AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Solid Earth

RESEARCH ARTICLE
10.1002/2016JB012981

Key Points:
- Agglomerative clustering of spectra and time signals rapidly finds repeating seismicity
- Successfully detects swarms in volcanic, subduction, and induced seismicity settings
- Detector replicates or improves upon existing catalogs in much less processing time

Supporting Information:
- Supporting Information S1

Correspondence to:
R. J. Skoumal,
skoumarj@miamioh.edu

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

Received 9 MAR 2016
Accepted 25 JUL 2016
Accepted article online 29 JUL 2016
Published online 15 AUG 2016

An efficient repeating signal detector to investigate earthquake swarms

Robert J. Skoumal[1], Michael R. Brudzinski[1], and Brian S. Currie[1]

[1]Department of Geology and Environmental Earth Science, Miami University, Oxford, Ohio, USA

Abstract Repetitive earthquake swarms have been recognized as key signatures in fluid injection induced seismicity, precursors to volcanic eruptions, and slow slip events preceding megathrust earthquakes. We investigate earthquake swarms by developing a Repeating Signal Detector (RSD), a computationally efficient algorithm utilizing agglomerative clustering to identify similar waveforms buried in years of seismic recordings using a single seismometer. Instead of relying on existing earthquake catalogs of larger earthquakes, RSD identifies characteristic repetitive waveforms by rapidly identifying signals of interest above a low signal-to-noise ratio and then grouping based on spectral and time domain characteristics, resulting in dramatically shorter processing time than more exhaustive autocorrelation approaches. We investigate seismicity in four regions using RSD: (1) volcanic seismicity at Mammoth Mountain, California, (2) subduction-related seismicity in Oaxaca, Mexico, (3) induced seismicity in Central Alberta, Canada, and (4) induced seismicity in Harrison County, Ohio. In each case, RSD detects a similar or larger number of earthquakes than existing catalogs created using more time intensive methods. In Harrison County, RSD identifies 18 seismic sequences that correlate temporally and spatially to separate hydraulic fracturing operations, 15 of which were previously unreported. RSD utilizes a single seismometer for earthquake detection which enables seismicity to be quickly identified in poorly instrumented regions at the expense of relying on another method to locate the new detections. Due to the smaller computation overhead and success at distances up to ~50 km, RSD is well suited for real-time detection of low-magnitude earthquake swarms with permanent regional networks.

1. Introduction

1.1. Repetitive Seismic Events

Relationships between earthquakes are observed by the clustering of seismic events in space and time. This clustering commonly occurs as main shock-aftershock sequences, which are generally interpreted to contain the initial rupture of a fault (the main shock) and a decaying cascade of smaller ruptures on or very near to the initial rupture plane (aftershocks) [*Lay and Wallace*, 1995]. Clustering of earthquakes in space and time can also occur as earthquake swarms, which are empirically defined as an increase in seismicity rate above the background rate without a clear triggering main shock earthquake [*Mogi*, 1963]. Earthquake swarms are often associated with volcanic regions and are studied because of their relationship to eruptions [*McNutt*, 1996]. Earthquake swarms have also been correlated with subduction zone slow slip events [*Hirose et al.*, 2014], including a case that led into the 2011 Tohoku earthquake [*Kato et al.*, 2012]. Earthquake swarms are also well associated with many induced ("human-influenced") earthquake sequences [e.g., *Horton*, 2012]. Understanding the mechanisms that lead to earthquake swarms and the rapid detection of these events is key factor in reducing the hazard posed by these events.

Recent advances in seismic waveform template matching techniques have improved the detection of similar earthquakes, such as those that occur in swarms [e.g., *Kato et al.*, 2012; *Shelly et al.*, 2015]. Two earthquakes from a given swarm are likely to have a similar source mechanism and be located in close proximity such that the radiated energy from both events is expected to produce waveforms that have the same polarity, similar seismic wave arrival times, and interact with the same rock layers while follow a common pathway to a seismic station. The resulting ground motions for both events would be similar, enabling detection through correlation-based processing.

1.2. Induced Seismicity Overview

As enhanced oil and gas operations have proliferated over the past decade in North America, the potential to generate induced seismicity has also grown [*National Research Council* (*NRC*), 2013; *Ellsworth*, 2013]. New

industrial technologies have broadened the geographic area over which hydrocarbons can be economically recovered, meaning that the areas at risk for associated induced seismicity are broadening as well. In response, regulations have been implemented in some states and countries in an attempt to limit the occurrence of induced earthquakes [*Ground Water Protection Council and Interstate Oil and Gas Compact Commission*, 2015]. In Ohio, for example, if a hydraulic fracturing operation (HF) is associated with a magnitude (M) >1 earthquake, operations at the well may be terminated [*Ohio Department of Natural Resources* (*ODNR*), 2014]. This magnitude threshold is below the detection capabilities of traditional earthquake detection methods that utilize regional seismic data. To distinguish whether a well is associated with low-magnitude seismicity, local seismic deployments and/or new earthquake detection algorithms are required.

A monitoring strategy for induced seismicity is needed that can rapidly analyze this growing geographic region that may experience induced seismic events. Since 2010, induced seismicity associated with enhanced recovery has been observed in Arkansas [*Horton*, 2012], Colorado [*Rubinstein et al*., 2014], Kansas [*Buchanan*, 2015], New Mexico [*Rubinstein et al*., 2014], Ohio [*Skoumal et al*., 2015c], Oklahoma [*Walsh and Zoback*, 2015], Texas [*Frohlich et al*., 2011], as well as British Columbia [*British Columbia Oil and Gas Commission (BCOGC)*, 2012] and Alberta [*Schultz et al*., 2015]. Most of this seismicity is attributed to deep disposal of large volumes of leftover water following oil and gas well stimulation or recovery [*McGarr et al*., 2015; *Walsh and Zoback*, 2015; *Weingarten et al*., 2015]. Although less common, there are also a number of cases where HF itself has been well correlated to earthquakes [*Nicholson and Wesson*, 1990; *BCOGC*, 2012, *Holland*, 2013; *Clarke et al*., 2014; *Schultz et al*., 2015; *Skoumal et al*., 2015a]. In both types, the induced seismicity tends to occur as seismic swarms that can be effectively detected with advanced correlation algorithms [*Frohlich et al*., 2011; *Benz et al*., 2015; *Schultz et al*., 2015; *Holtkamp et al*., 2015; *Huang and Beroza*, 2015; *Skoumal et al*., 2015c]. Multichannel cross correlation using waveforms from a previously cataloged seismic event (template matching) is well suited for the identification of similar, small-magnitude events [e.g., *Gibbons and Ringdal*, 2006; *Shelly et al*., 2007]. Due to the swarm-like nature that is commonly associated with induced sequences, template matching is ideally suited for lowering the magnitude detection threshold, increasing the confidence in classifying a sequence as natural or induced [e.g., *Skoumal et al*., 2015c]. The question remains whether other induced sequences exist below the catalog detection threshold.

The swarm-like nature of induced seismicity creates an opportunity to develop a new algorithm to identify these repeating signals at lower magnitude than traditional earthquake detection methods. We first develop our algorithm by applying it to previously studied natural earthquake swarms at Mammoth Mountain, California, and Oaxaca, Mexico. We then applied our algorithm by investigating seismicity associated with HF, as the shorter time of fluid injection (typically weeks) relative to wastewater disposal (typically years) helps to discern whether seismicity is related to HF. Our study focuses on Ohio where three sequences of induced seismicity have been attributed to wastewater disposal [*Kim*, 2013; *Skoumal et al*., 2015b, 2015c] and three associated with HF [*Friberg et al*., 2014; *Skoumal et al*., 2015a, 2015c]. One advantage of this study region is that each case was spatially or temporally separated from other industrial operations, which allows us to simplify the investigation by examining a single area with HF seismicity. Since the previously identified sequences associated with HF were characterized via catalog-based template matching, each had to have at least one earthquake with $M>2$, but none were larger than M 3. These seismic sequences were all shown to occur at or the below the crystalline basement-sedimentary cover contact with hypocentral distributions and focal mechanism orientations consistent with reactivation of preexisting faults [*Kim*, 2013; *Friberg et al*., 2014; *Skoumal et al*., 2015b, 2015c]. Operationally induced microseismicity typically produces $M<0$ in the stimulated formation [*Warpinski*, 2013], suggesting that sequences with seismicity in the M 0–2 range should be the target for improving detection of induced seismicity on preexisting faults.

Another region with prevalent HF and correlated seismicity that was investigated in this study is the Western Canada Sedimentary Basin Duvernay play in Alberta [*Eaton and Mahani*, 2015; *Schultz et al*., 2015]. It was previously thought that HF posed little risk of inducing felt seismicity [*NRC*, 2013], but the Kaybob area near Fox Creek has experienced several earthquakes with $M>4$ in 2015 near HF operations. Induced subsequences found in this region had remarkable interevent similarity with average normalized correlation coefficients $>$0.8 that aided detection via template matching [*Schultz et al*., 2015]. The nearest publicly available seismometer is ~30–50 km away, providing an opportunity to test whether a new algorithm can successfully detect small, repeating seismic events at regional distances.

1.3. Previous Repeating Seismic Event Detection Algorithms

Template matching has been demonstrated as a valuable tool to characterize low-magnitude induced seismic events, but it is inherently reliant on another algorithm to identify the template signals. A short-term/long-term average amplitude (STA/LTA) earthquake detection algorithm is commonly used to detect seismic wave arrivals, and then a data acquisition and processing system (e.g., Earthworm and Antelope) will determine if enough stations recorded arrivals at appropriate times to be from a single location in the Earth. Traditionally, the application of these algorithms to regional networks results in a magnitude detection threshold of M ~2.5. Consequently, catalog-based template matching relies on the detection of a $M > 2$ event before smaller-magnitude events could be identified. If an induced sequence has not yet produced a $M > 2$ earthquake (or if it is not cataloged), the sequence would go undetected unless a local network is present. Induced seismicity often contains swarms of smaller-magnitude events that precede larger events [e.g., *Friberg et al*., 2014; *Huang and Beroza*, 2015; *Skoumal et al*., 2015c], increasing the importance of rapidly detecting $M < 2$ seismicity. Ideally, a real-time detection algorithm would not rely on the identification of a $M > 2$ event that may occur later in the sequence.

Autocorrelation has the potential to identify repeating signals below the magnitude detection thresholds from multistation STA/LTA detection, and it is also not reliant on the prior identification of a seismic event. However, autocorrelation is computationally intensive when large time windows are considered, with N $(N-1)/2$ correlations required. A previously proposed solution to this computational challenge is to implement a time step between windows used in autocorrelation [*Brown et al*., 2008]. While this will reduce the number of correlations to N $(N-1)/(2 \times \text{step})$, it fails to address the quadratically increasing nature of the computations. Additionally, a significant step size would fail to detect repeating signals that have windows that are not aligned and is therefore not be an ideal detection tool. Due to the computationally intensive nature of autocorrelation, it is currently not practical to implement the algorithm on a regional scale for many months or years, which are common durations for induced seismicity.

