JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH001194

Transfer Learning and Benchmarking for Induced Seismic Event Detection: Insights From Oklahoma

Hongyu Xiao[1], Jacob I. Walter[1], Paul Ogwari[1], Long M. Ho[2], Andrew D. Thiel[1], Nicholas Gregg[1], Brandon Mace[1], and Isaac Woelfel[1]

[1]Oklahoma Geological Survey, University of Oklahoma, Norman, OK, USA, [2]The University of Alabama, Tuscaloosa, AL, USA

Key Points:

- A high quality, manually curated data set provides a benchmark for machine learning detection of induced earthquakes
- Fine-tuned models with this data set significantly improve detection of small, induced earthquakes, even with limited localized data
- The best fine-tuned model recovered almost all cataloged events in 2022 and identified thousands of additional earthquakes

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

H. Xiao,
Hongyu.Xiao-1@ou.edu

Citation:

Xiao, H., Walter, J. I., Ogwari, P., Ho, L. M., Thiel, A. D., Gregg, N., et al. (2026). Transfer learning and benchmarking for induced seismic event detection: Insights from Oklahoma. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH001194. https://doi.org/10.1029/2025JH001194

Received 20 DEC 2025
Accepted 19 JUN 2026

Author Contributions:

Conceptualization: Hongyu Xiao, Jacob I. Walter
Data curation: Hongyu Xiao, Jacob I. Walter, Andrew D. Thiel, Nicholas Gregg, Brandon Mace, Isaac Woelfel
Formal analysis: Hongyu Xiao, Jacob I. Walter, Paul Ogwari
Funding acquisition: Jacob I. Walter

Abstract Machine learning models for microseismicity detection are often limited by the scarcity of large and high-quality labeled data sets in many regions. To address this need, we introduce the Oklahoma Labeled AI Dataset (OKLAD), a manually curated data set compiled by the Oklahoma Geological Survey (OGS). OKLAD is designed to support studies of induced seismicity and serves as a benchmark for evaluating deep-learning detection models in local and regional monitoring contexts. Using OKLAD, we fine-tuned several established phase-picking models and observed substantial improvement in local and regional detection. The best performing model achieved recalls of 91.1% for first arrival P- detection and 89.8% for first arrival S-wave detection. Validating this model on continuous data in Oklahoma, we recovered 96.8% of the OGS-cataloged events and identified 146.8% more events after associative comparison with the events reported by routine network operations. Comparable improvements were also observed when applying the best performing models to other induced seismicity settings, such as west Texas. These results establish OKLAD as a benchmark data set for induced seismicity and demonstrate the effectiveness of transfer learning for improving regional seismic monitoring. This approach provides a replicable framework for enhancing microseismicity detection in challenging environments and can be extended to other regions where generalized deep-learning pickers may underperform.

Plain Language Summary Human activities such as wastewater injection and hydraulic fracking can trigger earthquakes. In the southern midcontinent of the U.S., such induced earthquakes have become more prominent in the last decade. However, the ground vibration signals of these from small and induced earthquakes are often buried in environmental noise, making them difficult to detect using traditional methods or by through visual inspection from seismic analysts. Recently, machine learning models have shown great promise in improving the ability to detect small earthquakes. In this study, we fine-tuned machine learning models using an Oklahoma-focused data set (OKLAD) and compared their performance to existing models. Our results show that even with a relatively small amount of localized data, the models' ability to detect small earthquakes can be substantially improved. Applying the best model to 2022 data, we recovered nearly all previously cataloged events while identifying more than 4,400 additional events. These results show that OKLAD can serve as a benchmark for earthquake detection and that transfer learning is an effective strategy for applying machine learning across regions. We are releasing both our trained models and the data set to the public to support future research on induced earthquakes.

1. Introduction

Over the past decade, Oklahoma and surrounding regions have experienced a significant increase in seismic activity, primarily due to human activities such as wastewater injection, hydraulic fracturing, and other industrial operations (e.g., Ellsworth, 2013; Frohlich et al., 2016; Van der Baan & Calixto, 2017). Specifically, the seismicity rate has risen sharply since 2010 (Walter et al., 2020). In 2015, the highest recorded monthly total of earthquakes exceeded 1,000, including 903 events of magnitude 3.0 or greater (Figure 1). Although the seismicity rate declined after the peak, seismicity remained elevated relative to pre-2010 levels.

Most induced earthquakes in Oklahoma occur at shallower depths, mainly in the sedimentary layers or the upper crust, and are caused by human activities (e.g., Keranen et al., 2014; Rubinstein & Mahani, 2015). These events tend to be lower-magnitude, but some moderate earthquakes have occurred that caused damage to nearby structures, such as the 2011 Prague M5.7 and 2016 Pawnee M5.8 earthquakes (Walter et al., 2020). Although induced earthquakes are often small and difficult to detect and identify, they pose notable risks due to their

Investigation: Hongyu Xiao, Jacob I. Walter, Paul Ogwari, Andrew D. Thiel, Nicholas Gregg
Methodology: Hongyu Xiao, Jacob I. Walter, Long M. Ho
Project administration: Hongyu Xiao, Jacob I. Walter
Resources: Hongyu Xiao, Jacob I. Walter, Paul Ogwari
Software: Hongyu Xiao, Jacob I. Walter, Long M. Ho
Supervision: Jacob I. Walter
Validation: Hongyu Xiao, Jacob I. Walter
Visualization: Hongyu Xiao
Writing – original draft: Hongyu Xiao
Writing – review & editing: Hongyu Xiao, Jacob I. Walter, Paul Ogwari, Long M. Ho, Andrew D. Thiel, Nicholas Gregg, Brandon Mace, Isaac Woelfel

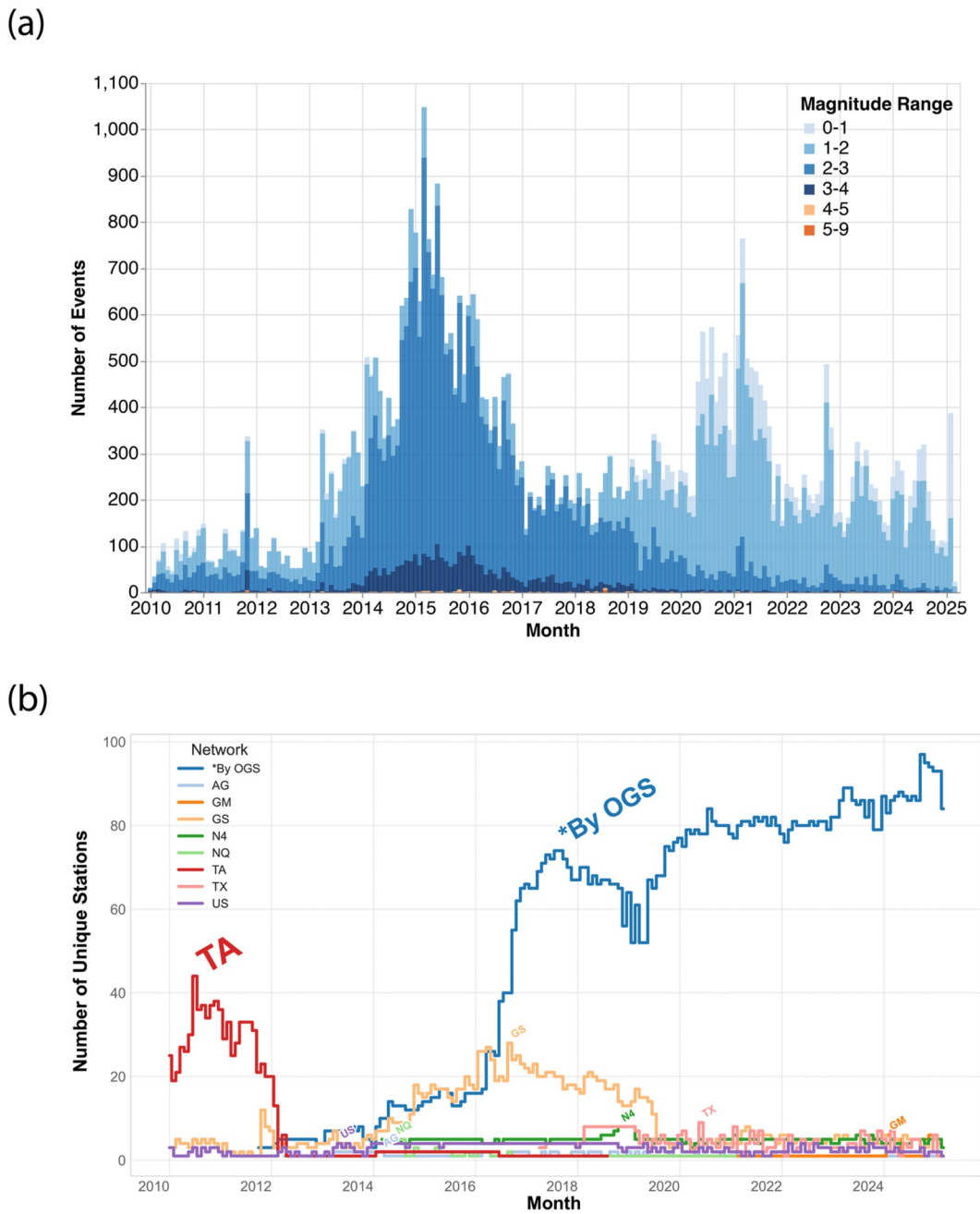

Figure 1. (a) Monthly earthquake event distribution in OKLAD from 2010 to 2024. Each bar represents the total count of events within that month. Each color presents the corresponding magnitude ranges. (b) Number of monitoring stations in OKLAD from 2010 to 2024, categorized by network.

proximity to the surface, which increases potential for infrastructure damage and amplifies public concern (Walter et al., 2017; Yeck et al., 2017). In some cases, induced seismicity can also occur at distances of 10–40 km from the injection zone (Goebel, Walter, et al., 2017, Goebel, Weingarten, et al., 2017).

