Bulletin of the Seismological Society of America, Vol. 109, No. 4, pp. 1451–1468, August 2019, doi: 10.1785/0120190006

Ⓔ

Unsupervised Large-Scale Search for Similar Earthquake Signals

by Clara E. Yoon, Karianne J. Bergen,* Kexin Rong, Hashem Elezabi, William L. Ellsworth, Gregory C. Beroza, Peter Bailis, and Philip Levis

Abstract Seismology has continuously recorded ground-motion spanning up to decades. Blind, uninformed search for similar-signal waveforms within this continuous data can detect small earthquakes missing from earthquake catalogs, yet doing so with naive approaches is computationally infeasible. We present results from an improved version of the Fingerprint And Similarity Thresholding (FAST) algorithm, an unsupervised data-mining approach to earthquake detection, now available as open-source software. We use FAST to search for small earthquakes in 6–11 yr of continuous data from 27 channels over an 11-station local seismic network near the Diablo Canyon nuclear power plant in central California. FAST detected 4554 earthquakes in this data set, with a 7.5% false detection rate: 4134 of the detected events were previously cataloged earthquakes located across California, and 420 were new local earthquake detections with magnitudes $-0.3 \le M_L \le 2.4$, of which 224 events were located near the seismic network. Although seismicity rates are low, this study confirms that nearby faults are active. This example shows how seismology can leverage recent advances in data-mining algorithms, along with improved computing power, to extract useful additional earthquake information from long-duration continuous data sets.

Supplementary Content: Figures comparing performance of the improved Fingerprint And Similarity Thresholding (FAST) earthquake-detection algorithm to a previous version of the FAST algorithm on a synthetic data set, examples of detected signals that represent vibrations in the earth but are not the local earthquakes of primary interest to this study, and historical catalog seismicity in the area near the Diablo Canyon power plant (DCPP).

Introduction

Seismology has large data sets of continuous ground-motion measurements, and the rate of data accumulation is accelerating with the development and deployment of cheap, capable sensors. For example, the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC) archive has over 500 TB of seismic time-series data as of May 2019 (IRIS-DMC Archive, 2019). This data growth in seismology is occurring in two complementary directions (e.g., Lindsey *et al.*, 2017). In one direction, the number of seismic stations is rapidly growing, and there are examples of dense temporary seismic networks with hundreds to thousands of stations that can record the full seismic wavefield. Recordings from such deployments are sometimes referred to as large-N data sets (Li *et al.*, 2018; Meng and Ben-Zion, 2018). In the other direction, a single-permanent seismic station can have recorded continuous data for many years or decades. We refer to these long-duration data as large-T data sets.

Advances in computing technology, including increased computing power, parallel and distributed processing, cheaper memory, and archival data storage, together with the development of new scalable data mining and machine-learning algorithms, make it feasible to search for hidden patterns in massive data sets. Seismology can benefit by leveraging these advances to extract useful information. In this study, we seek to detect more earthquakes from a large-T data set near a critical facility: a nuclear power plant in central California.

We focus on the problem of earthquake detection, which is a fundamental first step in observational seismology. Given existing seismic data, we would like to detect, locate, and characterize earthquakes as completely and accurately as possible. Earthquake catalogs generated either manually or automatically by seismic networks are incomplete at lower magnitudes, and the smallest earthquake signals remain hidden in continuous seismic data or are detected at too few

*Now at Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, Massachusetts 02138 U.S.A.

stations to associate and locate. Seismologists have found that waveform similarity between earthquakes is an effective criterion for earthquake detection. Waveform matching does not require impulsive high-energy arrivals as do short-term average/long-term average (STA/LTA) ratio detectors (Allen, 1982; Withers *et al.*, 1998; Velasco *et al.*, 2016). It can be a very sensitive detector of small earthquakes, especially when applied across a network of multiple stations (Gibbons and Ringdal, 2006). Seismic sources that repeat in time have very similar waveforms when recorded at the same station, even if they occur several years apart (Geller and Mueller, 1980; Schaff and Beroza, 2004), because the velocity structure of the Earth is essentially constant over time periods of seismological observations (Poupinet *et al.*, 1984; Schaff and Beroza, 2004). Template matching, which cross-correlates known catalog template waveforms with continuous data to detect previously unknown low-magnitude events, is an example of informed similarity search, because it requires prior knowledge of a waveform already known to be an earthquake signal to detect other previously unknown earthquake signals similar to it. Informed similarity search has been applied to large-T continuous data sets, from 8- to 15-yr long (Shelly, 2017; Skoumal *et al.*, 2018; Ross *et al.*, 2019) and has successfully detected hundreds of thousands to millions of smaller events. A limitation of this informed search is that it requires prior knowledge of the desired signal.

We seek to search systematically for earthquakes using similarity as a detection criterion in the absence of prior signal information, which is the case if template waveforms are unknown or incomplete with respect to earthquake signals of interest. In this situation, we need to perform uninformed, or blind, search for similar earthquakes, checking pairs of short-time windows from all possible times in the continuous data for similarity. This is an example of an unsupervised machine-learning approach to discover similar patterns in an unstructured data set (James *et al.*, 2017). Autocorrelation, a brute-force uninformed search in which every time-window pair within the continuous data is cross-correlated to determine their similarity (Brown *et al.*, 2008; Aguiar and Beroza, 2014), is useful for detecting similar earthquakes within short-duration data. However, it is computationally impractical for large-T data sets because the runtime scales quadratically with the duration of continuous data. The Repeating Signal Detector (Skoumal *et al.*, 2016) accomplished a large-T similarity search for earthquakes in 3 yr of continuous seismic data, overcoming this computational limitation by first setting a low STA/LTA threshold to focus on promising signals and then using a combination of features in the frequency and time domains to identify earthquakes. The Fingerprint And Similarity Thresholding (FAST) earthquake detection algorithm (Yoon *et al.*, 2015) takes a data-mining approach to achieve an uninformed large-T similarity search for earthquakes, avoiding comparisons between the vast majority of times in the continuous data that are likely to have low similarity. FAST detected a wealth of microearthquakes induced by hydraulic fracturing when applied to three months of continuous data at one seismic station in Guy–Greenbrier, Arkansas (Yoon *et al.*, 2017). Previous efforts to extend the uninformed similarity search beyond six months of data with FAST (Yoon *et al.*, 2015) failed due to memory and runtime limitations in the original software.

Here, we demonstrate both the detection capability and scalability of FAST on a large data set, enabled by several improvements to the algorithm (Bergen *et al.*, 2016; Bergen and Beroza, 2018b) and software implementation (Rong *et al.*, 2018). FAST successfully detects small earthquakes missing from the catalog with an uninformed search for similar earthquake signals on a large-T data set, with 6–11 yr of data at a given station. FAST searches for times of similar earthquake signals independently on one channel of seismic data at a time, but detecting with multiple stations plays an essential role in reducing false-positive detections. Therefore, we used a new association algorithm (Bergen and Beroza, 2018a), originally designed to work with FAST outputs, to associate pairs of similar earthquakes across multiple stations within a seismic network. FAST software is widely applicable to continuous data of any duration, including large-T data, from any seismic network.

Methods and Results

Data

The Diablo Canyon Power Plant (DCPP), located in the central California coast (Fig. 1, star), is a nuclear power plant operated by Pacific Gas and Electric Company (PG&E) to generate electricity for northern and central California. The presence of several Quaternary active faults near the DCPP raises concerns about the possibility of an earthquake that might damage this critical facility (Pacific Gas and Electric Company [PG&E], 2011, 2014, 2015). PG&E installed a relatively dense seismic network near DCPP, resulting in a high-quality earthquake catalog, but the low-seismicity rate limits the available information on the character of nearby faults. To improve this situation, we perform an uninformed search for similar earthquakes on up to 11 yr of continuous data on an 11-station network (Fig. 1, inverted triangles) within 30 km of the DCPP to detect small earthquakes missing from the catalog.

These 11 stations are operated by PG&E (network code PG) and the Northern California Seismic Network (NCSN, network code NC), with a total of 27 channels of continuous seismic data, each sampled at 100 Hz. Eight of these stations have three components (east, north, and vertical), whereas the other three stations (PG.SH, NC.PABB, and NC.PPB) have one vertical component. The size of the raw time-series data is about 500 GB (Table A1). For each channel, we search for similar earthquakes, starting from the first available date of continuous recording and ending on 24 October 2017 00:00:00 coordinated universal time (UTC). The earliest starting dates are 1 September 2006 at Station PG.SH and 1 June 2007 at Station PG.DCD (resulting in over 10 yr

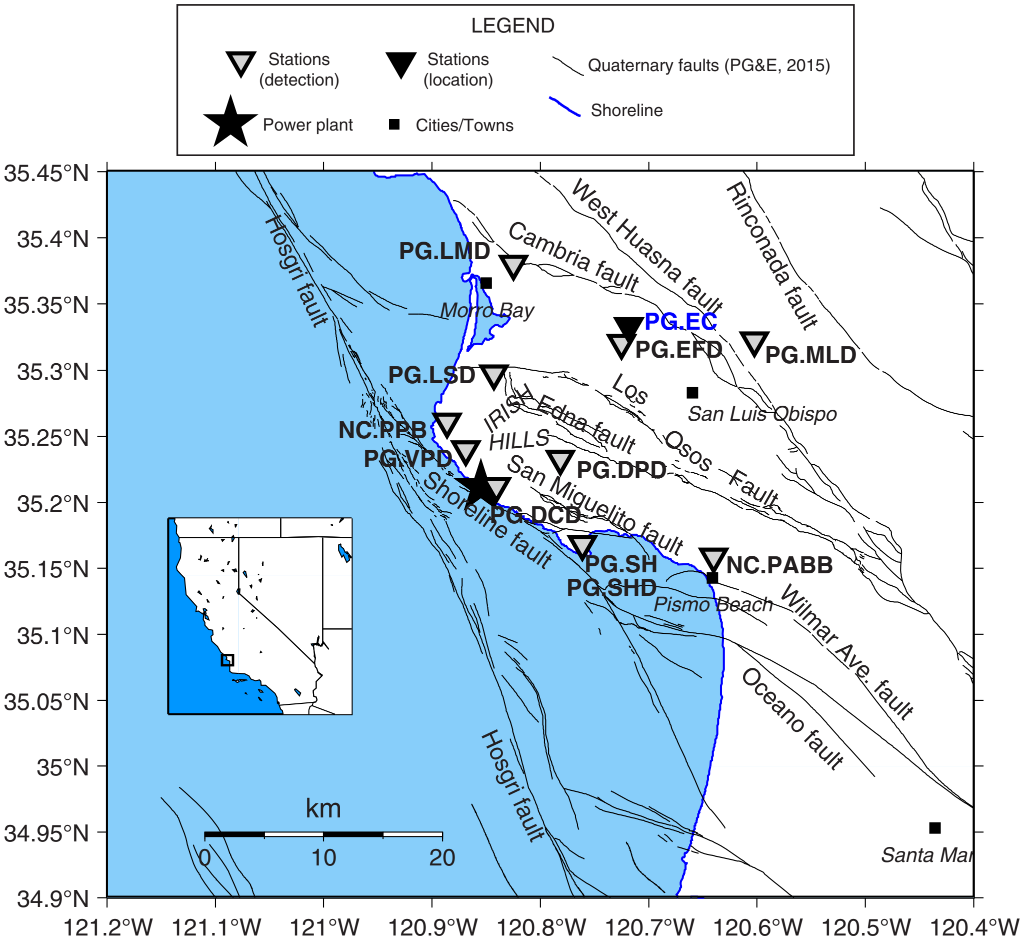

Figure 1. Map of area near Diablo Canyon Power Plant (DCPP) (star) on the central California coast (box in inset). Inverted triangles denote labeled seismic stations with continuous data used in earthquake detection and an additional Station PG.EC used only for location. Quaternary fault traces are from Pacific Gas and Electric Company (PG&E, 2015). The color version of this figure is available only in the electronic edition.

of continuous data at these stations), whereas the latest starting dates are 1 October 2011 at Station NC.PABB and 1 November 2011 at Station NC.PPB (resulting in 6 yr of continuous data at these stations).

