A Deep Earthquake Catalog for Oklahoma and Southern Kansas Reveals Extensive Basement Fault Networks

Yongsoo Park[1], **Gregory C. Beroza**[1], **William L. Ellsworth**[1]

[1]Department of Geophysics, Stanford University, Stanford, CA 94305, USA

Key Points:

- We report on a deep earthquake catalog of over 300,000 precisely located earthquakes.
- Numerous fault structures are revealed at high-resolution.
- The rich dataset provides new opportunities for data-driven analyses of induced earthquakes.

Corresponding author: Yongsoo Park, ysp@stanford.edu

Abstract

The successful application of deep learning for seismic phase arrival time picking has increased the efficacy of earthquake catalog development workflows. Earthquake catalogs with lower magnitude of completeness and better locational precision than current standard practice can now be generated with very limited need for human review and without the need for earthquake templates, which are not always available. Here, we report on a 'Deep Earthquake Catalog' with over 300,000 events from a geographically extensive region spanning Oklahoma and Southern Kansas from January 2010 to December 2020 developed using a workflow that leverages deep learning for phase picking. The increased number of events and improved spatial resolution compared to the previous statewide catalogs reveals numerous discrete faults and both broad trends and localized patterns of seismicity. This rich dataset provides new opportunities for data-driven analyses of induced earthquakes.

Plain Language Summary

Oklahoma and Southern Kansas have experienced unprecedented rates of seismicity for over a decade as a result of unconventional hydrocarbon development. The region did not have a history of frequent earthquake activity, and little was known about the location or nature of the faults that came to host this seismicity. We reanalyzed the seismological data from January 2010 to December 2020 using an advanced workflow and produced a map of over 300,000 earthquakes, most of which were previously unknown. These earthquakes clearly illuminate the hidden fault structures throughout the region and can be used to better understand the regional seismicity.

1 Introduction

The workflow for developing an earthquake catalog with lower magnitude of completeness and increased location precision, i.e., a high-resolution earthquake catalog, compared to standard procedures has become faster and easier through the application of deep learning (DL) for seismic phase arrival time picking (Zhu & Beroza, 2019). The workflow (referred to as DL-assisted workflow hereafter) has been successfully used to efficiently create high-resolution earthquake catalogs (referred to as deep earthquake catalogs hereafter) in both anthropogenic (Park et al., 2020) and tectonic (Liu et al., 2020; Tan et al., 2021) settings. These studies demonstrate that we can now produce a cat-

alog with resolution and sensitivity approaching that of template matching, but without the need for the prior information in the form of a set of templates.

One of the clearest use cases of high-resolution earthquake catalogs is to map hidden fault structures and derive fault attributes such as orientations and dimensions (Schoenball & Ellsworth, 2017a; Skoumal et al., 2019). The resulting information can be used for geomechanical analysis, for example, to evaluate how likely faults are to slip (Walsh III & Zoback, 2016), to resolve fault plane ambiguities in focal mechanism solutions and to constrain local stress fields (Angelier, 1979), or to create fault models in numerical simulations (Yehya et al., 2018). This is especially useful for Oklahoma and Southern Kansas as almost all earthquakes have occurred on previously unmapped faults (Schoenball & Ellsworth, 2017b).

Previous studies improved the statewide earthquake catalog developed by the Oklahoma Geological Survey (OGS) (Walter et al., 2020) by precisely relocating earthquakes using a well-established workflow (Schoenball & Ellsworth, 2017b, 2017a) and by detecting and locating more events with template matching (Skoumal et al., 2019). These studies identified fault structures based on the epicentral distance among earthquake events, but the difference in spatial resolution and earthquake location methods led to some disagreements between the two results.

In this study, we use a DL-assisted workflow to process continuous waveform data in Oklahoma and Southern Kansas from January 2010 to December 2020 and report on a deep earthquake catalog with over 300,000 earthquakes that illuminates regional fault structures at both a broader scale and in more detail than previous catalogs. We show that the template-independency of the DL-assisted workflow can lead to earthquake catalogs with greater spatial resolution than template matching.