Looking ahead, the anticipated growth in subsurface injection activities for applications such as carbon capture, utilization, and storage (CCUS) raises additional concerns about seismic hazard (Cappa & Rutqvist, 2011; Shapiro et al., 2007). In the continental US, many of the regional seismic network infrastructure relies upon a patchwork of regional networks that contribute catalog entries to the USGS Comprehensive Catalog (ComCat) or rely upon the NEIC to monitor areas that lack such an infrastructure for earthquake M2.5 or greater (Guy et al.,

2015). Previous studies demonstrate that timely operational adjustments can significantly mitigate earthquake hazards, including through the use of small magnitude earthquakes that are often unavailable in standard catalog products (Murray et al., 2023). For example, reducing injection rates following elevated seismicity has been shown to lower the aftershock production rates of major induced seismic events (Goebel et al., 2019). Meanwhile, shallower injection depth and reduced volume can decrease the seismicity rate (Murray et al., 2023; Skoumal et al., 2024). These findings highlight that rapid earthquake detection is not simply a scientific necessity but a key component of proactive seismic risk management. Consequently, regions like Oklahoma require more efficient and accurate monitoring systems to support both hazard mitigation, public safety, and regulatory compliance.

Traditional earthquake detection methods often struggle to identify and detect small, induced events in the southern midcontinent. Seismic records of these events are often masked by environmental and anthropogenic noise, resulting in low signal-to-noise ratios that hinder event detection and increase the likelihood of misclassification. In addition, the transient and spatially shifting nature of induced seismicity complicates long-term monitoring. Permanent seismic networks are often sparse and absent in induced regions and thus rely upon temporary seismic network deployments (e.g., Walter et al., 2018). However, these deployments are typically in an ad hoc and reactive manner—often after seismic activity has already begun—rather than through systematic, proactive planning. As a result, the temporary networks often have uneven spatial coverage and limited station density. Compared to tectonic events that occur around a dense network of sensors, induced earthquakes are typically recorded by fewer stations, further reducing the network's detection and location capability. Taken together, these factors make the reliable detection of induced seismicity a uniquely challenging problem.

Given the challenges associated with detecting small, induced earthquakes, there is growing interest in applying advanced detection methods that can overcome the limitations of traditional approaches. In recent years, machine learning (ML) models have shown substantial promise in improving the detection of low-magnitude seismic events, particularly those that are often missed by conventional techniques. Most recent advances in ML have largely been driven by the development of convolutional neural networks (LeCun & Bengio, 1995; Perol et al., 2018). Popular picker models such as Generalized Seismic Phase Detection (GPD) (Ross et al., 2018), PhaseNet (Zhu & Beroza, 2019), and EQTransformer (Mousavi et al., 2020) have achieved high precision in detecting seismic phases and events, particularly when trained on large, high-quality data sets from tectonically active regions such as California or Japan. However, systematic evaluations of these three models in the induced seismicity-dominant region remain limited. Several studies have indicated that ML models trained on global data sets experience degraded performance when applied to different geological settings (Jiang et al., 2021; Zhao & Chen, 2021). Factors including crustal velocity structures, station distributions, environmental noise situations, and instrument variations can alter waveform characteristics in ways that challenge the generalizability of pre-trained models (Mousavi & Beroza, 2022; Zhao et al., 2023). Consequently, induced earthquakes, which frequently occur in regions with sparse seismic networks and noisier environments, are particularly challenging to detect using off-the-shelf models trained on other regional or global data.

Transfer learning, in which knowledge learned from one data set or region is adapted (fine-tuned) to another, has emerged as a promising strategy to address this limitation of pre-trained models in new geological settings. Studies have demonstrated that fine-tuning pre-trained models using region-specific data sets can substantially improve detection accuracy in local geological settings (Chen et al., 2024; Saad et al., 2023). Chen et al. (2024) used the Texas Earthquake Data set for AI (TXED) to improve phase-picking performance in Texas. By leveraging fine-tuning, they significantly increased accuracy for P-phase and S-phase picks, achieving notable improvements over baseline models trained on a global data set (STEAD, Stanford Earthquake Dataset by Mousavi et al. (2019)). Despite these advances, few studies to date have evaluated the effectiveness of fine-tuning for induced seismicity in Oklahoma, where seismicity characteristics and network conditions differ from those in tectonically active regions.

In this work, we benchmark three widely used machine learning models—PhaseNet, EQTransformer, and GPD—using our newly curated data set, the Oklahoma Labeled AI Dataset (OKLAD), derived from the Oklahoma Geological Survey (OGS) earthquake catalog (Figure 1). We then fine-tune these models using OKLAD and evaluate their detection performance relative to their pre-trained counterparts using standard statistical metrics against the analyst-reviewed OGS catalog. The best-performing model is subsequently applied to continuous waveform data from 2022 in Oklahoma to assess its operational potential.

10.1029/2025JH001194

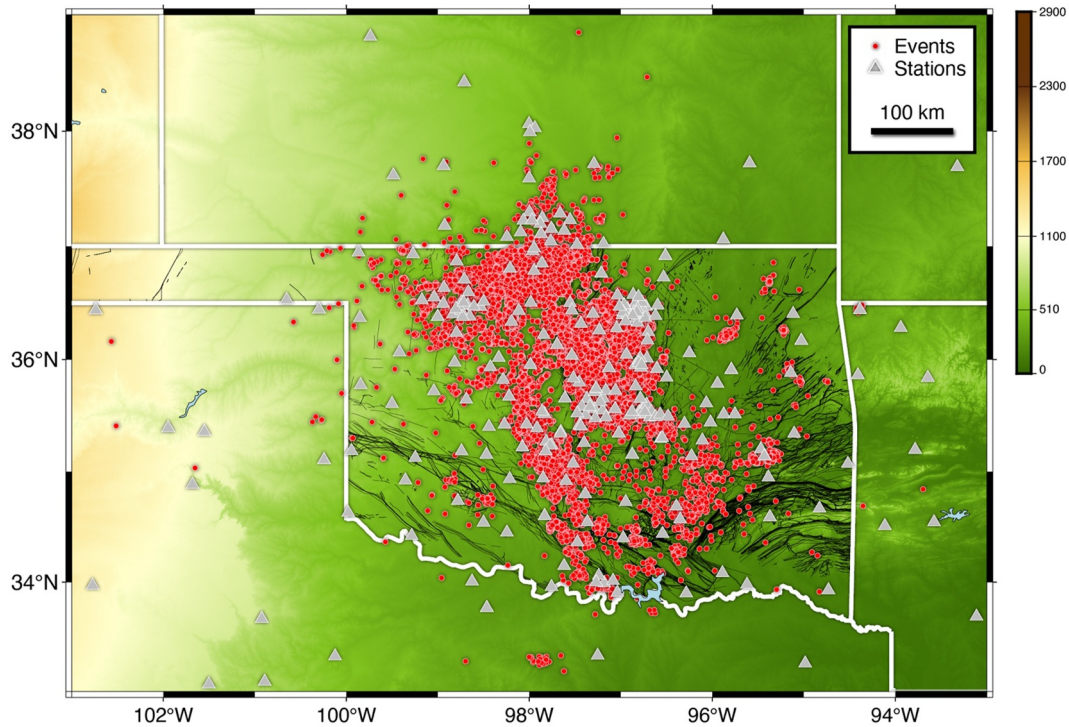

Figure 2. Spatial distributions of seismic events and stations included in the Oklahoma Labeled AI Dataset (OKLAD). Seismic stations are indicated by grey triangles, and earthquake epicenters are marked by red dots. The black line delimits major fault lines in Oklahoma (Marsh & Holland, 2016). The white solid line delimits the state boundary.

To support model evaluation in an induced seismicity-dominated region, OKLAD provides a quality-controlled, region-specific data set spanning 2010–2024. It contains a high proportion of small magnitude, shallow events, offering a benchmark for training and assessing machine-learning models in settings with elevated noise and rapid temporal variability.

To promote reproducibility and facilitate future research on induced earthquakes, we publicly released OKLAD, our fine-tuned PhaseNet, EQTransformer, and GPD models, and all associated training and evaluation code. Our results show that even modest amount of localized labeling can substantially improve model performance, highlighting the value of fine-tuning for induced seismicity monitoring. More broadly, this work provides a scalable pathway for adapting machine learning detection systems to regions affected by fluid injection, hydraulic fracturing, or emerging subsurface activities such as carbon capture, utilization, and storage (CCUS) projects.

2. Data

The Oklahoma Labeled AI Dataset was curated from the Oklahoma Geological Survey earthquake catalog from 2010 to 2024. Figure 2 shows the locations of 52,193 events. In total, 1,139,808 traces were used to construct this data set over 14 years (2010–2024). Events were selected from an analyst-reviewed earthquake catalog, which provides high-quality hypocenter locations, magnitudes, and manual phase picks. Each event underwent consistent quality control, and events with poorly constrained hypocenters or uncertain magnitudes were excluded.

While most data were from stations managed by the OGS or located in Oklahoma, the data set is further enhanced by the inclusion of data from other nearby networks. In total, 311 seismic stations (comprising temporary, long-term and permanent seismometers) distributed across Oklahoma and neighboring regions of Kansas, Missouri, Texas and Arkansas contributed to this data set (Figure 2). Nearby stations were included to improve event detection, association, and location accuracy. Waveforms in OKLAD are rotated into the ENZ coordinate system, and all the continuous data are also archived at the Earthscope SAGE Data Management Center (formerly IRIS -

Table 1
Summary of P-Phase and S-Phase Picks in the Oklahoma Labeled AI Dataset (OKLAD)

Category	Count	Percentage of total data set (%)
Total P Picks	1,075,795	94.38
Total S Picks	1,061,352	93.12
Both P & S Picks	997,339	87.50
Only P Picks	78,456	6.88
Only S Picks	64,013	5.62

Incorporated Research Institutions for Seismology). As data collection is ongoing, this data set will be regularly updated by the OGS. The data set sampling rate is resampled to 100 Hz as needed, and the length of each trace is constrained to 120 s, which includes a window of 60 s before and 60 s after the cataloged event origin time, resulting in standardized input lengths suitable for machine learning models. The data set predominantly contains earthquake events (local induced seismicity) and no noise-only traces are included.