Large-*T* Earthquake Detection

The FAST earthquake-detection algorithm (Yoon *et al.*, 2015) finds small earthquakes hidden in continuous seismic data through an uninformed, or blind, search for similar signals at all times. FAST performs this similarity search over long-duration continuous data by adapting data-mining techniques originally developed for audio and image search within massive databases. FAST converts seismic waveforms into compact binary features called fingerprints, based largely on time–frequency information from short-overlapping segments of the spectrogram. The fingerprints are designed to be discriminative: similar earthquake waveforms have highly similar fingerprints, whereas fingerprints extracted from noise have low similarity. FAST then uses locality-sensitive hashing (LSH; Andoni and Indyk, 2006) to organize the fingerprints into a database and efficiently search for similar fingerprints with high probability by avoiding unnecessary comparison of dissimilar fingerprints. Postprocessing software condenses the similar fingerprint information into a list of times with earthquake detections. Before this study, the longest duration of data processed by FAST was six months (Yoon *et al.*, 2015). Attempts to process longer data sets were hindered by memory limitations and lack of parallel processing capability.

FAST Software and Algorithm Improvements. We introduce the first demonstration of FAST earthquake detection on a long-duration continuous data set, with 6–11 yrs of data for each channel. This capability is enabled by a new FAST software implementation in Python and C++ (Rong *et al.*, 2018). The improved FAST software contains the following new capabilities and optimizations, missing from the original version used in Yoon *et al.* (2015, 2017):

- Parallel processing capability for fingerprint, similarity search, and postprocessing. We ran FAST on a Linux server with 512 GB memory and two 28-thread Intel Xeon E5-2690 v4 2.6 GHz central processing units (Rong *et al.*, 2018), using up to 56 processes.

- The LSH database size now can handle arbitrarily large data sets without being limited by the amount of available memory. It is now possible to divide the LSH database into a user-defined number of partitions (Table A2), so that each partition fits into memory. The runtime increases by ~20% with each additional partition, mainly from the overhead of initializing and deleting each partition. Fingerprints from one partition are compared with fingerprints from other partitions, so the similarity search results should be the same regardless of how many partitions are used.
- An occurrence filter is capable of removing frequently repeating wideband nonearthquake signals (Rong *et al.*, 2018) that cannot be removed with a simple band-pass filter but would significantly degrade detection performance and increase runtime. This nonearthquake repeating noise is often anthropogenic (e.g., Velasco *et al.*, 2016). To use this occurrence filter, the user-input number of partitions should be set such that when the continuous data duration is divided by the number of partitions, each partition has a relatively short duration of data (such as a day or a month). Then a threshold on the frequency of occurrence is set as a fraction between 0 and 1, so that a fingerprint that matches at least this fraction of the total number of fingerprints during each short partition is excluded from the similarity search (Table A2). The increased runtime from using many partitions is negligible relative to the vast speedup achieved by avoiding comparisons between repeating nonearthquake signals.
- Flexibility in input data formats: time gaps in the continuous data are permitted, and the continuous data can be distributed within an arbitrary number of Seismic Analysis Code (SAC) or miniSEED files. A global index assigns consistent time-stamp information to fingerprints from different channels and stations of continuous data that may start and end at different times, with variable time gaps.

The new FAST software also includes the following algorithm changes, compared to the original version used in Yoon *et al.* (2015, 2017):

- During fingerprint generation, each wavelet coefficient is normalized by its median and median absolute deviation (MAD) across the continuous data set. Bergen *et al.* (2016) and Bergen and Beroza (2018b) showed that fingerprints generated with median and MAD normalization are more discriminative than those generated by normalizing each wavelet coefficient by its mean and standard deviation, as used in Yoon *et al.* (2015, 2017). For large data sets, it is sufficient to calculate the median and MAD for each wavelet coefficient from a smaller random sample of the continuous data set; Rong *et al.* (2018) found that calculating median and MAD with a 1% random sample of the continuous data reproduced the fingerprints with 98.7% accuracy. The sample is selected with two user-input parameters (Table A2): the fraction of the entire continuous data set to include in the sample and the sampling frequency as a time interval.
- The spectrogram is cutoff at the corners of the band-pass filter applied to the continuous data, before fingerprints are generated, because including frequencies outside the target filter band can hurt detection performance (Bergen and Beroza, 2018b). Previously, Yoon *et al.* (2015, 2017) used the entire spectrogram, even after it was filtered.
- During spectrogram generation, time windows are tapered by a Hanning window before taking the Fourier transform. The resulting spectrogram has lower background noise levels because a Hanning window has less broadband spectral leakage than a Hamming window, which was previously used by Yoon *et al.* (2015, 2017).
- During spectral-image generation, when resizing and downsampling the image to a power of 2 for the wavelet transform, the new FAST software runs `scipy.misc.imresize()` with bilinear interpolation, which includes an antialias filter (SciPy, 2019). Previously, this step used a proprietary algorithm in MATLAB's `imresize()` (see Data and Resources).
- During database generation, min-max hash (Ji *et al.*, 2013) is used instead of min-hash (Broder *et al.*, 2000). Min-max hash estimates Jaccard similarity with comparable or better accuracy than min-hash, while halving the number of hash function calculations (Rong *et al.*, 2018). FAST uses a multiple hash function implementation of min-max hash, and the number of hash functions r is an important input parameter.

These algorithm changes significantly improve the precision–recall performance on synthetic data (Ⓔ Fig. S1, available in the supplemental content to this article) relative to the original FAST version from Yoon *et al.* (2015), meaning that there are not only fewer false detections but also fewer missed detections. The fingerprints generated by the new FAST software in Rong *et al.* (2018) are discriminative and robust to noise, even when the signal-to-noise ratio is less than 1.

FAST Processing Pipeline: Continuous Time-Series Data to Candidate Events. Preprocessing the continuous data set is an essential first step before running FAST. This study is explicitly focused on detecting high-frequency local earthquakes of tectonic origin. We apply a station-specific band-pass filter (Table A1) to each channel of continuous data to select the frequency band that is likely to contain our desired small local earthquakes and to remove repeating noise that is correlated in time. The amplitude and frequency band of this repeating noise can vary significantly between different stations. The minimum frequency of the band-pass filter is set to 3 or 4 Hz, because most repeating noise is present at lower frequencies, and we can also avoid detecting teleseismic earthquake signals that are not of primary interest to our study. To remove repeating noise within the passband, we apply the occurrence filter later in FAST to selected channels (Table A2), for which signals that repeat too often are assumed not to be from earthquakes. If repeating noise is not removed with a band-pass filter, both detection performance and runtime would significantly degrade, because the vast majority of detections from FAST would be pairs of similar noise signals. For the same reason, we do not run FAST on continuous data with time gaps that are filled with zeros, from when the station was not recording data, because time windows with zeros are identical; we either removed or replaced these zero sections with uncorrelated random noise. After filtering, we decimate each channel of continuous data to 25 samples per second, because the maximum frequency of the band-pass filter over all channels was 12 Hz. The resulting decimated, filtered time-series data have a total size of about 100 GB (Table A1).

FAST runs independently on one channel of continuous time series data at a time. We used the same FAST parameters in Table A2 on every channel. For event detection using continuous data from multiple stations, we advise using the same fingerprint parameters (first eight parameters in Table A2) across all channels and stations for consistency. For the similarity search parameters (last five parameters in Table A2), it is easiest to use the same values for all channels and stations. On the other hand, it is also possible to use different similarity search parameters on different channels and stations, especially if some stations are noisier than others, to adjust the threshold for a successful search. Fingerprint generation on one channel of data, split into month-long miniSEED files, took between 3 and 8 hr to run in parallel on 56 processes, depending on the duration of available continuous data. For one channel, the number of fingerprints ranged from 150 million to almost 300 million, and the fingerprint file size was between 75 and 137 GB; the total size of fingerprint files from all 27 channels was ~2.6 TB (Table A1). For similarity search, the runtime for one channel was 3–16 hr when run in parallel on 48 processes (Rong *et al.*, 2018). The output of FAST is a list of pairs of times within the continuous data with their associated FAST similarity (defined as the number of hash tables with this fingerprint pair in the same bucket), when the fingerprints (and therefore waveforms) are similar. We can visualize this output as an extremely sparse similarity matrix (Bergen and Beroza, 2018a; Rong *et al.*, 2018). This similarity search output file is written to disk as a binary file; this file size was between 38 and 549 GB for a given channel, indicating that some stations had more repeating noise signals than others, and the total size of these output pairs files from all 27 channels was ~4.7 TB (Table A1). The similarity search parameter r for the number of hash functions per hash table has the most significant effect on both detection performance and runtime. Through trial and error, we found that $r = 6$ was the best compromise for this long-duration data set (Table A2); setting $r = 7$ would result in too few earthquake detections, whereas setting $r = 5$ would increase false detections, make the output pairs files too large, and unacceptably increase the runtime. At this point, we have FAST output for one channel of continuous data: candidate pairs of times with similar signals, in binary format.

We perform additional postprocessing to obtain candidate pairs of times with similar signals at a given station by combining the similarity information from each channel. The binary FAST output from each channel is converted to text format using a parallel sort-merge-reduce procedure (Rong *et al.*, 2018) that increases the size of the output files between 94 and 1500 GB for an individual channel and a total of 12.4 TB for all 27 channels (Table A1). For the eight stations with three components, we combine the FAST similarity outputs in the text files by adding the similarity matrix from each component at that station (with an initial-pair threshold of $v = 2$ on the FAST similarity, Table A2) (Yoon *et al.*, 2017; Rong *et al.*, 2018), then setting a higher station-pair threshold $\tau_0 = (v = 2) \times (3 \text{ components}) = 6$ on the FAST similarity, which significantly reduces the size of the detection results (Table A3). For the three remaining one-component stations (NC.PPB, NC.PABB, and PG.SH), we multiplied the FAST similarity value from every pair by 3, so that their FAST similarity values would be weighted equally to those from the three-component stations. (Alternatively, one could use only the vertical component of continuous data from each station for detection, which would require running FAST on only 11 channels and converting the similarity search output from binary to text format, without any need to add the similarity matrix across components or reweight the similarity at single-component stations. Users might consider this option if the data quality is poor at some but not all channels at a station or for shorter-duration data sets in which reducing the detection results to a more manageable size is not as essential.) At this point, we have the FAST output from each station, with a total size of 867 GB for 11 stations: a list of candidate pairs of detections as $[dt = t_1 - t_2, t_1, \text{sim}]$ values, in which t_1 and t_2 are two different times in the continuous data with similar fingerprints at that station, and sim is their FAST similarity, with sim $\geq \tau_0$.

Finally, we use a network-detection algorithm (Bergen and Beroza, 2018a; Rong *et al.*, 2018), with parameters in Table A4, to perform event association of FAST outputs from each station. This algorithm has three main steps: first, it summarizes the FAST output into similar event-pair information for each station (reducing the data size from 867 to 119 GB); second, it associates candidate event pairs of similar signals across multiple stations in a sparse seismic network; and third, it extracts a list of detected events. This algorithm performs the association by recognizing that the interevent time $dt = t_1 - t_2$ between two earthquakes at different times t_1 and t_2, even if dt is many years long, remains the same at different stations. This step is essential in suppressing false detections from repeating noise signals that occur at a single station, whereas retaining signals that are likely to be from an earthquake. Bergen and Beroza (2018a) introduced the network-detection algorithm on a data set with five stations, and Rong *et al.* (2018) applied it on a five-station data set. This study extends network detection to a larger 11-station data set that required about 9 hr of serial processing. Network detection led to a list of 33,383 candidate events that were detected on at least two stations (Table A4), with an approximate event-detection time at every station that registered each event. For each detected event, the network-detection algorithm also outputs a start time and event duration over all stations that detected it; duplicate events often occur, in which the duration of one event overlaps that of another event. We identified and removed these duplicate events by aggregating the events with overlapping duration and keeping only the one event with the highest number of stations that detected it and the highest total FAST similarity over all stations. After this automatic pruning step, 29,623 candidate events remained.