2 Earthquake Catalog Development

We used data from 17 publicly available seismic networks (codes 4H, 9L, GS, N4, NP, NQ, NX, O2, OK, TA, US, XR, Y7, Y9, ZD, ZP, and ZQ), comprising 422 stations. The geographic distribution of stations is shown in Figure 1 and the operational time of each station is shown in the supporting information (Figure S1).

We used a pre-trained neural network (Zhu & Beroza, 2019) to detect earthquakes and pick P- and S-phase arrival times. Because the picks and the prediction scores vary

depending on where the arrival is in the 30-second input window, we used a small stride of 2 seconds and used the picks that were consistently predicted in at least 5 windows with a prediction score of 0.5 or greater. These picks were then associated using a grid search algorithm similar to (Johnson et al., 1997; Zhang et al., 2019) where theoretical travel times were computed from the velocity model developed by OGS (Darold et al., 2015). We used the theoretical travel times to constrain further the possible time window ranges between a pair of phases. This allows us to restrict the length of the time window between the first phase pick and the following phase picks during association, which removes potential ambiguities that can be introduced with longer time windows. To address seismic network variability, we used an adaptive association score as an event criterion, defined as

$$\text{Association score} = \frac{\sum_{i=1}^{N} w\left(r_i\right) \mathbf{1}\left\{\phi_i \in \mathcal{A}\right\}}{\sum_{i=1}^{N} w\left(r_i\right)} \tag{1}$$

which avoids thresholding on some constant number of required stations and/or phases. Here, N is the maximum number of phases that can be observed, i.e., twice the number of the stations that were operating at the time of the event, and $w(r_i)$ is a weighting function. The indicator in the numerator ($\mathbf{1}$) evaluates to 1 if the i'th phase (ϕ_i) is in the association result ($\mathcal{A}$). To downweight stations more distant from the epicenter of the grid search solution, we used the weighting function:

$$w\left(r_i\right) = \min\left(\frac{1}{r_i}, \frac{1}{R}\right) \tag{2}$$

where we set the cutoff radius (R) for constant weight to 10 km. We required the association results to have an association score of 0.3 or greater and to have phase picks from at least 3 stations where 2 of them have both P- and S-phases picked. This is a bare minimum requirement to create an overdetermined system while resolving any phase confusion by the picking algorithm, i.e., mis-identifying a P-phase as an S-phase or vice versa. We determined initial hypocentral locations with HypoInverse (Klein, 2002) and refined them with HypoDD (Waldhauser & Ellsworth, 2000). For the final locations, we used the differential travel times calculated from the phase picks supplemented with cross-correlation measurements. For the latter, we followed the approach described in (Shelly et al., 2013, 2016) and used the three-point quadratic interpolation for subsample precision and a weighting function that considers both the largest and the second largest cross-correlation coefficients to capture confidence in the measurements. We determined local magnitudes using the procedure and distance correction function reported in (Walter