A total of 1,075,795 traces (94.38% of the data set) contain P arrival picks, and a total of 1,061,352 traces (93.12%) contain S arrival picks. Both P and S arrivals were present in 997,339 traces (87.5%). Traces with only P arrival picks are 78,456 (6.88%), and those with only S picks are 64,013 (5.62%). Figure S1 in Supporting Information S1 and Table 1 summarize the distribution of P- and S- arrival picks in the OKLAD. For consistency, we use the simplified "P" and "S" denotation, and most of the P arrival picks are Pg phase rather than Pn in the data set. OKLAD includes only labeled P and S arrival traces without any noise-only traces. However, noise traces can be easily generated by extracting waveform segments preceding the first arrival as needed.

OKLAD data collection primarily uses high-broadband (HH), broadband (BH), and extremely short-period (EH) channel data from stations mostly within the state of Oklahoma. Specifically, 78.3% of the data comes from HH channels, 13.4% from EH channels, 6.8% from BH channels, and 1.6% from other types of channels. The majority of the data set originates from high-rate high-gain broadband channels (Figure 4d). Metadata in OKLAD follows a format similar to STEAD, with a detailed description of the data attributes, along with one example provided in Table S1 of Supporting Information S1.

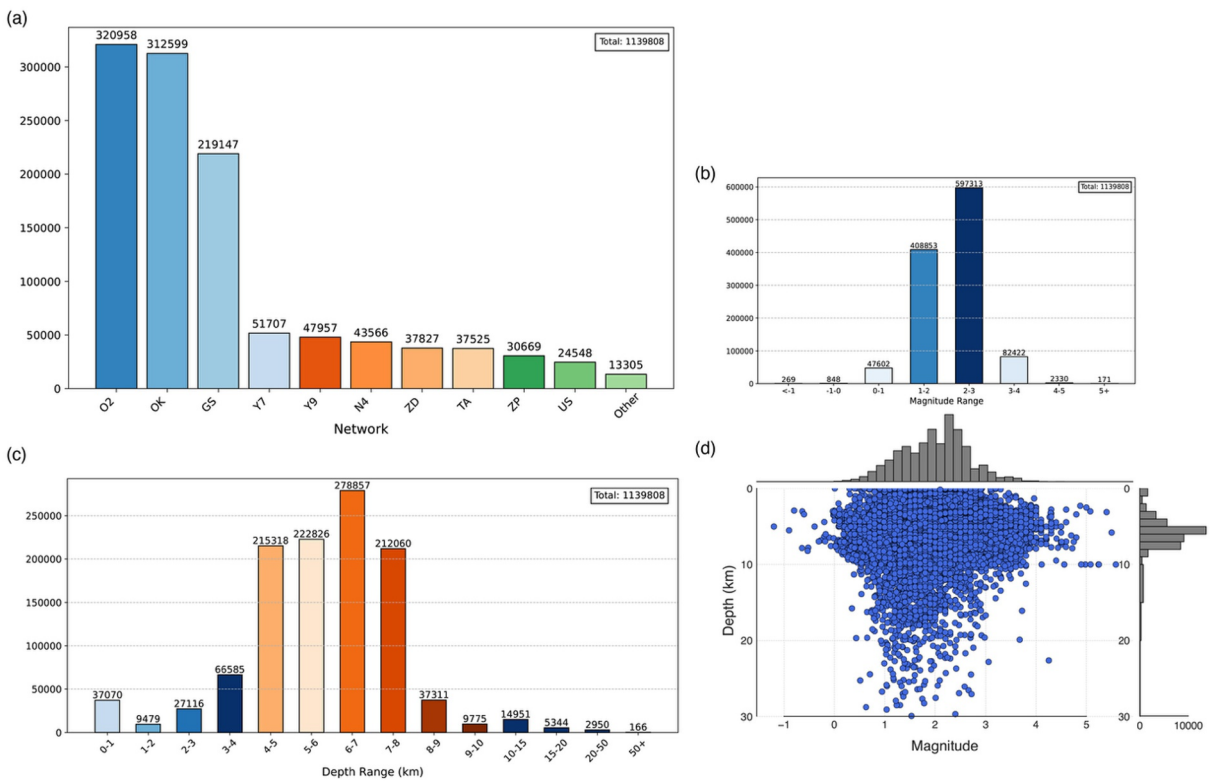

Figure 3. OKLAD network, magnitude and depth distribution. For (a, b, c), the vertical axis denotes the total trace count. (a) Network distribution of traces contributing to the OKLAD. (b) Distribution of magnitude associated with the traces in the OKLAD. (c) Depth distribution of traces comprising the OKLAD. (d) Joint distribution of depth and magnitude of traces in the OKLAD, with corresponding marginal histograms shown along each axis.

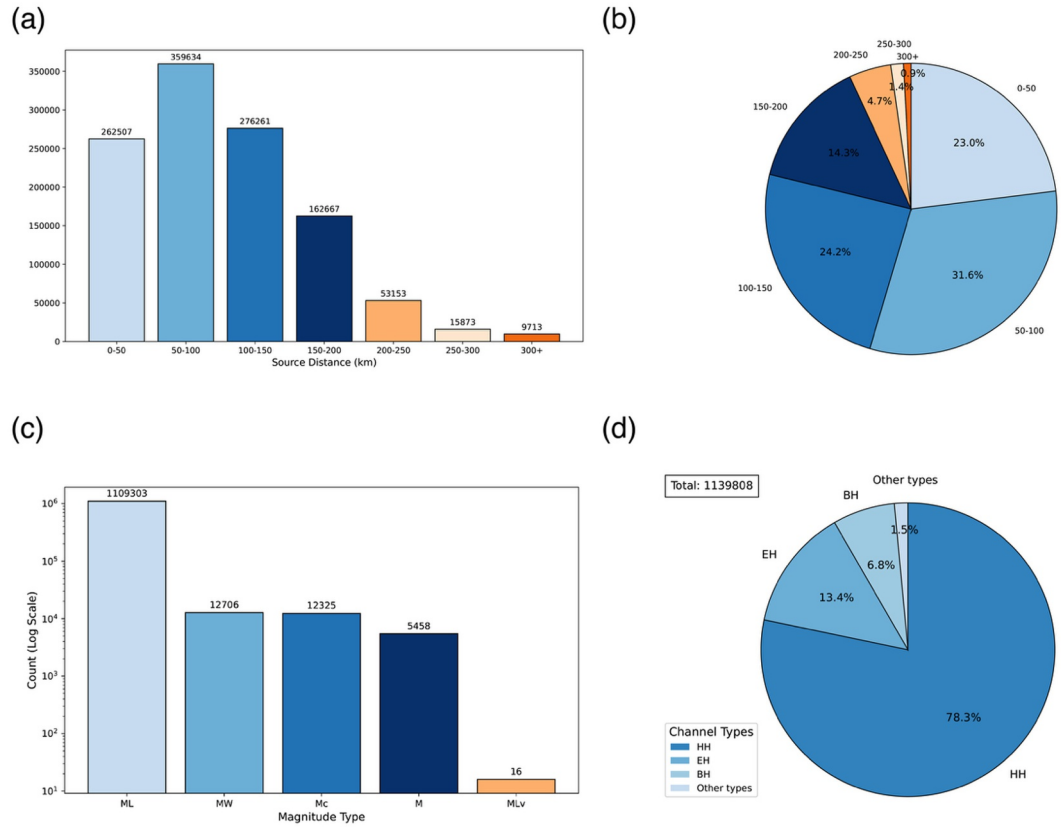

Figure 4. OKLAD epicentral distance, magnitude type and channel distribution. (a) Epicentral distance of traces contributing to the OKLAD; the vertical axis denotes the total trace count. (b) Percentage distribution of epicentral distance associated with the traces in the OKLAD. (c) Magnitude type distribution of traces in the OKLAD; the vertical axis denotes the total trace count in log scale. (d) Percentage distribution of channels of traces in the OKLAD.

In total, data in OKLAD were collected across 15 networks spanning Oklahoma and neighboring regions. Major contributors include the Oklahoma Consolidated Temporary Seismic Networks, USGS-operated networks, the USArray Transportable Array, Oklahoma Seismic Network and several smaller regional deployments. (Albuquerque Seismological Laboratory/USGS, 1980, 1990, 2013; U.S. Geological Survey, 1989, 2016; Bureau of Economic Geology, 2016; Chang, 2016; Chen et al., 2016; Darold, 2014; IRIS Transportable Arry, 2003; Nakata, 2016; Ogwari & Walter, 2023; Oklahoma Geological Survey, 1978, 2018) The Y9, ZD, ZP, and Y7 network represented temporary deployments originally established and maintained by OGS. These deployments were later merged under the O2 network code as a consolidation of temporary networks; O2, which continues to stream public data. Detailed station information, including station, network codes and locations, is provided in Table S2 of Supporting Information S1.

The traces in OKLAD originate from the networks described above (Figure 3a). Specifically, major contributions come from O2 (320,958 traces), OK (312,599), and GS (219,147). Additional contributions include Y7 (51,707), Y9 (47,957), N4 (43,566), ZD (37,827), TA (37,525), ZP (30,669), and US (24,548) traces.

In terms of magnitude distribution (Figure 3b), most traces in OKLAD correspond to earthquakes with magnitudes between 2 and 3 (597,313 traces, 52.4%), followed by magnitudes between 1 and 2 (408,853 traces, 35.9%) and magnitudes between 3 and 4 (82,422 traces, 7.2%). Smaller subsets include 47,602 traces with magnitudes between 0 and 1 (4.2%), 2,330 traces between 4 and 5, and 171 traces exceeding magnitude 5; the latter two categories together make up less than 0.2% of the data set. Very small earthquakes (<0) are also included. The relatively small number of these very low-magnitude events reflects the challenges in detecting and labeling microseismicity due to both analyst limitations and higher background noise obscuring body-wave onsets. This characteristic makes OKLAD particularly valuable for training and fine-tuning machine-learning models, as it captures the subtle signals necessary for improving microseismicity detection across regions.