Another proposed solution for the detection of repeating seismic events is the Fingerprint and Similarity Thresholding (FAST) algorithm [*Yoon et al*., 2015a]. FAST utilizes locality-sensitive hashing, which groups similar waveforms together in hash buckets based on the discriminative features of the signals. Correlations are only performed on similar events, reducing the number of required calculations. In a week-long trial in Central California, the algorithm successfully identified 21 of the 24 cataloged earthquakes in addition to detecting 68 uncataloged earthquakes while also identifying 12 false positives [*Yoon et al*., 2015a]. With computations that scale near linearly instead of quadratically, FAST has been demonstrated as having significant runtime advantages over autocorrelation, with 3 months of single-channel data processed in 16 days [*Yoon et al*., 2015b]. While FAST is significantly more efficient than autocorrelation, its utilization for real-time earthquake detection on a regional scale may still be limited by its computational overhead.

Another technique that addresses the challenge of identifying a template event targeted for low-frequency earthquakes was proposed by *Frank and Shaprio* [2014]. Using a three-dimensional grid of potential sources and a velocity model to estimate phase moveouts, recorded seismic energy peaks are summed to create templates in a near-automated manner similar to the source-scanning algorithm [*Kao and Shan*, 2004]. For a 2.5 year data set using 10 stations spaced ~10 km apart, 8 days of computation time were required to identify template events and perform template matching when used in parallel with 8 geopotential units [*Frank et al*., 2014]. The required number of seismometers, creation of accurate velocity models, and computation time may potentially impede this algorithm from being rapidly utilized for real-time monitoring over large regions with significant seismometer spacings where induced seismicity has been observed, like the Midcontinent United States, Alberta, and British Columbia.

1.4. Waveform Family Sorting

Clustering data is a method of organizing objects with similar properties into groups [e.g., *Kaufman and Rousseeuw*, 1990]. There are many different methods available to cluster data, and the creation of these groups (or "families") has been utilized to explore data sets in a wide variety of disciplines. For example, previous work in the seismology discipline has utilized the cross-correlation coefficients produced by correlating one or more earthquakes against a catalog in conjunction with hierarchical clustering techniques to distinguish families of waveforms. This approach has been applied to a variety of applications including studying

aftershock sequences [*Slinkard et al*., 2013], volcanic low-frequency earthquakes [*Neuberg et al*., 2006], distinguishing nuclear testing from mining blasts [*Carr et al*., 1999], and refining phase arrivals in large seismic catalogs [*Rowe et al*., 2002]. Sorting families using correlation coefficients have followed different approaches, some studies following a template matching approach [*Stephens and Chouet*, 2001] while others correlate each cataloged earthquake with the rest of the catalog, resembling an autocorrelation approach [*Green and Neuberg*, 2006]. This catalog autocorrelation approach has been frequently used in conjunction with a single-station STA/LTA detector to detect low-frequency volcanic earthquakes [*Petersen*, 2007].

In this study, we introduce a new family sorting approach. Rather than using cross correlation as a linkage metric, we identify families by determining the Euclidian distances between signals of interest. While prior studies have focused on sorting waveforms according to characteristics in the time domain, we apply family sorting to both the frequency and time domains. Additionally, rather than sorting cataloged events into families like many of the prior studies, we use the hierarchal clustering as a means of identifying characteristic repeating seismic signals that had previously gone undetected. Once these repeating signals have been detected, families are stacked and used in a template matching routine in order to detect previously unidentified earthquake swarms.

2. Methods

2.1. The Repeating Signal Detector Algorithm

In an effort to make the initial identification of small-magnitude, repetitive events more efficient and effective, we propose a new algorithm, referred to as a Repeating Signal Detector (RSD) (Figure 1). The initial step of RSD is identifying signals of interest (SoI) using continuous seismic data for a single station in a manner similar to the traditional STA/LTA detection algorithms. We employ a relatively low signal-to-noise threshold where the average absolute amplitude in the previous 1 s is at least 5 times larger than the average amplitude in the previous 30 s. When the threshold is exceeded using 5–15 Hz band-passed data, we reserve a 30 s unfiltered window of the SoI centered on the peak amplitude during the 1 s window. At this stage, many of these SoI are expected to consist of nonearthquake signals. To distinguish the repeating earthquakes, these SoI are grouped into families. For each SoI, the fast Fourier transforms of each component ($[E_{\mathrm{freq}}]$, $[N_{\mathrm{freq}}]$, and $[Z_{\mathrm{freq}}]$) considering frequencies above 5 Hz each consisting of 95 samples are concatenated ($[E_{\mathrm{freq}}\ N_{\mathrm{freq}}\ Z_{\mathrm{freq}}]$) to represent each SoI as a single point in n-dimensional space. In the implementation presented here, $n = 285$, the number of points in the concatenated array which is dependent on the frequency sampling interval and the frequency range being considered. We then determine the Euclidean distances between each of these 285-point arrays. The Euclidean distance formula is given as

$$\overline{ab} = \sqrt{\sum_i (a_i - b_i)^2}$$

where a and b are arrays that represent two SoI and i is the number of samples in the concatenated array. For this step, the arrays a and b are two of the 285-point concatenated spectra arrays described previously. Agglomerative clusters are then created using an average-linkage clustering approach by calculating the unweighted Euclidean distances between these arrays and then considering the smallest average distances between clusters. For this "bottom-up" form of hierarchical clustering, each SoI is initially contained in its own family. With each iteration, families with the smallest average Euclidean distances between the members are grouped together, and the family is represented by the average of all members. This method of average-linkage clustering can be described by

$$D(A,B) = \frac{1}{(n_A)(n_B)} \sum_{a \in A} \sum_{b \in B} \overline{ab}$$

where A and B are two families, n_A and n_B are the number of members in each of these families, and a and b are the arrays that belong to A and B. Distances between members are calculated, and families are combined until the total number of generated families reaches the total number of SoI divided by 5, an experimentally derived coefficient that provided the optimal balance between grouping visually similar waveforms and a manageable numbers of output families.

While this initial family clustering step is effective at sorting like signals into families, variations in waveforms may still be present (Figure 2). The small temporal offsets between different family members would create

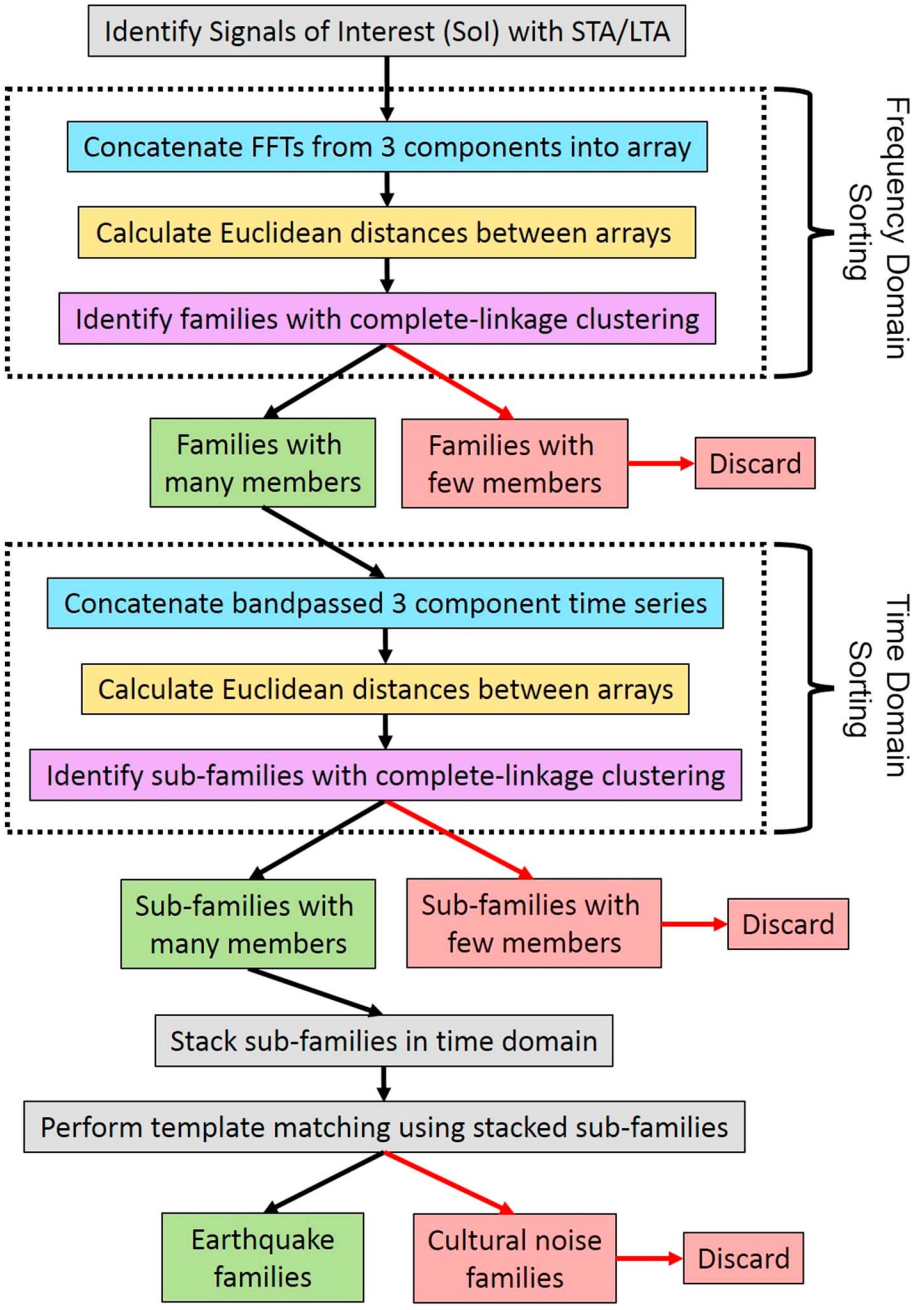

Figure 1. Algorithm flowchart describing the Repeating Signal Detector (RSD) technique. Further description and illustrations of the method can be found in the supporting information.

destructive interference if stacked at this stage. To account for these temporal arrival variations, the signals are band-pass filtered between 5 and 15 Hz and the largest amplitude signal in each family is identified and then cross correlated against the other members in the family to adjust for any temporal offsets. For each family, the agglomerative clustering is then performed on all members of a given family using the same average-linkage clustering process, combining similar members until the number of subfamilies reaches the total number of family members divided by 5. However, this time the clustering operates on the time domain signals by taking the medial 10 s of the adjusted band-passed waveforms from each of the three components ($[E_{time}]$, $[N_{time}]$, and $[Z_{time}]$) and concatenating them ($[E_{time}\ N_{time}\ Z_{time}]$) so that each SoI can once again be represented as a single point in n-dimensional space. This time, $n = 1200$, the number of points in the concatenated