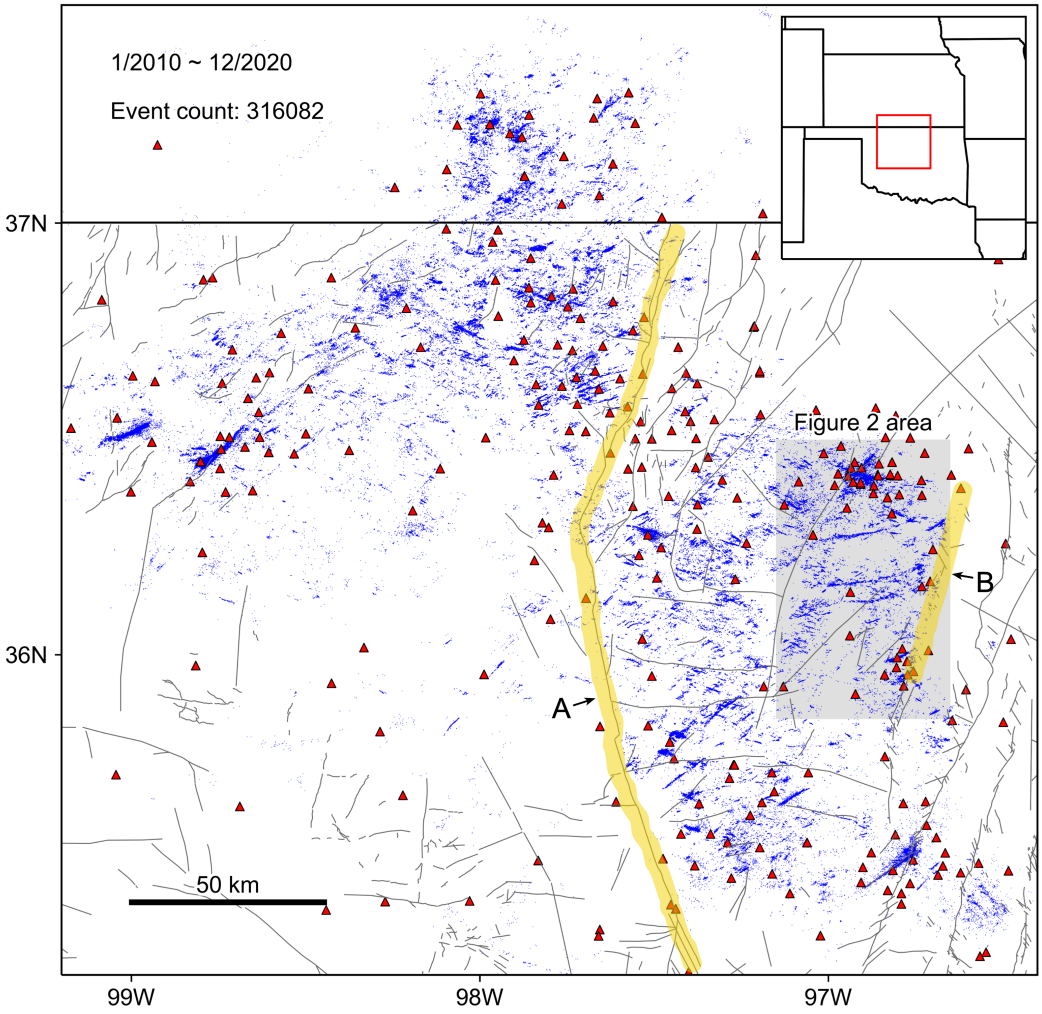

Figure 1. Map of the study area showing the faults compiled by OGS (Marsh & Holland, 2016) in black lines, seismic stations in red triangles, and epicenters of the events in our catalog in blue dots. The highlighted lines mark apparent seismicity boundaries. The area for Figure 2 is shown in grey box.

et al., 2020) but adopted moment magnitudes produced by St. Louis University (Herrmann et al., 2011) when available.

3 Results

3.1 Regional map of earthquakes

Figure 1 shows the epicenters in our catalog and the mapped faults compiled by OGS (Marsh & Holland, 2016). Equivalent maps with the events in other catalogs (Guy

et al., 2015; Schoenball & Ellsworth, 2017b; Skoumal et al., 2019) and the magnitude-frequency distributions of the catalogs are shown in the supporting information (Figures S2 through S6). A high-resolution map of the events colored by their origin time is attached as a separate file (Figure S8). As noted previously, most of the earthquakes are not associated with mapped faults (Schoenball & Ellsworth, 2017b). Instead, the event epicenters define hundreds or even thousands of discrete fault structures, as seen here in high spatial resolution. Note that the fault data were produced in 2016 and some of the faults were defined by the earthquakes that had occurred on those faults by that time. The fault that hosted the 2011 M5.6 Prague earthquake is an example.

One aspect that emerges in our catalog are clear boundaries of seismicity. The Nemaha Ridge, which is highlighted and marked with A in Figure 1, has been hypothesized as a barrier to flow (Weingarten et al., 2015) and our result strengthens the case. Specifically, the seismicity in eastern Oklahoma tends to be constrained to the east of the ridge south of 36.25 N, but to the west of the ridge farther north. Another example is marked with B in Figure 1 where the seismicity is confined to the west of the highlighted line. The origin of this boundary is not clear; however, we suspect that it too is geologically controlled since there are mapped faults with similar orientations in the vicinity of line B.

