number of true nearest neighbors obtained by searching the
243 template data set using standard full linear search. The standard full linear search method
244 provides the ground truth; and recall measures the fraction of that ground truth returned
245 by the ANN algorithm. Note that anything less than 100% recall implies that not all of the
246 most similar items were found. ANN methods nearly always exhibit recalls of less than
247 100%, so the challenge in using ANN methods is in balancing improved speed provided by
248 the approximation against decreased recall. If not stated otherwise, recall in this paper
249 refers to the average recall of all the query waveforms in the May 2010 test data set.

250

251 Because ANN methods seek to balance speed and recall, we introduce a performance score
252 (*PS*) as a linear combination of both the query time (*T* in seconds) and recall (*R* in percent)
253 as indicated in equation (2). *PS* expresses the tradeoff that exists between the
254 computational efficiency and the accuracy for ANN methods.

255 $$PS = \alpha R - \beta T \qquad (2)$$

256 The values of 0.01 for α and 3×10^{-4} sec^{-1} for β are chosen so that the standard full linear
257 search, because of its long query time (~1.4 seconds per query on average), has a low *PS* of
258 0, while a hypothetical scenario with a query time of 0.009 seconds per query and a perfect
259 recall of 100% would have a *PS* of 1. 0.009 seconds per query is about the shortest average
260 query time we have observed when searching matches for all the query waveforms in our

archive of templates. Positive (negative) value for *PS* means that ANN is performing better (worse) than the full linear search, and a perfect score corresponds to *PS* of 1.

Our LAFS Application: Rapid Signal Identification

For this study, we chose to apply our LAFS method to the problem of identifying a query waveform that has been flagged by a signal detection algorithm, e.g. an STA/LTA energy detector (e.g. Withers et al., 1998). In traditional event detection systems, continuous waveform data from a network of stations are processed with signal detection algorithms to identify possible signals that may correspond to events of interest. Each signal corresponds to a query waveform, but once the signal detections have been marked and a standard set of parameters measured for each (e.g. travel time, azimuth, slowness, signal-to-noise ratio), events are detected and located based on the signal detections, not the query waveforms. To detect an event, the set of signal detections for that event recorded at each station must be associated with the event and this implies that the phase of each signal must be identified. The accuracy of this phase identification is critical to making event detection work properly and is generally done by comparing the measured signal parameters with modeled values from the hypothesized event location. Errors in either the measurement of the observed signal parameters or in the modeled values for the hypothesized event location can result in improperly built events (e.g. wrong locations) or missed events. For signals corresponding to repeating events, correlating the query waveform with an archive of labeled waveform templates provides an excellent way to

reliably identify the signal and thereby can substantially improve the way the corresponding event is detected and located by the system.

Using waveform correlation to identify signals, however, has heretofore been impractical because the method will have very limited impact unless the archive of signals for each station is large (ideally spanning many years or decades), and correlating each query waveform against a large archive is computationally expensive. Our LAFS method provides a means to overcome this problem and exploit a very large waveform archive with computation performance sufficient to meet the needs of an automatic event monitoring system, and without requiring use of a complex dedicated distributed computing system. All of our computations were done using a single commodity desktop computer with 12 processors and 16 GB RAM. Our LAFS algorithm is implemented in Java.

Data Set and Data Processing

The IDC produces an automatic bulletin, which is then reviewed and corrected by analysts to produce the LEB, which records all of the analyst-reviewed events and corresponding signal detections prior to applying event screening criteria. We selected station MKAR as the focus of our study, because it is the most active IMS seismic station on the Asian continent, contributing to 314,328 of the 563,587 total LEB events (~56%) between 2002, the year MKAR became part of the IMS, and July 2016 (Fig. 1).

303 The template library consists of 248,237 analyst-reviewed MKAR signal detections for
304 2002 to July 2016 (excluding our May 2010 test set). The phase distribution of the template
305 data set is shown in Fig. 3, sorted in decreasing order. By far the most common type of
306 signal (73%) is teleseismic *P*. Next most common is *Pn*, but it is only 9% of the total. The
307 other regional phases are similarly represented: *Lg* at 4%, *Sn* at 3%, and *Pg* at 1%. This is
308 significant, because if in fact waveform correlation is only effective at regional distances, as
309 has been suggested in previous studies (e.g. Walter and Dodge, 2015), then only 17% of our
310 data set is relevant for waveform correlation. Testing that assumption is one of the
311 fundamental goals of this study. Except for *pP*, we did not include secondary phases for
312 two reasons. First, because there are relatively few secondary phases present in the LEB
313 catalog, and hence identification using waveform correlation tends to work less well for
314 these due to the limited catalog of template waveforms available to search. Second, because
315 secondary phase waveforms tend to be more generic due to the attenuation of high
316 frequencies along the longer paths. We tested our LAFS methodology by attempting to
317 identify the full set of 2,302 analyst-reviewed MKAR signal detections for May 2010 using
318 our template library.
319
320 Although MKAR is a 9-element short-period array, we chose to use only a single vertical
321 channel, MK05 SHZ. Waveform matching can be done using all the elements of an array, but
322 the intent of this study is to evaluate the general applicability of our methodology to the
323 vast majority of seismic stations in the world, and most stations are not arrays. After some
324 experimentation, we settled on using a 19-second waveform window length, starting 2

seconds before each analyst-reviewed signal detection. Choosing a shorter window speeds up both the KPCA data projection as well as the calculations involved in the search step, but we found that using a window shorter than 19 seconds resulted in an obvious degradation in the accuracy of the matches returned. Using a variable length window based on phase/distance probably makes sense, but building a single KPCA-indexed archive for the full set of MKAR signals requires us to use the same window length for all waveform templates. Similarly, we recognize that narrower band filters would be more appropriate for some phases/distances, but our LAFS method requires that we apply the same filter to all templates to build the KPCA-indexed archive, so a broad band filter made sense. Hence, we chose to use a generic 0.5-5.0 Hz bandpass Butterworth 3-pole filter as a good compromise for the wide variety of signal types and distances present in the MKAR archive.

Our LAFS-Based Signal Identification System

We begin by building an ANN index for our MKAR waveform template library, as described previously. This index is kept in memory, as are all of the corresponding waveform templates, which are needed for the pruning step. Hence, our method actually requires more memory usage than a full linear search method, which would only require the templates, but this increased memory requirement is offset by the improved search speed provided by the index. When a new query waveform is submitted for matching, in the search step we apply the LAFS algorithm to search the in-memory ANN index for the set of potential matches. In the pruning step, after the ANN indices have been returned, the

results are sorted by the value of the correlation coefficient, and only those with values higher than a specified static threshold are returned.

LAFS allows us to search a much bigger template library by down selecting from the full waveform archive the set of waveforms that are then being correlated against the input waveform to find the correlation coefficients and the optimal time lags. The performance increase is directly proportional to the reduction in the number of waveforms returned from the search step. For example, if the LAFS request specifies to return only 5% of the full waveform archive, then the search should be ~20 times faster than a full linear search. The smaller the size of the ANN set returned by LAFS, the better the computational efficiency, but at the potential cost of accuracy if the LAFS down selection step misses similar waveforms. This tradeoff should always be kept in mind when using any ANN method.

Configuring LAFS Waveform Searches

We tested our LAFS approach using different sets of parameters, each time comparing the returned results against those from a full linear search by evaluating both the recall and performance score (*PS*) introduced above. The purpose of this testing exercise was to find the set of optimum parameters that yields the best performance. Results from the full linear search were screened beforehand using the detection screening algorithm described in section "Detection Screening Algorithm" to ensure that only correct identifications were considered in the ANN performance evaluation. Fig. 4a shows recall for LAFS as a function of correlation threshold for different numbers of requested nearest neighbors. For

correlation threshold values of 0.80 and larger and for numbers of returns of 8,000 and 10,000, LAFS recall values are essentially 100%. LAFS query time increases linearly from ~22 seconds for 1,000 returns to ~163 seconds for 10,000 returns, while the search time for the full linear search method remains nearly constant and high at ~3,333 seconds across the range of returns (Fig. 4b). In general, the higher the number of requested returns, the better the recall, but the longer the search time. Fig. 4c shows LAFS performance score as function of correlation threshold for different numbers of requested returns. The full linear search is also shown with *PS* of 0 throughout the range of correlation threshold values. Even for a number of returns as low as 1,000, LAFS-yielded *PS* values for the range of threshold values considered are above those of the full linear search. At correlation threshold values higher than 0.8, LAFS *PS* values are all close or equal to the optimal score (i.e. *PS* of 1) for 1,000-10,000 returns. The cross-overs between the performance curves at higher threshold values in Fig. 4c result from the increase in search time for larger returns. Based on the arguments provided above, 8,000 was chosen as the preferred number of returns to achieve an optimal balance between recall and speed of query.

As mentioned in section "Proposed ANN Approach", LAFS is an extension of KDTF, and as such, LAFS involves tree forests. Requesting 8,000 returns each time, we investigated LAFS performance for forests of 1, 5, 10, 20 and 30 trees (Fig. 4d). The *PS* curves for the different number of trees do not differ that much; starting from a threshold value of 0.8, they all converge to and flatten near the *PS* value of ~1. The curves for 20 and 30 trees appear to be

almost identical. We chose 20 as the preferred number of trees to limit the computational cost associated with using more trees.

It is important to note that the choice of proper LAFS parameter values is very much dependent on the size and character of the template archive being used, hence application to a different signal archive (e.g. for a different IMS station) would imply a reassessment of the optimal parameter values.

Comparison of LAFS with KDTF and KDTF-1 Methods

We introduced KDTF and KDTF-1 above. This section focuses on the comparison of LAFS with these known ANN methods. Fig. 5 shows *PS* curves for different number of trees in the KDTF method. Also shown in the same figure are *PS* curves for KDTF-1 and the full linear search. For 1 to 125 trees, the overall performance for KDTF improves with increasing number of trees. For larger number of trees, however, the performance appears to deteriorate dramatically. In fact, for 200 trees, KDTF performance across the entire range of the correlation threshold values is far below that of the full linear search. The observed deterioration of performance for larger number of trees is the direct consequence of exploding query times. Between 1 and 125 trees the search time increases only slowly from about 115 to 137 seconds. The query time jumps to ~808 seconds for 150 trees, and reaches ~6,035 seconds for 200 trees, almost twice the average search time for the full linear search. KDTF-1 performs surprisingly well. Its scores appear to be as good as those of KDTF with 25 to 50 trees.

Using their respective optimum parameters, the ANN approaches (LAFS, KDTF, and KDTF-1) were each applied to the test data set. In Fig. 6a the histograms of correlation coefficient values obtained for the ANN approaches are plotted together with that of the full linear search method. The area enclosed by each histogram is a measure of the overall recall for the corresponding method. Among the ANN approaches, LAFS has the highest overall recall, followed by the KDTF. The difference between LAFS and the full linear search comes mostly from correlation coefficient values below about 0.80, i.e. the similar waveforms missed by ANN are lower correlating ones. This implies that for correlation threshold values of 0.80 or higher, LAFS would be as accurate as the full linear search method. The performance curves obtained for the respective optimum parameters are shown in Fig. 6b. As expected, when the search times are taken into account, all three ANN approaches score far better than the full linear search method. At correlation threshold value of 0.60, KDTF performance appears to be slightly below than that of KDTF-1. However, a cross-over occurs at 0.65, as the former method starts to outperforms the latter. LAFS is superior to both KDTF and KDTF-1 for all threshold values below 0.85.