The depth distribution in OKLAD (Figure 3c) shows that the data set is primarily shallow earthquakes, particularly within the 4–8 km range. The 6–7 km depth interval contains the largest number of traces (>227,000), followed by the 5–6 km (222,826 traces), 4–5 km (215,318 traces), and 7–8 km (213,060 traces) intervals. Figure 3d displays the depth–magnitude distribution. Event counts decrease substantially below 8 km, and earthquakes deeper than 20 km are very limited in the data set. Many earthquakes in this data set also predate unconventional drilling and wastewater disposal, indicating a mix of natural and induced activities. Because the data set is primarily composed of shallow crustal events, it provides a detailed representation of seismic characteristics relevant to induced earthquakes.

Figures 4a and 4b show the distribution of epicentral distances in our data set. Notably, the data set is composed mostly of events with epicentral distances ranging from 0 to 200 km (Figure 4b). The 50–100 km epicenter distance range has the highest count, with 359,634 traces, accounting for 31.6% of the data set. The 0–50 km and 100–150 km ranges follow closely, with 262,507 (23.0%) and 276,261 (24.2%) traces, respectively. Beyond 150 km, the number of traces decreases, but the 150–200 km range still contains 162,667 traces, making up 14.3% of the data set. For epicenter distances greater than 200 km, the total number of traces decreases further. There are 53,153 traces in the 200–250 km range (4.7% of the total), 15,873 traces in the 250–300 km range, and 9,713 traces at distances greater than 300 km. Overall, this data set focuses on shorter epicentral distances, particularly within 200 km (Figures 4a and 4b). This distribution reflects the strong regional character of the data set, with most events recorded at local to near-regional distances.

In May 2019, the OGS became a Tier 1 member of the USGS Advanced National Seismic System (ANSS). Since then, all events from routine operations of the OGS network have been the authoritative network within Oklahoma, and all earthquake products that originate from routine operations of the network are cataloged in the national ComCat database. OGS routine monitoring utilizes the SeisComP realtime monitoring software with an automatic trigger for the rapid initial detection. While some high-quality automatic or preliminary events are posted to ComCat, preliminary picks are reviewed and verified by experienced seismic analysts who manually confirm each event (Walter et al., 2020). This two-step process serves as a quality control measure, ensuring consistency and accuracy in labeling and reporting. Although the labeling procedures have remained consistent over the years, slight variations in labeling decisions may occur due to differences in individual analyst interpretations.

In 2020, OGS implemented the easyQuake workflow, incorporating the GPD deep-learning picker to augment routine monitoring (Walter et al., 2021). While the implementation of easyQuake augmented the automatic detection of seismic events, all events were manually reviewed by analysts before public release, and easyQuake doubled the number of seismic events. This system of catalog augmentation through offline deep learning picker detection has been in operation since May 2020. This led to an increase in the number of smaller events in the data set starting around mid-2020 in (Figure 1, Figure S2 in Supporting Information S1). Network geometry augmentation and a growth in state support to the network, including a $4 million capital investment in 2024, has allowed our group to significantly expand network coverage and install new equipment, so the OKLAD data set will continue to expand in size and geographic coverage.

After preparing event waveform and curation of associated metadata, the entire OKLAD is compiled in HDF5 format using the H5Py Python library (Collette et al., 2023). Each three-component waveform trace is structured as a 12,000 × 3 NumPy array under the ‘data' key of the HDF5 group, allowing for further trimming to accommodate different model input requirements. Metadata, including station and event information, are stored as HDF5 attributes. Examples of attribute details are listed in Table S1 of Supporting Information S1 and closely follow the STEAD practice.

3. Method

We evaluated three widely used machine learning models for seismic phase detection: GPD, PhaseNet, and EQTransformer. Each model employs a different architecture and approach to phase picking, reflecting different trade-offs in precision, computational cost, and temporal resolution.

Generalized Phase Detection (GPD) (Ross et al., 2018) employs a convolutional neural network (CNN) with a fully connected output layer, trained on millions of labeled waveforms primarily from southern California. It operates on short, fixed length 4-s windows (400 samples at 100 Hz), producing a single probability for each

Table 2
Summary of PhaseNet, EQTransformer and GPD Model Details and Training Configurations

	PhaseNet	EQTransformer	GPD
Architecture	U-Net CNN	CNN + LSTM + attention	CNN + FCNN
Original training data	Northern California	Global	Southern California
Input trace length	3,001 samples	6,000 samples	400 samples
Sampling rate	100 Hz	100 Hz	100 Hz
Pre-trained weights used	Yes (5 sets)	Yes	Yes
Loss function	Cross-entropy	Binary cross-entropy (weighted)	Vector cross-entropy
Batch size	64	256	64
Learning rate	0.01	0.001	0.001
Early stopping	Yes (5 epochs patience)	Yes (5 epochs patience)	Yes (5 epochs patience)
Learning rate scheduler	ReduceLROnPlateau	ReduceLROnPlateau	ReduceLROnPlateau
Data augmentation	Windowing + normalization	Windowing + normalization	Windowing + normalization
Labeling strategy	Probabilistic Gaussian	Probabilistic + event label	ProbabilisticPointLabeller
Output classes	P and S probability curves	P, S, probability curves, event curve	3-class probability (P, S, noise)
Evaluation metric	Residuals ± 0.6 s precision	Residuals ± 0.6 s precision	Residuals ± 0.6 s precision

class—P-phase, S-phase, or noise—per window. Its lightweight design makes it highly computationally efficient, enabling rapid training and inference, and its robustness to high noise levels makes it suitable for real-time processing of large catalogs.

PhaseNet (Zhu & Beroza, 2019) utilizes a U-Net-based (Ronneberger et al., 2015) fully convolutional neural network, analyzing a longer 30-s window (3,001 samples at 100 Hz) to generate per-sample probability curves for P, S, and noise classes. This approach enables highly precise arrival picking and strong performance even in noisy data. PhaseNet was trained on a diverse data set of Northern California events (NCEDC, 2014) and generalizes reasonably well to unseen stations and provides good temporal resolution.

EQTransformer (Mousavi et al., 2020) combines convolutional layers, attention mechanisms (Vaswani et al., 2017), and LSTM layers to capture both local waveform features and long-term temporal dependencies. EQTransformer takes 60-s input windows (6,000 samples at 100 Hz), enabling simultaneous event detection and phase picking. Its hybrid architecture is well-suited to handle complex, overlapping signals. However, EQT is the most computationally demanding among the three models, with longer training and inference times.

All three models process three-component seismograms and output per-sample phase probabilities. Aggregated P-phase and S-phase arrivals can be used to construct earthquake event catalogs. Benchmarking these models on the OKLAD allows an objective assessment of their region-specific performance and the potential improvements achievable through fine-tuning. Table 2 summarizes the models' architectures, input/output format, and their original corresponding training data sets.

We used the SeisBench package (Woollam et al., 2022) for data preparation, and to set up model (all 3 architectures) training, validation, and benchmarking workflow. SeisBench is an open-source Python package designed to streamline access to seismic data sets and models, supporting benchmarking, training, and deployment workflows as needed. Initially, we benchmarked several pre-trained weights for PhaseNet, EQTransformer, and GPD on the OKLAD. The best performing pre-trained models of each architecture were then fine-tuned using the OKLAD training subset and used for evaluation and deployment tests. True positive rate (also referred as "recall", or how many true phase arrivals were identified and labeled correctly compared to ground truth label) was the primary performance metric, calculated based on the residuals between model-predicted and ground-truth arrival times for both P- and S-phase picks in the test data set. This residual analysis provided a direct measure of the performance of different seismic phase arrival predictions.

We begin with PhaseNet architecture. The following section describes the training setup, fine-tuning procedure, and performance evaluation of the OKLAD. For benchmarking and transfer learning, the OKLAD was randomly

split into 70% training, 15% validation, and 15% test subsets. The same splitting strategy was also applied to EQTransformer and GPD fine-tuning to ensure consistency across models. All hyperparameters used during training were loaded from a configuration file.

PhaseNet was initialized with several pretrained weight sets available from the SeisBench repository, including ETHZ weights, trained on seismic data from the Switzerland region (Hetényi G et al., 2018); the INSTANCE weights, trained on the Italian Seismic network data set (Michelini, A et al., 2021); SCEDC weights, trained on the Southern California Earthquake Data Center data set (SCEDC, 2013); and STEAD weights: trained on the Stanford Earthquake Dataset (Mousavi et al., 2019).

The data set was loaded with a sampling rate of 100 Hz and component order as "ENZ." Phase dictionaries were then constructed based on "trace_p_arrival_sample" and "trace_s_arrival_sample" attributes, corresponding to "P" and "S" phases, respectively. These dictionaries guided the probabilistic phase labeling consequently.

We loaded data into the SeisBench generator and applied augmentations. For model training, we implemented a two-step waveform extraction procedure around labeled seismic phases. First, 6,000-sample windows were generated with the labeled "P" or "S" phase at the center, providing 30 s of pre-phase and post-phase data at the sampling rate. Second, from each 6,000-sample window, a 3,001-sample segment—required as input for PhaseNet—was randomly extracted as the final waveform for training. Therefore, each segment contains sufficient pre- and post-phase context, allowing the neural network to accurately detect phase arrivals while introducing variability in the exact segment presented during training. Normalization procedures included demeaning to center the data, detrending to remove linear trends, and amplitude normalization using peak values to standardize amplitude levels across different seismic records. All data were converted to 32-bit floating-point (float32) format to improve the computational efficiency. Finally, probabilistic labeling was applied using a Gaussian distribution with a standard deviation of 30 samples to account for uncertainties in manual phase picks. With the previously described workflow, we benchmarked all the mentioned pre-trained weights of PhaseNet models and compiled the benchmark results. Each model was evaluated on the OKLAD test subset using consistent preprocessing, windowing, and performance metrics.