3.2 Comparing the earthquake catalogs

Figure 2 compares event epicenters between January 2010 and December 2016 in the area highlighted in Figure 1 from our study with the two previous statewide catalogs (Schoenball & Ellsworth, 2017b; Skoumal et al., 2019). We refer the catalog from Schoenball and Ellsworth (Schoenball & Ellsworth, 2017b) as SNE2017, and that from Skoumal and coworkers (Skoumal et al., 2019) as SEA2019 hereafter. The epicenters belonging to the clustered structures in SNE2017, which were identified in a separate study (Schoenball & Ellsworth, 2017a), are shown in different colors. The faults that were identified and fit in SEA2019 are plotted beneath the event epicenters. The spatial resolution of SNE2017 and SEA2019 limited the structures that can be confidently identified. The number of events is larger in SEA2019 compared to SNE2017, but they are more tightly clustered, which resulted in many small-scale structures. Our deep earthquake catalog reveals that most of these separated structures are actually located within more extensive and continuous structures.

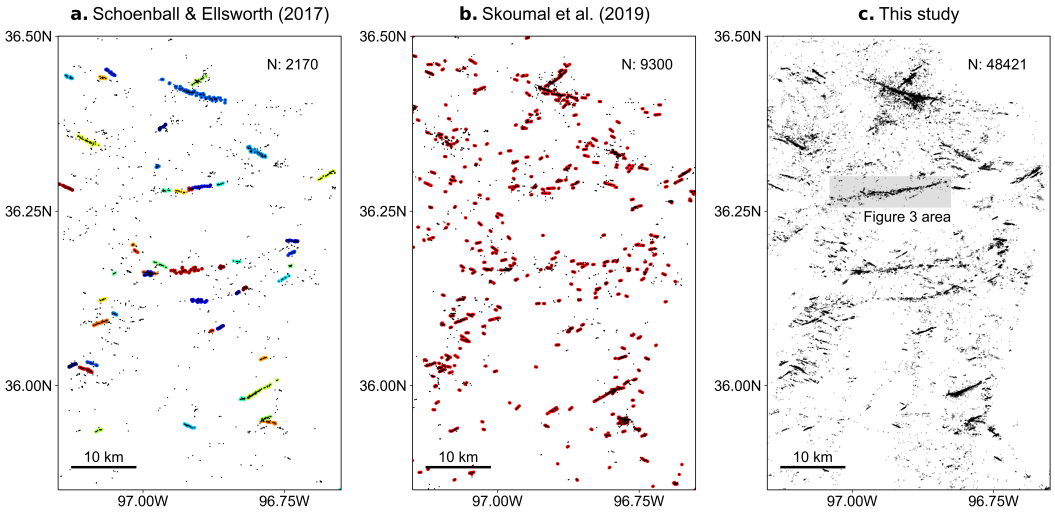

Figure 2. Comparisons of the event epicenters (black dots) among the three catalogs in the area highlighted in Figure 1. Events up to December 2016 are shown in **c** to match the date range with the other two catalogs. Events that were assigned to the identified fault structures in **a** are highlighted with colored backgrounds and the identified faults in **b** are plotted beneath the events in red. The area for Figure 3 in shown in **c**.

The spatial distribution of earthquakes seems more similar between SNE2017 and our study than SNE2017 and SEA2019. Note that SNE2017 was produced by relocating the events from a routine catalog without increasing the event count while both SEA2019 and this study increased the number of earthquakes detected to help resolve active structures. To compare the distribution of locations quantitatively we used the Chamfer Distance (CD), which in this context is defined as

$$CD(X,Y) = \frac{1}{N_x}\sum_{x\in X}\min_{y\in Y}||x-y||_2^2 + \frac{1}{N_y}\sum_{y\in Y}\min_{x\in X}||y-x||_2^2 \tag{3}$$

where x is each event epicenter in catalog X, y is each event epicenter in catalog Y, and N_x and N_y are event counts in catalog X and Y, respectively. While the absolute size of this distance does not convey much meaning, we can make relative comparisons among distances as smaller distances translate to higher similarity. When comparing a low-resolution catalog with a high-resolution catalog, the distance should decrease with increasing resolution of the high-resolution catalog because 'resolving' in this context refers to resolving the same underlying structures. Figure S7 gives more details on this logic. Using the events shown in Figure 2, the CD between SNE2017 and SEA2019, SNE2017 and our catalog, and SEA2019 and our catalog were 0.86, 0.54, and 0.66, respectively. The fact