Candidates for Earthquakes: FAST Output Analysis

FAST detects and builds a list of candidate earthquake events. Most of these candidates are seismic signals, but further analysis with additional automatic or manual methods is needed to determine what specific types of seismic signals were found. We investigated the 29,623 candidate events in further detail to determine which local earthquakes are of interest. We sorted these events in descending order of the number of stations that detected them and then in descending order of the peaksum similarity value, which is the total FAST similarity over all stations for the strongest similarity value involving this event (Bergen and Beroza, 2018a). We visually inspected the waveforms of these 29,623 candidate events, plotted at all stations (up to 11) that were available at their detection time. We displayed 3-min time windows around each event, starting 1 min before the earliest detection time at the nearest station, to capture the full range of possible waveform durations. This manual inspection process guided our choices for additional thresholds and decision rules, which we then used to automatically discard more nonearthquake signals. We kept all candidate events that were detected at five or more stations. For candidate events detected at three or four stations, we excluded events that occurred between 28 September 2011 and 1 November 2011, which were dominated by regularly repeating signals from a seismic reflection survey of the Shoreline fault (PG&E, 2011, 2014, 2015) with prominent recordings on Stations DCD, DPD, VPD, and SHD (Ⓔ Fig. S2). For candidate events detected at two stations, we only included events that happened before 4 September 2011, when fewer continuously recording stations were available (Table A1). We also excluded candidate events during 1 December 2009 to 16 April 2010, which mostly consisted of repeating signals on stations SH and SHD from another seismic-reflection survey in the area (PG&E 2011, 2014, 2015). After removing these nonearthquake signals, 5048 candidate events remained.

4134 of the 5048 candidate events were earthquakes in the existing catalog, with magnitudes $0.3 \le M \le 6.8$ (Fig. 2a, white circles). Because these stations are located in central California, near the boundary of the NCSN and the Southern California Seismic Network (SCSN), we looked for matching events in earthquake catalogs from both networks. 2080 events were only in the NCSN catalog; 453 events were only in the SCSN catalog; and 1601 events were in both catalogs. In addition, 18 of the 5048 candidate events were quarry blasts from the earthquake catalogs; four blasts were only in the NCSN catalog; seven blasts were only in the SCSN catalog; and seven blasts were in both catalogs.

We visually inspected the 3-min waveforms of the remaining 1073 candidate events that are not accounted for in the NCSN and SCSN catalogs. 30 candidate events were duplicate detections of earthquakes that were previously detected with higher similarity; we do not consider these as separate detections, because we do not want to double-count earthquakes. Five candidate events were deep teleseismic earthquakes, with longer duration and lower frequency waveforms compared to local earthquakes (Ⓔ Fig. S3); we would have detected many more teleseismic earthquakes had we not filtered out frequencies below 3–4 Hz (Table A1) before running FAST. 62 candidate events were infrasound signals (sound waves with frequencies <20 Hz) that are similar across the network used for detection but propagate at the speed of sound (~0.33 km/s) rather than at seismic velocities (Ⓔ Fig. S4); they are most likely sound waves from human activity, such as sonic booms from aircraft or artillery explosions at nearby military bases (Cates and Sturtevant, 2001; Cochran and Shearer, 2006; Walker *et al.*, 2011). 379 out of the 5048 candidate events (~7.5%) were manually interpreted to be low-amplitude background noise that we consider false detections.

The remaining 420 candidate events were newly detected local earthquakes. These small earthquakes have local magnitudes in the range $-0.3 \le M_L \le 2.4$ (Fig. 2a, dark circles). Figure 3 shows their magnitude distribution that we calculated using peak amplitudes on synthesized Wood–Anderson seismograms with a distance correction, calibrated to catalog-event magnitudes (see details in Appendix).

Table 1 summarizes the 5048 candidate events detected by FAST in this multiyear continuous seismic data set, categorized by event type. ~90% of the detected events were earthquakes. Only ~8% of the events were newly detected local earthquakes, indicating that the existing seismic network is generating reasonably complete catalogs in this low-seismicity area. Because we set our detection threshold relatively low, to include early events detected on only two stations, ~7.5% of our detections are false positives. We discuss false negatives (catalog earthquakes not detected by FAST) in the next section, only within the context of a limited local area near the network. It is not realistic to expect FAST to detect the smallest catalog earthquakes that were detected by other nearby stations not processed by FAST, at greater distances from the stations used for detection; FAST is intended to supplement, not entirely replace, existing earthquake-detection methods.

Earthquake Location

The 4134 catalog earthquakes detected by FAST using continuous data from an 11-station local seismic network are

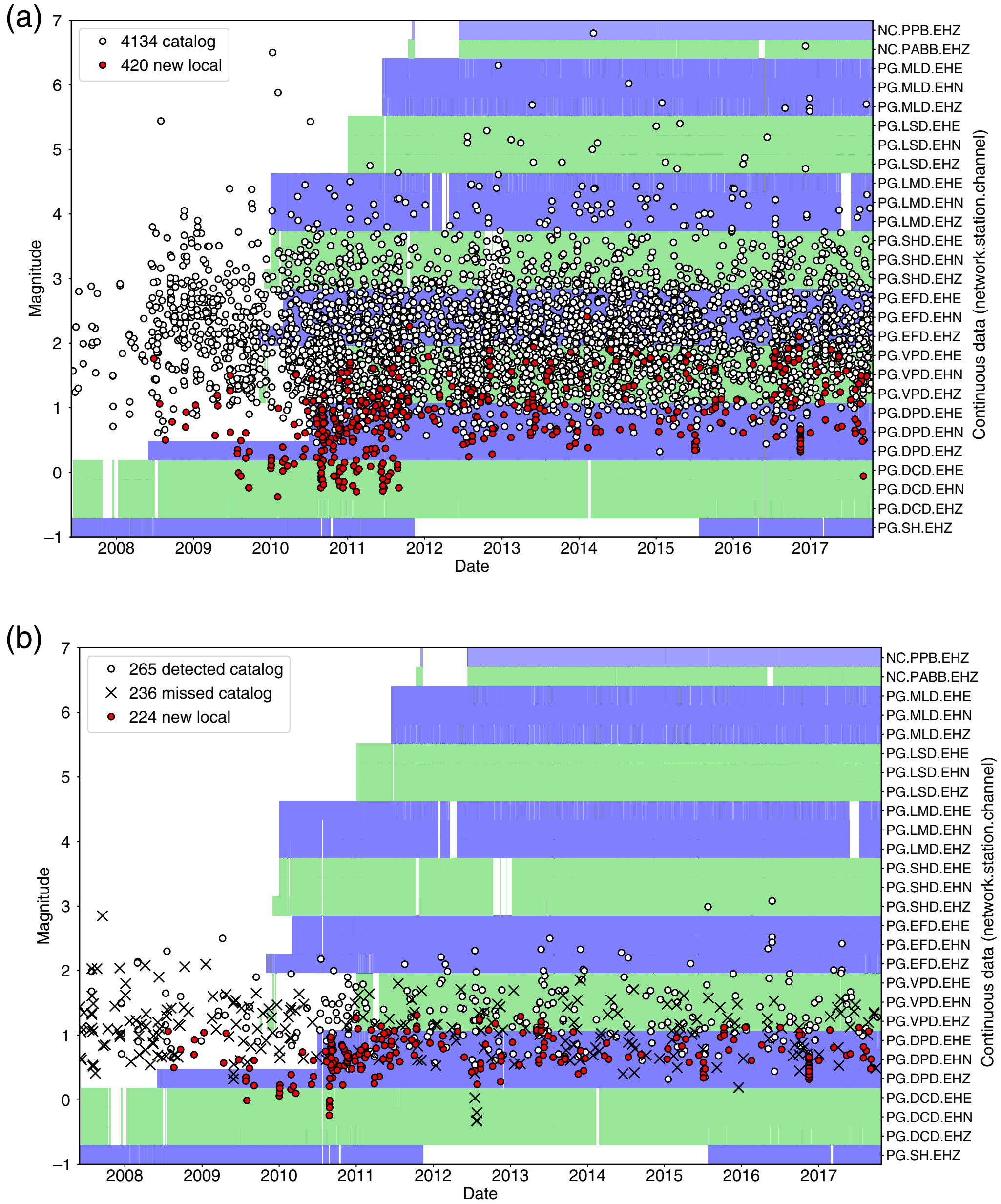

Figure 2. Magnitude–time plot for earthquakes detected by Fingerprint And Similarity Thresholding (FAST) from June 2007 (when PG.DCD started recording continuous data, providing two stations for detection) to October 2017. Earthquakes (circles and X symbols) are plotted according to their magnitude (left axis) and time. Background horizontal bands indicate the data availability of each channel of continuous seismic data used for detection (Table A1), labeled by the network, station, and channel (right axis); alternating light and dark bands represent different stations, whereas time gaps are white. Long-duration time gaps on channels NC.PPB.EHZ, NC.PABB.EHZ, and PG.SH.EHZ indicate noisy time periods purposely excluded from detection. Tick marks on the horizontal axis indicate 1 January of the labeled year. (a) All events detected by FAST, regardless of location: 4134 catalog earthquakes (white circles) and 420 new local earthquakes (dark circles). (b) 725 earthquakes near the DCPP and 11-station network, located within the box in Figure 6: 265 catalog earthquakes detected by FAST (white circles), 236 catalog earthquakes missed by FAST (X symbols), and 224 new local earthquakes detected by FAST (dark circles). The color version of this figure is available only in the electronic edition.

located across California (Fig. 4). Plotting their magnitudes against their epicentral distance allows us to estimate a detection limit (solid black line) for a given magnitude and distance. The maximum epicentral distance from the network in which magnitude 1, 2, and 3 events are no longer detected (dotted lines) are 100, 320, and 1000 km, respectively.

We locate 351 of the 420 newly detected earthquakes using *P*- and *S*-wave arrival time picks on up to 12 stations as inputs into VELEST (Kissling *et al.*, 1994) (see parameters in Appendix), and a 1D *P*-wave velocity model for this region from McLaren and Savage (2001) (Table A5) with $V_P/V_S = 1.66$ (Hardebeck, 2010). We were unable to locate

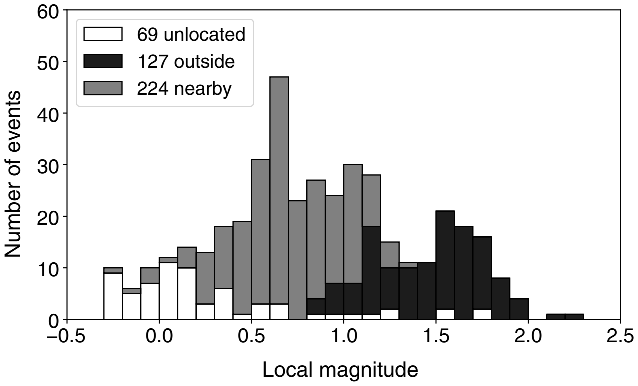

Figure 3. Magnitude distribution of 420 new local earthquakes detected by FAST, with $-0.3 \leq M_L \leq 2.4$. The 69 unlocated earthquakes (white) have lower magnitudes. Out of the 351 located earthquakes, 224 lower-magnitude events (gray) are located near the network within the box boundaries in Figure 6, whereas 127 other events (black) are located farther from the network, outside this box (Fig. 5).

69 predominantly low-magnitude new events (Fig. 3, white bars), because they lacked sufficient high-quality arrival-time measurements.

Figure 5 displays a regional overview of the earthquakes detected by FAST with the 11-station network near the DCPP. There are 3106 catalog earthquakes (white circles sized by magnitude) within the area shown in Figure 5 that were used to calibrate the local magnitude calculation for the new detected earthquakes (see details in Appendix). Most of the catalog earthquakes are located far from the DCPP, in areas known to have higher rates of seismicity: long-lived aftershocks of the 2003 M_w 6.5 San Simeon earthquake (McLaren *et al.*, 2008) and earthquakes on the creeping section of the San Andreas fault (e.g., Nadeau and McEvilly, 2004) and on the Kettleman Hills blind-thrust fault system (Ekstrom *et al.*, 1992; Stein and Ekstrom, 1992). Out of the 351 new detected events we were able to locate, 224 of them are located near the network within the black box in Figure 5; most of their magnitudes are between 0.0 and 1.5 (Fig. 3, gray bars). The remaining 127 new events, many of them over 50 km away from the network, are located offshore to the south where the SCSN coverage is sparse; they have slightly higher local magnitudes between 1 and 2 (Fig. 3, black bars).