During fine-tuning, we performed hyperparameter optimization before further evaluation. To optimize, we tested key training hyperparameters, focusing on batch size and learning rate. We evaluated multiple parameter combinations via grid search to identify settings that achieved low validation loss and high recall for P and S phases. This tuning process was essential to balance training stability and model generalization. The final configuration has a batch size of 64 and a learning rate of 0.01. (Figures S3a and S3b in Supporting Information S1)

The maximum training epoch limit was set to 50. To prevent overfitting and promote efficient training, we implemented an Early Stopping mechanism that monitors validation loss and halts training if no improvement is observed for five consecutive epochs. Model checkpoints were saved throughout the training, including both the best-performing and final models. We used the Adam optimizer (Kingma & Ba, 2014) and incorporated an adaptive learning rate strategy via the ReduceLROnPlateau scheduler from the PyTorch library to encourage convergence. The ReduceLROnPlateau scheduler automatically lowers the learning rate during training when a monitored metric stops improving, helping the model escape plateaus and continue learning effectively. In this work, the scheduler monitors the loss and reduces the learning rate by a factor of 0.5 if there is no improvement over three consecutive epochs.

During training, model performance is evaluated using a loss function. The loss function provides a quantitative measurement of the difference between the model's predictions and the ground-truth labels. Loss indicates how well the model's predicted phase arrival time aligns with the expected phase arrival time (as determined by manual analyst picking). A smaller loss indicates that the model's predictions are more accurate. By monitoring the loss, we can evaluate whether the model is learning effectively. For PhaseNet fine-tuning, we used the standard cross-entropy loss with a small numerical-stability constant, following the approach of Zhu and Beroza (2019). This loss function ensures stable training when computing probabilities. After each training epoch, the model was evaluated on the validation subset without updating any model weights. The validation loss is calculated using the same loss function. The validation loss was accumulated across batches and averaged to obtain a single epoch-level value. All loss values were recorded and plotted as loss curves across epochs (Figure S10 in Supporting Information S1).

Final model evaluation was conducted on the test subset of the original split data set, which was withheld from both training and validation. To evaluate the recall of predicted seismic phase arrivals, we calculated the residuals between model-predicted and ground-truth arrival times for both P and S arrival picks. For each trace, ground-truth P and S arrival samples were extracted from the target labels using a peak detection algorithm with a minimum probability trigger threshold of 0.5 and a minimum peak distance of 100 samples. When multiple candidate picks existed for the same phase, we selected the residual with the smallest absolute value to represent the closest time match. Although the ground-truth P and S arrival times are provided in seconds, PhaseNet outputs a probability time series for each phase. A peak detection algorithm is applied to identify the time sample with the maximum probability, which is then used as the predicted phase arrival for evaluation against the catalog. Model precision was quantified by calculating the proportion of residuals falling within ± 0.6 s of the annotated ground-truth arrivals. The standard deviation of the residuals served as a measure of overall pick errors. Model predictions were obtained by passing the input waveform through the final model. The model generated probability distributions indicating the presence of P and S phases. Arrival times were extracted from these distributions using the same detection parameters applied to the ground truth.

Having established the PhaseNet training pipeline, we then fine-tuned EQTransformer to evaluate its performance under similar conditions. For EQTransformer, we adopted the same training, validation, and testing splits as PhaseNet. We carried the same benchmark workflow. Hyperparameter tuning for EQTransformer also followed the same principles as PhaseNet: we searched combinations of batch sizes and learning rates and selected the configuration that yielded the lowest validation loss and highest classification true positive rate for P and S phases (Figures S3c and S3d in Supporting Information S1) The final hyperparameters chosen for EQTransformer were a batch size of 256 and a learning rate of 0.001.

The training and evaluation process for EQTransformer followed a similar workflow to PhaseNet, with adaptations to accommodate the model's architectural requirements. EQTransformer requires input traces of 6,000 samples; therefore, we doubled the random extraction window lengths accordingly to meet this input size requirement. In addition to waveform and phase labels, EQTransformer also requires an "event" label. We constructed this label using the SeisBench DetectionLabeller function.

During training, we followed Mousavi et al. (2020)'s approach and used the binary cross-entropy (BCELoss) function as the loss function. To account for the model's multi-task output (P-phase, S-phase, and event detection), similarly, we adopted Mousavi et al. (2020)'s custom loss weights. These weights balance the contribution of each component to the total loss. The training process also included early stopping and learning rate scheduling, following the strategies used for PhaseNet. All evaluations, including residual analysis and performance measurements, were conducted using the same benchmarking procedures as described previously for consistency.

We then fine-tuned the GPD model using a similar pipeline, while adapting the process to fit GPD's unique architecture and data requirements. GPD was designed to process input traces of 400 samples per segment and to classify windows into three classes: P-phase, S-phase, and noise. To generate training data, we randomly sampled 400-sample windows centered at labeled P or S phases, ensuring each window included 200 samples preceding the phase arrival. For noise traces, we extracted 400-sample segments starting 7 s (700 samples) prior to the P arrival pick to avoid contamination from the signal waveform. The class labels were generated using the SeisBench ProbabilisticPointLabeller, with the peak class probability aligned at the 0.5 position (the 200th sample point).

During evaluation, we applied a sliding window of 400 samples with a stride of 1 sample to generate continuous probability curves, following the same phase-picking procedure used for PhaseNet and EQTransformer. This maintained consistency across all models and ensured comparability of results. GPD training employed the same early stopping, learning rate scheduling, and performance tracking strategies as the other models. All training, validation, and benchmarking scripts used in this study are openly available in the repositories listed in the Data Availability section.

In summary, we benchmarked and fine-tuned PhaseNet, EQTransformer, and GPD on the OKLAD using cautiously aligned training, validation, and evaluation protocols. Figure 5 shows examples of waveform input, ground truth labels and final models annotating their input. Across all models, we applied consistent data splits, labeling strategies, and performance metrics to ensure a fair comparison of seismic phase picking performance.

The authors used ChatGPT (OpenAI) to assist with grammar checking and wording refinement for clarity. The AI tool did not contribute to the scientific content, data analysis, interpretation of results, or figure generation.

10.1029/2025JH001194

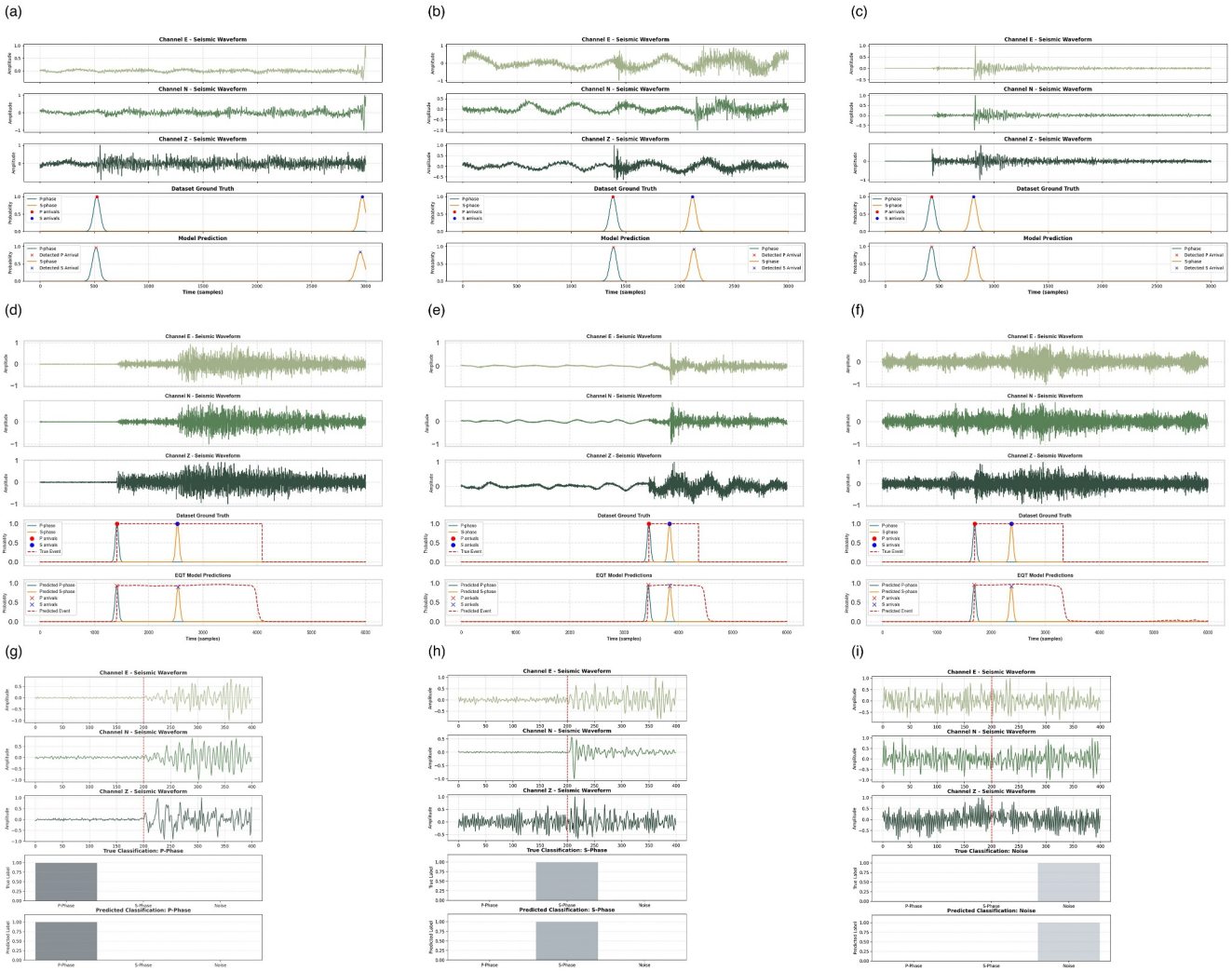

Figure 5. Examples of waveform data, labels and model predictions. (a–c) Examples of trace samples constructed for PhaseNet training and Model prediction. The upper 3 panels show the waveform from Channel E, Channel N, and Channel Z. The data set ground truth panel shows the manual P and S arrival labels from OKLAD. The bottom panel shows the model prediction of the P and S possibilities. Red dots or crosses denote the P arrival time picks from labels and predictions, respectively. Blue dots or crosses denote the S arrival time picks from labels and predictions. (d–f) Examples of trace samples constructed for EQTransformer training and Model prediction. The upper 3 panels show the waveform from Channel E, Channel N, and Channel Z. The data set ground truth panel shows the manual P and S arrival labels from OKLAD. The red dashed line denotes the event label. The bottom panel shows the model prediction of the P and S possibilities. Red dots and crosses denote the P arrival time picks from labels and predictions, respectively. The red dashed line denotes the event prediction from trained EQTransformer models. (g–i) Examples of trace samples constructed for GPD training and Model prediction. The upper 3 panels show waveform from Channel E, Channel N, and Channel Z. The data set ground truth panel shows the manual labels of P, S and noise labels from OKLAD. The red dashed line denotes the probability position of the label. The bottom 2 panels show the model's ground-truth label and the trained model's prediction.