Results of LAFS Signal Detection for the May 2010 MKAR Query Data Set

An Example of ANN Results from the Search Step

Fig. 7 shows an example of the full ANN set returned by our LAFS method for an *Lg* phase from a May 2010 regional event in the region of Kemerovo, in south-central Russia at a distance of 8.2^{o} from MKAR. The query waveform is shown in the bottom. To the top left is a plot of the distances between the query event and the events of the ANN set. Evidently the

ANN set includes signals from events all over the world, not just events in the source region of the query event, thus affirming the need for the pruning step. Dark vertical bands in this plot indicate many returned archive signals at a variety of correlation coefficient values from particular distances. To the top right is shown a plot of the actual correlation coefficient values corresponding to the returned ANN set. The shape of this curve is approximately Gaussian and the mean value is 0.47, indicating that the vast majority of the ANN waveforms have low similarity with the query waveform and will be rejected in the subsequent pruning step. To explore the portion of the ANN set with high correlation coefficient values, in the middle left plot we again plot the set of distances between the query event and the events of the matching templates sorted in decreasing order of correlation coefficient values (i.e. the same as the plot directly above), but only for the members of the ANN set with correlation coefficient larger than 3 standard deviations above the mean, i.e. 0.75. This value is indicated with a vertical dashed line in the top right plot of correlation coefficient values. This reduces the total number of correlation coefficients values plotted to 21, and in this much more limited set of the ANN values, 7 members are from events co-located with the query event (i.e. 0^{o} distance), as can be seen in a histogram of distances to the right. In fact, all of the top 5 correlation coefficient values come from this distance. Using our pruning algorithm with a static threshold of 0.6, our LAFS method found ~800 potential matches for this query waveform, but as can be seen in the figure most of these do not come from the correct source region.

Signal Identification Screening

As the above example illustrates, an additional screening step is necessary to robustly identify a query waveform, hence the introduction of the screening methodologies discussed earlier. Fig. 8 shows another example that further motivates our approach. This is a *P* phase from a magnitude (m_b) 3.0 event 49° away off the cost of Italy. Of the pool of potential matches identified by LAFS, 35 had a correlation coefficient larger than 0.6. All the 10 best correlating matches are *P* signals. The corresponding events for three of these are in fact from the region off the cost of Italy (the second, fourth and ninth waveforms up from the query waveform at the bottom), but neither of these are the most similar. If we had identified the query waveform based on the most similar ANN signal returned, it would have been identified as a *P* from the Java Sea 96.8° away from the actual location. This example, clearly shows that identifying a signal based solely on the best correlating waveform can sometimes be problematic. Fig. 9 illustrates a case for another *P* signal from a magnitude (m_b) 4.1 event 45.2° away near the Kuril Trench. Again, all the LAFS matched and pruned ANN waveforms look similar and have high correlation coefficients (up to 0.90). In this case, there were 4,354 matches with correlation coefficient larger than 0.6; the figure shows only the top 10. Obviously, the returned waveforms must be similar for a successful match, but it is also clear that there is an additional factor in this case that was not present in the previous example: as indicated by their respective distances to the query event, the template events for the top 3 potential matches are located close to each other, suggesting that these waveforms are coming from events in a limited region. In other words, the top matches are not just similar looking waveforms, but they are similar looking

476 waveforms predominantly from events in the same source region. The detection screening
477 approach previously described in section "Detection Screening Algorithm" was developed
478 with that in mind. For an actual case, the location of the query event is not known a priori,
479 hence the spatial relationship between the template events is assessed in the screening
480 method by evaluating their inter-event distances through a cluster algorithm. One might
481 well ask why we did not choose to use phase label consistency in addition to spatial
482 relationship. We found phase labels to be much less reliable because phase names can
483 change at transition distances (e.g. *Pn* to *P*) and because they are not always applied
484 consistently between analysts (e.g. *PKiKP* vs. *PKPcd*).

485 ***Examples of Signal Detections***

486 Fig. 10 shows an accepted *Pn* signal identification from an event at distance of 2.8°. As has
487 been shown in numerous other studies, waveform correlation tends to perform very well
488 for regional distance signals due to the high time-bandwidth product of the waveforms (the
489 signals have long durations and are spectrally complex). Clearly, a signal with this
490 temporally extensive complexity is encoding very path specific information, so it is not
491 surprising that all the events for the 10 best matches shown (there were 25 in total with
492 correlation coefficient larger than 0.6) are from the same region.

493

494 We could show many more examples of successful regional phase identifications beyond
495 those shown in Fig. 10, but choose to instead focus on teleseismic identifications because
496 that is a unique aspect of this study.

497

Fig. 11 shows one such match for a *P* phase from a magnitude (m_b) 3.4 event at a distance of 59.8° in the Morotai Island, Indonesia. The background noise level is moderately high, and the signal is apparently short in duration and fairly simple (compared to the regional distance signals in Figure 10), all of which would suggest that this signal would be difficult to identify. Yet our algorithm robustly identifies it, returning 837 matches with correlation coefficient larger than 0.6, and with short inter-event distances for the top 7 matches.

Fig. 12 shows another accepted match for a magnitude (m_b) 3.7 event at a distance of 51.7° in the Sumatra subduction zone. Again, the background noise level is moderately high relative to the signal. In this case the signal has a longer temporal extent and exhibits more complexity, though it is still much more generic looking than a typical regional distance signal. Once again, our system was able to robustly identify this signal based on inter-event distances for the top 4 matches. In fact, events for all the top 10 matches except one are from the same region. We also note that the matches extend back to 2002, illustrating the value of using a long baseline template library when matching earthquake waveforms. Except during aftershock sequences, earthquakes have very low recurrence rates.

Successful signal identifications were also found at much further distances. Fig. 13 shows an accepted identification for a magnitude (m_b) 3.7 event at a distance of 145.0° in northern Chile. The signal embedded in moderate background noise shows some degree of complexity, hence it is not surprising that the identification is robust: events for all the top 10 matches are from the same source region, as inferred by their inter-event distances.

With our screening process, we were able to identify this far-teleseismic distance waveform correctly as a *PKPbc* from northern Chile.

Summary of Signal Detection Results

The search results for 3 static threshold values are summarized in Fig. 14. For the chosen threshold value of 0.6, about half of the 2,302 MKAR May 2010 detections have no matches, and about half had one or more matches. Our screening algorithm rejected 41% of the total number of detections (i.e. false alarms), which reduces the final number of accepted matches to 11% (Fig. 14a). None of the 248 accepted matches involves the single match variety. Increasing the static correlation threshold to 0.7 reduces both the percentages of false alarms and the number of accepted matches to 13% and 7%, respectively (Fig. 14b). For a static threshold of 0.8, the percentage of accepted matches drops dramatically to 4% (Fig. 14c). This is due to the fact that, for such a high threshold value, the pruning process returns no matches for most (94%) of the queries.

The achieved proportion of 11% for accepted detections (for the 0.6 correlation threshold) is decidedly lower than some other waveform correlation studies (e.g. Waldhauser & Schaff, 2008), but we believe it makes sense. Those studies typically focus on regional or near teleseismic distances where template time-bandwidth products are much better. Also, at those distances, the template sets have much lower magnitude of completeness than the global value of ~3.7 (m_b) we calculated for the LEB, hence we believe those studies had a much more complete library of templates to compare signals against. Being able to build a template library with a long time line (15 years in our case) helps make a more complete

542 template library, but we are still limited by the completeness of the LEB event catalog,
543 which in turn reflects the sparse spacing of the IMS network. The LEB is predominantly a
544 teleseismic distance catalog; there are very few events that consist of only regional distance
545 signal detections. Hence many smaller events that might fill in missing templates on faults
546 are not in our archive and hence cannot detect new events occurring in those same areas.
547 Given those considerations, an 11% identification rate is not surprising.
548

549 **Discussion**

550 The identification results categorized by phase are shown in Fig. 15 (only phases with 5 or
551 more screened detections are shown). For each phase, we show the unscreened and the
552 screened results. Because we know the true event information for the May 2010 signals, we
553 can also further check the screened detections against the actual detections by examining
554 the locations. We characterized a screened detection as validated if at least one event of the
555 top 4 matching templates is located within 2.5° from the query event. Validated detections
556 are shown as the "Validated" bars in Fig. 15. Our detection screening can fail if the top
557 match(es) returned by LAFS satisfy the criteria described in section "Detection Screening
558 Algorithm", but the associated events have no actual spatial relationship with the query
559 event. The degree to which this happens depends on the phase. For the regional phases (*Pn*,
560 *Pg*, *Sn*, and *Lg*), we would expect the complex waveform signatures to be very path-specific
561 so this sort of misidentification should rarely happen and indeed for *Lg* and *Sn* almost all of
562 the accepted identifications were validated (98-100%). For telseismic first *P*, the
563 proportion validated is lower (84%) but still very good, and this is important because as

564 we pointed out earlier, the majority of signal detections made at MKAR are teleseismic first
565 *P*. For the far-teleseismic phases (*PKP*, *PKPab*, *PKPbc*) the percentage of validated
566 detections is expectedly low (17-90% with an average of 55%) as consequence of
567 attenuation of high frequencies along the longer paths. For *pP*, the numbers of screened
568 detections of 2-3 were too low for sound statistics.
569
570 Using a static threshold of 0.60 coupled with our screening methodology, a total of 205 out
571 of 2302 MKAR May 2010 signal identifications were accepted and validated. A map
572 showing the locations of events associated with all the validated signal identifications is
573 shown in Fig. 16. Comparing with the total set of events recorded at MKAR from 2002 to
574 July 2016 (see Fig. 1), it is apparent that the majority of successful phase identifications
575 were for telesismic distant events all over the world (~70%). Of these, the majority are
576 from the western Pacific, which we think is directly related to the fact that this area is very
577 seismically active and is close enough to MKAR such that many of the events produce
578 detectable signals. Hence the archive of template waveforms in the ANN index for this
579 region for 2002 to July 2016 is larger than for any other area, and also the number of
580 phases to identify in May 2010 for this region is larger than for any other.
581
582 Another factor in this unexpectedly high ratio of teleseismic to regional phase
583 identifications is the inherent under-sampling of smaller events in the LEB catalog. As
584 mentioned earlier, the spacing of the IMS network is sparse enough that for most locations
585 it cannot detect smaller events that can only be seen at regional distances (LEB events must

include signal detections from at least three stations). Hence the LEB catalog of events includes far fewer events at regional distances from MKAR than would be recorded in a catalog developed from a true regional network. Thus there are fewer available regional distance templates in the 2002-2016 catalog. Also, there are fewer regional distance detections in our May 2010 test set to try. Quite likely there were a lot more regional signals recorded at MKAR, but those never made it into events, hence weren't included in the LEB arrivals.