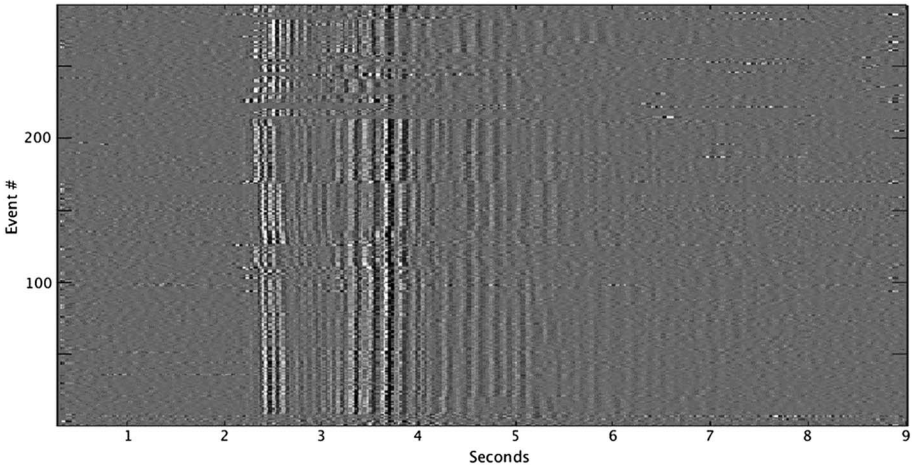

Figure 2. Example waveforms for a family with 292 members identified by the initial frequency domain sorting step of RSD. Gray scale shows normalized waveform amplitudes from black to white. Each horizontal line in the grid image shows the *P* and *S* wave amplitudes over time for 1 of the 292 SoI in this family. Waveforms were recorded at TA O53A BHZ in Harrison County. Note that while the frequency contents for these signals were similar such that they were clustered in the same family, differential phase arrival times (and the absence of phases for some members) are present, highlighting the importance of the secondary sorting step in the time domain.

arrays which is based on the 40 Hz sampling rate and the three 10 s windows. The agglomerative clustering calculates the unweighted Euclidean distances (using equation $\overline{ab}$) between these concatenated arrays of time domain signals and then identifies the smallest average distances to progressively form clusters (using equation $D(A, B)$). This time domain clustering step results in aligned subfamilies that can be stacked, improving the signal-to-noise ratio and providing characteristic repetitive signals that can then be used as templates in a cross-correlation routine. In an effort to remove families identified by RSD that consisted of harmonic noise, we removed stacks with low-frequency ($<$3 Hz) normalized amplitudes above 0.3, an experimentally determined threshold that sufficiently removed the undesired signals in our study areas.

To remove nonrepetitive signals, we excluded identified families (following the frequency domain sorting step) and subfamilies (following the time domain sorting step) from further analysis if they had fewer than five members in cases where the station was farther ($>$30 km) from the source (Oaxaca and Central Alberta) or 20 members in cases where the station was near ($<$10 km) the source (Mammoth Mountain and Harrison County). These differences highlight how the thresholds are likely both station and region specific and should be expected to vary depending on the target. The objective of this study is to describe the promise of RSD and encourage the community to modify the algorithm to suit their individual needs. For further description of the RSD algorithm used in this study, see the supporting information.

Computational overhead for the identification of SoI depends on the type of data compression and sample rate. Using a single core of an Intel Xeon Processor E5520 (2.26 Ghz), SoI from a day of 40 Hz MiniSEED data can be identified in ~1–2 s, which largely consists of reading data from the disk. Computation time for the family clustering process is dependent on the number of SoI identified, but it is capable of producing stacked subfamilies from 10,000+ SoI in ~1–2 min. Subsequent template matching using the stacked waveform requires ~15 min per template through each year of data. These relatively unintensive computational requirements allow for a year of continuous seismic data from a single station to be processed well within 1 h in serial, but the calculations are also readily parallelizable. As a result, the computational overhead of RSD is a fraction of the amount required by autocorrelation or FAST, RSD does not rely on a velocity model or a relatively dense regional network required by an automated beamforming algorithm, and RSD has the potential to be employed in real-time processing of seismic data on a regional scale. Next, we demonstrate how successfully RSD can identify repetitive, small-magnitude earthquakes below the detection threshold of traditional STA/LTA.

2.2. Investigating Signals Identified by the Repeating Signal Detector

For the remaining subfamily stacks, we use the medial 10 s as a template in a three-component normalized cross-correlation routine. We perform template matching over the same time window that RSD was run

(ranging between a few weeks and ~3 years, depending on the study region) in order to identify the temporal patterns of the repetitive signals. To ensure that duplicate event detections by multiple templates were excluded, detected events that were analyzed were a minimum of 30 s apart, consistent with previous studies [e.g., *Skoumal et al*., 2014]. We use a detection threshold of 15 times the daily median absolute deviation of the stacked correlation coefficients, a previously demonstrated conservative detection threshold [e.g., *Skoumal et al*., 2014]. Using a relatively high threshold ensures that a single station can discern events from different source regions. Using hundreds of stations in the Northern California Seismic Network, *Schaff and Waldhauser* [2005] analyzed the cross-correlation coefficients (CC) of earthquakes at each station and found that the CC values decrease precipitously as interevent distances exceed 1 km, due to the breakdown in waveform similarity with increasing differences in path.

To further investigate the results of template matching using RSD-derived templates, we performed manual inspection of the matched waveforms recorded by the station used for RSD in addition to other seismometers in the region to look for appropriate seismic wave characteristics. In particular, we looked for evidence of different *P* and *S* wave signatures on the vertical and horizontal components that would indicate a noncultural source that can be located. The remaining signals were bursts of energy recorded by the station used for RSD, but with no coherent signals on other nearby stations. Moreover, there was typically no time separation between noise signals on the vertical and horizontal channels at the RSD station and hence no indication of separate *P* and *S* wave arrivals. We interpret these signals to be cultural noise near the seismometer, but the lack of seismic wave information prevents further investigation. Due to the vast number of matches to investigate in the larger sequences we identified, we utilized a grid plot of the events to help discriminate seismicity from noise (see examples in the supporting information).

We approximated local magnitudes through a simplified Richter scale approach:

$$M_L = \log_{10}[A / A_0]$$

The median scale factor (A_0) was calculated using the peak-to-peak filtered *S* waveform amplitude and utilized existing catalog magnitudes of events in the corresponding catalog whenever present. For each matched event, we calculated a magnitude from the scale factor and *S* waveform amplitude at each station and component and took the median value as our final magnitude. The largest-magnitude events in each of the temporal clusters were then located using *elocate* [*Herrmann*, 2013] on manually picked *P* and *S* times using seismic data that were available at the time of the events. In Harrison County, we used a locally derived 1-D velocity model [*Friberg et al*., 2014] adjusted to a basement depth of 3.4 km. In the Kaybob region, we derived a velocity model from sonic logs in the area (see supporting information). For the Mammoth Mountain and Oaxaca, Mexico, regions, cataloged earthquakes were well characterized and located by previous studies [*Shelly et al*., 2015; *Fasola et al*., 2016]. For these regions, we utilize the cataloged locations provided by the earlier work.

2.3. Data Sets

We utilized continuous seismic data from four regions in North America: (1) Mammoth Mountain, California, (2) Oaxaca, Mexico, (3) Central Alberta, Canada, and (4) Harrison County, Ohio. For the Mammoth Mountain study area, data were obtained from station NN OMMB (40 Hz) during 2–18 February 2014, and we compared our results to an earthquake catalog generated by *Shelly et al*. [2015] created during the same time window. In Oaxaca, Mexico, we used data from station MU OXTT (40 Hz) during 10–30 July 2006 and compared our results to the earthquakes identified by *Fasola et al*. [2016]. For Central Alberta, we applied RSD to station BRLDA of the Raven network from 15 August 2014 to 27 November 2015 after downsampling the data from 100 Hz to 40 Hz. Catalog earthquakes in Alberta were obtained from Natural Resources Canada (NRCAN). In Harrison County, we utilized station O53A (40 Hz) in Harrison County, Ohio, originally part of the EarthScope Transportable Array, from 14 December 2012 to 27 November 2015. To aid in characterization of the detected signals, we utilized a Freedom of Information Act request to Ohio Department of Natural Resources (ODNR) to obtain data from two three-component short-period stations, and we also deployed a broadband seismometer on 13 September 2015. Catalog earthquakes in Ohio were obtained from ODNR and Advanced National Seismic System (ANSS). Times and locations of HF for the Harrison County and Central Alberta regions were obtained from ODNR, FracFocus, and FracFocus.ca.

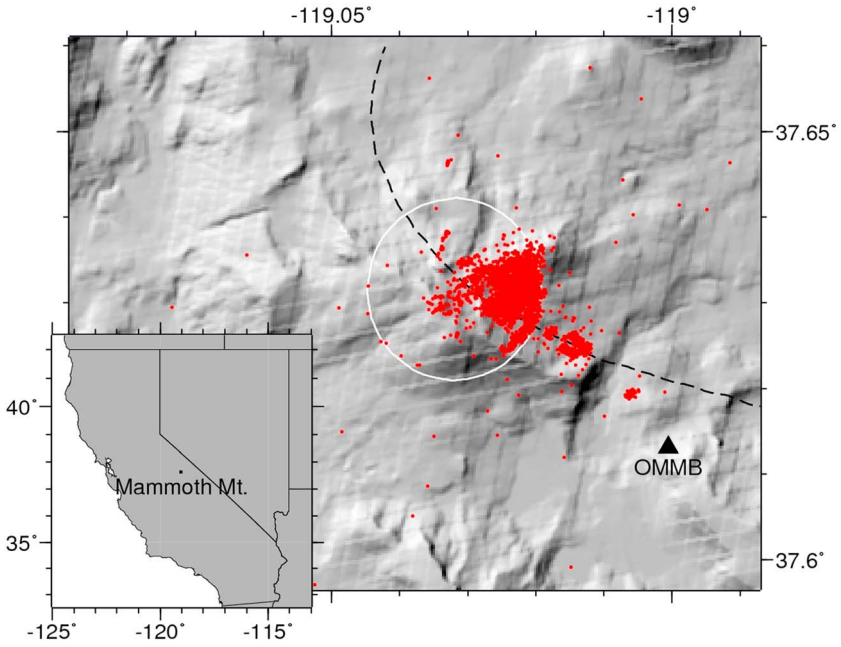

Figure 3. Map showing shaded topographic relief of the Mammoth Mountain study region at the southwestern edge of Long Valley Caldera (dashed line). White circle shows the deep ring-shaped fault structure centered on the summit of Mammoth Mountain [*Prejean et al.*, 2003], which probably reflects an inherited structure from the time the mountain was actively erupting, ~90–50 ka. The February 2014 swarm described by *Shelly et al.* [2015] (red dots) is compared with the results of applying RSD to station NN OMMB (black triangle). Small black rectangle in the map inset shows the location and size of the local area map.

3. Results

We first describe our comparisons to three documented swarms based on earthquake catalogs in volcanic, subduction, and induced seismicity cases. We then apply RSD to an induced seismicity case of interest where there is limited documentation of prior swarms but indications that it could be more pervasive.

3.1. Mammoth Mountain, California

Located in the Long Valley Caldera, Mammoth Mountain has been host to numerous earthquake swarms and outgassing events within the past few decades [*Hill et al.*, 1990; *Sorey et al.*, 1998; *Lin*, 2013; *Lewicki et al.*, 2014]. A previous large-scale template matching effort at Mammoth Mountain that utilized 1545 cataloged earthquakes identified 6179 events during 2–18 February 2014 from a swarm associated with fluid migration (Figure 3) [*Shelly et al.*, 2015]. The detection of this earthquake swarm was aided by the presence of a dense seismic network, including six seismometers within a few kilometers horizontally from the swarm [*Shelly et al.*, 2015]. Using three-component waveforms from a single broadband station (NN OMMB) between 2 and 18 February 2014, RSD identified 6026 SoI that were grouped into 1206 families following the frequency domain grouping step. When these families were used in the time domain grouping, 18 subfamilies (~1.5% of the families) were found that contained more than 20 members. Using stacks of these 18 subfamilies as templates in a template matching routing, the temporal earthquake patterns identified by *Shelly et al.* [2015] were duplicated. Nearly identical numbers of events were detected (6197 versus the 6121 from RSD) between the methods (Figure 4). Approximately 87% (5397) of the events from *Shelly et al.* [2015] were detected by RSD. Considering the reliance on an existing earthquake catalog, the use of a network of seismometers, and the large number of computations required by the large-scale template matching study, the relative ease by which RSD produced similar results with a single station and 2 orders of magnitude less templates to process shows the promise for rapid earthquake identification.