Table 1
Summary of the 5048 Candidate Events Detected by Fingerprint and Similarity Thresholding (FAST) Categorized by Event Type after Visual Inspection and Catalog Comparison

Number of Events	Percentage of Events	Category
4134	81.90	Catalog earthquakes
420	8.32	New local earthquakes
18	0.35	Catalog quarry blasts
5	0.10	Teleseismic earthquakes
30	0.59	Duplicate earthquake detections
62	1.23	Infrasound signals (sound waves)
379	7.51	Background noise (false positive detections)
5048	100.00	Total candidate event detections

We conducted additional detailed analysis of seismicity for the region near the DCPP and seismic network, within the box defined by 34.9°–35.45° N, 121.2°–120.4° W (Fig. 6). This area contains 265 out of the 4134 catalog earthquakes that were detected by FAST (Fig. 6, hollow circles sized by magnitude), as well as 224 new detected local earthquakes (Fig. 6, dark diamonds sized by magnitude). Also, 6 of the 18 detected catalog quarry blasts are located inside this area. 236 catalog earthquakes within this region, with magnitudes −0.3–2.9, were missed detections that FAST failed to identify (Fig. 6, X symbols sized by magnitude); 210 missed events were only in the NCSN catalog, 23 only in the SCSN catalog, and three in both catalogs. Most of the missed catalog events are located outside and to the north of the 11-station network used for detection, in which there are several other stations in the PG and NC networks that we did not use for detection. In addition, about one-third of the missed catalog events occurred from 2007 to 2009, when only three of our selected stations had available continuous data to use for detection (Fig. 2b); there were other triggered stations in the network that detected catalog events during this time. Figure 2b displays the magnitude–time information for the 725 known earthquakes that occurred between June 2007 and October 2017 within this box: 265 detected catalog earthquakes (hollow circles), 236 missed catalog earthquakes (X symbols), and 224 new detected earthquakes (dark circles). The catalog and new events detected by FAST in Figure 2b are a subset of the events shown in Figure 2a.

For 715 out of the 725 local earthquakes near the DCPP, we located both the catalog and newly detected events with VELEST (see details in Appendix) using a consistent procedure: we ignore poor quality *P* and *S* picks, allow slight changes to the velocity model (Table A6), and calculate station corrections to reduce errors from near-surface variations in the velocity model. Figure A2 shows the starting locations for this second VELEST run, whereas Figure 6 displays our resulting preferred locations for these earthquakes, especially for those located inside the network. In Figure 6, some of the new detected earthquakes (dark diamonds) have different locations from the catalog events, which demonstrates that a blind search for similar earthquakes in continuous data is capable of finding unknown sources of low-magnitude seismicity. Large-*T* earthquake detection with FAST revealed many uncataloged earthquakes within the seismic network, especially in the Irish Hills (Fig. 1), where the NCSN and SCSN catalogs between 1967 and June 2007, before the 10-yr time period in this study, show low levels of seismicity (Ⓔ Fig. S5). These earthquake locations confirm that the faults located near DCPP are seismically active, even though the overall rate and magnitude of seismicity in the area was relatively low during 2007–2017 (Fig. 2b), and reinforces the importance of continued earthquake monitoring in this area.

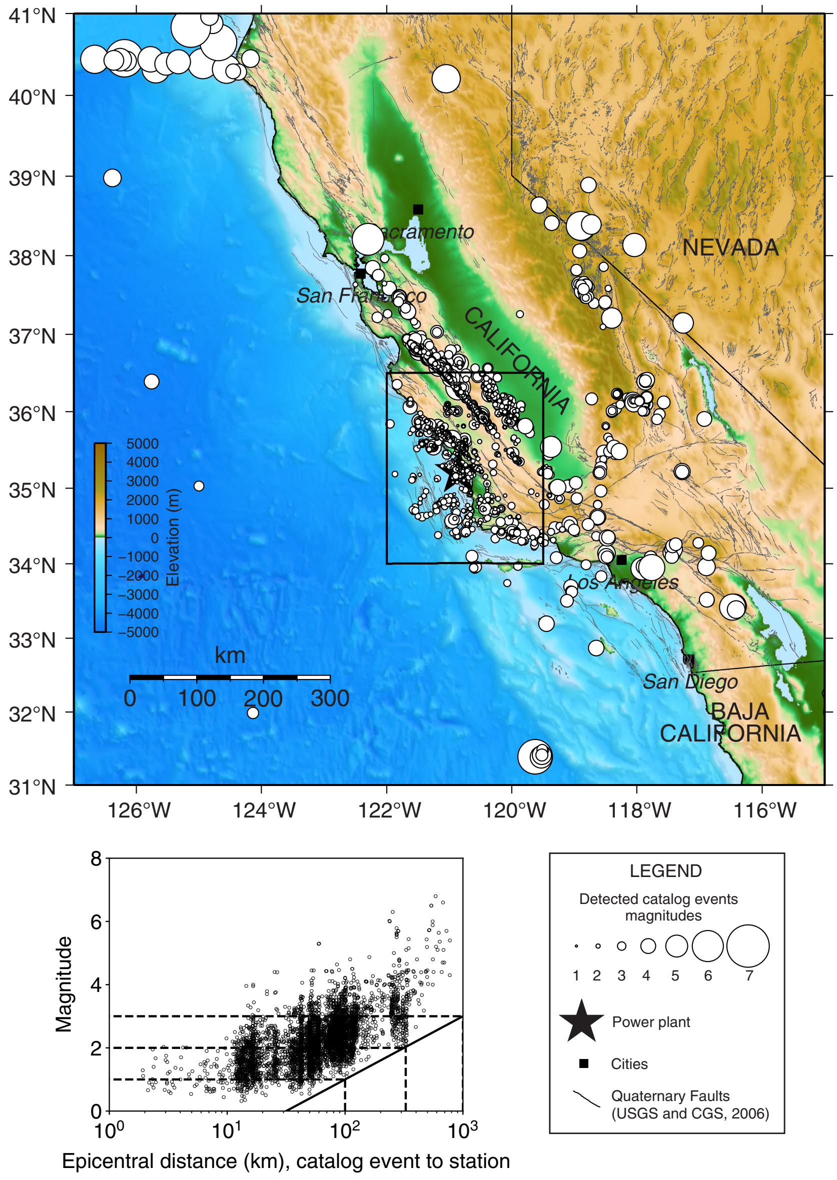

Figure 4. Locations of 4134 catalog earthquakes detected by FAST (white circles sized by magnitude), with $0.3 \le M \le 6.8$. The 11-station seismic network is located within 30 km of the DCPP (the star). The box indicates the region shown in Figure 5. Topography data are from Amante and Eakins (2009). Bottom left: magnitude of these catalog events, as a function of their epicentral distance; the solid black line denotes an approximate detection limit for a given magnitude and distance beyond which no events are detected, and dashed lines denote the maximum distance from the station where magnitude 1, 2, 3 events can be detected. The color version of this figure is available only in the electronic edition.

Discussion and Conclusions

An uninformed search for earthquake signals within continuous seismic data over long time periods, using waveform similarity as a detection criterion, is possible with the enhanced FAST software (Rong *et al.*, 2018). The improved algorithms overcame many of the challenges and limitations described in the initial implementation of FAST (Yoon *et al.*, 2015), while also demonstrating improved detection performance on synthetic data (Ⓔ Fig. S1). Detecting earthquakes with FAST on decadal data sets requires us to leverage significant computational resources: several Linux clusters with 512 GB memory, over 20 TB of disk space, and ample computing power with the ability to run up to 56 processes in parallel. For best detection results and shorter runtime using FAST, it is essential to eliminate or mitigate the presence of correlated noise that repeats in time, which can be done by an informed choice of band-pass filter for each channel of continuous data and by applying an occurrence filter for frequently repeating nonearthquake signals. Sometimes it makes sense to exclude time periods dominated by repeating noise from the detection process (Fig. 2a, channels PG.SH.EHZ, NC.PPB.EHZ, and NC.PABB.EHZ), or even exclude an entire station with poor quality data (PG.EC). Also, an appropriate choice of the similarity search parameter r (number of hash functions per hash table) will maximize the number of detected earthquakes, without becoming overwhelmed with false positive detections that increase both runtime and output file size. Different choices of r and other similarity search parameters (Table A2) can result in nearly identical detection probabilities with very different runtimes (Rong *et al.*, 2018). Finally, detection and association of similar earthquakes across a seismic network (Bergen and Beroza, 2018a) is essential to condensing and reconciling the outputs from FAST at each station, while removing many false positive detections, into a list of likely earthquake candidates.

This list of earthquake candidates provides a starting point for further analysis of these signals, including additional automatic and manual inspection, classification, removal of duplicate events, phase picking, location, magnitude estimation, and comparison with existing earthquake catalogs. One limitation of our study was the manual inspection of 29,623 candidate events; this step could ideally be replaced with automatic classification of different categories of seismic signals. We find that FAST will detect earthquakes located at different distances from the seismic network used for detection, although only larger earthquakes can be detected farther away from the network (Fig. 4). Because the detection criterion is similarity, a pair of low-magnitude earthquakes are often detected with very high similarity, whereas the coda from a larger earthquake often matches a

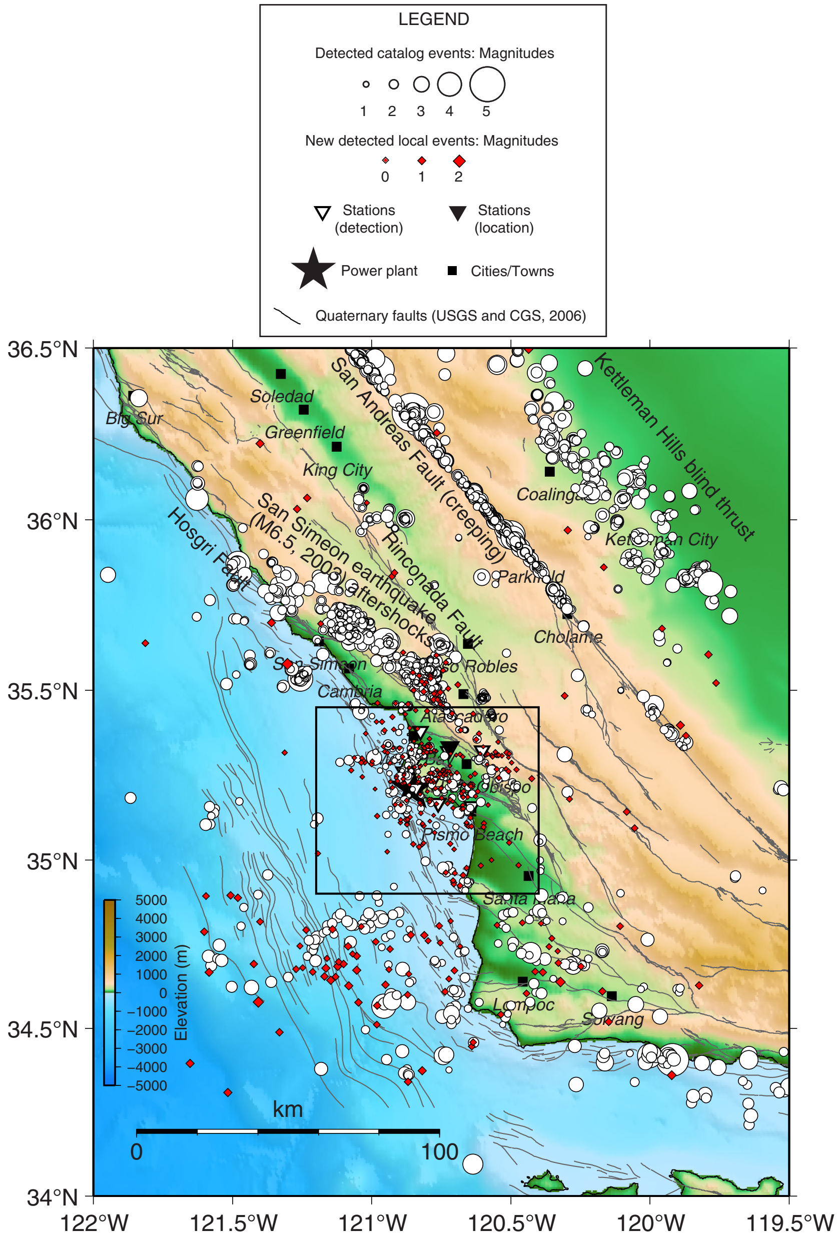

Figure 5. Regional view of earthquakes detected by FAST with continuous data from an 11-station network: 3106 catalog earthquakes within this area (white circles sized by magnitude) that were used to calibrate the local magnitude calculation (see the Appendix), and 351 new local earthquakes located with VELEST (dark diamonds sized by magnitude). The box indicates the zoomed region near the DCPP, plotted in Figure 6. Topography data are from Amante and Eakins (2009). The color version of this figure is available only in the electronic edition.