4. Results and Discussion

We benchmarked the PhaseNet trained weights from different data sets, including ETHZ, INSTANCE, SCEDC, and STEAD, and tested their performance on the OKLAD test data set. Residuals (Δt) between model-predicted and analyst-labeled phase picks were calculated for each trace, with the smallest residual kept when multiple picks were detected. The standard deviation of these residuals provides a measure of overall pick errors. This metric quantifies the bias or shift in phase arrival predictions from models such as PhaseNet, EQTransformer, or GPD with different pre-trained weights (see Methods for calculation details).

Figure 6a shows a summary of all the PhaseNet pretrained weights' performance measured by the overall true positive rate. We also included our best-trained OKLAD PhaseNet model alongside these pretrained models. As shown in Figure 6a, different pretrained PhaseNet models' performance on the OKLAD test data set varies

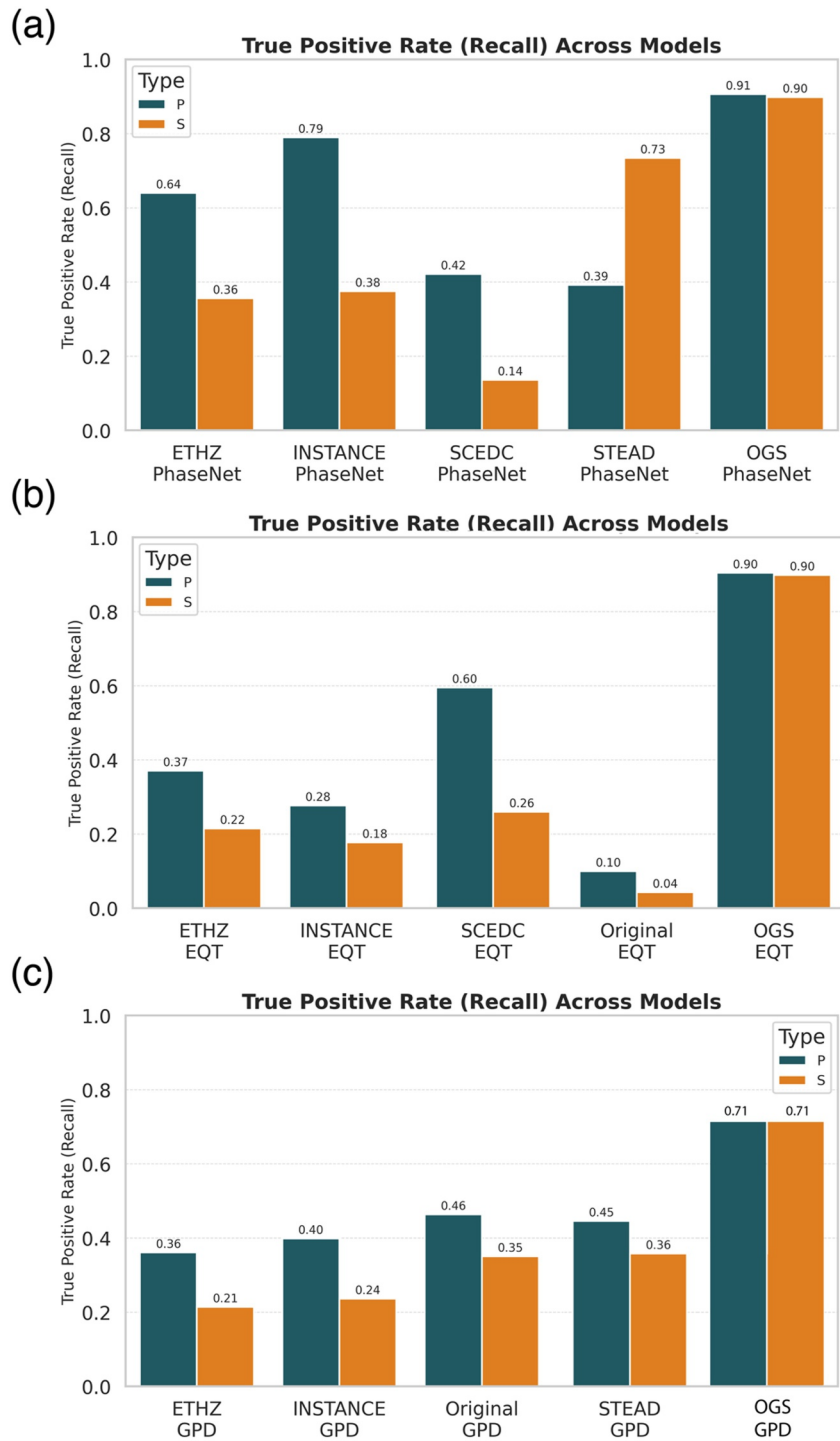

Figure 6. Benchmark results of different PhaseNet/EQT/GPD pretrained models on OKLAD. A prediction is considered a true positive if the absolute time difference between the predicted pick and the labeled pick is less than 0.6 s, and the subsequent true positive rate is defined as the total number of true positive picks divided by the total number of labels for different phases. (a) Benchmark results showing the true positive rates of different pre-trained PhaseNet models evaluated on OKLAD. The rightmost collumn shows the fine-tuned PhaseNet model. (b) Benchmark results showing the true positive rates of different pre-trained EQT models evaluated on OKLAD. The rightmost collumn shows the fine-tuned EQT model. (c) Benchmark results showing the true positive rates of different pre-trained GPD models evaluated on OKLAD. The rightmost collumn shows the fine-tuned GPD model.

significantly. The ETHZ model, while achieving a true positive rate of 64% for P-phase events, performs poorly in S-phase detection, with a detection rate of only 36%. Similarly, the INSTANCE model performs moderately better for P-phase detection (79%) but struggles significantly with S-phase (38%). The SCEDC model performs poorly for both phases, with detection rates as low as 42% for the P-phase and 14% for the S-phase. The STEAD model displays strength in S-phase detection, with a detection rate of 73% but fails in P-phase detection (39%). Overall, the STEAD pretrained PhaseNet model outperforms the other models in S-phase detection.

We then benchmarked EQTransformer pre-trained weights, including ETHZ, INSTANCE, SCEDC, and STEAD, on the OKLAD test data set. Residual calculation, peak detection, and true positive rate metrics were performed consistently with the PhaseNet evaluation. As shown in Figure 6b, the performance of different pre-trained EQTransformer models on OKLAD varies considerably. The ETHZ model achieved a true positive rate of 37% for P-phase but only 22% for S-phase detection. The INSTANCE model performed worse, with 28% for P-phase and 18% for S-phase. The original EQTransformer, trained on STEAD, achieved 10% for P-phase and 4% for S-phase. The SCEDC model performed best among EQTransformer weights, reaching 60% for P-phase and 26% for S-phase. Overall, the SCEDC pre-trained EQTransformer outperforms the other EQTransformer models in S-phase detection, but pre-trained EQTransformer models generally perform worse than PhaseNet pre-trained models on the OKLAD test data set.

In addition, we benchmarked the GPD pre-trained weights on OKLAD test subset in a similar fashion. Unlike PhaseNet or EQT, the GPD model takes 400 samples as input and outputs a tensor that indicates P, S or "noise" with a probability variation from 0 to 1 in each dimension. Thus, when benchmarked, probability curves were constructed using a sliding window with a window length of 400 samples and stride of 1 sample. The converted sliding-window curves were then subjected to the same evaluation rubric as the PhaseNet and EQTransformer models.

As shown in Figure 6c, the performance of different pre-trained GPD models on OKLAD varies. The ETHZ model achieved a true positive rate of 36% for P-phase and 21% for S-phase. The INSTANCE model achieved 40% for the P-phase and 24% for the S-phase. The original GPD model achieved 46% for P-phase and 35% for S-phase, while the STEAD model achieved 45% for P-phase and 36% for S-phase. Overall, performance differences among pre-trained GPD models are consistent with expectations, reflecting variations in training data and regional applicability.

GPD models were evaluated based on reconstructed probability curves and the minimum residual point was used to calculate the true positive rate. Over the same data window, GPD model predictions would yield more peak detections than the PhaseNet and EQT models. Fine-tuned GPD model prediction still contains more detections than EQT or PhaseNet. Examples of GPD models' predictions with sliding windows are shown in Figure S4 of Supporting Information S1. To better evaluate GPD models during transfer learning and fine-tuning, we used accuracy for the GPD model during fine-tuning. Accuracy in fine-tuning is a classification metric that measures the proportion of correct predictions out of the total number of predictions. Mathematically, it is defined as:

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of prediction}} \quad (1)$$

For GPD fine-tuning, accuracy was calculated based on P, S and N classification prediction instead of probability curve calculations.