3.2. Oaxaca, Mexico

Earthquake swarms, episodic slow slip, and tectonic tremor have previously been observed in Oaxaca, Mexico, associated with the subduction of the Cocos Plate [e.g., *Brudzinski et al.*, 2007, 2010; *Fasola et al.*,

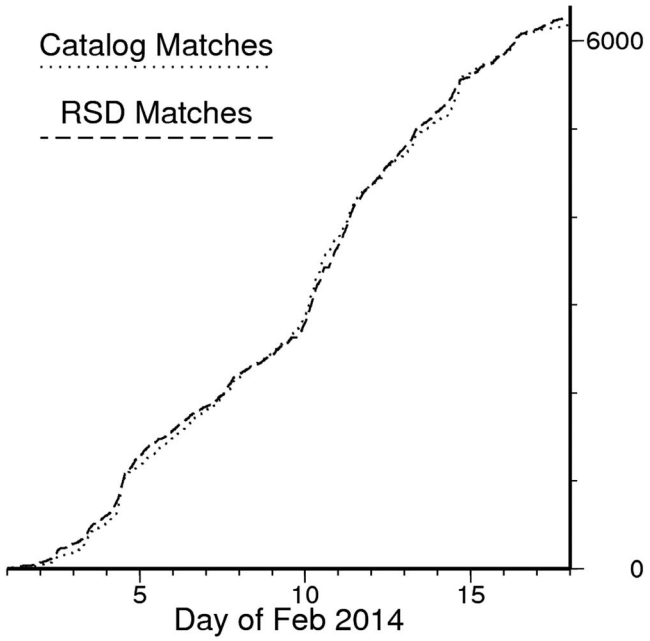

Figure 4. Cumulative number of seismic events detected using template matching in the Mammoth Mountain study region. The catalog of detections identified by *Shelly et al*. [2015] using 1545 earthquakes cataloged by the Northern California Seismic Network as templates is shown as a dotted line. The total number of unique events using each of the 18 stacked subfamilies identified by RSD as a template (dashed line) closely matches the temporal pattern of detections from *Shelly et al*. [2015]. For individual subfamily performance, see Figure S14.

2016]. In mid-July 2006, a swarm with $M_w \leq$~3.5 occurred ~20 km east of the city of Pinotepa Nacional [*Fasola et al*., 2016]. While this swarm has not been studied in detail, previous work on subduction zone swarms indicates that they are potentially driven by slow slip and/or fluid flux [*Holtkamp and Brudzinski*, 2011], both of which are possible in this region [*Brudzinski et al*., 2007, 2010; *Song et al*., 2009].

Using a single station (MU OXTT) of data between 10 and 23 July 2006 that was located ~30 km southeast from the July 2006 swarm (Figure 5), RSD identified 1586 SoI that were grouped into 318 families following the frequency domain grouping step. Using these 318 families in the time domain grouping step, 33 subfamilies (~10.4% of the families) were found that contained more than five members. Using stacks of these 33 subfamilies in a template matching routine, nine stacks each matched with $>$15 real earthquakes (average of ~89 detections per stack) during expected seismically active periods, nine stacks matched fairly continuously with noise, and 15 stacks matched with $<$15 events and were discarded (Figure 6). A similar temporal seismicity pattern was observed to the catalog created by *Fasola et al*. [2016], with two primary bursts of activity on 12 and 15–16 July. Moreover, RSD identified a tenfold increase in the amount of detected events. Of the 43 earthquakes included in the catalog of *Fasola et al*. [2016] that constituted the swarm, RSD detected 35 events (~81%).

3.3. Central Alberta, Canada

RSD was also applied to the Kaybob area in Alberta, Canada, where some of the largest earthquakes associated with HF have occurred, including a M_w 3.8 in August 2014 and a M_L 4.4 in January 2015 [*Eaton and Mahani*, 2015; *Schultz et al*., 2015]. Unconventional oil and gas operations in this region are targeting the Duvernay Formation, with many of the wells focused in the zone of hydrocarbon thermal maturity where natural gas condensate is expected to be most prevalent (Figure 7). Station RV BRLDA is located in Central Alberta, ~250 km northwest of Edmonton and ~40 km south of the 2015 M_L 4.4 event (Figure 7) [*Schultz et al*., 2015]. While there has been no record of a major, destructive earthquake in Alberta, moderate ($M < 6$) historical seismicity is common, occurring principally along the Rocky Mountain foreland belt [e.g., *Stern et al*., 2013]. Between 1985 and 2012, there were three earthquakes with $M_L > 3$ in the NRCAN catalog in our study region (118–116.5°W, 53.5–55°N), compared with the 25 cataloged earthquakes with $M_L > 3$ in 2013–2015.

Using continuous seismic data recorded by station RV BRLDA between 15 August 2014 and 27 November 2015, a total of 9984 SoI were grouped into 1997 families following the frequency domain grouping, with 120 families (~6.0% of the families) containing more than five members. Following the time domain grouping, 88 subfamilies were identified that contained more than five members. Single-station normalized cross correlation revealed 49 productive subfamilies that each matched with more than 100 events (Figure 8). Of these productive subfamilies, 42 were found to have matched consistently with real seismic events, while 7 matched with cultural noise fairly continuously throughout the sampling period. Template matching using the 42 productive subfamilies identified 4411 unique events. Following manual inspection, ~93% of these

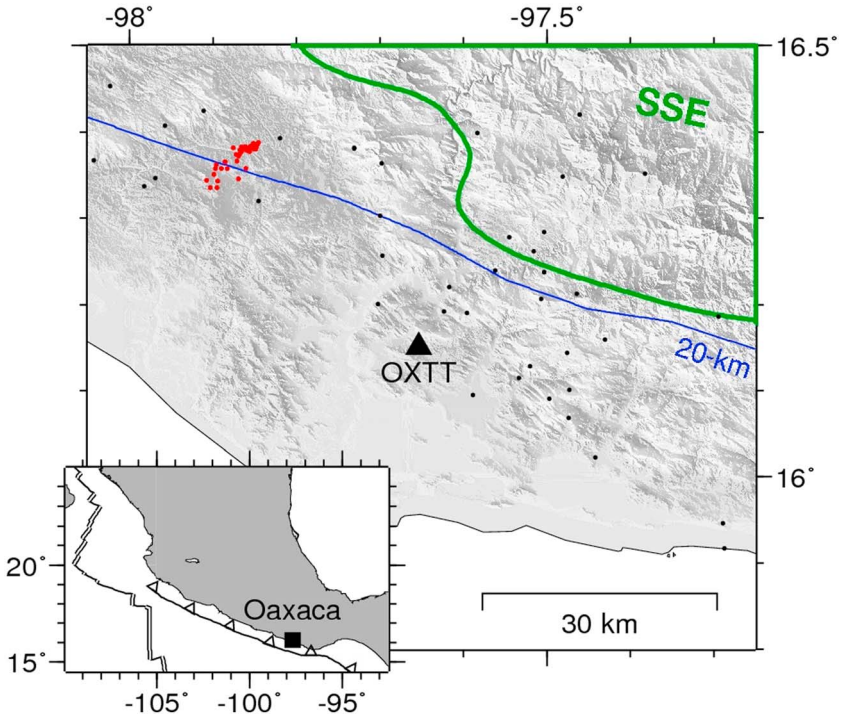

Figure 5. Map showing shaded topographic relief of the Oaxaca region of Mexico. Seismic station MU OXTT (black triangle) was used in the RSD analysis. Black dots show earthquakes in July 2006 identified by *Fasola et al.* [2016] with a swarm highlighted in red. Blue contour indicates depth of the subducting plate interface in the vicinity of the swarm [*Fasola et al.*, 2016]. Green curve outlines the 15 mm slip contour for the best determined slow slip event (SSE) in this region that occurred in the months leading up to the 2012 M_w 7.4 Ometepec earthquake [*Graham et al.*, 2014] with an epicenter ~30 km west of our study region. Map inset shows the nearby plate boundaries and outlines the extent of the local area map (black rectangle).

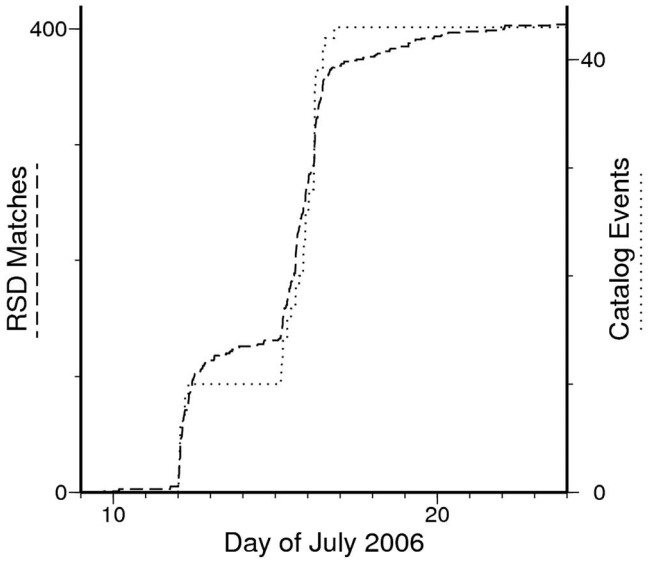

Figure 6. Cumulative number of seismic events detected using template matching in the Oaxaca, Mexico, study region, showing earthquakes identified by RSD (dashed line) relative to the earthquakes cataloged by *Fasola et al.* [2016] in the July 2006 swarm (dotted line). For example of earthquake waveforms and individual subfamily performance, see Figures S11 and S15.

events were confirmed as real seismic events, although distinguishing earthquakes from cultural noise was more difficult without a local network.