Conclusions

We have developed a fast and efficient ANN method (LAFS) based on KPCA projected data and randomized KDTF indexes to search a very large archive of signals to identify the source location and phase of a query waveform. Our method consists of a "search" step and a "pruning" step. In the search step, LAFS is used to return a set of approximate nearest neighbors. In the pruning step, only those members with correlation coefficients values above a specified static threshold value are returned as potential matches. To achieve robust phase identifications, we added a post-LAFS step that screens results based on inter-event distances between the events associated with the best-correlating LAFS matches for multiple LAFS matches and high correlation value for single LAFS matches. Detection is declared only when the potential match(es) satisfy the screening criteria. We tested our method by creating an ANN index of IDC LEB analyst-reviewed signal detections for station MKAR from when the station first became operational in 2002 through July 2016 (excluding May 2010). This archive was then used to try to identify the analyst-

608 reviewed signal detections for May 2010 from the same station. The analyses performed on
609 a standard desktop computer shows that LAFS performs the search of the large template
610 library about 25 times faster than the standard full linear search, while achieving recall
611 rates greater than 80%, with the recall rate increasing for higher correlation values. For the
612 optimal set of parameters, LAFS typical search time of the template database is as low as
613 ~60 milliseconds per query. For a static threshold of 0.6, 49% of the May 2010 waveforms
614 returned no potential matches and hence could not be identified; about 40% did not pass
615 the detection screening algorithm and hence also could not be identified. Thus, 11% of the
616 May 2010 test waveforms were assigned phase identifications by our algorithm. About
617 83% of these were validated, i.e. matched the known analyst-assigned phases. Successful
618 phase identifications were made for events all over the world, from near regional to far
619 teleseismic distances. By far the most common phase identified (~70%) was teleseismic
620 first *P*, reflecting the dominance of this phase in the LEB historic catalog of phases for
621 MKAR. Our results suggest that waveform correlation can be used effectively for identifying
622 some teleseismic phases if the catalog of template waveforms is sufficiently large.
623
624 This study was made as simple as possible to test the effectiveness of the LAFS method as it
625 might be applied to a typical seismic station (i.e. not an array), hence we used only a single
626 channel from the MKAR array. Efforts are underway to take advantage of the signal
627 coherency and move-out across the array by building the ANN index using data from
628 multiple elements of the MKAR array. This is expected to dramatically reduce the number
629 of false positives in the ANN set of potential matches, because in addition to matching the

shape of the waveforms at each element, the move-out of the signal across the array must be matched. In future efforts, we will also explore frequency dependence by building separate ANN indexes for several overlapping narrower bands to try to better enhance the signals from particular event locations. We will also investigate using LAFS directly as a signal detector rather than as an identifier of signals detected by another algorithm. Already the LAFS algorithm performs better than real-time and so could be used as an operational detector.

Data and Resources

MKAR is an IMS array located in Kazakhstan. MKAR waveform data and LEB catalog data (events and arrivals) were obtained from IDC, a part of the Comprehensive Test Ban Treaty Organization (CTBTO) in Vienna, Austria.

Acknowledgements

This research was funded by the U.S. Department of Energy. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Andoni, A., and Indyk, P. (2008). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, *Comm. ACM* **51,** no. 1, 117-122.

Ballard, S., C. Young, A. Gonzales, A. Encarnacao, R. Tibi, and R. Brogan (2017). Seismic event bulletin construction for a global sparse network using waveform correlation, manuscript in preparation.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching, *Comm. Am. Assoc. Comp. Mach.* **18,** no.9, 509–517.

Dodge, D. A., and W. R. Walter (2015). Initial global seismic cross-correlation results: Implications for empirical signal detectors, *Bull. Seismol. Soc. Am.* **105,** no. 1, 240–256, doi: 10.1785/0120140166.

Dong, W., C. Moses, and K. Li (2011). Efficient k-nearest neighbor graph construction for generic similarity measures, in *Proceedings of the 20th International Conference on World Wide Web*, ACM, New York, NY, 577–586.

Friedman, J. H., J. L. Bentley, and R. A. Finkel (1975). An algorithm for finding best matches in logarithm expected time, *ACM Trans. Math. Softw.* **3,** 209-226.

Gibbons, S. J., and F. Ringdal (2006). The detection of low magnitude seismic events using array-based waveform correlation, *Geophys. J. Int.* **165,** 149-166.

Gionis, A., P. Indyk, and R. Motwani (1999). Similarity search in high dimensions via hashing, in *Proceeding of the 25th VLDB Conference* **99,** 518–529.

Gonzales and Blazier (2016). Enhanced approximate nearest neighbor via local area focused search, *Computer Vision and Pattern Recognition Conference*, Submitted.

Harris, D. B., and D. A. Dodge (2011). An autonomous system for grouping events in a developing aftershock sequence, *Bull. Seismol. Soc. Am.* **101,** no. 2, 763–774, doi: 10.1785/0120100103.

Henzinger, M. (2006). Finding near-duplicate Web pages: a large-scale evaluation of algoritms, *Proceedings of the 29th SIGIR Conference*, Seattle, WA, 284-291.

Kulis, B., and K. Grauman (2009). Kernelized locality-sensitive hashing for scalable image search, in *Computer Vision, IEEE 12th International Conference*, 2130–2137.

Muja, M., and D. G. Lowe (2009). Fast approximate nearest neighbors with automatic algorithm configuration, in *VISAPP International Conference on Computer Vision Theory and Applications*, 331–340.

692 Peng, Z., and P. Zhao (2009). Migration of early aftershocks following the 2004 Parkfield
693 earthquake, *Nat. Geosci.* **2,** 877–881, doi: 10.1038/NGEO697.
694
695 Schaff, D. (2010). Improvements to detection capability by cross correlating for similar
696 events: A case study of the 1999 Xiuyan, China, sequence and synthetic sensitivity tests,
697 *Geophys. J. Int.* **180,** no. 2, 829– 846, doi 10.1111/j.1365-246X.2009.04446.x.
698
699 Schaff, D. P. (2009). Broad-scale applicability of correlation detectors to China seismicity,
700 *Geophys. Res. Lett.* **36,** L11301, doi 10.1029/2009GL038179.
701
702 Schaff, D. P., and F. Waldhauser (2010). One magnitude unit reduction in detection
703 threshold by cross correlation applied to Parkfield (California) and China seismicity, *Bull.*
704 *Seismol. Soc. Am.* **100,** no. 6, 3224–3238, doi: 10.1785/0120100042.
705
706 Silpa-Anan, C., and R. Hartley (2008). Optimized KD-trees for fast image matching. In. *CVPR*
707 *IEEE Computer Society*.
708
709 Slaney, M., and M. Casey (2008). Locality-sensitive hashing for finding nearest neighbors,
710 IEEE Signal Process. Mag. **128,** doi: 10.1109/MSP.2007.914237.
711

712 Slinkard, E. M., D. B. Carr, and C. J. Young (2013). Applying waveform correlation to three
713 aftershock sequences, *Bull. Seismol. Soc. Am.* **103,** no. 2A, 675–693, doi:
714 10.1785/0120120058
715
716 Slinkard, M., S. Heck, D. Schaff, N. Bonal, D. Daily, C. Young, and P. Richards (2016).
717 Detection of the Wenchuan aftershock sequence using waveform correlation with a
718 composite regional network, *Bull. Seismol. Soc. Am.* **106,** no. 4, 1371–1379, doi:
719 10.1785/0120150333.
720
721 Waldhauser, F., and D. P. Schaff (2008). Large-scale relocation of two decades of Northern
722 California seismicity using cross-correlation and double-difference methods, *J. Geophys. Res.*
723 **113**, B08311, doi:10.1029/2007JB005479.
724
725 Wang, A. (2003). An industrial-strength audio search algorithm, *Proceedings of the*
726 *International Conference on Music Information Retrieval*, Baltimore, MD, 713-718.
727
728 Wang, Q. (2014). Kernel principal component analysis and its applications in face
729 recognition and active shape models, arXiv preprint arXiv:1207.3538v3[cs.CV], 1–9
730
731 Withers, M., R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, S., and J. Trujillo (1998). A
732 comparison of select trigger algorithms for automated global seismic phase and event
733 detection, *Bull. Seismol. Soc. of Am.* **88,** no. 1, 95–106.

Yoon, C. E., O. O'Reilly, K. J. Bergen, and G. C. Beroza (2015). Earthquake detection through computationally efficient similarity search, *Sci. Adv.* **1,** e1501057.

Zhang, M., and L. Wen (2015). Seismological evidence for a low-yield nuclear test on 12 May 2010 in North Korea, *Seismol. Res. Lett*. **86,** no. 1, 138–145, doi: 10.1785/02201401170.

Zhang, J., H. Zhang, E. Chen, Y. Zhen, W. Kuang, and X. Zhang (2014). Real-time earthquake monitoring using search engine method, *Nat. Comm.* 5, 5664, doi: 10.1038/ncomms6664.

Full mailing addresses

Rigobert Tibi
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0404

Christopher Young
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0401

Antonio Gonzales
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0974.

Sanford Ballard
Sandia National Laboratories
P.O. Box 5800,
Albuquerque, NM 87185-0404

765 Andre Encarnacao

766 Sandia National Laboratories

767 P.O. Box 5800

768 Albuquerque, NM 87185-0401

List of Figure Captions

Figure 1. Map showing the locations (black dots) of 249,237 IDC LEB analyst-reviewed events with associated MKAR signal detections between 2002 and July 2016 (excluding May 2010). The white triangle indicates the location of the seismic station MKAR.

Figure 2. Plot of distances between each query event and the events of the matching templates against the correlation coefficient values. The data points were generated by performing a standard full linear search of our archive of 249,237 templates for each query from our data set of 2,302 queries. (Left) entire data set. (Right) zoom in of the area within the black box in the plot to the left. The chosen maximum cluster radius and minimum correlation threshold for the detection screening algorithm are indicated with the horizontal and vertical dashed lines, respectively.

Figure 3. Histograms of the 249,237 IDC LEB analyst-reviewed signals detections for station MKAR from 2002 to July 2016 (excluding May 2010), grouped by phase and sorted by decreasing count. The percentage of detections for each phase is indicated at the top of the corresponding bar.

Figure 4. (a and c) Recall and performance score of LAFS and the full linear search method (FLS) as function of correlation threshold value for the number of requested ANN returns of 1000, 2000, 4000, 6000, 8000 and 10,000 using 15 trees. (b) Search time of LAFS and FLS as function of number of ANN returns for 15 trees. (d) Performance score of LAFS and

FLS as function of correlation threshold value for 1, 5, 10, 15, 20, and 30 trees while requesting 8000 returns for each case.