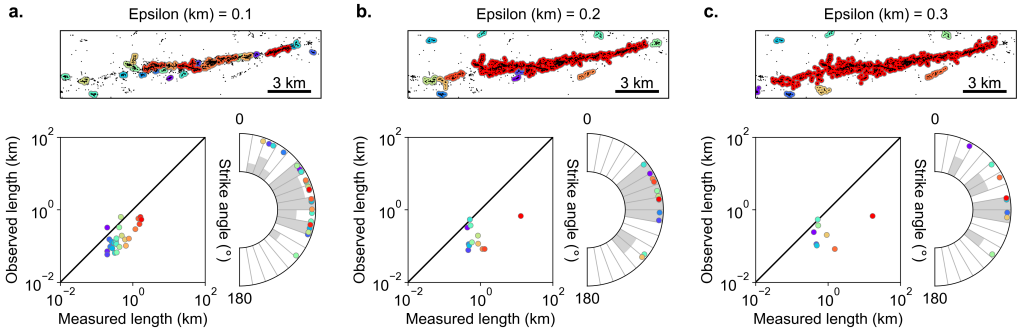

Figure 3. Clustering results of the earthquake events (up to December 2020) within the highlighted area in Figure 2-**c**. Events with the same background color belong to the same identified structure in the first row. The second row shows the distributions of measured lengths and strike angles of the identified structures. The lengths are plotted against the observed lengths, which are derived from the circular fault model (Eshelby, 1957) using the largest magnitude in each structure. Normalized histograms of the strike angles are shown in grey bars with a bin size of 10 degrees. The colors of the identified structures and the data points match in each column.

that the CDs between ours and each of the two previous catalogs are lower than the CD between the two previous catalogs indicates that the events that were added in our catalog resolved structures common to both, but previously less well illuminated. It is interesting to note that the CD between ours and SNE2017 was lower than the CD between ours and SEA2019 even though SEA2019 had over 4 times more events than SNE2017. This means that the additional events included in SEA2019 did not necessarily resolve the same underlying structures in SNE2017.

3.3 Identifying the structures

The increased spatial resolution of earthquake events in our catalog poses a challenge when defining faults from earthquake locations. Some earthquakes form distinctive clusters that make the task easy, but it becomes less obvious for other earthquakes like those inside the highlighted box in Figure 2-**c**. Previous studies defined multiple structures from these earthquakes (Figure 2-**a** and **b**), but one could possibly group all the earthquakes to form a linear trend in the box as a single structure when using our dataset (Figure 2-**c**).

To evaluate this quantitatively, we ran the DBSCAN algorithm (Ester et al., 1996) on the event epicenters as was done in the previous studies. Figure 3 shows the clustering result under three different epsilon parameters, which control the distance cutoff used to determine the neighboring events of each event. We used a fixed value of 3 for the parameter that controls the minimum number of neighbors while clustering. After clustering, we selected the clusters with aspect ratio of equal or greater than 3 and derived the length of each structure by calculating the greatest distance between any two points within the structure and the orientation using the largest eigenvector of the points within the structure. The orientations were measured in strike angles, i.e., the angle from north in the clockwise direction. For the lengths, we used the circular fault model (Eshelby, 1957) to roughly translate the largest magnitude in each cluster into a corresponding length scale. We used a constant stress drop value of 3 MPa, following (Huang et al., 2016), and defined twice the radius as the observed length.

With the lowest epsilon value among the three, the earthquakes were grouped into multiple small-scale structures, and the measured structure lengths followed the trend of the observed lengths (Figure 3-**a**). With increasing epsilon, however, separated groups began to merge, and the measured lengths became much larger than the observed lengths (Figure 3-**b**, **c**). The distribution of strike angles also varied under different epsilon values. The data occupied more than half of the angle bins when epsilon was 0.1 km while macroscopically, the events seem to form a structure with a strike angle close to 90 degrees as shown in the results with epsilon of 0.2 and 0.3 km.