The Canadian National Earthquake Database (CNED) catalog identified 58 earthquakes during this time frame in this region (118–116.5°W, 53.5–55°N), with magnitudes between M_L 2.0 and M_w 4.6. Most earthquakes were clustered ~30 km away from station BRLDA (Figure 7). During the time frame considered, HF occurred as close as ~5 km from station BRLDA, but most operations were >30 km away (Figure 8a). These CNED catalog earthquakes were used in a template matching routine in a manner similar to the stacked subfamilies identified by RSD (Figure 8b). The cumulative number of event detections found using RSD were similar to the results obtained from using the catalog, and 91% of the template matched

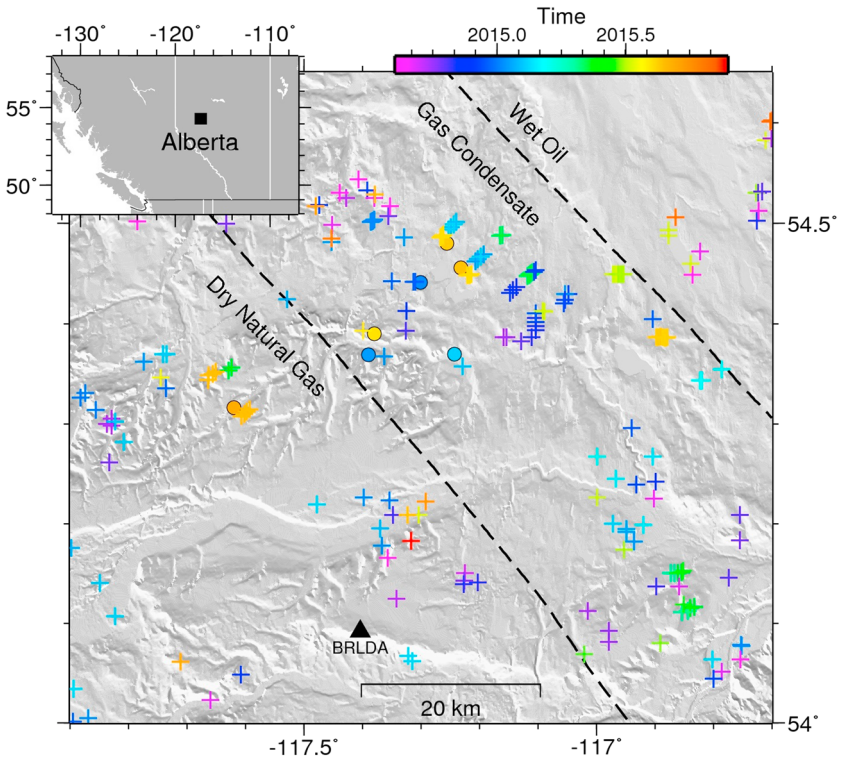

Figure 7. Map showing shaded topographic relief of the Kaybob region of west central Alberta, Canada. Crosses show locations of unconventional well pads targeting the Duvernay Formation, colored by time of hydraulic fracturing, restricted to the time range of our seismic analysis. Dashed lines indicate the approximate divisions of thermal maturity for hydrocarbons in the Duvernay Formation [*Rokosh et al.*, 2012], with many operations focused in the condensate region. Circles colored by time show epicenters determined in this study for seven sequences with large enough seismicity to be located with the regional network. Triangle shows the RV BRLDA seismic station for which RSD was performed. Small black rectangle in the map inset shows the location and size of the local area map.

events using the catalog were also found using our RSD method. These results demonstrate the potential ability of RSD to duplicate results obtained using traditional earthquake detection over moderate (~50 km) source-receiver distances (Figures 8b and 8c).

3.4. Harrison County, Ohio

In September 2013, an earthquake sequence with magnitudes up to M_w 2.2 was induced by HF in Harrison County, Ohio [*Friberg et al.*, 2014]. Two earthquakes in this sequence were included in the ODNR catalog following the published study. Despite the ongoing industry activity near the vicinity of the sequence, the only other documented earthquake in the region was a M_L 2.1 on 30 September 2015, initially reported in the ANSS catalog.

Transportable Array (TA) station O53A is located only a few kilometers from this seismicity in an area of Harrison County where unconventional wells are actively being drilled and stimulated to target the Utica Shale/Point Pleasant Formation (Figure 9). Between 14 December 2012 and 27 November 2015, a total of 30,157 SoI were detected and grouped into 6032 families following the frequency domain grouping, with 138 families (2.3% of the families) containing more than 20 members. This step eliminated a large nonrepetitive. Following the time domain grouping step using the 138 families, 85 subfamilies were identified that contained more than 20 members. Using the 85 subfamilies, single-station normalized template matching revealed 38 productive subfamilies that each matched with more than 600 matched events, while 47 less productive subfamilies were removed (solid gray lines, Figure 10).

We grouped the productive subfamilies into four categories based on inspection of waveforms. The first group of 17 productive subfamilies was found to have matched with real seismic events nearly exclusively (solid black lines, Figure 10). A second group of two subfamilies matched with cultural noise at a small rate

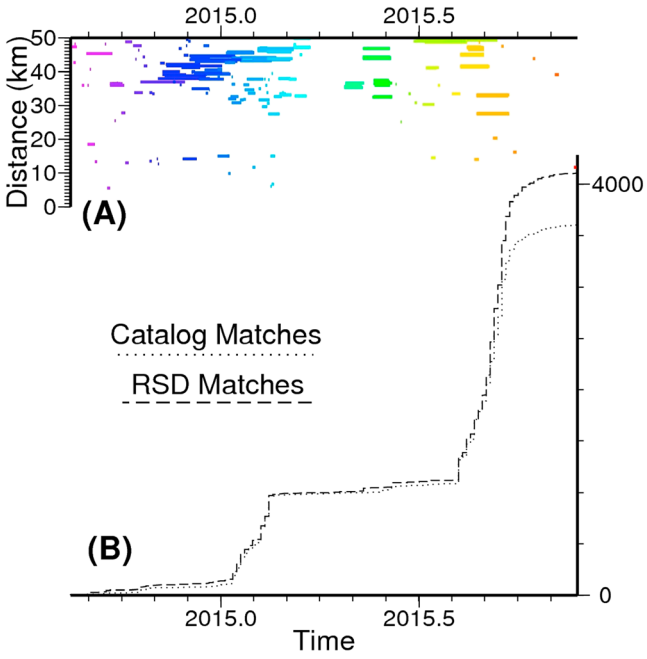

Figure 8. Cumulative number of seismic events detected using template matching compared with hydraulic fracturing operations (HF) in the Kaybob region of west central Alberta, Canada. (a) Distance between station BRLDA and the well heads of HF, with the line length indicating the duration of HF. Colors denote mean time of HF (color scale is shown in Figure 7) to aid in comparisons with symbols in Figure 7. (b) Cumulative number of events detected with a template matching routine using 57 cataloged earthquakes and 88 stacked subfamilies created by RSD. For example of waveforms and individual subfamily performance, see Figures S12 and S16.

(averaging ~2 detections/day), but also had detection rate increases during times of real seismic events, although those detections essentially reproduced the events found with the first group of 17 subfamilies (dotted black lines, Figure 10). A third group of two subfamilies matched with cultural noise fairly continuously throughout the sampling period averaging ~5 detections/day (dashed gray lines, Figure 10). The fourth group of 17 subfamilies matched with cultural noise but had a large increase in March–April 2013 that occurred while the Boy Scout 4-33H well was being drilled less than 1 km away from station O53A (dotted gray lines, Figure 10). Intriguingly, we detected very few events during the HF of this well (23–28 May 2013). We focused our analysis on the first group of 17 productive subfamilies with the most limited cultural noise detections (solid black lines, Figure 10), and matching using the stacked subfamily templates identified 14,270 unique events. Using manual inspection of individual waveforms for smaller sequences and correlated waveform grid plots for larger sequences (see supporting information), we estimate that >95% of these events were confirmed as real seismic events. The Harrison County example appears to have a large amount of noise that is both repetitive and nonrepetitive. The initial 20-member threshold eliminated a large amount of nonrepetitive families, while the subfamily investigation eliminated a large amount of repetitive noise subfamilies. In contrast, the Mammoth Mountain example has similar percentage of nonrepetitive families but far fewer repetitive noise subfamilies.

Our culled seismicity catalog reveals four primary time periods of seismic activity, each with at least 500 events and an average of 3300 events (Figure 10). To investigate whether these four sequences were induced, we compared the seismicity with the time and location of HF based on reports available from ODNR and FracFocus. In each of the four cases, the locations correlate with unconventional wells (Figure 9), and the timing of detections occurs during the HF of those wells or immediately after (Figure 11). We further examined these four sequences in the supporting information (Figure S15). The September–December 2013 sequence was previously reported as induced [*Friberg et al*., 2014]. While the temporal patterns are similar to previous studies, the number of detections have dramatically increased when compared to previous work; *Friberg et al*. [2014] and *Skoumal et al*. [2015c] reported 698 and 2788 events, respectively, while this analysis brings the number of events in the sequence to 7231. The other three primary sequences (September 2014, December 2014 to January 2015, and August–November 2015) had previously gone unreported. Note that the lack of M_L 0–1.5 events in the September 2014 sequence is not erroneous. Rather, location analysis of this sequence suggests that the majority of the $M_L \leq 0$ events clustered along a separate feature ~1 km north of where the $M_L > 1.5$ events occurred [*Brudzinski et al*., 2015; *Friberg et al*., 2015]. Another intriguing observation is the identification of a M_L 2.7 event in the August–November 2015 sequence that was not reported in any of the available catalogs but was well recorded on regional seismometers. We were able to confirm that this event was felt by the local population (J. Butler, personal communication, 2015). This underscores the importance of utilizing a technique like RSD that does not rely on a catalog to search for seismicity.

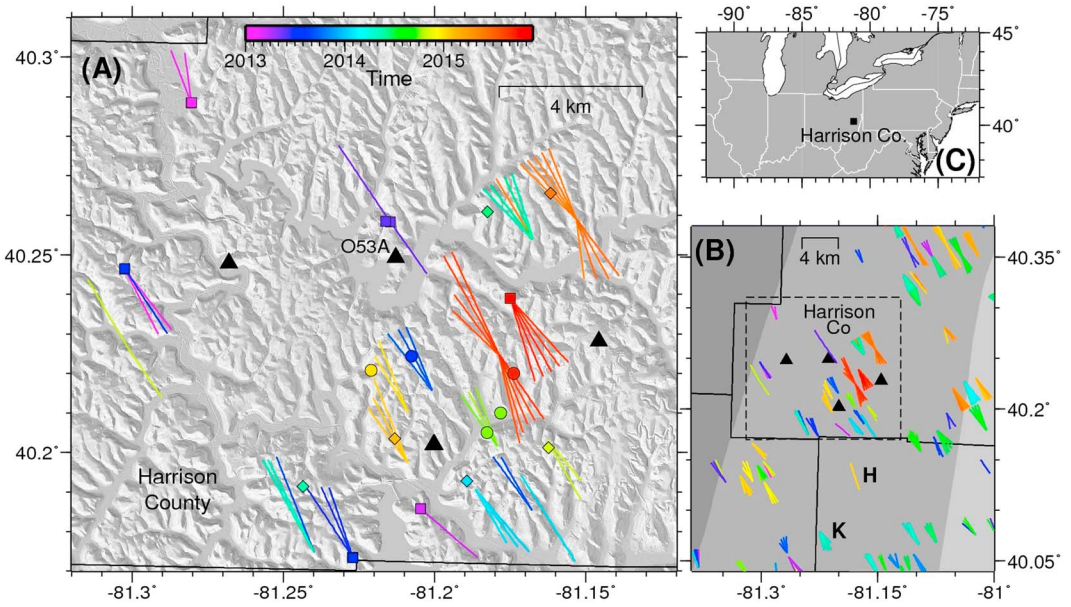

Figure 9. Maps of the Harrison County, Ohio, study region. (a) Map showing shaded topographic relief of southwestern Harrison County. Lines show locations of unconventional well head and toe, colored by time of hydraulic fracturing (HF), restricted to the time range of our seismic analysis. Triangles show local seismic stations. Symbols are also colored by time and show locations of large (>750 events, circles) and small (diamonds) earthquake sequences. Squares mark well heads where stimulation occurred before the local network was installed, but HF correlates temporally with seismic activity, and *S-P* times recorded by O53A are consistent with earthquakes occurring near these wells. (b) Spatial and temporal distribution of HF in the region surrounding that shown in Figure 9a (dashed line). There are two wells south of our main study region (H: Horseshoe and K: Kirkwood) that also show evidence of seismic activity. Shading indicates the transition from dry gas (light) to condensate (medium) to oil (dark) based on predicted thermal maturity of the Utica Shale-Point Pleasant Formation [*Riley*, 2015]. (c) Broader regional map with a black rectangle showing the location of Figure 9b.