Figure 5. Performance score of KDTF as function of correlation threshold value for 1, 5, 10, 25, 50, 100, 125, 150, and 200 trees while requesting 8000 returns for each case. Performance curves for KDTF-1 and the full linear search method (FLS) for the same number of returns are also shown.

Figure 6. (a) Histogram of correlation coefficient values of the returned matches for LAFS (red), KDTF (blue), KDTF-1 (green), and the full linear search method (FLS) (white). The area enclosed by each histogram is a measure of the overall recall for the corresponding method. (b) Performance score as function of correlation threshold value for LAFS, KDTF, KDTF-1, and FLS. Both the histograms in (a) and the performance curves in (b) were generated using the respective optimum parameters for the methods.

Figure 7. ANN results for an example query waveform. (Bottom) the query waveform, a regional *Lg* phase. (Top left) distances between the query event and the events of the ANN set returned. (Top right) the corresponding correlation coefficient values. Dashed line indicates threshold of 3 standard deviations above the mean. (Middle left) zoomed in version of the plot of distances between the query event and the events of the ANN set showing only correlation coefficients values greater than the threshold. (Middle right)

812 histogram of distances between the query event and the events of the matching templates
813 for correlation coefficients values greater than the threshold.
814
815 **Figure 8**. Waveform profile showing the 10 best correlating matches for the 806th analyst-
816 reviewed MKAR signal detection of May 2010, a *P* signal from an event 49^{o} away. Julian
817 date, phase, and station-to-event distance (delta in degrees) for the query are indicated at
818 the top of the plot. The query waveform is shown at the bottom of the waveform section
819 (thick line). The signal detection is at 2 seconds. Matched waveforms (shifted by the
820 optimum time lag for best correlation) are plotted above in order of decreasing correlation
821 coefficient (cc) moving upward through the section, i.e. the most similar waveform is
822 directly above the query waveform. For each matching waveform, correlation coefficient,
823 Julian date, phase, and distance to the query event are indicated.
824
825 **Figure 9.** Waveform profile for an accepted identification: a *P* signal from an event 45.2^{o}
826 away. Organization of plot is the same as in Fig. 8. There were 4,354 matches with
827 correlation coefficient larger than 0.6, but we only show the 10 best. For each matching
828 waveform, correlation coefficient, Julian date, phase, and distance to the query event are
829 indicated.
830
831 **Figure 10.** Waveform profile for an accepted *Pn* identification from an event 2.8^{o} away.
832 Organization of plot is the same as in Fig. 8. The high time-bandwidth product of the query

833 waveform (lowermost) effectively guarantees that all potential matches with high
834 correlation coefficient values will be from the same source region.
835
836 **Figure 11.** Waveform profile for an accepted teleseismic first *P* identification from an event
837 59.8° away in Morotai Island, Indonesia. Organization of plot is the same as in Figure 8.
838
839 **Figure 12.** Waveform profile for an accepted teleseismic first *P* identification from an event
840 51.7° away in the Sumatra subduction zone. Organization of plot is the same as in Figure 8.
841
842 **Figure 13.** Waveform profile for an accepted *PKPbc* identification from an event 145° away
843 in northern Chile. Organization of plot is the same as in Figure 8.
844
845 **Figure 14.** Pie charts showing the percentages of the 2,302 detections processed with
846 LAFS. "No Matches" indicates the percentage of detections for which LAFS found no
847 potential matches; "Accepted Matches" is the percentage of detections that passed the
848 screening process (i.e. true detections), and "Rejected Matches" the percentage of
849 detections that was rejected by the screening process (i.e. false alarms). (a-c) For static
850 correlation threshold values of 0.6, 0.7, and 0.8, respectively.
851
852 **Figure 15:** Histograms showing LAFS results for MKAR May 2010 analyst-reviewed signals
853 matched by one or more archived signals, grouped by phase and sorted by decreasing
854 count. Only phases with 5 or more screened signals are shown. "Unscreened" shows the

855 total number of phases matched. “Screened” shows the number that passed the screening
856 process. “Validated” shows the number of “screened” that were validated using the known
857 distances between the query event and events of the top matching signals. The percentages
858 of validated signals with respect to the number of screened signals are indicated. A static
859 correlation threshold of 0.6. was used.
860
861 **Figure 16:** Map showing locations of events associated with the 205 validated May 2010
862 MKAR detections. Dashed circle around the location of MKAR shows approximate
863 regional/teleseismic boundary (20^{o} from the station).

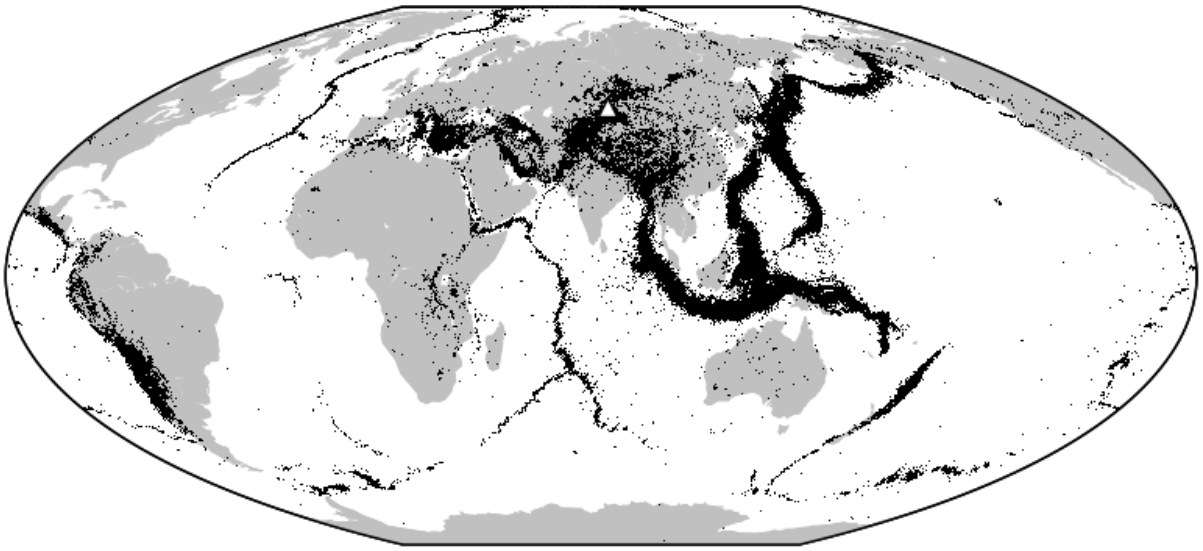


Figure 1. Map showing the locations (black dots) of 249,237 IDC LEB analyst-reviewed events with associated MKAR signal detections between 2002 and July 2016 (excluding May 2010). The white triangle indicates the location of the seismic station MKAR.

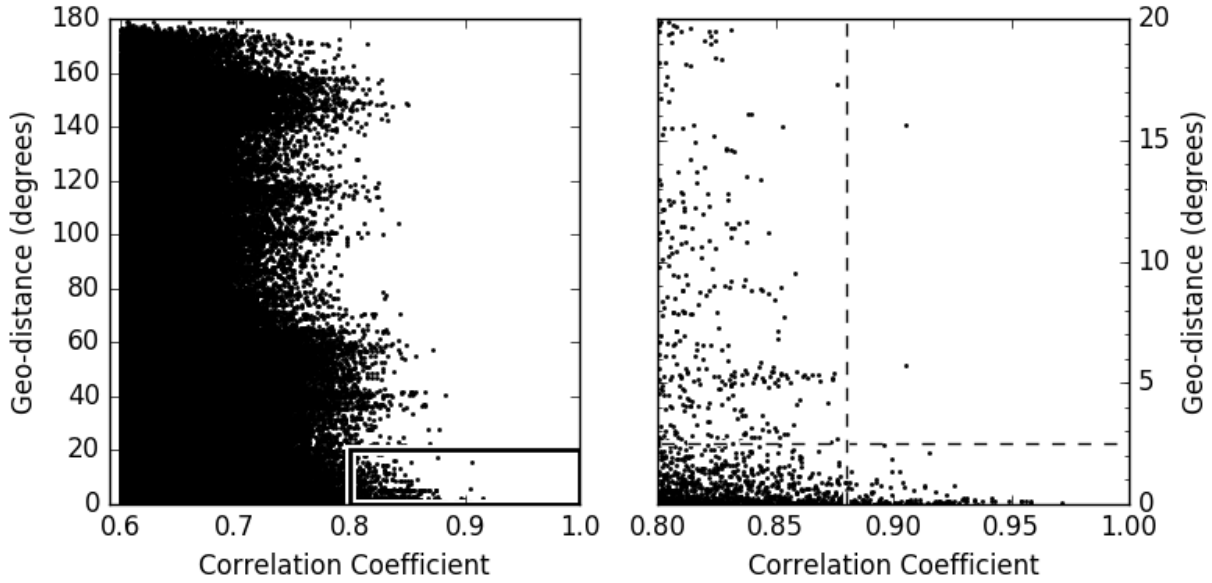
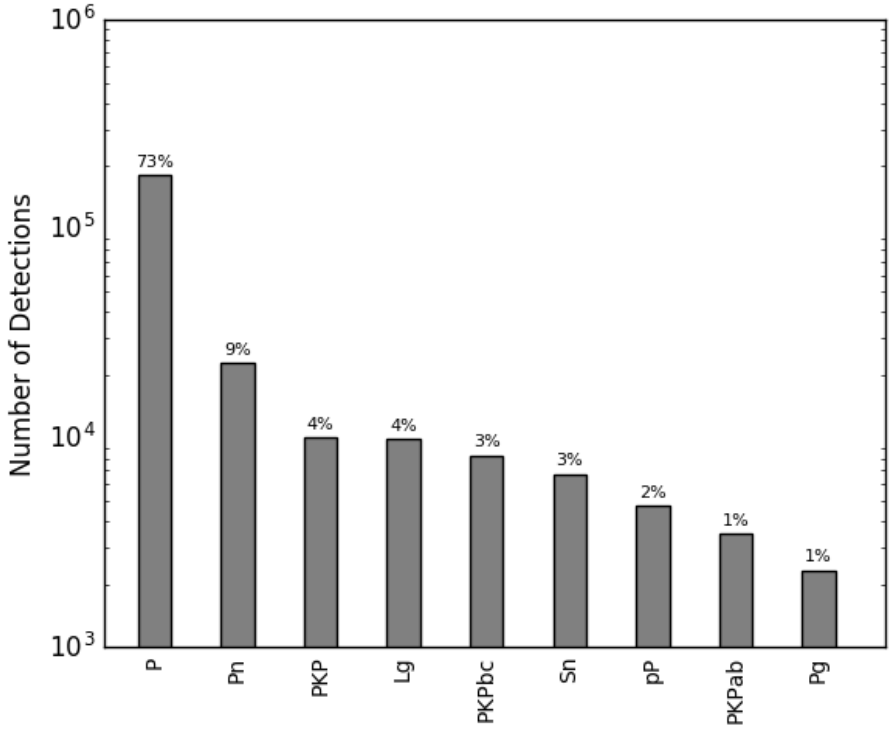


Figure 2. Plot of distances between each query event and the events of the matching templates against the correlation coefficient values. The data points were generated by performing a standard full linear search of our archive of 249,237 templates for each query from our data set of 2,302 queries. (Left) entire data set. (Right) zoom in of the area within the black box in the plot to the left. The chosen maximum cluster radius and minimum correlation threshold for the detection screening algorithm are indicated with the horizontal and vertical dashed lines, respectively.