smaller earthquake with lower similarity. FAST can also detect teleseismic earthquakes (Ⓔ Fig. S3), although the number of detected teleseismic earthquakes depends on the band-pass filter applied; decreasing the minimum frequency of the band-pass filter to 1 Hz would result in many more teleseismic earthquake detections. In addition, FAST identifies other repeating signals that are not earthquakes, such as quarry blasts, vibrations from a seismic-reflection survey (Ⓔ Fig. S2), and infrasound signals that travel more slowly than seismic waves (Ⓔ Fig. S4). Most importantly, FAST is capable of detecting small ($M_L \leq 2$) local earthquakes (Fig. 3) missing from existing catalogs, even in a well-instrumented region near the DCPP where the seismic network is relatively dense and a carefully compiled catalog is complete to lower magnitudes. Some of the newly detected events appear to represent previously unknown earthquake sources. FAST is especially useful in situations in which the seismic network is sparse, but continuously recorded data are available. For example, this study found new earthquakes ~60 km away from the network, located offshore where there are no SCSN stations (Fig. 5), and single-station FAST detected over 100 times the number of catalog events in Guy–Greenbrier, Arkansas, where the seismic network is sparse (Yoon *et al.*, 2017). The detection performance of FAST is comparable to that of supervised detection methods such as template matching; Yoon *et al.* (2017) found ~90% overlap between events detected by FAST and events detected by template matching, with a small minority of events that were detected by either FAST or by template matching but not by both methods. Rong *et al.* (2018) found that when the ConvNetQuake supervised detection method (Perol *et al.*, 2018) was trained only on existing catalog events, it failed to identify ~30% of the new events detected by FAST. A limitation of FAST is that a unique earthquake signal occurring only once during the duration of continuous data processed, which is not similar to another earthquake signal that occurs at a different time, will not be detected. FAST is not intended to replace existing earthquake detection algorithms such as STA/LTA or template matching but instead offers a complementary method to identify additional earthquakes that would otherwise be overlooked.

The source code for the new FAST software introduced by Rong *et al.* (2018) is available on GitHub (see Data and Resources). For large-T data sets, ranging from months to years, FAST should be run on clusters with at least a few hundred GB memory, several TB of disk space, and parallel processing, which are also accessible through cloud computing services such as Amazon Web Services, Microsoft Azure, or Google Cloud Platform; it is not meant to run on a laptop or desktop computer, except perhaps for data sets of modest duration, ranging from hours to weeks. The FAST software can be a useful tool for seismologists who have

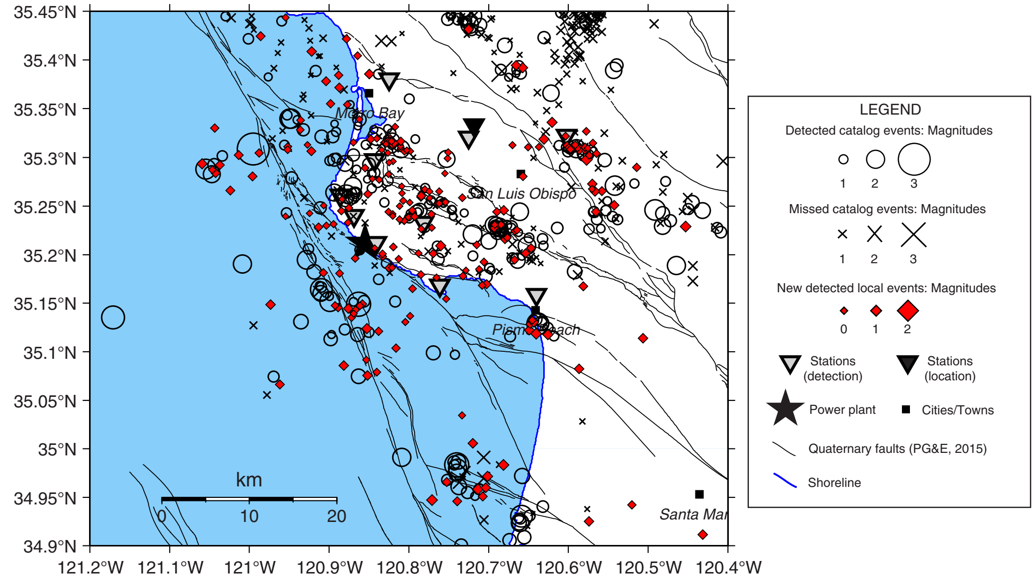

Figure 6. 715 local earthquakes (sized by magnitude) near the DCPP (star): 265 catalog earthquakes detected by FAST (hollow circles), 226 catalog earthquakes missed by FAST (X symbols), 224 new earthquakes detected by FAST (dark diamonds). These locations were computed by VELEST with station corrections, using a consistent procedure for all 715 earthquakes in this box (see the Appendix) and starting locations from Figure A2. The seismic network used for event detection (hollow triangles) and an additional station used for location (solid triangle) is also shown. The color version of this figure is available only in the electronic edition.

a collection of continuous data already recorded at several seismic stations from either a permanent seismic network or a temporary deployment but who do not have prior knowledge about earthquakes or possible template waveforms and are interested in scanning it for small local earthquakes. The software does not yet have the capability for real-time earthquake detection based on similarity to other existing earthquakes in the database, which would require significant changes to the implementation. FAST can be helpful for tracking changes in low-level seismicity over long-time periods and for detecting missing events in earthquake sequences such as repeating earthquakes, swarms, foreshocks, aftershocks, and induced earthquakes, in which the existence of many events with similar waveforms enhances detection capability.

FAST has been applied to only a few data sets to date (Yoon *et al.*, 2015, 2017; Bergen and Beroza, 2018a; Rong *et al.*, 2018). Future applications of FAST could include moving toward large-N applications of large-T seismology. This study showed the successful application of FAST on an 11-station network, but there were some catalog events that were not detected by FAST, particularly on the periphery of the stations used for detection. Running FAST on the entire archive of continuous seismic data from ~500 seismic stations in a regional seismic network, such as the NCSN or SCSN, would require significant computational resources and would encounter additional unforeseen challenges when scaling up to a larger network, but the resulting catalogs would be more complete and might reveal unexpected new sources of earthquake activity. FAST also might detect repeating weak and unusual events that are not typical earthquakes but still interesting to geophysicists, such as volcanic drumbeat earthquakes (e.g., Bell *et al.*, 2017), glacial icequakes (e.g., Helmstetter *et al.*, 2015), and tectonic tremor and low-frequency earthquakes (e.g., Shelly, 2017); however, successful detection of these weak events would require careful treatment of the data and might require changes in feature selection.

Data-mining techniques are just beginning to have useful impacts on earthquake seismology, and we can anticipate future discoveries enabled by FAST and related methods. FAST is an example of an unsupervised machine-learning algorithm (James *et al.*, 2017) that finds patterns (similar signals) in an unstructured data set (continuous time-series data without prior knowledge of earthquake activity). On the other hand, there are supervised machine-learning algorithms that take labeled data with known characteristics (e.g., waveforms of template earthquakes or known phase arrivals) to train a model that can classify, or predict a label for, new unlabeled data. Seismologists used supervised machine-learning methods, such as artificial neural networks, to analyze earthquake signals for over 20 yr (see Mousavi *et al.*, 2016 for a thorough review). Recent machine-learning developments created new opportunities for seismologists to extract useful information, with advances in algorithms, computational resources, and open-source software accessible to nonexperts. For example, deep learning and convolutional neural networks, which are supervised machine-learning methods that require massive labeled data sets to train a model, were originally developed for image recognition, but seismologists are starting to harness these powerful techniques to automatically detect earthquakes (Perol *et al.*, 2018; Ross, Meier, and Hauksson, 2018) and pick phase-arrival times (Ross, Meier, Hauksson, and Heaton, 2018; Zhu and Beroza,

2019) with high accuracy and few errors. Unsupervised methods like FAST can be used to generate more complete training data sets for supervised earthquake-identification methods. Data mining and machine-learning techniques are poised to have more prominent impacts on seismology in the near future (Bergen *et al.*, 2019; Kong *et al.*, 2019), and we look forward to new developments that surpass the earthquake-detection capabilities presented in this study.

Data and Resources

Continuous waveform data and earthquake catalogs for this study were last accessed in October 2017 through the Northern California Earthquake Data Center (NCEDC), doi: 10.7932/NCEDC (Northern California Earthquake Data Center [NCEDC], 2014), operated by the UC Berkeley Seismological Laboratory and the U.S. Geological Survey (USGS). Earthquake catalogs were last accessed October 2017, provided by the the Caltech/USGS Southern California Seismic Network (SCSN), doi: 10.7914/SN/CI, operated by the Caltech Seismological Laboratory and the USGS, which is archived at the Southern California Earthquake Data Center (SCEDC) (2013). Comprehensive Earthquake Catalog (ComCat) data for teleseismic earthquakes were downloaded from the USGS website https://earthquake.usgs.gov/data/comcat/ (last accessed October 2017). We ran Fingerprint And Similarity Thresholding (FAST) on Linux clusters provided by the Data Analytics for What's Next (DAWN) project (https://dawn.cs.stanford.edu) and Future Data Systems group (https://futuredata.stanford.edu) in the computer science department at Stanford University. We used Seismic Analysis Code (SAC; Helffrich *et al.*, 2013) to manually pick *P* and *S* arrivals as needed, ObsPy for downloading continuous waveform data, seismological data processing, and visualization (Beyreuther *et al.*, 2010), and Generic Mapping Tools (GMT) to generate maps (Wessel *et al.*, 2013). The source code for the new FAST software (Rong *et al.*, 2018) is available at https://github.com/stanford-futuredata/FAST. Several references can be accessed on the web. Bormann (2012) is available at https://nmsop.gfz-potsdam.de (last accessed July 2017). Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC) (2019) is available at https://ds.iris.edu/files/stats/data/archive/Archive_Growth.jpg (last accessed May 2019). James *et al.* (2017) is available at https://www-bcf.usc.edu/~gareth/ISL/ (last accessed May 2019). PG&E (2011) is available at https://www.pge.com/mybusiness/edusafety/systemworks/dcpp/shorelinereport/index.shtml (last accessed May 2019). PG&E (2014) is available at https://www.pge.com/en_US/safety/how-the-system-works/diablo-canyon-power-plant/seismic-safety-at-diablo-canyon/seismic-report.page (last accessed May 2019). PG&E (2015) is available at https://www.pge.com/en_US/safety/how-the-system-works/diablo-canyon-power-plant/seismic-safety-at-diablo-canyon/sshac.page (last accessed May 2019). Rong *et al.* (2018) is available at https://www.vldb.org/pvldb/vol11/p1674-rong.pdf (last accessed May 2019). The scipy.misc.imresize documentation SciPy (2019) is available at https://docs.scipy.org/doc/scipy-1.1.0/reference/generated/scipy.misc.imresize.html (last accessed June 2019). USGS and CGS (2006) is available at https://earthquake.usgs.gov/hazards/qfaults/ (last accessed March 2018). MATLAB is avaialble at https://www.mathworks.com/help/matlab/ref/imresize.html (last accessed June 2019).

Acknowledgments

Serkan Bozkurt of Lettis Consultants International provided the detailed fault map in Figure 1 (PG&E, 2015). The authors thank Stuart Nishenko and Marcia McLaren from Pacific Gas and Electric Company (PG&E) for helpful discussion. C. Y. was funded by a Chevron Stanford Graduate Fellowship, and this work was funded by National Science Foundation (NSF) Grant EAR-1818579 and Pacific Gas and Electric. This study was part of the Mining Seismic Wavefields project in the Southern California Earthquake Center (SCEC), Contribution Number 8971. SCEC is funded by NSF Cooperative Agreement EAR-1600087 and U.S. Geological Survey (USGS) Cooperative Agreement G17AC00047. Constructive reviews from Phil Dawson, an anonymous reviewer, and Associate Editor Eric Chael improved the quality of this article.