Fourteen additional smaller sequences that contained between 2 and 210 identified earthquakes (~30 events on average) were also found to correlate temporally with separate HF (Figure 11). Events from eight of these additional sequences were located using *P* and *S* arrivals from at least three stations, and each was found to be in close proximity (<1 km) to the wells being stimulated during the seismicity (Figure 10). Accurate locations could not be determined for the six sequences that occurred either prior to the deployment of additional stations or after the period covered by the data obtained via FOIA request. Although data were only available for O53A for these six sequences, *S-P* times are consistent with distances between the station and well locations with active HF when the seismicity was detected. Two sequences were detected >10 km from O53A, both of which occurred in Belmont County, south of Harrison County (Figure 9b). This includes the Horseshoe well (~15 km from O53A) in addition to the previously reported May 2014 sequence associated with the Kirkwood well, near the border of Belmont/Guernsey Counties (~19 km from O53A) [*Skoumal et al*., 2015c]. These two wells are the closest wells to the south of the Harrison County sequences in an area that has not been as heavily targeted. If wells are stimulated in this area in the future, the risk of generating induced seismicity may be higher than previously recognized.

Considering that only three of the 18 detected sequences were previously reported [*Skoumal et al*., 2015c], this case study demonstrates the potential of RSD to identify characteristic, repeating seismic signals below the detection threshold of traditional earthquake detection methods. It is remarkable that >99% of seismicity in the culled catalog correlates in space and time with HF, but a more detailed study of each sequence is needed to characterize the nature of how seismicity is related to industry operations.

4. Discussion

4.1. Computation Time

The processing time for RSD depends on the length of the time period being considered and the number of SoI identified. The process of identifying SoI and reserving waveforms is done at a rate of ~1.4 s/d of 40 Hz

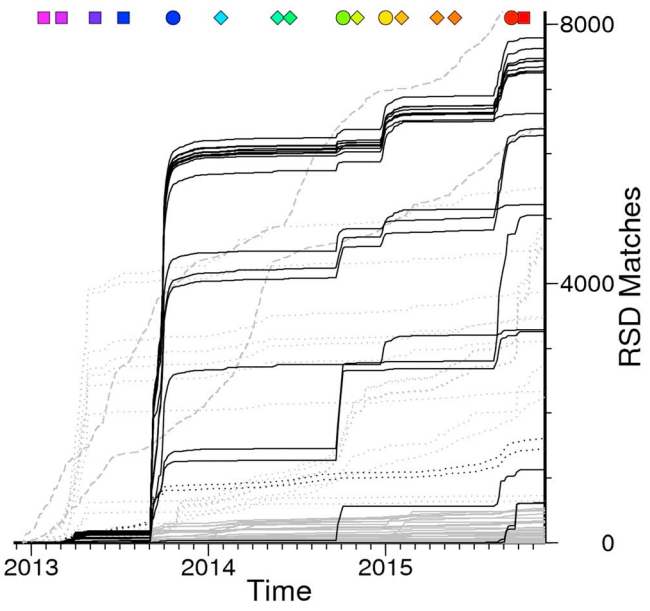

Figure 10. The cumulative number of seismic events detected using in the Harrison County, Ohio, study region using each of the 85 stacked subfamilies identified by RSD as a template. Solid gray lines show less productive subfamilies that each matched with less than 600 events and were removed. Solid black lines denote productive subfamilies that matched with real seismic events nearly exclusively. Dashed gray lines show subfamilies that matched with cultural noise fairly continuously throughout the sampling period. Dotted gray lines mark subfamilies that matched with cultural noise but had a large increase in March–April 2013 that occurred while the Boy Scout 4-33H well was being drilled less than 1 km away from station O53A. Dotted black lines indicate subfamilies that matched with cultural noise but also had detection rate increases during times of real seismic events. Colored symbols are consistent with Figures 9 and 11 and denote times of earthquake sequences that correlate in space and time with hydraulic fracturing. For example of earthquake waveforms, see Figure S13.

data, a rate consistent for all four of our case studies. Identifying which SoI were repetitive and appropriate for use as templates depended on the number of SoI. The largest number of SoI (and the longest runtimes) was observed in the Harrison County case. Clustering of the 30,157 SoI in the frequency domain took 304 s, and the following time domain step took 241 s (total of 546 s). Had only the time domain sorting step been used to characterize repetitive sequences, computation time would have been 2173 s, roughly 4 times as long. Additional delays from only using the time domain approach occur as it produced nearly twice as many subfamilies (157), which would nearly double the time required for template matching. We also found the time domain-only approach to be less effective as most of the identified subfamilies were repetitive noise signals such that 30% fewer earthquakes were identified.

True autocorrelation for the same 1079 day long window in Harrison County would have been impossible given the computational resources at our disposal, requiring $\sim 7 \times 10^{18}$ correlations with an estimated runtime of 20 million computation days. Template matching only takes 91,715 s once the 85 subfamily templates were produced by our two-step frequency and time domain clustering method. For the Harrison County case, total runtime for RSD from raw data to template matched catalog was 93,739 s, which takes about 2 h distributed across 12 nodes on a cluster. This highlights how template matching is the longest-duration step, so the clustering and stacking of repetitive signals via RSD has the biggest impact in reducing the processing time by limiting the

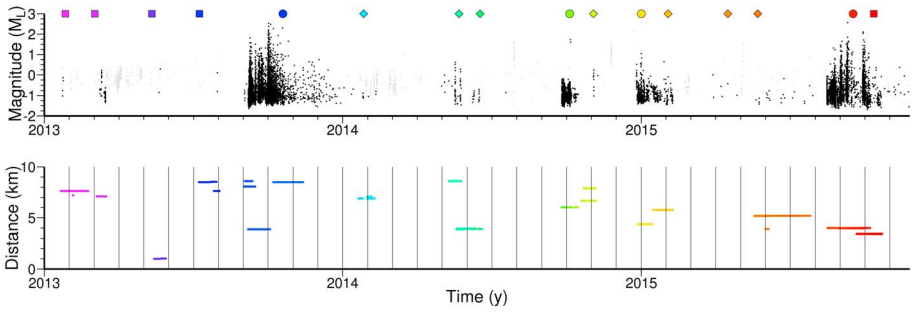

Figure 11. Temporal relationship between detected seismicity and hydraulic fracturing (HF) in Harrison County, Ohio. (a) Magnitudes of seismic events detected using RSD and template matching (black dots) and spurious cultural noise signals discriminated by manual waveform inspection (gray dots). Colored symbols are consistent with Figures 9 and 10. (b) Vertical position indicates distance from station O53A to nearby HF well heads, and the extent of the horizontal line indicates the timing of HF. Note that HF at the Kirkwood [*Skoumal et al.*, 2015c] and Horseshoe wells in Belmont County occurred >10 km away from O53A and are not shown (K and H in Figure 9b).

number of templates that need to be scanned. For the Mammoth Mountain case, *Shelly et al*. [2015] applied template matching to 1531 cataloged events, while we needed only 18 stacked subfamilies as templates to generate a similar temporal history of seismicity, reducing the longest-duration processing step by 2 orders of magnitude.

4.2. Limitations and Future Work

This study principally seeks to show the results that can be obtained by applying RSD to a single seismometer with limited potential human interaction to demonstrate its capability for monitoring smaller seismicity in a near-real-time environment. To determine the effectiveness of the algorithm, detected events were manually investigated to classify them as earthquakes or noise. More detailed investigation of the frequency and time domain characteristics of the noise signals could provide a means to automate the discrimination of these cases in the future. Such a discriminant would be particularly effective at the stage where stacked subfamilies are formed, as subfamily stacks formed primarily from cultural noise tend to find more noise than earthquakes.

In its current form, RSD can only utilize a single station for earthquake swarm detection. While we have demonstrated that the algorithm is capable of identifying repeating earthquakes in a variety of settings with a single station, the algorithm is still reliant on records at additional stations and an external method, such as *elocate* [*Herrmann*, 2013], to locate detected earthquakes. A multistation form of RSD is currently in development, but determining how to handle different detections at multiple stations is beyond the scope of this study. The goal with multistation RSD would be to create an algorithm that retains the computational efficiency while also providing relative station lags appropriate for determining source locations. It also has the potential to improve our identification and removal of repetitive cultural noise signals that would be unique to a single station.

Further analysis is also needed to clarify the connection between the identified seismicity detected in the Harrison County region and the HF operations to determine if these sequences were induced. For example, one could utilize an empirical subspace detector on the events from each smaller sequence to increase the number of detections [*Barrett and Beroza*, 2014; *Skoumal et al*., 2015b]. Considering that many of the smaller sequences also have small magnitudes, a detailed location study utilizing improved velocity models, correlated arrival times, and/or double difference relocation could potentially improve the precision of hypocentral depths to differentiate whether these are operationally induced microseismicity in the target formation or earthquakes in the Precambrian basement. The other recent documented induced seismicity sequences in Ohio with $M > 2$ have all occurred near or below the basement contact [*Kim*, 2013; *Friberg et al*., 2014; *Skoumal et al*., 2014, 2015a, 2015b, 2015c]. The lack of locations for newly detected events also limited the accuracy of magnitudes calculated in this study. Event magnitudes in the Ohio region farther from O53A than the prominent seismic sequence in 2013 are likely to be underestimated with the single-station calculation we used. The most extreme example are the events associated with the Kirkwood wells that were estimated ~1 magnitude unit below their reported values [*Skoumal et al*., 2015c].

While we focused on two regions known to have induced seismicity from HF, previous work has indicated that induced seismicity in the Midcontinent United States is largely attributed to disposal of large volumes of wastewater or produced water [*McGarr et al*., 2015; *Walsh and Zoback*, 2015; *Weingarten et al*., 2015]. Considering that the life span of wastewater disposal wells often lasts years or decades relative to HF operations that may last less than a couple of months, seismicity patterns between the two sources of induced seismicity are expected to differ. In this study, we did not initially determine whether RSD stacked subfamily templates were real seismic signals or noise. Rather, after template matching, we found that signals that found matches "continuously" throughout the time of the study were cultural noise and discarded them. While this may be appropriate to identify seismicity induced by HF, this approach would not be viable for long-lived wastewater disposal wells that may have associated seismicity for many years. For these cases, confirming a stacked subfamily represents an earthquake through additional frequency analysis, identifying the characteristics of *P* and *S* waves and/or utilizing recordings on multiple seismometers may be required to ensure high detection fidelity and low false positive rate.