4 Discussion

The catalog we developed using the DL-assisted workflow resulted in significantly more events with high precision locations than the statewide template matching catalog (SEA2019), enabling the mapping of numerous fault structures only hinted at by both SNE2017 and SEA2019. Our workflow, which requires no prior knowledge of template events, was an important factor behind this increase. Because template matching requires template events, which are only available for some stations, the SEA2019 study was unable to use all the waveform data available in the public domain at the time of their study. This template-independency also allows us to apply the workflow in real-time to increase the resolution over that of a conventionally developed earthquake catalog. Precision of the earthquake hypocenters can be managed using the near real-time double-difference

approach (Waldhauser, 2009) and the neural network for picking phase arrivals can incrementally be trained for even better performance as we collect more labeled data (Chai et al., 2020).

The catalog illuminates previously hidden fault structures in Oklahoma and Southern Kansas and can be used to derive fault attributes such as dimensions and orientations for further analysis. However, we showed that a simple clustering algorithm such as DBSCAN that is purely based on earthquake locations has limitations and the statistics of the identified faults derived from it can be very sensitive to the parameters. Future work is needed on algorithms for identifying the faults from seismicity distributions. These algorithms should be robust against missing events and changes in relative event locations. Using a probabilistic approach such as sampling the events before clustering and quantifying the uncertainties of the derived fault attributes is another option.

5 Conclusion

Developing a high-resolution earthquake catalog has become faster and easier due to the application of deep learning algorithms for seismic phase arrival time picking. Through our case study of a decade of Oklahoma and Southern Kansas seismicity, we found this workflow provides significant improvements over the existing catalogs. The newly identified seismicity illuminates numerous previously unseen fault structures and sharpens the definition of those previously revealed.

Acknowledgments

This work is supported by the Stanford Center for Induced and Triggered Seismicity (SCITS) and the Department of Energy (Basic Energy Sciences; Award DE-SC0020445). Data for the 17 public seismic networks were downloaded through `https://www.iris.edu/`. The links to each seismic network data are 4H: `https://doi.org/10.7914/SN/4H_2014`, 9L: `https://doi.org/10.7914/SN/9L_2013`, GS: `https://doi.org/10.7914/SN/GS`, N4: `https://doi.org/10.7914/SN/N4`, NP: `https://doi.org/10.7914/SN/NP`, NQ: `https://doi.org/10.7914/SN/NQ`, NX: `https://doi.org/10.7914/SN/NX`, O2: `https://doi.org/10.7914/SN/O2`, OK: `https://doi.org/10.7914/SN/OK`, TA: `https://doi.org/10.7914/SN/TA`, US: `https://doi.org/10.7914/SN/US`, XR: `https://doi.org/10.7914/SN/XR_2016`, Y7: `https://doi.org/10.7914/SN/Y7_2016`, Y9: `https://doi.org/10.7914/SN/Y9_2016`, ZD: `https://doi.org/10.7914/SN/ZD_2014`, ZP: `https://`

doi.org/10.7914/SN/ZP_2016, and ZQ: https://doi.org/10.7914/SN/ZQ_2011. The earthquake catalog produced from this study is included in the supporting information (Table S1).

References

Angelier, J. (1979). Determination of the mean principal directions of stresses for a given fault population. *Tectonophysics*, *56*(3-4), T17–T26.

Chai, C., Maceira, M., Santos-Villalobos, H. J., Venkatakrishnan, S. V., Schoenball, M., Zhu, W., ... Team, E. C. (2020). Using a deep neural network and transfer learning to bridge scales for seismic phase picking. *Geophysical Research Letters*, *47*(16), e2020GL088651.

Darold, A. P., Holland, A. A., Morris, J. K., & Gibson, A. R. (2015). Oklahoma earthquake summary report 2014. *Okla. Geol. Surv. Open-File Rept. OF1-2015*, 1–46.

Eshelby, J. D. (1957). The determination of the elastic field of an ellipsoidal inclusion, and related problems. *Proceedings of the royal society of London. Series A. Mathematical and physical sciences*, *241*(1226), 376–396.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).

Guy, M. R., Patton, J. M., Fee, J., Hearne, M., Martinez, E. M., Ketchum, D. C., ... others (2015). *National earthquake information center systems overview and integration*. US Department of the Interior, US Geological Survey.