References

Aguiar, A. C., and G. C. Beroza (2014). PageRank for earthquakes, *Seismol. Res. Lett.* **85,** no. 2, 344–350, doi: 10.1785/0220130162.

Allen, R. (1982). Automatic phase pickers: Their present use and future prospects, *Bull. Seismol. Soc. Am.* **72,** no. 6, S225–S242.

Amante, C., and B. W. Eakins (2009). ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis, *NOAA Technical Memorandum NESDIS NGDC-24*, National Geophysical Data Center, NOAA, doi: 10.7289/V5C8276M.

Andoni, A., and P. Indyk (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, *Foundations of Computer Science, 2006. FOCS‘06. 47th Annual IEEE Symposium*, 459–468, doi: 10.1145/1327452.1327494.

Bell, A. F., S. Hernandez, H. E. Gaunt, P. Mothes, M. Ruiz, D. Sierra, and S. Aguaiza (2017). The rise and fall of periodic ‘drumbeat’ seismicity at Tungurahua volcano, Ecuador, *Earth Planet. Sci. Lett.* **475,** 58–70, doi: 10.1016/j.epsl.2017.07.030.

Bergen, K., C. Yoon, and G. C. Beroza (2016). Scalable similarity search in seismology: A new approach to large-scale earthquake detection, *Proc. of the 9th International Conference on Similarity Search and Applications*, 301–308, doi: 10.1007/978-3-319-46759-7_23.

Bergen, K. J., and G. C. Beroza (2018a). Detecting earthquakes over a seismic network using single-station similarity measures, *Geophys. J. Int.* **213,** no. 3, 1984–1998, doi: 10.1093/gji/ggy100.

Bergen, K. J., and G. C. Beroza (2018b). Earthquake fingerprints: Extracting waveform features for similarity-based earthquake detection, *Pure Appl. Geophys.* doi: 10.1007/s00024-018-1995-6.

Bergen, K. J., T. Chen, and Z. Li (2019). Preface to the focus section on machine learning in seismology, *Seismol. Res. Lett.* **90,** no. 2A, 477–480, doi: 10.1785/0220190018.

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

Bormann, P. (Editor) (2012). Seismic sources and source parameters, in *New Manual of Seismological Observatory Practice (NMSOP-2)*, Chapter 3, Second Ed., IASPEI, GFZ German Research Centre for Geosciences, Potsdam, Germany, doi: 10.2312/GFZ.NMSOP-2.

Broder, A. Z., M. Charikar, A. M. Frieze, and M. Mitzenmacher (2000). Min-wise independent permutations, *J. Comput. Syst. Sci.* **60,** no. 3, 630–659, doi: 10.1006/jcss.1999.1690.

Brown, J. R., G. C. Beroza, and D. R. Shelly (2008). An autocorrelation method to detect low frequency earthquakes within tremor, *Geophys. Res. Lett.* **35,** L16305, doi: 10.1029/2008GL034560.

Cates, J. E., and B. Sturtevant (2001). Seismic detection of sonic booms, *J. Acoust. Soc. Am.* **111,** no. 1, 614–628, doi: 10.1121/1.1413754.

Cochran, E. S., and P. M. Shearer (2006). Infrasound events detected with the Southern California Seismic Network, *Geophys. Res. Lett.* **33,** L19803, doi: 10.1029/2006GL026951.

Ekstrom, G., R. S. Stein, J. P. Eaton, and D. Eberhart-Phillips (1992). Seismicity and geometry of a 110-km-long blind thrust fault 1. The 1985 Kettleman Hills, California, earthquake, *J. Geophys. Res.* **97,** no. B4, 4843–4864, doi: 10.1029/91JB02925.

Geller, R. J., and C. S. Mueller (1980). Four similar earthquakes in central California, *Geophys. Res. Lett.* **7,** no. 10, 821–824, doi: 10.1029/GL007i010p00821.

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, doi: 10.1111/j.1365-246X.2006.02865.x.

Hardebeck, J. L. (2010). Seismotectonics and fault structure of the California Central Coast, *Bull. Seismol. Soc. Am.* **100,** no. 3, 1031–1050, doi: 10.1785/0120090307.

Helffrich, G., J. Wookey, and I. Bastow (2013). *The Seismic Analysis Code: A Primer and User's Guide*, First Ed., Cambridge University Press, Cambridge, United Kingdom.

Helmstetter, A., B. Nicolas, P. Comon, and M. Gay (2015). Basal icequakes recorded beneath an Alpine glacier (Glacier dArgentiere, Mont Blanc, France): Evidence for stick-slip motion?, *J. Geophys. Res.* **120,** 379–401, doi: 10.1002/2014JF003288.

Incorporated Research Institutions for Seismology Data Management Center IRIS -DMC Archive (2019). *IRIS DMC Archive as of May 2019*, available at https://ds.iris.edu/files/stats/data/archive/Archive_Growth.jpg (last accessed May 2019).

James, G., D. Witten, T. Hastie, and R. Tibshirani (2017). *An Introduction to Statistical Learning*, Springer, New York, New York.

Ji, J., J. Li, S. Yan, Q. Tian, and B. Zhang (2013). Min-max hash for Jaccard similarity, *2013 IEEE 13th International Conference on Data Mining*, 301–309, doi: 10.1109/ICDM.2013.119.

Kissling, E., W. L. Ellsworth, D. Eberhart-Phillips, and U. Kradolfer (1994). Initial reference models in local earthquake tomography, *J. Geophys. Res.* **99,** 19,635–19,646.

Kong, Q., D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft (2019). Machine learning in seismology: Turning data into insights, *Seismol. Res. Lett.* **90,** no. 1, 3–14, doi: 10.1785/0220180259.

Li, Z., Z. Peng, D. Hollis, L. Zhu, and J. McClellan (2018). High-resolution seismic event detection using local similarity for Large-N arrays, *Nature Sci. Rept.* **8,** 1646, doi: 10.1038/s41598-018-19728-w.

Lindsey, N. J., E. R. Martin, D. S. Dreger, B. Freifeld, S. Cole, S. R. James, B. L. Biondi, and A.-J. B. Franklin (2017). Fiber-optic network observations of earthquake wavefields, *Geophys. Res. Lett.* **44,** 11,792–11,799, doi: 10.1002/2017GL075722.

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

McLaren, M. K., and W. U. Savage (2001). Seismicity of south-central coastal California: October 1987 through January 1997, *Bull. Seismol. Soc. Am.* **91,** no. 6, 1629–1658, doi: 10.1785/0119980192.

McLaren, M. K., J. L. Hardebeck, N. van der Elst, J. R. Unruh, G. W. Bawden, and J. L. Blair (2008). Complex faulting associated with the 22 December 2003 M_w 6.5 San Simeon, California, earthquake, aftershocks, and postseismic surface deformation, *Bull. Seismol. Soc. Am.* **98,** no. 4, 1659–1680, doi: 10.1785/0120070088.

Meng, H., and Y. Ben-Zion (2018). Detection of small earthquakes with dense array data: Example from the San Jacinto fault zone, southern California, *Geophys. J. Int.* **212,** 442–457, doi: 10.1093/gji/ggx404.

Mousavi, S. M., S. P. Horton, C. A. Langston, and B. Samei (2016). Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression, *Geophys. J. Int.* **207,** 29–46, doi: 10.1093/gji/ggw258.

Nadeau, R. M., and T. V. McEvilly (2004). Periodic pulsing of characteristic microearthquakes on the San Andreas fault, *Science* **303,** 220–222, doi: 10.1126/science.1090353.

Northern California Earthquake Data Center (NCEDC) (2014). *UC Berkeley Seismological Laboratory, Dataset*, doi: 10.7932/NCEDC.

Pacific Gas and Electric Company (PG&E) (2011). Report on the analysis of the Shoreline fault zone, central coastal California, *Report to the U.S. Nuclear Regulatory Commission*.

Pacific Gas and Electric Company (PG&E) (2014). Report on the Central Coastal California Seismic Imaging Project (CCCSIP), *Report to the U.S. Nuclear Regulatory Commission*.

Pacific Gas and Electric Company (PG&E) (2015). Seismic source characterization for the Diablo Canyon Power Plant, San Luis Obispo County, California, *Report on the results of a SSHAC level 3 study, Rev. A*.

Perol, T., M. Gharbi, and M. Denolle (2018). Convolutional neural network for earthquake detection and location, *Sci. Adv.* **1,** e1700578, doi: 10.1126/sciadv.1700578.

Poupinet, G., W. L. Ellsworth, and J. Frechet (1984). Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California, *J. Geophys. Res.* **89,** no. B7, 5719–5731, doi: 10.1029/JB089iB07p05719.

Rong, K., C. E. Yoon, K. J. Bergen, H. Elezabi, P. Bailis, P. Levis, and G. C. Beroza (2018). Locality-sensitive hashing for earthquake detection: A case study scaling data-driven science, *Proc of the VLDB Endowment*, Vol. 11, 1674–1687, doi: 10.14778/3236187.3236214.

Ross, Z. E., M. A. Meier, and E. Hauksson (2018). *P*-wave arrival picking and first-motion polarity determination with deep learning, *J. Geophys. Res.* **23,** 5120–5129, doi: 10.1029/2017JB015251.

Ross, Z. E., M. A. Meier, E. Hauksson, and T. H. Heaton (2018). Generalized seismic phase detection with deep learning, *Bull. Seismol. Soc. Am.* doi: 10.1785/0120180080.

Ross, Z. E., D. T. Trugman, E. Hauksson, and P. M. Shearer (2019). Searching for hidden earthquakes in Southern California, *Science* **364,** 767–771, doi: 10.1126/science.aaw6888.

Schaff, D. P., and G. C. Beroza (2004). Coseismic and postseismic velocity changes measured by repeating earthquakes, *J. Geophys. Res.* **109,** no. B10302, doi: 10.1029/2004JB003011.

SciPy (2019). *SciPy, Version 1.1.0*, Open source scientific tools for Python, scipy.misc.imresize documentation, available at https://docs.scipy.org/doc/scipy-1.1.0/reference/generated/scipy.misc.imresize.html (last accessed June 2019).

Shelly, D. R. (2017). A 15 year catalog of more than 1 million low-frequency earthquakes: Tracking tremor and slip along the deep San Andreas Fault, *J. Geophys. Res.* **122,** 3739–3753, doi: 10.1002/2017JB014047.

Skoumal, R. J., M. R. Brudzinski, and B. S. Currie (2016). An efficient repeating signal detector to investigate earthquake swarms, *J. Geophys. Res.* **121,** 5880–5897, doi: 10.1002/2016JB012981.

Skoumal, R. J., P. B. Dawson, S. H. Hickman, and J. O. Kaven (2018). Microseismic events associated with the Oroville Dam Spillway, *Bull. Seismol. Soc. Am.* **109,** 387–394, doi: 10.1785/0120180255.

Southern California Earthquake Center (2013). *Caltech. Dataset*, doi: 10.7909/C3WD3xH1.

Stein, R. S., and G. Ekstrom (1992). Seismicity and geometry of a 110-km-long blind thrust fault 2. Synthesis of the 1982–1985 California earthquake sequence, *J. Geophys. Res.* **97,** no. B4, 4865–4883, doi: 10.1029/91JB02847.

U.S. Geological Survey and California Geological Survey (USGS and CGS) (2006). *Quaternary Fault and Fold Database for the United States*, available at https://earthquake.usgs.gov/hazards/qfaults/ (last accessed March 2018).

Velasco, A. A., R. Alfaro-Diaz, D. Kilb, and K. L. Pankow (2016). A time-domain detection approach to identify small earthquakes within the continental United States recorded by the USArray and regional networks, *Bull. Seismol. Soc. Am.* **106,** no. 2, 512–525, doi: 10.1785/0120150156.

Walker, K. T., R. Shelby, M. A. H. Hedlin, C. de Groot-Hedlin, and F. Vernon (2011). Western U.S. infrasonic catalog: Illuminating infrasonic hot spots with the USArray, *J. Geophys. Res.* **116,** no. B12305, doi: 10.1029/2011JB008579.

Wessel, P., W. H. F. Smith, R. Scharroo, J. F. Luis, and F. Wobbe (2013). Generic mapping tools: Improved version released, *Eos Trans. AGU* **94,** 409–410, doi: 10.1002/2013EO450001.