4.3. Induced Seismicity Regulations

In April 2014, new permit conditions set forth by the ODNR stated that new HF wells within ~4.8 km (3 miles) of a known fault or earthquake with $M > 2$ will be required to install seismic monitors [*ODNR*, 2014]. If the

seismometers detect a $M > 1$ event associated with the HF, the operations would be suspended. Had the four primary sequences in Harrison County been permitted following this regulation was enacted, operations may have been ceased. In March 2015, a new regulatory system was established by the Alberta Energy Regulator in the Kaybob area near Fox Creek that would require operators to adjust operations if a $M > 2$ event was associated with HF and suspend operations if it was responsible for a $M > 4$ event. HF operators in and around Harrison County and the Kaybob area should be aware of the increased likelihood of inducing seismicity based on the prevalent prior activity, and we propose implementing an algorithm such as RSD to maximize the detection capabilities of existing seismic networks.

4.4. Application to Other Repetitive Seismic Signals

In addition to detecting induced seismic sequences, we have sought to demonstrate that RSD also has the potential to improve characterization of natural earthquake swarms in areas with limited station coverage. Template matching approaches have been successfully applied to natural earthquake swarms in a variety of tectonic settings, identifying low-frequency earthquake swarms [*Shelly et al*., 2007] including imaging of fault fluid interactions in volcanic settings [*Shelly et al*., 2015] and detecting slow slip in advance of megathrust earthquakes [*Kato et al*., 2012]. While the use of earthquake swarms for forecasting subduction zone behavior is still only a hypothesis [*Holtkamp and Brudzinski*, 2014], earthquake swarms are considered to be a key component of eruption forecasting [*McNutt*, 1996]. Much like in the induced seismicity examples, RSD would improve upon the previous studies that utilized catalog-based template matching by detecting natural earthquake swarms below the catalog detection level. This should improve the forecasting capabilities, particularly in areas without dense seismic networks.

5. Conclusion

While the commonly used multistation STA/LTA earthquake detection algorithms are ideally suited for identifying $M > 2$ seismicity using existing regional networks in the United States, this approach struggles to effectively identify smaller-magnitude earthquakes. When traditional detection techniques are employed using many regional networks in the U.S., swarms of $M < 2$ events typically go undetected until a larger-magnitude event is observed. The RSD algorithm was created to efficiently identify these small-magnitude, repeating earthquakes, such as in the increasingly more prevalent cases of induced seismicity. This algorithm identifies signals of interest on a single regional station using a low-threshold STA/LTA amplitude detector and then performs agglomerative clustering to sort these signals into families according to their frequency and time domain characteristics.

When RSD was applied to a seismometer near Mammoth Mountain during the time of a volcanic swarm, it produced a remarkably similar temporal pattern of seismicity to that from catalog-based template matching [*Shelly et al*., 2015]. In the Oaxaca portion of the Mexico subduction zone, RSD identified ~10 times more events than a traditionally constructed catalog [*Fasola et al*., 2016]. In Central Alberta, Canada, template matching using RSD applied over moderate distances (~50 km) from the station was able to independently duplicate event detections using template matching based on cataloged events. When this algorithm was applied in Harrison County, Ohio, 18 potentially induced earthquake swarms were identified between 2013 and 2015, 15 of which were previously unreported. Although $M > 1$ seismicity induced by HF is likely rare and typically restricted to small magnitudes, our study indicates that it may be more common than previously thought, a new interpretation that resulted from the application of RSD.

RSD currently focuses on a single seismometer for earthquake detection, which allows for seismicity to be quickly identified in poorly instrumented regions, but means that it is reliant on another method to locate the newly detected events. While autocorrelation and FAST have also been proven effective at identifying repeating seismic events with single stations, the computationally intensive nature of these algorithms may impede the wide-scale application of these algorithms. Considering the minimal computational requirements of RSD, this algorithm may be better suited for the real-time detection and temporal characterization of small-magnitude earthquake swarms. As more observations typically lead to improved interpretations, techniques like RSD are important tools for the future understanding of earthquake swarms in a variety of settings.

Acknowledgments

Support for this work was provided by ODNR, NSF grant EAR-0847688 (M.B.), and USGS NEHRP grant 2015-0176 (M.B. and B.C.). Seismic data were obtained through the Incorporated Research Institutions for Seismology Data Management Center (IRIS) (www.iris.edu; last accessed November 2015), through the Northern California Earthquake Data Center (NCEDC) (www.ncedc.org; last accessed February 2016), and by submitting Freedom of Information Act request for local seismic data collected by the Ohio Department of Natural Resources (ODNR). Waveforms from OXTT are available by contacting M. Brudzinski and will be available at the IRIS DMC under the network code MU in June 2016. Earthquake catalogs were obtained from the (ODNR) (http://geosurvey.ohiodnr.gov; last accessed December 2015), U.S. Geological Survey National Earthquake Information Center (NEIC) (http://earthquake.usgs.gov/data; last accessed December 2015), and the Canadian National Earthquake Database (http://earthquakescanada.nrcan.gc.ca; last accessed December 2015). Well data were retrieved from the Ohio Department of Natural Resources (https://gis.ohiodnr.gov/website/dog/oilgaswells; last accessed December 2015), FracFocus (https://fracfocus.org; last accessed December 2015), and FracFocus.ca (https://fracfocus.ca; last accessed December 2015). Plots were made using the Generic Mapping Tools version 4.2.1 (www.soest.hawaii.edu/gmt; last accessed December 2015). We would like to thank ODNR for the assistance, specifically D. Blake, S. Dade, J. Fox, M. Hansen, and D. Rush. The work in Harrison County was greatly aided by P. Friberg. We also benefitted from discussions with G. Beroza, D. Eaton, O. Kaven, and C. Yoon. Comments from two reviewers and an Associate Editor helped improve this manuscript. Our analysis relied heavily on Miami University's High Performance Computing, and we thank J. Mueller for his assistance.

References

Barrett, S. A., and G. C. Beroza (2014), An empirical approach to subspace detection, *Seismol. Res. Lett.*, *85*(3), 594–600.

Benz, H. M., N. D. McMahon, R. C. Aster, D. E. McNamara, and D. B. Harris (2015), Hundreds of earthquakes per day: The 2014 Guthrie, Oklahoma, earthquake sequence, *Seismol. Res. Lett.*, *86*(5), doi:10.1785/0220150019.

British Columbia Oil and Gas Commission (BCOGC) (2012), Investigation of observed seismicity in the Horn River basin, Technical Rep. [Available at http://www.bcogc.ca/investigation-observed-seismicity-horn-river-basin, (last accessed December 2015).]

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.

Brudzinski, M. R., E. Cabral-Cano, F. Correa-Mora, C. DeMets, and B. Márquez-Azúa (2007), Slow slip transients along the Oaxaca subduction segment from 1993 to 2007, *Geophys. J. Int.*, *171*(2), 523–538.

Brudzinski, M. R., H. R. Hinojosa-Prieto, K. M. Schlanser, E. Cabral-Cano, A. Arciniega-Ceballos, O. Diaz-Molina, and C. DeMets (2010), Nonvolcanic tremor along the Oaxaca segment of the Middle America subduction zone, *J. Geophys. Res.*, *115*, B00A23, doi:10.1029/2008JB00606.

Brudzinski, M. R., R. J. Skoumal, and B. S. Currie (2015), Multistation template matching to characterize frequency-magnitude distributions of induced seismicity in the Central and Eastern US, Abstract T43D-3047 presented at 2015 Fall Meeting, AGU, San Francisco, Calif., 14–18 Dec.

Buchanan, R. C. (2015), Increased seismicity in Kansas, *Leading Edge*, *34*(6), 614–617, doi:10.1190/tle34060614.1.

Carr, D., C. Young, R. Aster, and X. Zhang (1999), Cluster analysis for CTBT seismic event monitoring, Presented at 21st Annual Seismic Research Symposium on Monitoring a CTBT, 285–293.

Clarke, H., L. Eisner, P. Styles, and P. Turner (2014), Felt seismicity associated with shale gas hydraulic fracturing: The first documented example in Europe, *Geophys. Res. Lett.*, *41*, 8308–8314, doi:10.1002/2014GL062047.

Eaton, D. W., and A. B. Mahani (2015), Focal mechanisms of some inferred induced earthquakes in Alberta, Canada, *Seismol. Res. Lett.*, *86*(4), 1078–1085.

Ellsworth, W. L. (2013), Injection-induced earthquakes, *Science*, *341*, doi:10.1126/science.1225942.

Fasola, S., M. R. Brudzinski, N. Ghouse, K. Solada, S. Sit, E. Cabral-Cano, A. Arciniega-Ceballos, N. Kelly, C. DeMets, and K. Jensen (2016), New perspective on the transition from flat to steeper subduction in Oaxaca, Mexico based on seismicity, nonvolcanic tremor, and slow slip, *J. Geophys. Res. Solid Earth*, *121*, 1835–1848, doi:10.1002/2015JB012709.

Frank, W. B., and N. M. Shaprio (2014), Automatic detection of low-frequency earthquakes (LFEs) based on a beamformed network response, *Geophys. J. Int.*, doi:10.1093/gji/ggu058.

Frank, W. B., N. M. Shapiro, A. L. Husker, V. Kostoglodov, A. Romanenko, and M. Campillo (2014), Using systematically characterized low-frequency earthquakes as a fault probe in Guerrero, Mexico, *J. Geophys. Res. Solid Earth*, *119*, 7686–7700, doi:10.1002/2014JB011457.

Friberg, P. A., G. M. Besana-Ostman, and I. Dricker (2014), Characterization of an earthquake sequence triggered by hydraulic fracturing in Harrison County, Ohio, *Seismol. Res. Lett.*, *85*(6), 1295–1307.

Friberg, P. A., M. R. Brudzinski, B. S. Currie, and R. J. Skoumal (2015), Observations of a hydrofracture induced earthquake sequence in Harrison County Ohio in 2014, Abstract S11C-03 presented at 2015 Fall Meeting, AGU, San Francisco, Calif., 14–18 Dec.

Frohlich, C., C. Hayward, B. Stump, and E. Potter (2011), The Dallas-Fort Worth earthquake sequence: October 2008 through May 2009, *Bull. Seismol. Soc. Am.*, *101*, 327–340.

Gibbons, S. J., and F. Ringdal (2006), The detection of low magnitude seismic events using array-based waveform correlation, *Geophys. J. Int.*, *165*(1), 149–166.

Graham, S. E., C. DeMets, E. Cabral-Cano, V. Kostoglodov, A. Walpersdorf, N. Cotte, M. Brudzinski, R. McCaffrey, and L. Salazar-Tlaczani (2014), GPS constraints on the 2011–2012 Oaxaca slow slip event that preceded the 2012 March 20 Ometepec earthquake, southern Mexico, *Geophys. J. Int.*, *197*(3), 1593–1607.

Green, D. N., and J. Neuberg (2006), Waveform classification of volcanic low-frequency earthquake swarms and its implication at Soufrière Hills Volcano, Montserrat, *J. Volcanol. Geotherm. Res.*, *153*(1), 51–63.

Ground Water Protection Council, and Interstate Oil and Gas Compact Commission (2015), Potential Injection-Induced Seismicity Associated with Oil & Gas Development: A Primer on Technical and Regulatory Considerations Informing Risk Management and Mitigation.

Herrmann, R. B. (2013), Computer programs in seismology: An evolving tool for instruction and research, *Seismol. Res. Lett.*, *84*, 1081–1088, doi:10.1785/0220110096.