876

877 **Figure 3.** Histograms of the 249,237 IDC LEB analyst-reviewed signals detections for

878 station MKAR from 2002 to July 2016 (excluding May 2010), grouped by phase and sorted

879 by decreasing count. The percentage of detections for each phase is indicated at the top of

880 the corresponding bar.

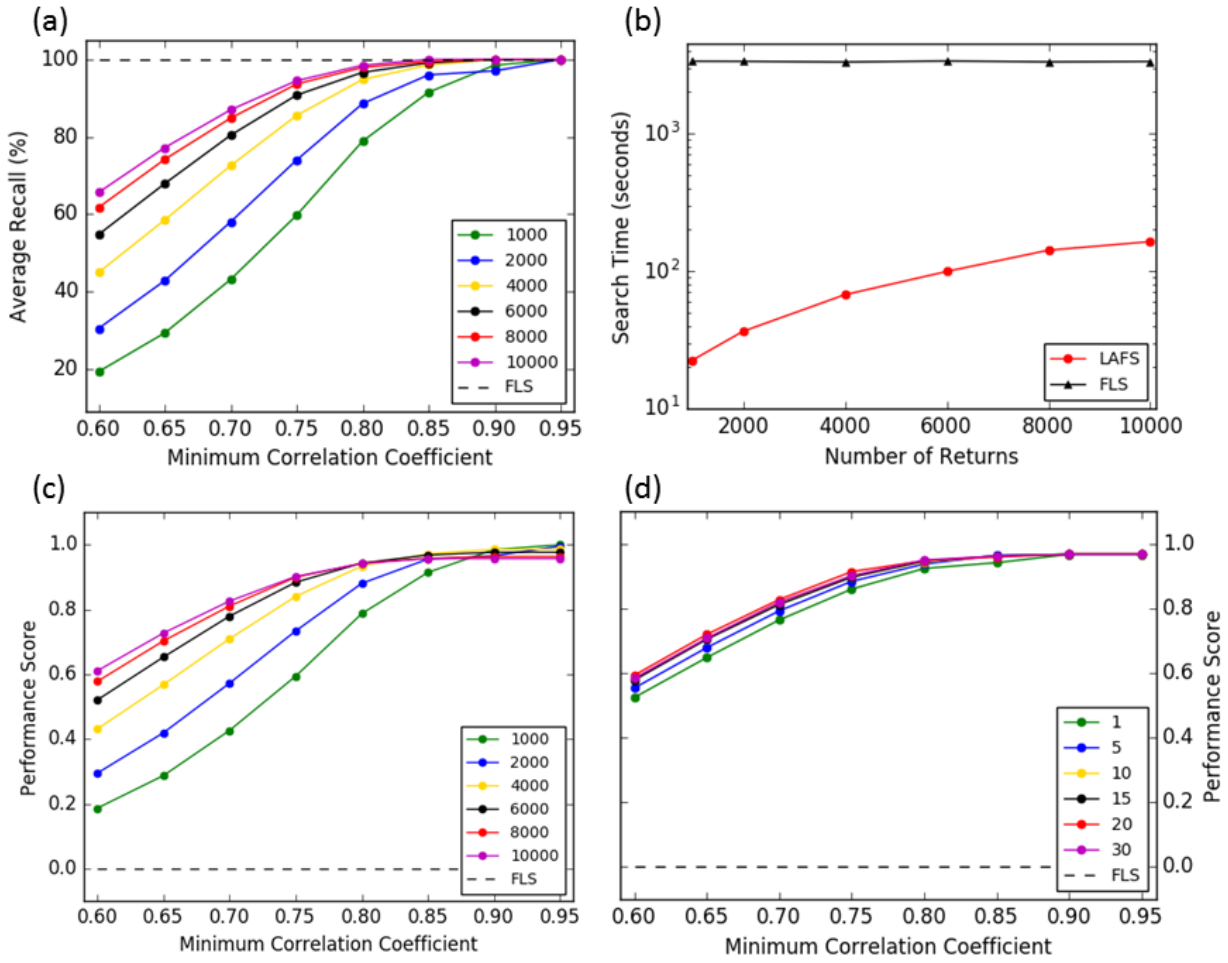


Figure 4. (a and c) Recall and performance score of LAFS and the full linear search method (FLS) as function of correlation threshold value for the number of requested ANN returns of 1000, 2000, 4000, 6000, 8000 and 10,000 using 15 trees. (b) Search time of LAFS and FLS as function of number of ANN returns for 15 trees. (d) Performance score of LAFS and FLS as function of correlation threshold value for 1, 5, 10, 15, 20, and 30 trees while requesting 8000 returns for each case.

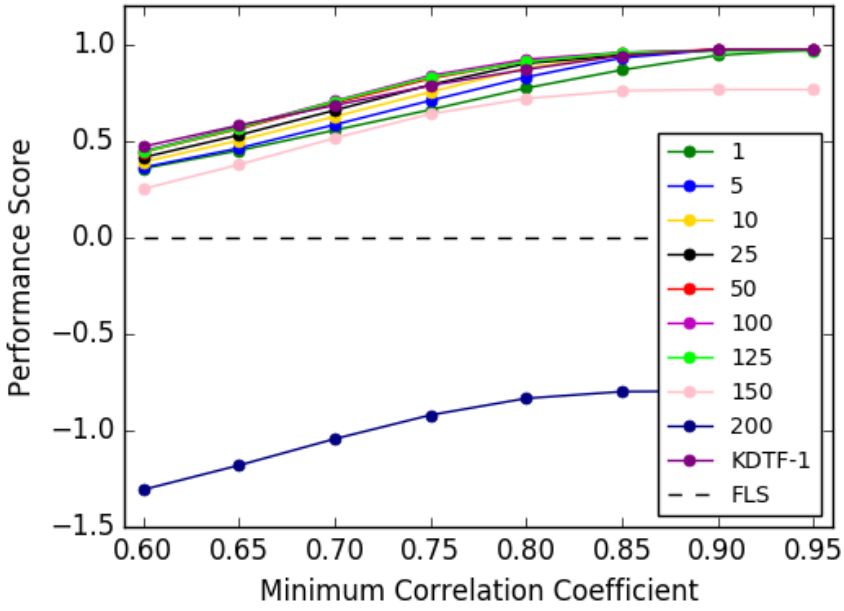
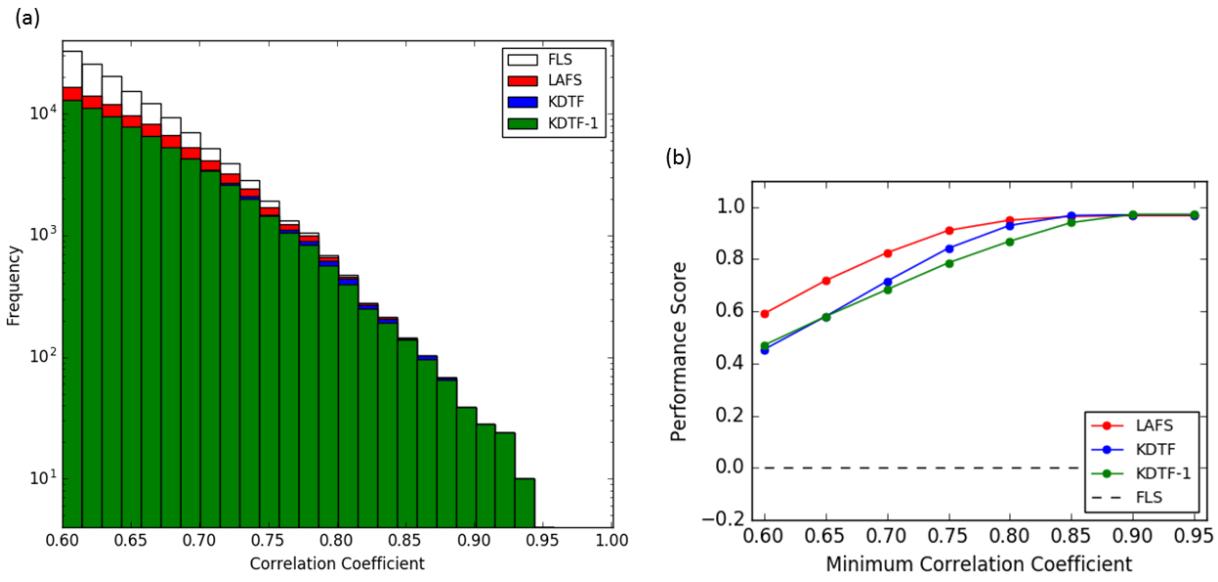


Figure 5. Performance score of KDTF as function of correlation threshold value for 1, 5, 10, 25, 50, 100, 125, 150, and 200 trees while requesting 8000 returns for each case. Performance curves for KDTF-1 and the full linear search method (FLS) for the same number of returns are also shown.



893

894 **Figure 6.** (a) Histogram of correlation coefficient values of the returned matches for LAFS
895 (red), KDTF (blue), KDTF-1 (green), and the full linear search method (FLS) (white). The
896 area enclosed by each histogram is a measure of the overall recall for the corresponding
897 method. (b) Performance score as function of correlation threshold value for LAFS, KDTF,
898 KDTF-1, and FLS. Both the histograms in (a) and the performance curves in (b) were
899 generated using the respective optimum parameters for the methods.