Following the workflow described in the Methods section earlier, we then fine-tuned the best-performing PhaseNet, EQTransformer, and GPD models on progressively larger subsets (1%–90%) of the OKLAD training data and evaluated each on the full test subset (Figure 7). For PhaseNet, P- and S-phase recall increases from 81.5% and 85.7%, respectively, when using 1% of OKLAD (~8,550 traces), to 90.3% and 88.7% with 30% of the data, and 90.7% and 89.8% with 90% of the data. EQTransformer shows a similar pattern, with P and S phase recall rising from 86.0% and 87.3% at 1% to 89.3% and 88.4% at 30% and 90.4% and 89.9% at 90%. For GPD, the overall accuracy increases from 94.26% at 1% to 96.91% at 30% and 97.52% at 90%. Across all three architectures, fine-tuning with localized OKLAD data substantially outperforms the corresponding pre-trained models on the OKLAD test subset.

10.1029/2025JH001194

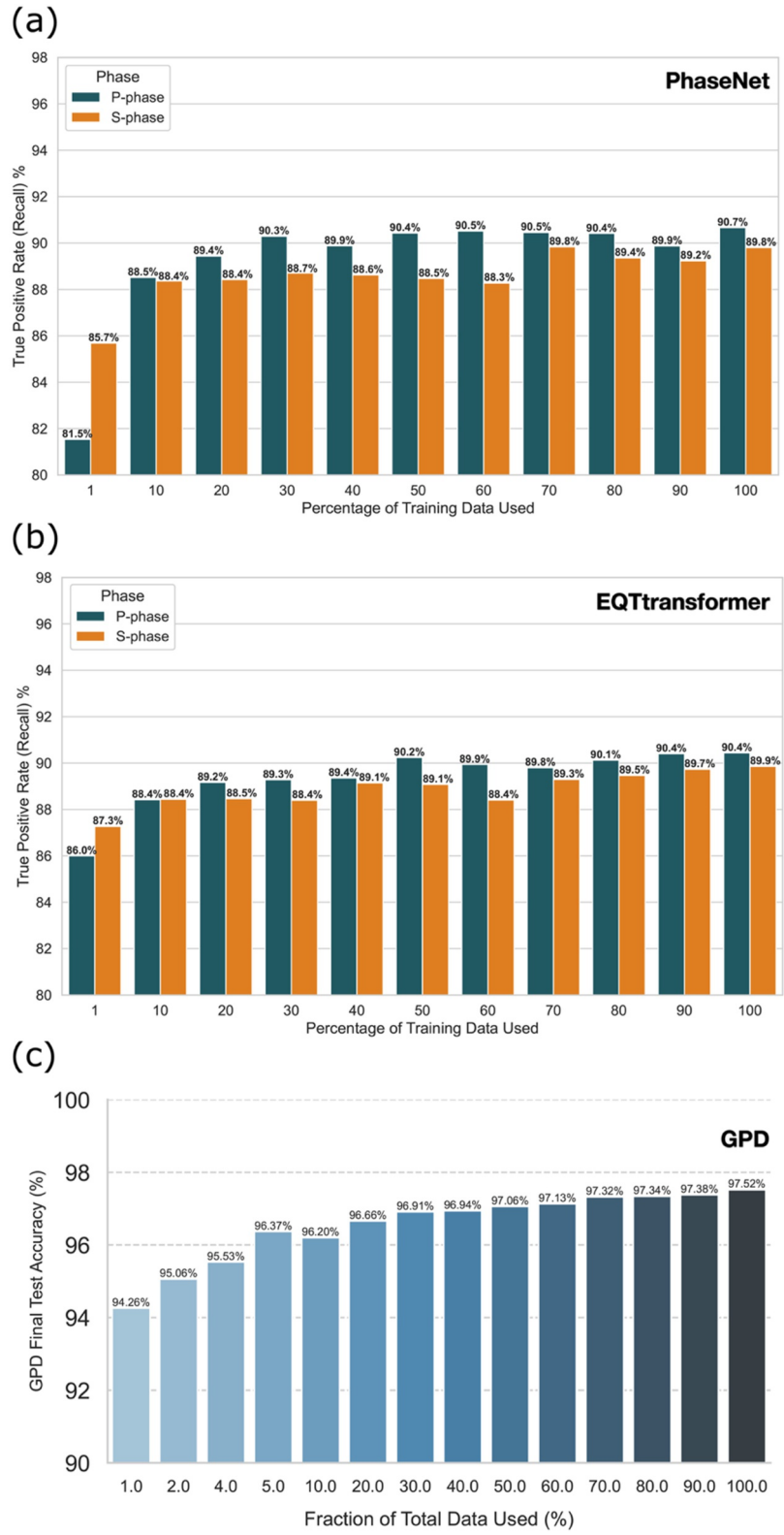

Figure 7. (a) True positive rate with increasing percentage of training data used for transfer learning for PhaseNet. (b) True positive rate with increasing percentage of training data used for transfer learning for EQT. (c) Accuracy with increasing percentage of training data used for transfer learning for GPD.

Table 3
True Positive Rate of P-Phase and S-Phase Detection Across Different Magnitude Ranges

Magnitude range	P-phase TPR	S-phase TPR
0–1	0.81	0.83
1–2	0.84	0.87
2–3	0.94	0.94
3–4	0.97	0.96
4–5	0.99	0.92
>5	0.85	0.75

Among all the fine-tuned models, our best fine-tuned OKLAD PhaseNet model's performance on the test data set demonstrates robust prediction capabilities for both P and S arrival times. Model performance was evaluated using precision, recall (true positive rate), and F1 score, where precision measures the fraction of predicted picks that are correct, recall quantifies the fraction of true arrivals that are successfully detected, and the F1 score represents the harmonic mean of precision and recall. For P-phase detection, the model achieved a precision of 0.9504, a recall of 0.9084, and an F1 score of 0.9290. For S-phase detection, the model achieved a precision of 0.9330, recall of 0.8989, and an F1 score of 0.9156.

We further calculated the true residuals between the ground truth and the model prediction. For the P arrival times (Figure S5a in Supporting Information S1), the mean residual was 0.009 s, indicating a minimal bias in the predictions relative to the ground truth values (seismic analyst's label), with a standard deviation (STD) of 0.861 s. For the S arrival times (Figure S5b in Supporting Information S1), the mean residual between model prediction and ground truth was −0.001 s, showing similarly minimal bias, with a smaller STD of 0.666 s, indicating tighter clustering of predictions around the ground-truth labeling.

The true positive rate for both P-phase and S-phase detection varies across different magnitude ranges. The total true positive rates across magnitude ranges are summarized in Table 3. These results indicate that the model performs effectively across a range of magnitudes, achieving particularly high recall for magnitudes between 2 and 5. For smaller magnitudes (0–2), which are typically more difficult to detect, the model demonstrates strong performance as well. These results indicate that the fine-tuned PhaseNet model is well-optimized for detecting microseismic events, particularly in detecting S-phase arrivals.

Using the best-performing fine-tuned model, OKLAD-PhaseNet, we first evaluated its performance on continuous waveform data from 92 stations spanning 1–30 September 2016. With a probability threshold of 0.5, the OGS PhaseNet model detected 480,185 picks, compared with 122,456 by the ETHZ PhaseNet model, 105,664 by the INSTANCE PhaseNet model, and 74,544 by the STEAD PhaseNet model. This represents a 2.5- to 6.4-fold increase in detections relative to the pre-trained models. We then assessed OKLAD-PhaseNet on continuous waveform data from 138 stations over the full year of 2022. Using PyOcto (Münchmeyer, 2023) for event association, the model detected a total of 7,445 associated events with a minimum of 6 picks per event, compared to 3,017 events identified by the current easyQuake workflow and routine network operations, representing a 1.5-fold increase in detected earthquakes. During this period of 2022, 96.8% of OGS-reported events were successfully identified (Figure S6 in Supporting Information S1), confirming the model's high reliability.

To further evaluate the generalization capability of our fine-tuned models beyond Oklahoma, we applied the best-performing model to the Midland, West Texas region, using data from 105 stations between 1 January and 31 January 2024 (Figure S7 in Supporting Information S1). This experiment is intended as a controlled external validation of model performance under cross-regional deployment conditions. The model recovered over 97% (634/648) of USGS-reported events. Using a conservative threshold of at least 15 picks per event, the model identifies a total of 2,427 events (Figure S8 in Supporting Information S1), substantially increasing the number of detected events relative to the routine catalog. To assess agreement with cataloged seismicity, we examined spatial and temporal proximity between transfer-learning catalog events and USGS catalog events. We find that 626/648 USGS events fall within 25 s and 40 km of the corresponding transfer-learning catalog associations (Figure S9 in Supporting Information S1), indicating strong catalog-level consistency between the two catalogs. This experiment is not intended to provide a geophysical interpretation of seismicity in West Texas, but rather to evaluate the robustness and transferability of our detection model in a different tectonic and operational setting.

Our results highlight how transfer learning improves the detection performance of pre-trained deep learning models in response to the specific challenges of induced seismicity in Oklahoma. As noted above, pre-trained PhaseNet, EQTransformer, and GPD models, which perform well in tectonically active regions, show degraded performance when applied to OKLAD. In contrast, the same model architectures fine-tuned on Oklahoma data close much of this performance gap, with recall and accuracy exceeding 90% for both P and S phases. This contrast suggests that the main limitation is not model capacity but rather a mismatch between the original training data and the induced, shallow seismic environment in Oklahoma. In line with earlier work on

smaller local arrays (Chai et al., 2020), our results demonstrate that this principle also holds at the regional-network scale. This study provides a practical path for improving monitoring in induced-seismicity settings with sparse station coverage and complex noise conditions.

A key advantage of this transfer learning approach is its data efficiency. Even when only 1% of OKLAD is used for fine-tuning, all three models outperform their pre-trained baselines, and performance continues to improve as additional local data is incorporated. This suggests that a small subset of carefully labeled data can improve the generalization of pre-trained models to new regions, which is consistent with findings from smaller local arrays and other regional transfer learning studies (Ho et al., 2024; Jozinović et al., 2022; Niksejel & Zhang, 2024).