Hill, D. P., W. L. Ellsworth, M. J. S. Johnston, J. O. Langbein, D. H. Oppenheimer, A. M. Pitt, P. A. Reasenberg, M. L. Sorey, and S. R. McNutt (1990), The 1989 earthquake swarm beneath Mammoth Mountain, California: An initial look at the 4 May through 30 September activity, *Bull. Seismol. Soc. Am.*, *80*, 325–339.

Hirose, H., T. Matsuzawa, T. Kimura, and H. Kimura (2014), The Boso slow slip events in 2007 and 2011 as a driving process for the accompanying earthquake swarm, *Geophys. Res. Lett.*, *41*, 2778–2785, doi:10.1002/2014GL059791.

Holland, A. (2013), Earthquakes triggered by hydraulic fracturing in south-central Oklahoma, *Bull. Seismol. Soc. Am.*, *103*(3), 1784–1792.

Holtkamp, S. G., and M. R. Brudzinski (2011), Earthquake swarms in circum-Pacific subduction zones, *Earth Planet. Sci. Lett.*, *305*, 215–225.

Holtkamp, S. G., and M. R. Brudzinski (2014), Megathrust earthquake swarms indicate frictional changes which delimit large earthquake ruptures, *Earth Planet. Sci. Lett.*, *390*, 234–243.

Holtkamp, S. G., M. R. Brudzinski, and B. S. Currie (2015), Regional detection and monitoring of injection-induced seismicity: Application to the 2010–12 Youngstown, Ohio seismic sequence, *AAPG Bull.*, *99*(9), 1671–1688, doi:10.1306/03311513194.

Horton, S. (2012), Disposal of hydrofracking waste fluid by injection into subsurface aquifers triggers earthquake swarm in central Arkansas with potential for damaging earthquake, *Seismol. Res. Lett.*, *83*, 250–260, doi:10.1785/gssrl.83.2.250.

Huang, Y., and G. C. Beroza (2015), Temporal variation in the magnitude-frequency distribution during the Guy-Greenbrier earthquake sequence, *Geophys. Res. Lett.*, *42*, 6639–6646, doi:10.1002/2015GL065170.

Kao, H., and S.-J. Shan (2004), The Source-Scanning Algorithm: Mapping the distribution of seismic sources in time and space, *Geophys. J. Int.*, *157*(2), 589–594, doi:10.1111/j.1365-246X.2004.02276.x.

Kato, A., K. Obara, T. Igarashi, H. Tsuruoka, S. Nakagawa, and N. Hirata (2012), Propagation of slow slip leading up to the 2011 M_w 9.0 Tohoku-Oki earthquake, *Science*, *335*, 705–708, doi:10.1126/science.1215141.

Kaufman, L., and P. J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Kim, W.-Y. (2013), Induced seismicity associated with fluid injection into a deep well in Youngstown, Ohio, *J. Geophys. Res. Solid Earth*, *118*, 3506–3518, doi:10.1002/jgrb.50247.

Lay, T., and T. C. Wallace (1995), *Modern Global Seismology*, pp. 521, Academic Press, San Diego, Calif.

Lewicki, J. L., G. E. Hilley, D. R. Shelly, J. C. King, J. P. McGeehin, M. Mangan, and W. C. Evans (2014), Crustal migration of CO_2-rich magmatic fluids recorded by tree-ring radiocarbon and seismicity at Mammoth Mountain CA, USA, *Earth Planet. Sci. Lett.*, doi:10.1016/j.epsl.2013.00.832.

Lin, G. (2013), Seismic investigation of magmatic unrest beneath Mammoth Mountain California, USA, *Geology*, *41*, 847–850.

McGarr, A., et al. (2015), Coping with earthquakes induced by fluid injection, *Science*, *347*(6224), 830–831.

McNutt, S. R. (1996), Seismic monitoring and eruption forecasting of volcanoes: A review of the state-of-the-art and case histories, *Monit. Mitigation Volcano Hazards*, 99–146, doi:10.1007/978-3-642-80087-0_3.

Mogi, K. (1963), Some discussions on aftershocks, foreshocks and earthquake swarms: The fracture of a semi-infinite body caused by an inner stress origin and its relation to the earthquake phenomena, *Bull. Earthquake Res. Inst.*, *41*, 615–658.

National Research Council (NRC) (2013), *Induced Seismicity Potential in Energy Technologies*, pp. 225, Natl. Acad. Press, Washington, D. C.

Neuberg, J. W., H. Tuffen, L. Collier, D. Green, T. Powell, and D. Dingwell (2006), The trigger mechanism of low-frequency earthquakes on Montserrat, *J. Volcanol. Geotherm. Res.*, *153*(1), 37–50.

Nicholson, C., and R. L. Wesson (1990), Earthquake hazard associated with deep well injection: A report to the U.S. Environmental Protection Agency, U.S. Geol. Surv. Bull. 1951.

Ohio Department of Natural Resources (ODNR) (2014), Ohio announces tougher permit conditions for drilling activities near faults and areas of seismic activity. [Available at http://ohiodnr.gov/news/post/ohio-announcesougher-permit-conditions-for-drilling-activities-near-faults-and-areas-ofseismic-activity, (last accessed December 2015).]

Petersen, T. (2007), Swarms of repeating long-period earthquakes at Shishaldin Volcano, Alaska, 2001–2004, *J. Volcanol. Geotherm. Res.*, *166*(3), 177–192.

Prejean, S., A. Stork, W. Ellsworth, D. Hill, and B. Julian (2003), High precision earthquake locations reveal seismogenic structure beneath Mammoth Mountain, California, *Geophys. Res. Lett.*, *30*(24), 2247, doi:10.1029/2003GL018334.

Riley, R. A. (2015), Regional drilling activity and production, in *A Geologic Play Book for Utica Shale Appalachian Basin Exploration, Final Report of the Utica Shale Appalachian Basin Exploration Consortium*, edited by D. G. Patchen, and K. M. Carter, pp. 11–19, West Virginia Geological and Economic Survey Morgantown. [Available at http://www.wvgs.wvnet.edu/utica, (last accessed April 2016).]

Rokosh, C. D., et al. (2012), Summary of Alberta's shale- and siltstone-hosted hydrocarbon resource potential, Energy Resources Conservation Board, *ERCB/AGS Open File Rep., 2012-06*, 327 p.

Rowe, C. A., R. C. Aster, B. Borchers, and C. J. Young (2002), An automatic, adaptive algorithm for refining phase picks in large seismic data sets, *Bull. Seismol. Soc. Am.*, *92*(5), 1660–1674.

Rubinstein, J. L., W. L. Ellsworth, A. McGarr, and H. M. Benz (2014), The 2001–present induced earthquake sequence in the Raton Basin of northern New Mexico and southern Colorado, *Bull. Seismol. Soc. Am.*, *104*(5), 2162–2181, doi:10.1785/0120140009.

Schaff, D. P., and F. Waldhauser (2005), Waveform cross-correlation-based differential travel-time measurements at the Northern California Seismic Network, *Bull. Seism. Soc. Am.*, *95*, 2446–2461.

Schultz, R., V. Stern, M. Novakovic, G. Atkinson, and Y. J. Gu (2015), Hydraulic fracturing and the Crooked Lake sequence: Insights gleaned from regional seismic networks, *Geophys. Res. Lett.*, *42*, 2750–2758, doi:10.1002/2015GL063455.

Shelly, D. R., G. C. Beroza, and S. Ide (2007), Non-volcanic tremor and low-frequency earthquake swarms, *Nature*, *446*, doi:10.1038/nature05666.

Shelly, D. R., T. A. Taira, S. G. Prejean, D. P. Hill, and D. S. Dreger (2015), Fluid-faulting interactions: Fracture-mesh and fault-valve behavior in the February 2014 Mammoth Mountain, California, earthquake swarm, *Geophys. Res. Lett.*, *42*, 5803–5812, doi:10.1002/2015GL064325.

Skoumal, R. J., M. R. Brudzinski, B. S. Currie, and J. Levy (2014), Optimizing multi-station earthquake template matching through re-examination of the Youngstown, Ohio sequence, *Earth Planet. Sci. Lett.*, *405*, 274–280.

Skoumal, R. J., M. R. Brudzinski, and B. S. Currie (2015a), Induced earthquakes during hydraulic fracturing in Poland Township, Ohio, *Bull. Seismol. Soc. Am.*, *105*(1), 189–197, doi:10.1785/0120140168.

Skoumal, R. J., M. R. Brudzinski, and B. S. Currie (2015b), Microseismicity induced by deep wastewater injection in Southern Trumbull County, Ohio, *Seismol. Res. Lett.*, *86*(5), doi:10.1785/0220150055.

Skoumal, R. J., M. R. Brudzinski, and B. S. Currie (2015c), Distinguishing induced seismicity from natural seismicity in Ohio: Demonstrating the utility of waveform template matching, *J. Geophys. Res. Solid Earth*, *120*, 6284–6296, doi:10.1002/2015JB012265.

Slinkard, M. E., D. B. Carr, and C. J. Young (2013), Applying waveform correlation to three aftershock sequences, *Bull. Seismol. Soc. Am.*, *103*(2A), 675–693.

Song, T.-R. A., D. Helmberger, M. R. Brudzinski, R. W. Clayton, P. Davis, X. Pérez-Campos, and S. K. Singh (2009), Subducting slab ultra-slow velocity layer coincident with silent earthquakes in southern Mexico, *Science*, *324*, 502–506.

Sorey, M. L., W. C. Evans, B. M. Kennedy, C. D. Farrar, L. J. Hainsworth, and B. Hausback (1998), Carbon dioxide and helium emissions from a reservoir of magmatic gas beneath Mammoth Mountain, California, *J. Geophys. Res.*, *103*, 15,303–15,323, doi:10.1029/98JB01389.

Stephens, C. D., and B. A. Chouet (2001), Evolution of the December 14, 1989 precursory long-period event swarm at Redoubt Volcano, Alaska, *J. Volcanol. Geotherm. Res.*, *109*(1), 133–148.

Stern, V. H., R. J. Schultz, L. Shen, Y. J. Gu, and D. W. Eaton (2013), Alberta earthquake catalogue, version 1.0: September 2006 through December 2010, Alberta Geological Survey Open File Report 2013-15.

Walsh, F. R., and M. D. Zoback (2015), Oklahoma's recent earthquakes and saltwater disposal, *Sci. Adv.*, *1*(5), doi:10.1126/sciadv.1500195.

Warpinski, N. R. (2013), Understanding hydraulic fracture growth, effectiveness, and safety through microseismic monitoring, in *Effective and Sustainable Hydraulic Fracturing*, chap. 6, pp. 123–135, Tech Croatia.

Weingarten, M., S. Ge, J. W. Godt, B. A. Bekins, and J. L. Rubinstein (2015), High-rate injection is associated with the increase in U.S. mid-continent seismicity, *Science*, *348*(6241), 1336–1340.

Yoon, C., O. O'Reilly, K. Bergen, and G. Beroza (2015a), Earthquake detection through computationally efficient similarity search, *Sci. Adv.*, *1*, doi:10.1126/sciadv.1501057.

Yoon, C., O. O'Reilly, K. Bergen, Y. Huang, W. Ellsworth, and G. Beroza (2015b), Searching for unknown earthquakes in the Guy-Greenbrier, Arkansas, earthquake sequence using efficient waveform similarity search, Abstract S13B-2850 presented at 2015 Fall Meeting, AGU, San Francisco, Calif., 14–18 Dec.