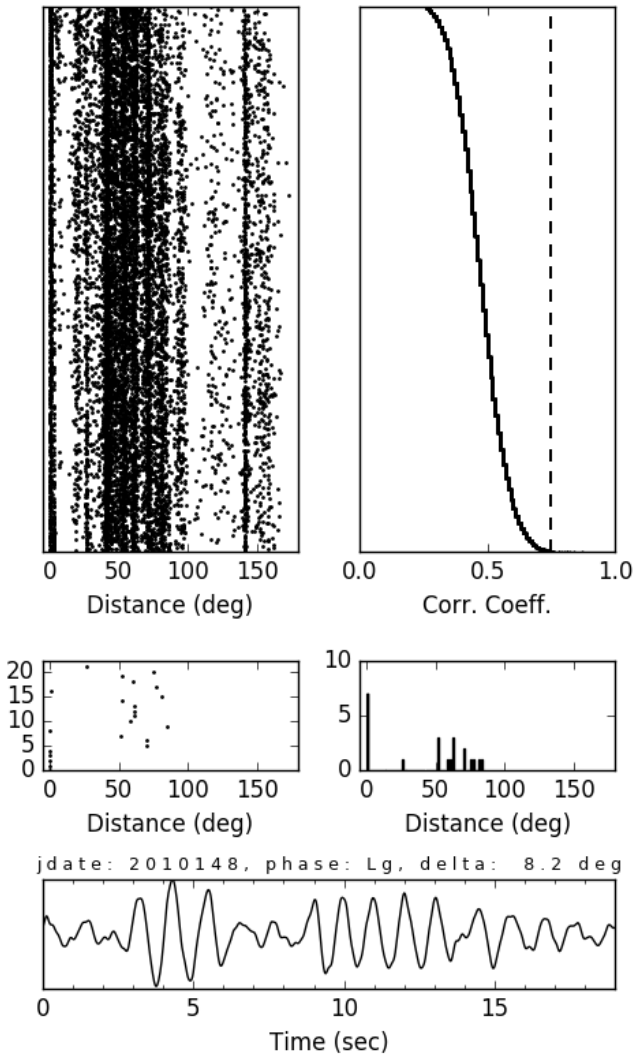


Figure 7. ANN results for an example query waveform. (Bottom) the query waveform, a regional *Lg* phase. (Top left) distances between the query event and the events of the ANN set returned. (Top right) the corresponding correlation coefficient values. Dashed line indicates threshold of 3 standard deviations above the mean. (Middle left) zoomed in version of the plot of distances between the query event and the events of the ANN set showing only correlation coefficients values greater than the threshold. (Middle right) histogram of distances between the query event and the events of the matching templates for correlation coefficients values greater than the threshold.

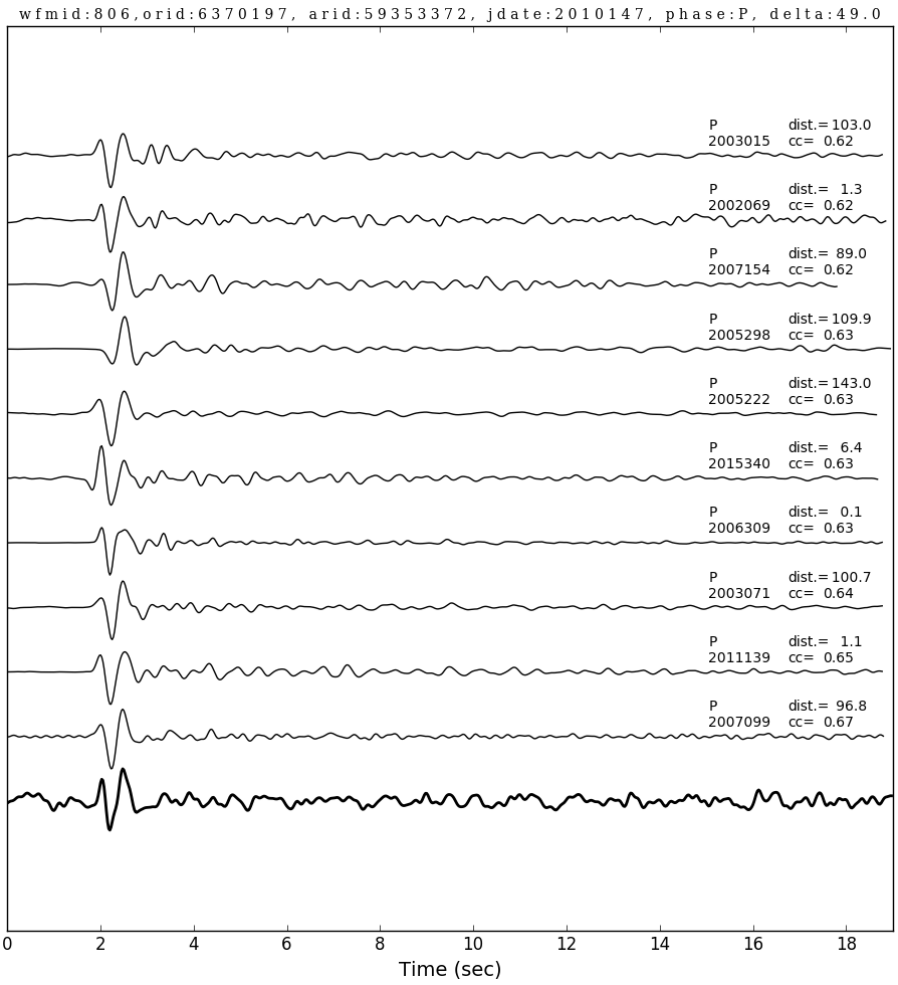


Figure 8. Waveform profile showing the 10 best correlating matches for the 806th analyst-reviewed MKAR signal detection of May 2010, a *P* signal from an event 49° away. Julian date, phase, and station-to-event distance (delta in degrees) for the query are indicated at the top of the plot. The query waveform is shown at the bottom of the waveform section (thick line). The signal detection is at 2 seconds. Matched waveforms (shifted by the optimum time lag for best correlation) are plotted above in order of decreasing correlation coefficient (cc) moving upward through the section, i.e. the most similar waveform is directly above the query waveform. For each matching waveform, correlation coefficient, Julian date, phase, and distance to the query event are indicated.

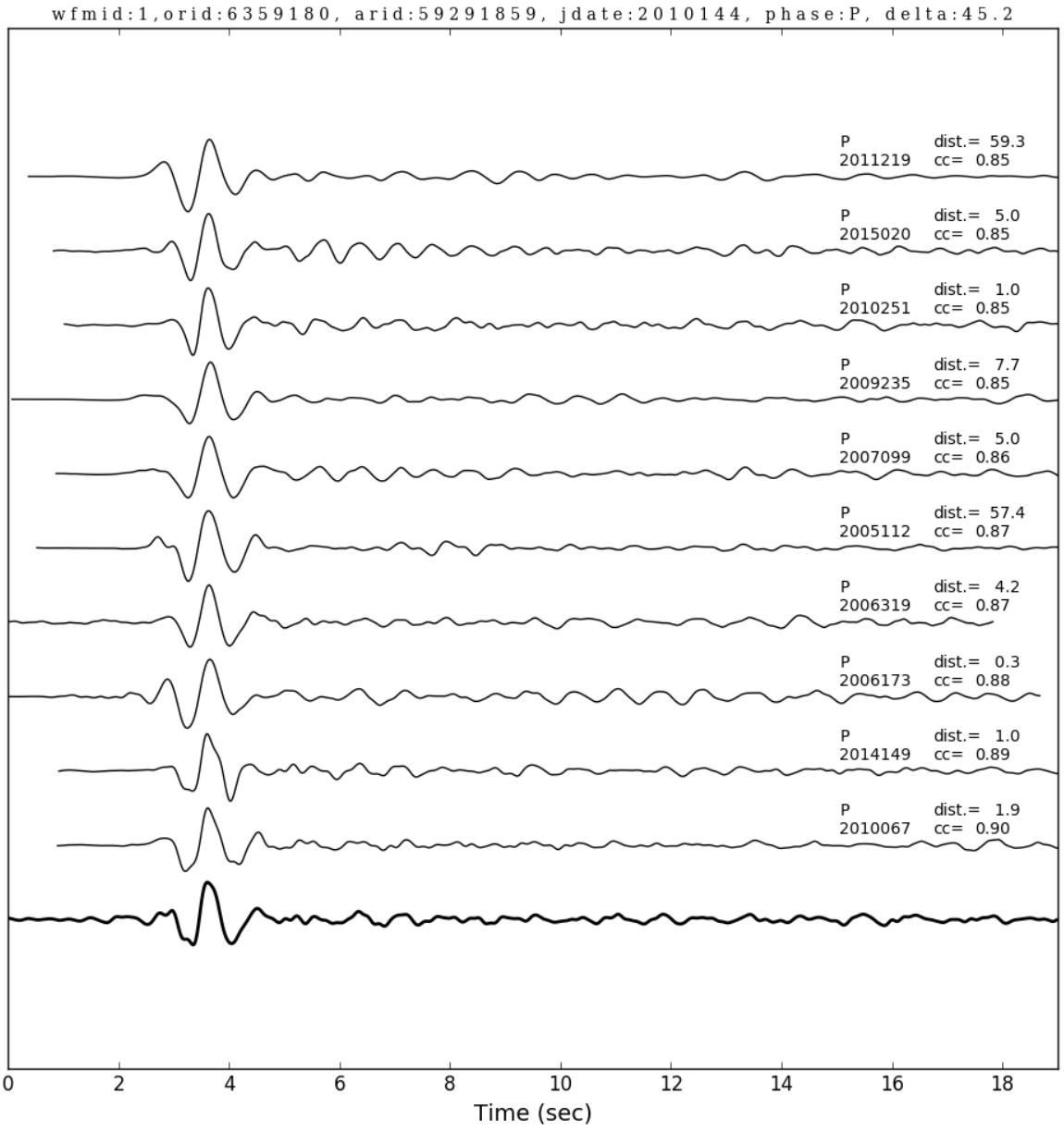
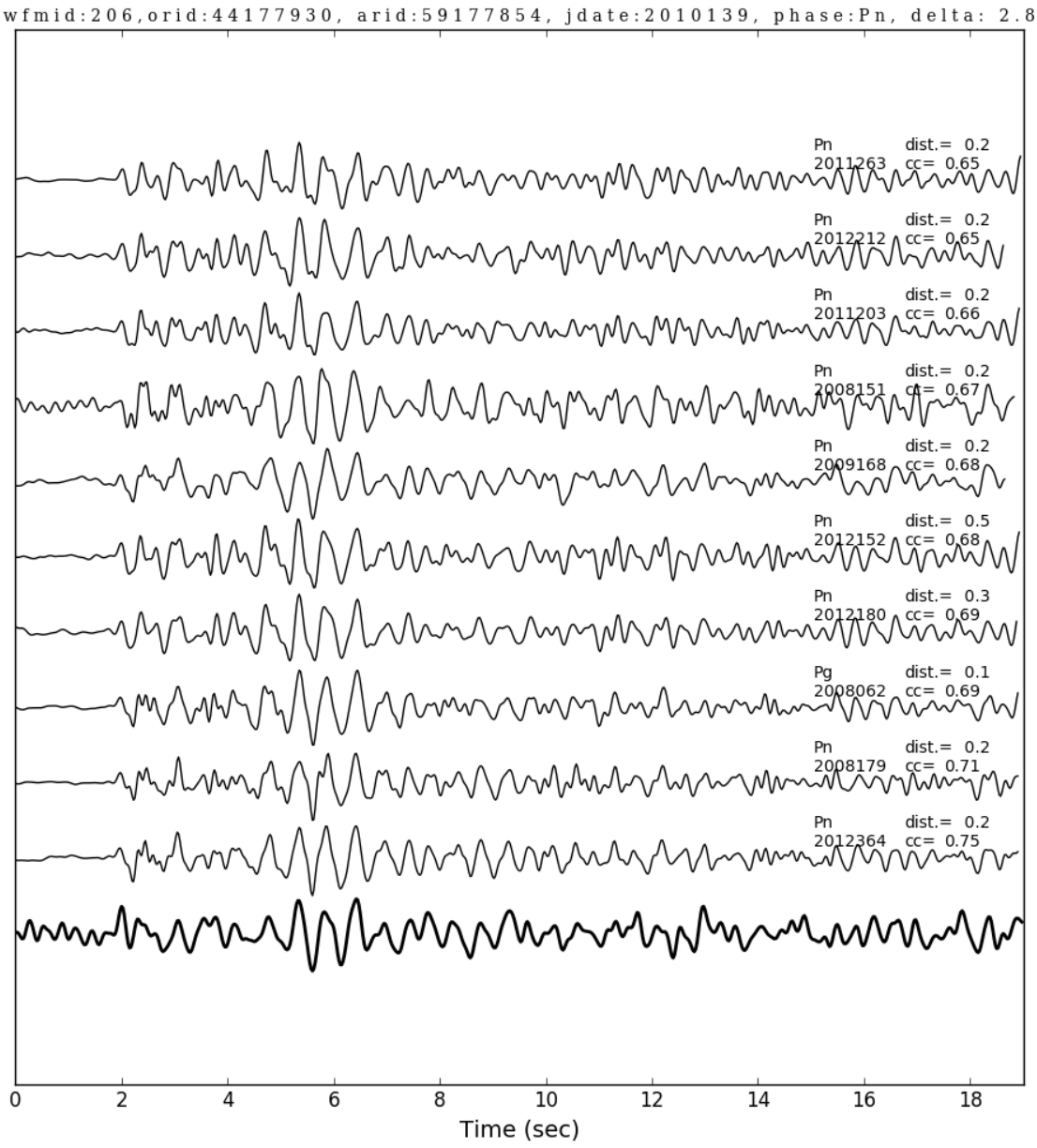