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

Yoon, C. E., Y. Huang, W. L. Ellsworth, and G. C. Beroza (2017). Seismicity during the initial stages of the Guy-Greenbrier, Arkansas, earthquake sequence, *J. Geophys. Res.* **122,** doi: 10.1002/2017JB014946.

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, doi: 10.1126/sciadv.1501057.

Zhu, W., and G. C. Beroza (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method, *Geophys. J. Int.* **216,** 261–273, doi: 10.1093/gji/ggy423.

Appendix

The appendix contains more detailed information about the data set and Fingerprint And Similarity Thresholding (FAST) input parameters (Tables A1–A4), local magnitude estimation for new detected earthquakes, and the earthquake location procedure including velocity models (Tables A5 and A6).

Local Magnitude Estimation for New Detected Earthquakes

We report local magnitude M_{L} for the 420 new detected local earthquakes in this study. We solve for the local-magnitude distance-correction parameters by calibrating M_{L} to the catalog magnitudes M_{cat} from the 3106 catalog earthquakes located within the region in Figure 5 (34°–36.5° N, 122°–119.5° W). We assume that the magnitude can be expressed as

$$M_{\mathrm{cat}} = \log_{10}(A_{\mathrm{peak}} R^k) + C$$
$$\Rightarrow M_{\mathrm{cat}} - \log_{10} A_{\mathrm{peak}} = k \log_{10} R + C, \qquad \text{(A1)}$$

with $\log_{10} A_{\mathrm{peak}}$ computed as

$$\log_{10} A_{\mathrm{peak}} = \frac{1}{2}(\log_{10} A_{\mathrm{peak,east}} + \log_{10} A_{\mathrm{peak,north}}), \qquad \text{(A2)}$$

in which $A_{\mathrm{peak,east}}$ and $A_{\mathrm{peak,north}}$ are peak Wood–Anderson seismogram amplitudes, from the east and north components, respectively (Bormann, 2012), at each of the 12 stations in Table A1. The Wood–Anderson seismograms were synthesized from the original data after applying a 1-Hz high-pass filter to remove low-frequency noise. For stations where only the vertical component was available, we used the peak Wood–Anderson amplitude on the 1-Hz high-pass filtered vertical component: $A_{\mathrm{peak}} = A_{\mathrm{peak,vertical}}$. In equation (A1), R is the epicentral distance, and there are two distance-correction parameters to estimate: k (representing the effect of geometric spreading and attenuation), and a constant C. Equation (A3) shows equation (A1) in matrix form ($\mathbf{d} = \mathbf{Gm}$). Here, $\mathbf{G}$ is the design matrix, in which the number of rows equals 3106 events times the number of stations that recorded each event (j is the station index from 1 to at most 12), and i is the event index between 1 and 3106

$$\begin{bmatrix} M_{\mathrm{cat}}[i=1] - \log_{10}(A_{\mathrm{peak},j=1}[i=1]) \\ M_{\mathrm{cat}}[i=1] - \log_{10}(A_{\mathrm{peak},j=2}[i=1]) \\ \vdots \\ M_{\mathrm{cat}}[i=1] - \log_{10}(A_{\mathrm{peak},j=12}[i=1]) \\ M_{\mathrm{cat}}[i=2] - \log_{10}(A_{\mathrm{peak},j=1}[i=2]) \\ M_{\mathrm{cat}}[i=2] - \log_{10}(A_{\mathrm{peak},j=2}[i=2]) \\ \vdots \\ M_{\mathrm{cat}}[i=2] - \log_{10}(A_{\mathrm{peak},j=12}[i=2]) \\ M_{\mathrm{cat}}[i=3106] - \log_{10}(A_{\mathrm{peak},j=1}[i=3106]) \\ M_{\mathrm{cat}}[i=3106] - \log_{10}(A_{\mathrm{peak},j=2}[i=3106]) \\ \vdots \\ M_{\mathrm{cat}}[i=3106] - \log_{10}(A_{\mathrm{peak},j=12}[i=3106]) \end{bmatrix}$$
$$= \begin{bmatrix} \log_{10}(R_{\mathrm{j}=1}[i=1]) & 1 \\ \log_{10}(R_{\mathrm{j}=2}[i=1]) & 1 \\ \vdots & \vdots \\ \log_{10}(R_{\mathrm{j}=12}[i=1]) & 1 \\ \log_{10}(R_{\mathrm{j}=1}[i=2]) & 1 \\ \log_{10}(R_{\mathrm{j}=2}[i=2]) & 1 \\ \vdots & \vdots \\ \log_{10}(R_{\mathrm{j}=12}[i=2]) & 1 \\ \log_{10}(R_{\mathrm{j}=1}[i=3106]) & 1 \\ \log_{10}(R_{\mathrm{j}=2}[i=3106]) & 1 \\ \vdots & \vdots \\ \log_{10}(R_{\mathrm{j}=12}[i=3106]) & 1 \end{bmatrix} \begin{bmatrix} k \\ C \end{bmatrix}. \qquad \text{(A3)}$$

Inverting for the best-fit distance-correction parameters in a least-squares sense, we get $k = 1.22$ and $C = 1.18$. Plugging in these parameters into equation (A1), and assuming $M_{\mathrm{cat}} = M_{\mathrm{L}}$, we calculate local magnitude M_{L} for every new detected earthquake, given peak amplitudes from 1-Hz high-pass filtered synthesized Wood–Anderson seismograms (equation A2) and epicentral distances, as the average of M_{L} estimates at each station j

$$(\log_{10} A_{\mathrm{peak,j}} + k \log_{10} R_j + C)$$
$$\Rightarrow M_{\mathrm{L}} = \sum_{j=1}^{12} (\log_{10} A_{\mathrm{peak,j}} + 1.22338 \log_{10} R_j + 1.17722). \qquad \text{(A4)}$$

For validation, we calculate M_{L} with equation (A4) for the 3106 catalog events used to calibrate the distance correction,

Table A1

Channel-Specific Fingerprint And Similarity Thresholding (FAST) Information for the 11 Stations (27 channels) of Continuous Seismic Data Used for Earthquake Detection, from the PG and NC Seismic Networks, Spanning a Time Period between 6 and 11 yr (Start Dates in Column 6)

Network	Station	Channel	Latitude (°)	Longitude (°)	Start Date (yyyy/mm/dd)	Band-pass Filter (Hz)*	Original Data Size (GB)[†]	Decimated Data Size (GB)[‡]	Number of Fingerprints[§]	Fingerprint Size (GB)[‖]	Similar Pairs Size (GB)—Binary[#]	Similar Pairs Size (GB)—Text**
NC	PPB	EHZ	35.261	−120.8862	2011/11/01	5–10	13	3.1	156,476,176	75	95	231
NC	PABB	EHZ	35.1589	−120.6399	2011/11/01	4–12	12	3.2	157,984,685	76	107	256
PG	MLD	EHE	35.3225	−120.6025	2011/06/01	4–10	20	3.6	165,651,918	79	198	510
PG	MLD	EHN	35.3225	−120.6025	2011/06/01	4–10	20	3.5	165,626,784	79	154	394
PG	MLD	EHZ	35.3225	−120.6025	2011/06/01	4–10	18	3.4	165,652,418	79	117	298
PG	LSD	EHE	35.2973	−120.843	2011/01/01	5–12	17	3.4	173,874,091	83	316	809
PG	LSD	EHN	35.2973	−120.843	2011/01/01	5–12	16	3.6	173,847,307	83	195	491
PG	LSD	EHZ	35.2973	−120.843	2011/01/01	5–12	15	3.6	173,821,481	83	102	256
PG	LMD	EHE	35.3803	−120.8247	2010/01/01	4–12	15	3.5	196,449,918	94	110	281
PG	LMD	EHN	35.3803	−120.8247	2010/01/01	4–12	15	3.5	196,489,413	94	68	170
PG	LMD	EHZ	35.3803	−120.8247	2010/01/01	4–12	15	3.5	196,512,756	94	38	94
PG	SHD	EHE	35.1682	−120.7613	2010/01/01	6–12	15	3.5	194,298,427	93	549	1500
PG	SHD	EHN	35.1682	−120.7613	2010/01/01	6–12	15	3.5	194,265,145	93	300	778
PG	SHD	EHZ	35.1682	−120.7613	2009/12/01	6–12	14	3.4	196,463,495	94	370	966
PG	EFD	EHE	35.3205	−120.725	2010/03/01	3–7	22	3.9	198,771,156	95	478	1500
PG	EFD	EHN	35.3205	−120.725	2010/03/01	3–7	23	3.8	198,713,221	95	222	571
PG	EFD	EHZ	35.3205	−120.725	2009/11/01	3–7	21	3.7	206,999,626	99	367	955
PG	VPD	EHE	35.2399	−120.869	2009/12/01	4–12	20	4.0	171,538,101	82	75	186
PG	VPD	EHN	35.2399	−120.869	2009/12/01	4–12	20	4.1	171,583,064	82	54	132
PG	VPD	EHZ	35.2399	−120.869	2009/10/01	4–12	17	3.9	173,924,002	83	94	238
PG	DPD	EHE	35.233	−120.7817	2010/07/01	4–12	21	3.8	189,686,360	91	231	586
PG	DPD	EHN	35.233	−120.7817	2010/07/01	4–12	21	3.6	189,691,188	91	177	451
PG	DPD	EHZ	35.233	−120.7817	2008/06/01	4–12	25	4.4	243,899,073	117	64	156
PG	DCD	EHE	35.2122	−120.8408	2007/06/01	4–8	25	5.6	261,202,942	125	83	203
PG	DCD	EHN	35.2122	−120.8408	2007/06/01	4–8	27	5.8	261,209,473	125	56	135
PG	DCD	EHZ	35.2122	−120.8408	2007/06/01	4–8	24	5.0	261,324,443	125	46	107
PG	SH	EHZ	35.1682	−120.7613	2006/09/01	4–10	28	6.2	286,876,081	137	61	139
PG	EC	EHZ	35.3333	−120.7182	2006/09/01	–	24	–	–	–	–	–

The end date and time of each channel of continuous data was 2017/10/24 00:00:00 UTC. Stations SHD and SH are at the same location, but we used both stations for detection because Station SH recorded continuously for a longer time period. The last station in the list (PG.EC) was not used for detection because of persistent repeating noise, but it was used to locate earthquakes.

*Station-specific band-pass filter (Hz) applied to each channel of data as a preprocessing step before running FAST.

[†] Original time-series data size (GB) sampled at 100 Hz; total size over all channels ~500 GB. Channels with the same duration may have different data sizes because of differences in time gaps.

[‡] Decimated time-series data size (GB) sampled at 25 Hz; total size over all channels ~100 GB.

[§] Number of fingerprints generated from continuous seismic data.

[‖] Size of fingerprint file (GB); total size over all channels is ~2.6 TB.

[#] Size of output binary file (GB) from similarity search, containing pairs of fingerprint indices (representing times in the continuous data) with similar signals and their associated FAST similarity; total size over all channels is ~4.7 TB.

**Size of output text file (GB) from similarity search (same information as in the binary file from column 8 but converted to text format after postprocessing); total size over all channels is ~12.4 TB.

which agree reasonably well with their original catalog magnitudes M_{cat} (Fig. A1).

For the 351 new located earthquakes, the epicentral distance R to each station is determined from the VELEST event location. The 69 new detected earthquakes that we were unable to locate almost always had one station (usually LMD or SHD) with both high-quality P- and S arrival-time picks (t_P and t_S respectively), so we estimate the epicentral distance at this station j as

$$R_j = \frac{V_P V_S}{V_P - V_S}(t_S - t_P), \quad \text{(A5)}$$

in which $V_P = 5$ km/s, $V_S = 3$ km/s and then calculate M_{L} with equation (A4) at this one station.