Herrmann, R. B., Benz, H., & Ammon, C. J. (2011). Monitoring the earthquake source process in north america. *Bulletin of the Seismological Society of America*, *101*(6), 2609–2625.

Huang, Y., Beroza, G. C., & Ellsworth, W. L. (2016). Stress drop estimates of potentially induced earthquakes in the guy-greenbrier sequence. *Journal of Geophysical Research: Solid Earth*, *121*(9), 6597–6607.

Johnson, C. E., Lindh, A., & Hirshorn, B. (1997). Robust regional phase association.

Klein, F. W. (2002). *User's guide to hypoinverse-2000, a fortran program to solve for earthquake locations and magnitudes* (Tech. Rep.). US Geological Survey.

Liu, M., Zhang, M., Zhu, W., Ellsworth, W. L., & Li, H. (2020). Rapid characterization of the july 2019 ridgecrest, california, earthquake sequence from raw seismic data using machine-learning phase picker. *Geophysical Research Letters*, *47*(4), e2019GL086189.

Marsh, S., & Holland, A. (2016). Comprehensive fault database and interpretive fault map of oklahoma. *Oklahoma Geol. Surv. Open-File Rep. OF2-2016, Oklahoma Geological Survey, Norman, OK*.

Park, Y., Mousavi, S. M., Zhu, W., Ellsworth, W. L., & Beroza, G. C. (2020). Machine-learning-based analysis of the guy-greenbrier, arkansas earthquakes: A tale of two sequences. *Geophysical Research Letters*, *47*(6), e2020GL087032.

Schoenball, M., & Ellsworth, W. L. (2017a). A systematic assessment of the spatiotemporal evolution of fault activation through induced seismicity in oklahoma and southern kansas. *Journal of Geophysical Research: Solid Earth*, *122*(12), 10–189.

Schoenball, M., & Ellsworth, W. L. (2017b). Waveform-relocated earthquake catalog for oklahoma and southern kansas illuminates the regional fault network. *Seismological Research Letters*, *88*(5), 1252–1258.

Shelly, D. R., Ellsworth, W. L., & Hill, D. P. (2016). Fluid-faulting evolution in high definition: Connecting fault structure and frequency-magnitude variations during the 2014 long valley caldera, california, earthquake swarm. *Journal of Geophysical Research: Solid Earth*, *121*(3), 1776–1795.

Shelly, D. R., Moran, S. C., & Thelen, W. A. (2013). Evidence for fluid-triggered slip in the 2009 mount rainier, washington earthquake swarm. *Geophysical Research Letters*, *40*(8), 1506–1512.

Skoumal, R. J., Kaven, J. O., & Walter, J. I. (2019). Characterizing seismogenic fault structures in oklahoma using a relocated template-matched catalog. *Seismological Research Letters*, *90*(4), 1535–1543.

Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., ... Segou, M. (2021). Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 central italy sequence. *The Seismic Record*, *1*(1), 11–19.

Waldhauser, F. (2009). Near-real-time double-difference event location using long-term seismic archives, with application to northern california. *Bulletin of the*

Seismological Society of America, *99*(5), 2736–2748.

Waldhauser, F., & Ellsworth, W. L. (2000). A double-difference earthquake location algorithm: Method and application to the northern hayward fault, california. *Bulletin of the seismological society of America*, *90*(6), 1353–1368.

Walsh III, F. R., & Zoback, M. D. (2016). Probabilistic assessment of potential fault slip related to injection-induced earthquakes: Application to north-central oklahoma, usa. *Geology*, *44*(12), 991–994.

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., Woelfel, I., Chang, J. C., ... Holland, A. A. (2020). The oklahoma geological survey statewide seismic network. *Seismological Research Letters*, *91*(2A), 611–621.

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

Yehya, A., Yang, Z., & Rice, J. R. (2018). Effect of fault architecture and permeability evolution on response to fluid injection. *Journal of Geophysical Research: Solid Earth*, *123*(11), 9982–9997.

Zhang, M., Ellsworth, W. L., & Beroza, G. C. (2019). Rapid earthquake association and location. *Seismological Research Letters*, *90*(6), 2276–2284.

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.