Figure 9. Waveform profile for an accepted identification: a *P* signal from an event 45.2° away. Organization of plot is the same as in Fig. 8. There were 4,354 matches with correlation coefficient larger than 0.6, but we only show the 10 best. For each matching waveform, correlation coefficient, Julian date, phase, and distance to the query event are indicated.



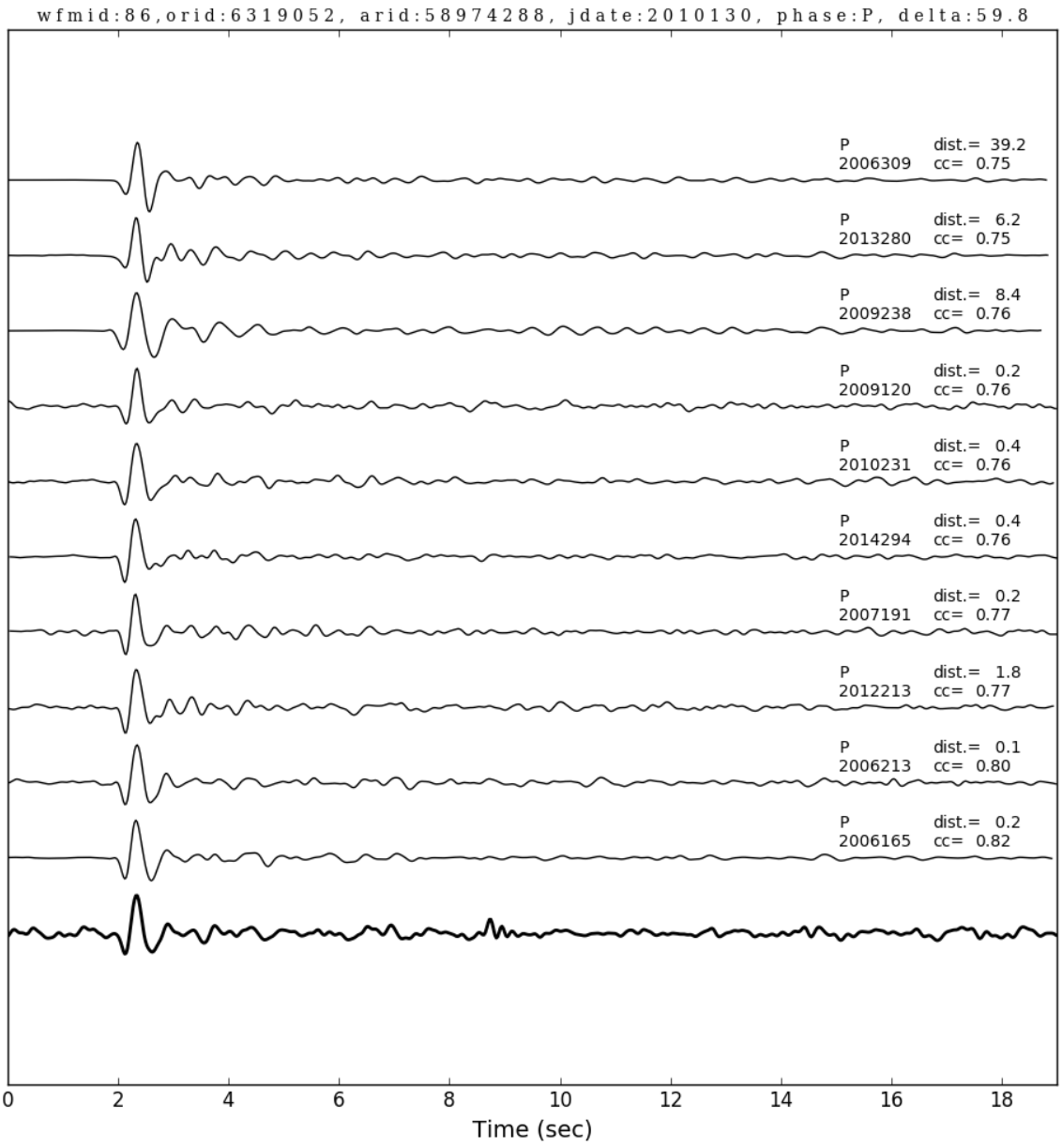
925

926 **Figure 10.** Waveform profile for an accepted *Pn* identification from an event 2.8° away.

927 Organization of plot is the same as in Fig. 8. The high time-bandwidth product of the query

928 waveform (lowermost) effectively guarantees that all potential matches with high

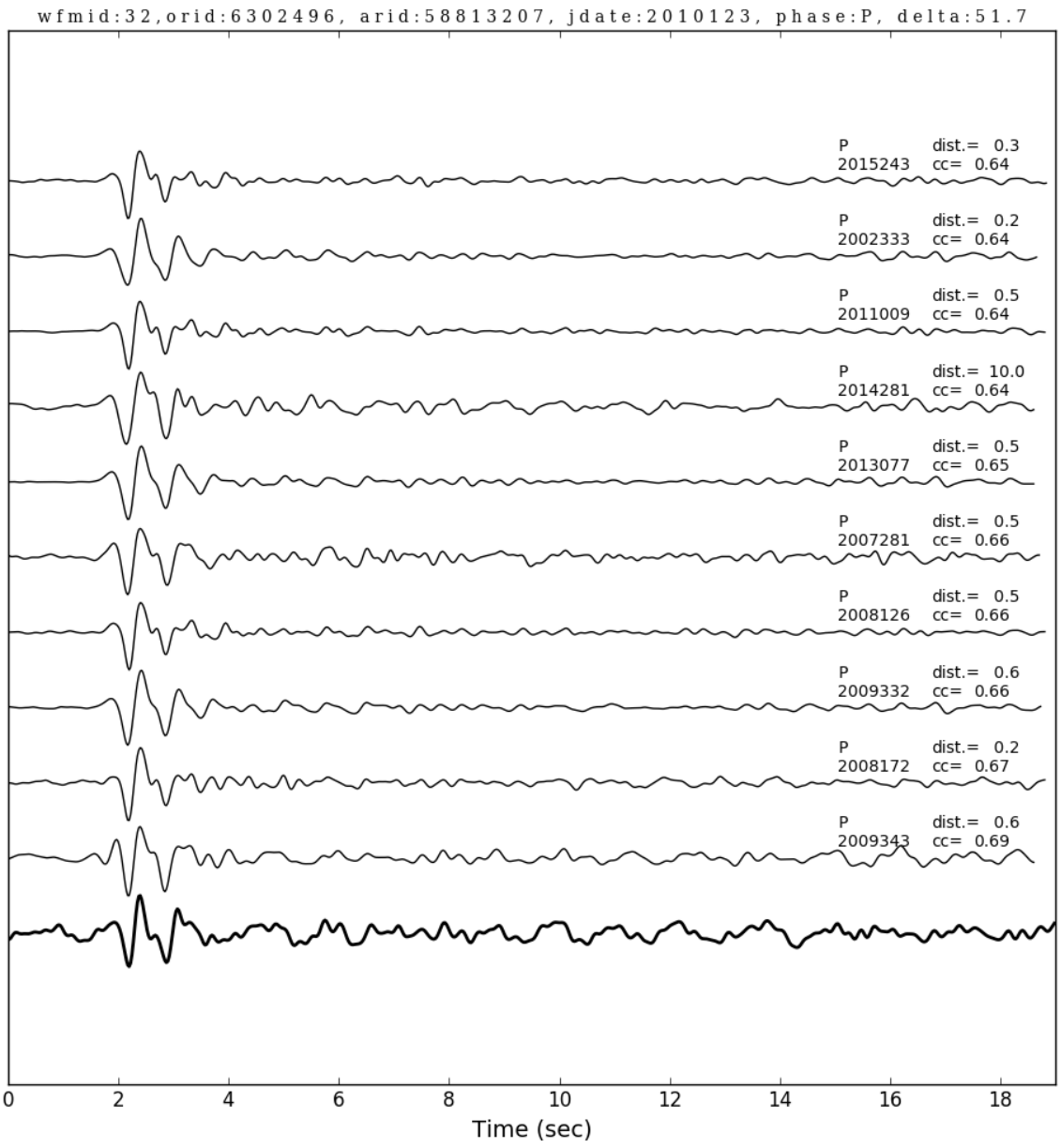
929 correlation coefficient values will be from the same source region.