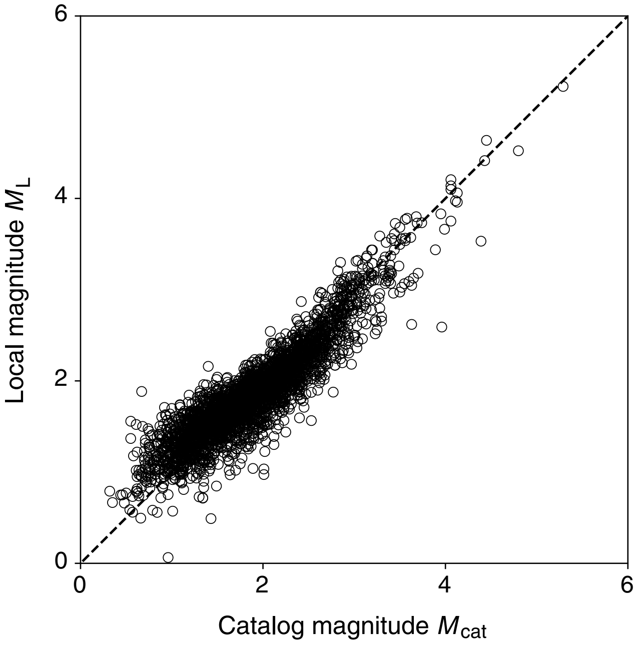

Figure A1. Magnitude calibration results: comparison of local magnitude M_{L} (equation A4) with catalog magnitude M_{cat} for the 3106 catalog earthquakes located within the region in Figure 5 (34°–36.5° N, 122°–119.5° W).

Earthquake Location Procedure

Initial VELEST Locations for 351 New Detected Earthquakes

We automatically pick P- and S-wave arrival times on 12 stations (Fig. 1, Table A1) with the Akaike information-criteria picker (Maeda, 1985), manually adjust them and remove noisy picks as needed, and assign integer weights for pick quality from 0 (best) to 3 (worst). We pick P phases on the vertical components and S phases on the horizontal components and estimate the origin time of each event using the S–P time on the station with either the earliest arrivals or the highest quality arrivals (usually station LMD or SHD).

These P- and S-wave picks are input into VELEST (Kissling *et al.*, 1994), with a 1D P-wave velocity model (Table A5) and $V_P/V_S = 1.66$ (Hardebeck, 2010). P- and S arrivals are equally weighted. We use the location-damping parameters *othet* = *xythet* = *zthet* = 0.03. We run VELEST for 50 iterations, without solving for a new velocity model and without station corrections. The dark diamonds in Figure 5 indicate the resulting initial VELEST locations for these 351 new detected earthquakes.

Refined VELEST Locations for 715 Earthquakes near Diablo Canyon Nuclear Power Plant

We perform a second VELEST run to carefully locate only the 715 earthquakes within the box defined by 34.9°–35.45° N, 121.2°–120.4° W (Fig. 6). This location procedure allows a consistent comparison between all earthquakes, regardless of whether they were in the NCSN catalog, SCSN catalog, both catalogs, or neither catalog. We locate the catalog events and new detected events using the same procedure, picking P- and S arrival times on the 12 stations (Fig. 1, triangles; Table A1), except that we already have origin times for the catalog events. In this second VELEST run, we do not use poor-quality picks weighted as a 3. Figure A2 shows the starting locations for this second VELEST run; we use the catalog locations for the 265 detected events (hollow circles) and 226 out of 236 missed events (X symbols), whereas we use locations from the initial VELEST run (Fig. 5) for the 224 new detected events (dark diamonds). We also calculate locations for the six catalog quarry blasts in this box, so we actually locate a total of 721 events. We were unable to locate 10 missed catalog events with VELEST, because they did not have any reliable picks.

We run VELEST with station corrections turned on (*nsinv* = 1), with damping-parameter *stathet* = 0.1, to reduce errors from near-surface heterogeneity in the velocity structure at each station. We allow a different station-correction value at each of the 12 stations. P- and S arrivals are equally weighted. We use the location-damping parameters *othet* = *xythet* = *zthet* = 0.01. For the initial velocity model, we use the 1D P-wave velocity model from McLaren and Savage (2001) (Table A5) and $V_P/V_S = 1.66$ (Hardebeck, 2010). We jointly estimate the velocity model, station corrections, and earthquake locations every *invertratio* = 3 iterations until we reach *ittmax* = 60 iterations. Although the damping parameter for the velocity model was very high (*vthet* = 1000), the output velocity model (Table A6) still changed relative to the initial model (Table A5). Figure 6 displays our resulting preferred locations for these 715 earthquakes from this second VELEST run.

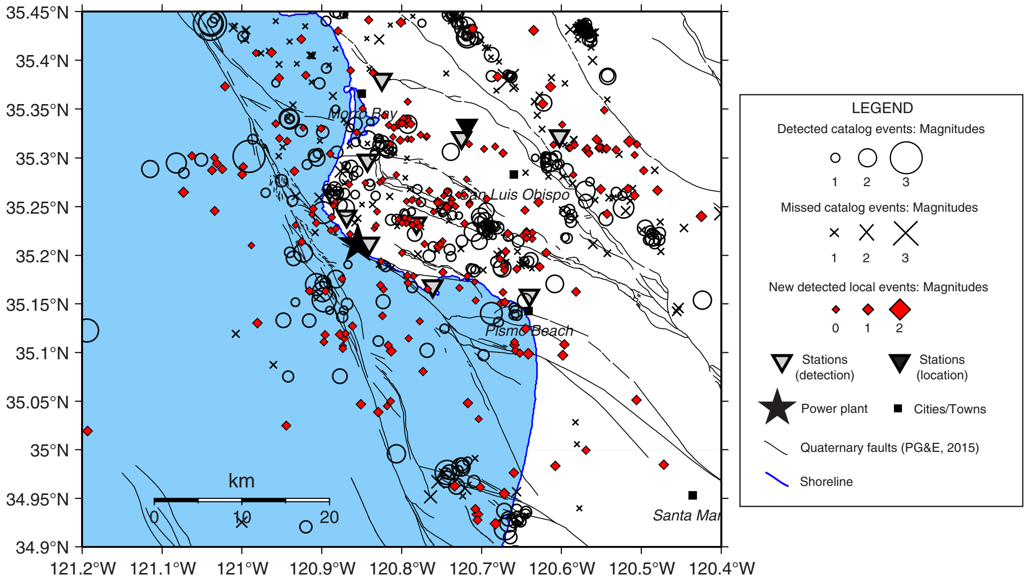

Figure A2. Initial locations for 715 out of 725 local earthquakes (sized by magnitude) near the DCPP (star), used as input into a second VELEST location run with station corrections; Figure 6 shows the output locations after the second VELEST run. For the 265 catalog earthquakes detected by FAST (hollow circles) and 226 out of 236 catalog earthquakes missed by FAST (X symbols), these are the catalog locations; if an event was in both the Northern California Seismic Network (NCSN) and Southern California Seismic Network (SCSN) catalogs, we used the NCSN catalog location. (10 missed catalog earthquakes did not have any reliable picks to use in VELEST.) For the 224 new earthquakes detected by FAST (dark diamonds), we used the locations calculated from the initial VELEST run that located 351 out of 420 new detected events. The seismic network used for event detection (hollow triangles) and an additional station used for location (solid triangle) are also shown. The color version of this figure is available only in the electronic edition.

Table A2

FAST Input Parameters (Yoon *et al.*, 2015; Bergen *et al.*, 2016; Bergen and Beroza, 2018b; Rong *et al.*, 2018) Used to Detect Earthquakes in Each Channel of Continuous Seismic Data (Band-Pass Filtered and Decimated to 25 Hz Sampling Rate) at Stations Listed in Table A1

FAST Algorithm Section	Parameter Description	Value*
Fingerprint	Time-window length (s) for spectrogram	6 s (150 [time] samples)
	Time-window lag (s) for spectrogram, between adjacent windows	0.12 s (3 [time] samples)
	Spectral-image length (samples)	64 (spectrogram time) samples (13.68 s)
	Spectral-image lag (samples) = fingerprint sampling period	10 (spectrogram time) samples (1.2 s)
	Final spectral-image width (samples) = number of frequency bins	32 (spectrogram frequency) samples
	Number of wavelet coefficients to keep	400 (out of 2048)
	Median/MAD sampling fraction of continuous data	0.01 (1%)
	Median/MAD sampling frequency	86,400 s
Similarity search	LSH: number of hash functions per hash table r	6
	LSH: number of hash tables b	100
	Initial pair threshold: number v (fraction) of tables, pair in same bucket	2 (2/100 = 0.02)
	Similarity search: near-repeat exclusion parameter	10 (fingerprint) samples (12 s)
	Number of partitions for LSH database†	10

LSH, locality-sensitive hashing; MAD, median absolute deviation.

*Values contained in the FAST software input file, with an equivalent calculated value in parentheses that may be more meaningful, whereas brackets indicate the specific type of "samples". LSH, locality-sensitive hashing; MAD, median absolute deviation.

†For the three single-component stations (NC.PPB, NC.PABB, PG.SH), the number of partitions was set to be 72, 73, 134, respectively (instead of 10), so that each partition would be about a month long. At these stations, we applied an occurrence filter to exclude frequently repeating nonearthquake signals (Rong *et al.*, 2018), in which a fingerprint that matched over 1% of the total number of fingerprints during a month-long partition was excluded from the similarity search.

Table A3

Combining FAST Similarity Search Outputs by Adding the Similarity Matrix from All Three Components at a Given Station Dramatically Reduces the Size of the Detection Results (Rong *et al.*, 2018)

Network	Station	Similar Pairs Size (GB) Three Components, $v = 2$	Similar Pairs Size (GB) Combined Station, $\tau_0 = 6$
PG	MLD	1202	60
PG	LSD	1556	0.13
PG	LMD	545	3.4
PG	SHD	3244	146
PG	EFD	3026	19
PG	VPD	556	3.7
PG	DPD	1193	4.0
PG	DCD	445	0.6

For each component, $v = 2$ is the initial-pair threshold (Table A2); column 3 shows the total size (GB) of the output text files from similarity search on all three components at the station. After adding the similarity matrix from the three components, we set a higher station-pair threshold $\tau_0 = (v = 2) \times (3 \text{ components}) = 6$; column 4 shows the size (GB) of the combined similarity output for the station.

Table A4

Input Parameters for Pair-Wise Association and Detection over the 11-Station Network (Bergen and Beroza, 2018a)

Network-Association Algorithm Section	Parameter Description	Value
Event-pair extraction	Time gap (along diagonal), g_L	10 samples (12 s)
	Time gap (adjacent diagonal), g_W	3 samples (3.6 s)
	Adjacent diagonal merge iterations, p	2
Event-pair pruning	Number of votes (station-pair threshold), τ_0	6
	Minimum fingerprint-pairs, $\lvert C \rvert_{\min}$	3
	Minimum total similarity, $v_{\min}^{(C)}$	18
	Maximum bounding box width	8 samples (9.6 s)
Pseudoassociation	Minimum number of stations for detection	2 (out of 11)
	Arrival-time constraint: maximum time gap, g_N	15 samples (18 s)

All samples in this table refer to fingerprint samples, in which the time lag between adjacent fingerprints, also called the fingerprint sampling period, is 1.2 s (Table A2).

Table A5

1D *P*-Wave Velocity Model (McLaren and Savage, 2001), with $V_P/V_S = 1.66$ (Hardebeck, 2010), Used to Locate New Detected Earthquakes

Depth (km)	*P* Velocity (km/s)
0.0	4.0
4.0	5.7
10.0	6.0
14.0	6.2
24.0	8.0

Table A6

New 1D Velocity Model, Output from Second VELEST Run with Station Corrections, for 715 Earthquakes near the Diablo Canyon Nuclear Power Plant (DCPP)

Depth (km)	*P* Velocity (km/s)	V_P/V_S
0.0	4.01	1.69
4.0	5.58	1.68
10.0	6.00	1.65
14.0	6.20	1.66
24.0	8.00	1.66

Clara E. Yoon
William L. Ellsworth
Gregory C. Beroza
Department of Geophysics
Stanford University
397 Panama Mall, Mitchell Earth Sciences Building
Stanford, California 94305 U.S.A.
claraeyoon@gmail.com

Karianne J. Bergen
Institute for Computational and Mathematical Engineering
Stanford University
475 Via Ortega, Huang Engineering Building
Stanford, California 94305 U.S.A.

Kexin Rong
Hashem Elezabi
Peter Bailis
Philip Levis
Department of Computer Science
Stanford University
353 Serra Mall, Gates Computer Science Building
Stanford, California 94305 U.S.A.

Manuscript received 10 January 2019;
Published Online 25 June 2019