930

931 **Figure 11.** Waveform profile for an accepted teleseismic first *P* identification from an event

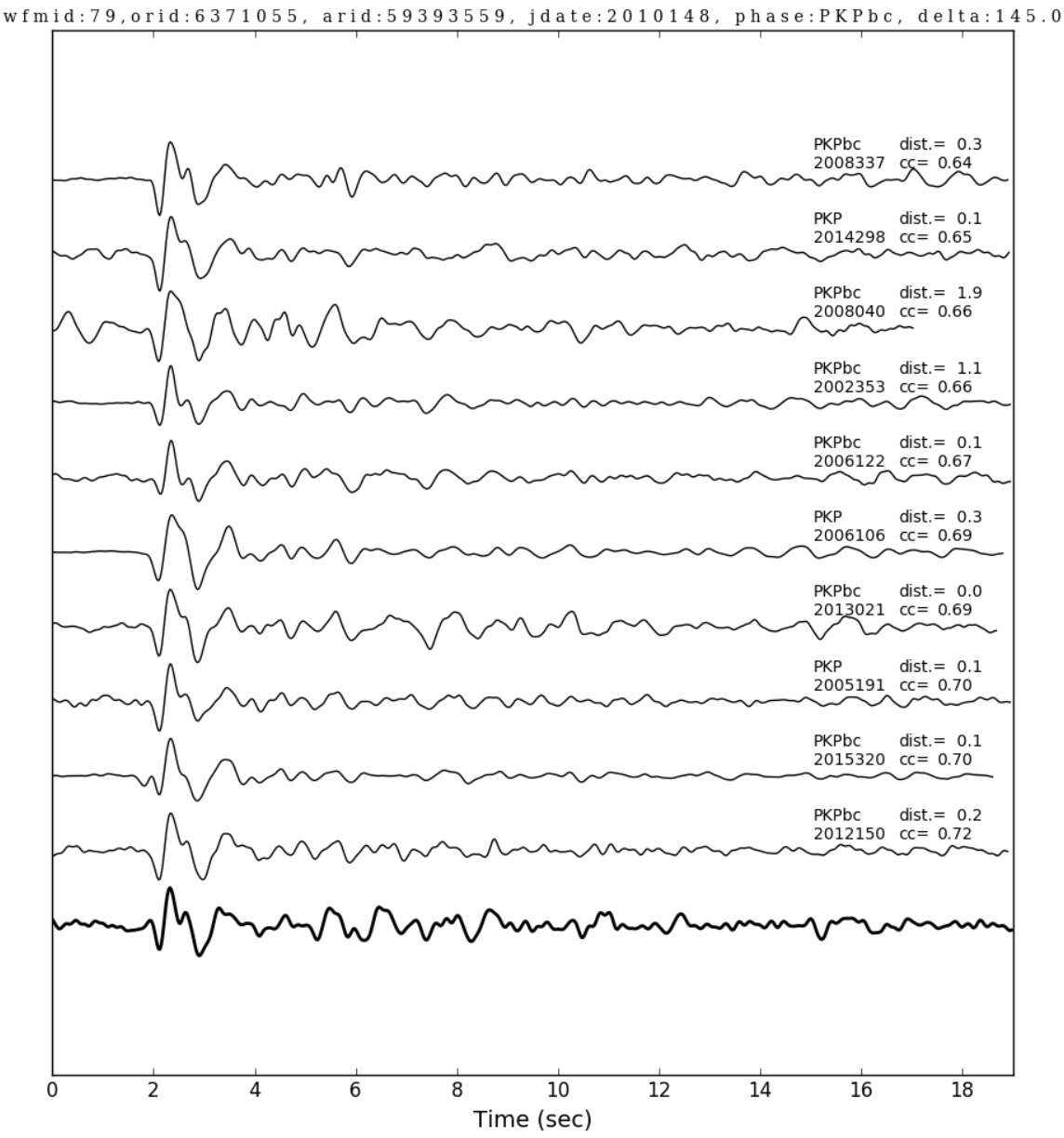
932 59.8° away in Morotai Island, Indonesia. Organization of plot is the same as in Figure 8.



933

934 **Figure 12.** Waveform profile for an accepted teleseismic first *P* identification from an event

935 51.7° away in the Sumatra subduction zone. Organization of plot is the same as in Figure 8.



936

937 **Figure 13.** Waveform profile for an accepted *PKPbc* identification from an event 145° away

938 in northern Chile. Organization of plot is the same as in Figure 8.

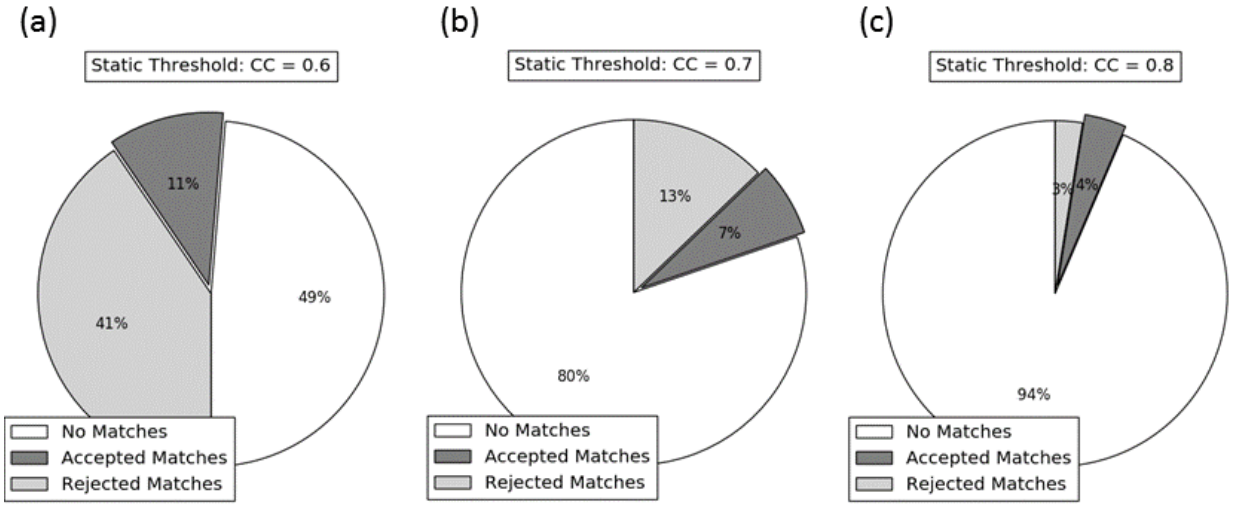
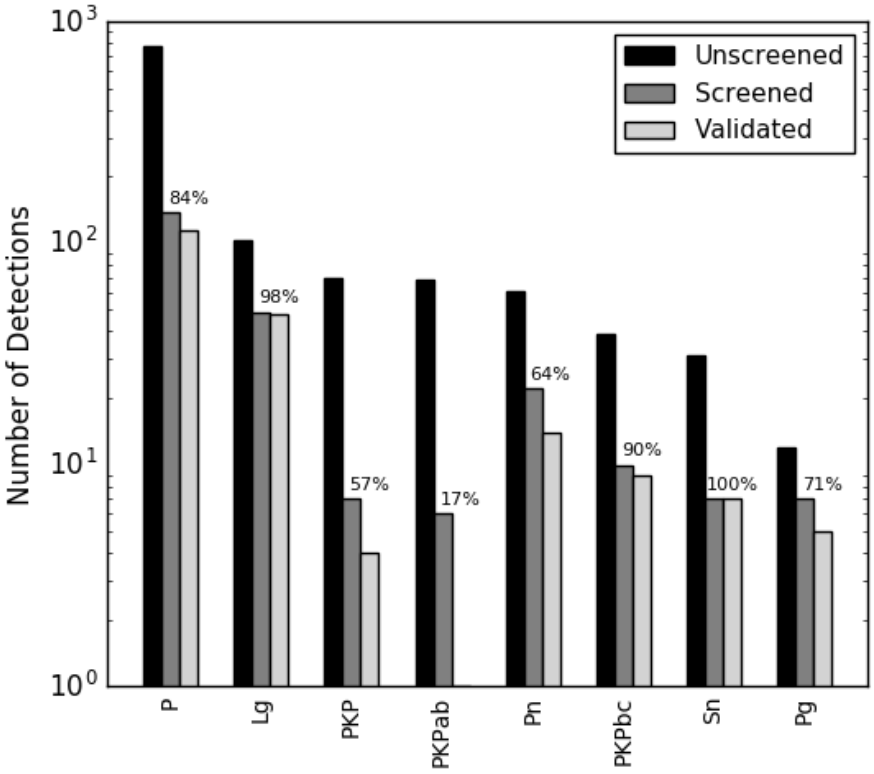


Figure 14. Pie charts showing the percentages of the 2,302 detections processed with LAFS. "No Matches" indicates the percentage of detections for which LAFS found no potential matches; "Accepted Matches" is the percentage of detections that passed the screening process (i.e. true detections), and "Rejected Matches" the percentage of detections that was rejected by the screening process (i.e. false alarms). (a-c) For static correlation threshold values of 0.6, 0.7, and 0.8, respectively.



946

947 **Figure 15:** Histograms showing LAFS results for MKAR May 2010 analyst-reviewed signals
948 matched by one or more archived signals, grouped by phase and sorted by decreasing
949 count. Only phases with 5 or more screened signals are shown. "Unscreened" shows the
950 total number of phases matched. "Screened" shows the number that passed the screening
951 process. "Validated" shows the number of "screened" that were validated using the known
952 distances between the query event and events of the top matching signals. The percentages
953 of validated signals with respect to the number of screened signals are indicated. A static
954 correlation threshold of 0.6. was used.

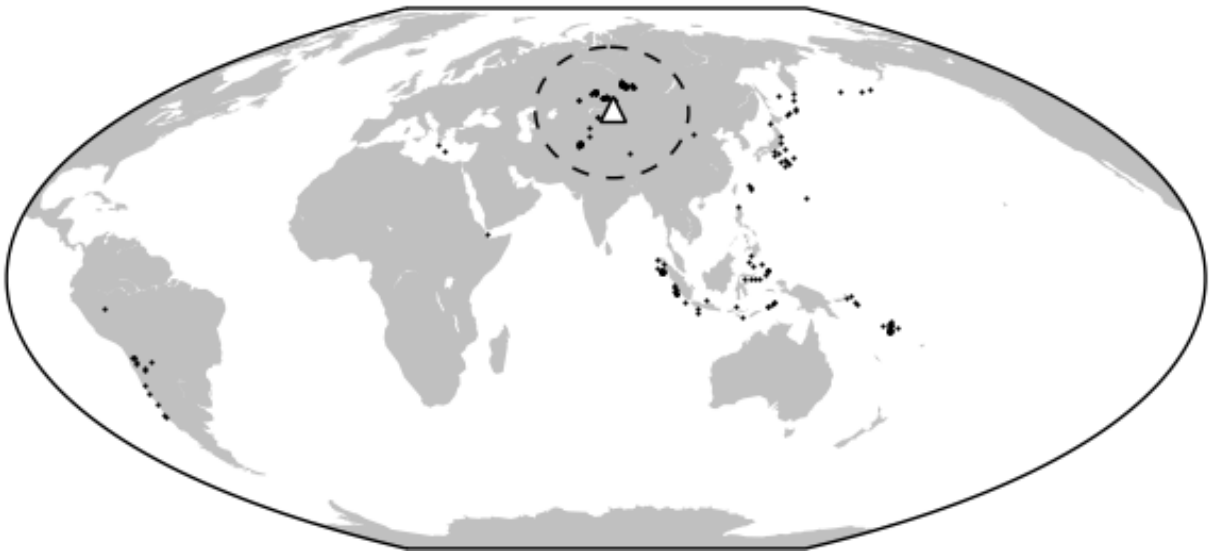


Figure 16: Map showing locations of events associated with the 205 validated May 2010 MKAR detections. Dashed circle around the location of MKAR shows approximate regional/teleseismic boundary (20^{o} from the station).