These gains provide a more complete earthquake catalog for Oklahoma. Our fine-tuned PhaseNet model recovers nearly all OGS reported events while identifying thousands of additional earthquakes that routine workflows and the current easyQuake implementation miss. The resulting 1.5-fold increase in associated events, together with a several-fold increase in detected picks relative to pre-trained models, demonstrates that transfer-learned pickers can lower the effective magnitude of completeness for regional monitoring. In induced settings where hazard mitigation relies on tracking changes in seismic rates, improved catalog completeness can inform timely decisions on injection parameters (Convertito et al., 2012) and support future applications such as carbon capture, utilization, and storage (CCUS) projects.

5. Conclusion

In this study, we introduced the Oklahoma Labeled AI Dataset (OKLAD), a high-quality, manually curated data set spanning 2010–2024 from the OGS, tailored for evaluating induced seismicity. Using OKLAD, we systematically benchmarked three state-of-the-art machine-learning models—PhaseNet, EQTransformer, and GPD—and applied transfer learning to fine-tune them for regional adaptation.

Our benchmarking analysis on the Oklahoma Labeled AI Dataset (OKLAD) shows that different pre-trained deep-learning models for seismic phase detection exhibit varying degrees of performance when applied to induced-seismicity dominated regions such as Oklahoma. This finding indicates that region-specific considerations are needed, as models originally trained on tectonic data sets may not generalize well to areas where seismicity is strongly influenced by anthropogenic activities.

By fine-tuning PhaseNet, EQTransformer (EQT), and Generalized Phase Detection (GPD) models with localized OKLAD data, we demonstrate consistent and substantial improvements in phase-picking performance. Even with as little as 1% of Oklahoma training data, recall increased significantly, while models fine-tuned with larger subsets achieved recall rates exceeding 90% for both P and S phases. These results highlight the power of transfer learning and fine tuning: incorporating relatively modest amounts of local data can effectively overcome generalization limitations of pre-trained models and adapt them to local monitoring environments.

Beyond controlled test data sets, our results showcased that fine-tuning with localized data substantially improves detection performance. The best-performing OKLAD-PhaseNet recovered nearly all cataloged events while detecting thousands of additional earthquakes previously unreported by standard methods. Evaluations on continuous waveform data confirmed the model's robustness and operational potential, highlighting its applicability for real-time monitoring and hazard assessment.

To promote reproducibility and accelerate future advances, we release this unique and curated OKLAD, our fine-tuned models, and associated training and evaluation workflows as openly available resources. These resources provide a new benchmark for studying induced seismicity in the southern midcontinent of the United States. Our work highlights the crucial role of localized data sets in advancing deep-learning based seismic monitoring and offers a scalable framework for regions facing similar challenges.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

Waveform data and associated metadata used in this study were obtained from the Oklahoma Geological Survey (OGS) data archive (Walter et al., 2020). The continuous waveforms were from the OGS data archive but were also archived at the NSF SAGE data archive operated by EarthScope Consortium (NSF award 1724509). The fine-tuned models, event catalogs, and associated source code used in this study are publicly available (Xiao, 2026). In addition, all curated waveform data and metadata in this study are publicly available (Xiao et al., 2026).

Machine-learning analyses were conducted using the open-source easyQuake package (Walter et al., 2021) and the SeisBench framework (Münchmeyer, 2023; Woollam et al., 2022; Xiao, 2026). The versions of the software, trained models, and associated source code used in this study have been archived in Zenodo and are cited in the References. Development versions of these software packages are maintained on GitHub. Some figures in this study were generated using the Generic Mapping Tools (Wessel et al., 2013).

Acknowledgments
This research was supported by the U.S. Department of Energy (DOE) award DE-FE0031837. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). Additional support was provided by the University of Oklahoma (OU) and the Oklahoma Geological Survey (OGS). Their contributions and resources were invaluable to this research. Parts of OKLAD were accessed from the NSF SAGE data archive operated by EarthScope Consortium (NSF award 1724509).

References

Albuquerque Seismological Laboratory (ASL)/USGS. (1980). US geological survey networks [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/GS

Albuquerque Seismological Laboratory (ASL)/USGS. (1990). United States national seismic network [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/US

Albuquerque Seismological Laboratory/USGS. (2013). Central and eastern US network [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/N4

Bureau of Economic Geology, The University of Texas at Austin. (2016). Texas seismological network [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/TX

Cappa, F., & Rutqvist, J. (2011). Impact of CO_2 geological sequestration on the nucleation of earthquakes. *Geophysical Research Letters*, *38*(17). https://doi.org/10.1029/2011gl048487

Chai, C., Maceira, M., Santos-Villalobos, H. J., Venkatakrishnan, S. V., Schoenball, M., Zhu, W., et al. (2020). Using a deep neural network and transfer learning to bridge scales for seismic phase picking. *Geophysical Research Letters*, *47*(16), e2020GL088651. https://doi.org/10.1029/2020GL088651

Chang, J. (2016). Seismic investigation of south central Oklahoma [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/ZP_2016

Chen, X., Peng, Z., & Chang, J. C. (2016). Rapid response for Fairview aftershock in Oklahoma [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/Y9_2016

Chen, Y., Savvaidis, A., Saad, O. M., Dino Huang, G. C., Siervo, D., O'Sullivan, V., et al. (2024). TXED: The Texas earthquake dataset for AI. *Seismological Research Letters*, *95*(3), 2013–2022. https://doi.org/10.1785/0220230327

Collette, A., Kluyver, T., Caswell, T. A., Tocknell, J., Kieffer, J., Jelenak, A., et al. (2023). h5py/h5py: 3.8.0 (version 3.8.0) [Computer software]. *Zenodo*. https://doi.org/10.5281/ZENODO.7560547

Convertito, V., Maercklin, N., Sharma, N., & Zollo, A. (2012). From induced seismicity to direct time-dependent seismic hazard. *Bulletin of the Seismological Society of America*, *102*(6), 2563–2573. https://doi.org/10.1785/0120120036

Darold, A. (2014). 4D integrated study using geology, geophysics, reservoir modeling & rock mechanics to develop assessment models for potential in [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/ZD_2014

Ellsworth, W. L. (2013). Injection-induced earthquakes. *Science*, *341*(6142), 1225942. https://doi.org/10.1126/science.1225942

Frohlich, C., DeShon, H., Stump, B., Hayward, C., Hornbach, M., & Walter, J. I. (2016). A historical review of induced earthquakes in Texas. *Seismological Research Letters*, *87*(4), 1022–1038. https://doi.org/10.1785/0220160016

Goebel, T., Rosson, Z., Brodsky, E. E., & Walter, J. I. (2019). Aftershock deficiency of induced earthquake sequences during rapid mitigation efforts in Oklahoma, Earth and planet. *Science Letter*, *522*, 135–143. https://doi.org/10.1016/j.epsl.2019.06.036

Goebel, T. H. W., Walter, J. I., Murray, K., & Brodsky, E. E. (2017a). *Comment on "how will induced seismicity in Oklahoma respond to decreased saltwater injection rates?* In C. Langenbruch & M. D. Zoback (Eds.) (Vol. 3, p. 8). https://doi.org/10.1126/sciadv.1700441

Goebel, T. H. W., Weingarten, M., Chen, X., Haffener, J., & Brodsky, E. E. (2017b). The 2016 Mw5. 1 Fairview, Oklahoma earthquakes: Evidence for long-range poroelastic triggering at> 40 km from fluid disposal wells. *Earth and Planetary Science Letters*, *472*, 50–61. https://doi.org/10.1016/j.epsl.2017.05.011

Guy, M. R., Patton, J. M., Fee, J., Hearne, M., Martinez, E., Ketchum, D., et al. (2015). *National Earthquake Information Center systems overview and integration*. Open-File Report. https://doi.org/10.3133/ofr20151120

Hetényi, G., Molinari, I., Clinton, J., Bokelmann, G., Bondár, I., Crawford, W. C., et al. (2018). The AlpArray seismic network: A large-scale European experiment to image the Alpine Orogen. *Surveys in Geophysics*, *39*(5), 1009–1033. https://doi.org/10.1007/s10712-018-9472-4

Ho, L. M., Walter, J. I., Hansen, S. E., Sánchez-Roldán, J. L., & Peng, Z. (2024). Evaluating automated seismic event detection approaches: An application to Victoria Land, East Antarctica. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(3), e2024JH000185. https://doi.org/10.1029/2024jh000185

IRIS Transportable Array. (2003). USArray transportable array [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/TA

Jiang, C., Fang, L., Fan, L., & Li, B. (2021). Comparison of the earthquake detection abilities of PhaseNet and EQTransformer with the Yangbi and Maduo earthquakes. *Earthquake Science*, *34*(5), 425–435. https://doi.org/10.29382/eqs-2021-0038

Jozinović, D., Lomax, A., Štajduhar, I., & Michelini, A. (2022). Transfer learning: Improving neural network based prediction of earthquake ground shaking for an area with insufficient training data. *Geophysical Journal International*, *229*(1), 704–718. https://doi.org/10.1093/gji/ggab488

Keranen, K. M., Weingarten, M., Abers, G. A., Bekins, B. A., & Ge, S. (2014). Sharp increase in central Oklahoma seismicity since 2008 induced by massive wastewater injection. *Science*, *345*(6195), 448–451. https://doi.org/10.1126/science.1255802

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Marsh, S., & Holland, A. (2016). Comprehensive fault database and interpretive fault map of Oklahoma. *Oklahoma Geological Survey Open-File Rept. OF2-2016*, *2*.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., & Lauciani, V. (2021). INSTANCE–The Italian seismic dataset for machine learning. *Earth System Science Data*, *13*(12), 5509–5544. https://doi.org/10.5194/essd-13-5509-2021

Mousavi, S. M., & Beroza, G. C. (2022). Deep-learning seismology. *Science*, *377*(6607), eabm4470. https://doi.org/10.1126/science.abm4470

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake Transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1), 3952. https://doi.org/10.1038/s41467-020-17591-w

Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). STanford EArthquake dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, *7*, 179464–179476. https://doi.org/10.1109/access.2019.2947848

Münchmeyer, J. (2023). PyOcto: A high-throughput seismic phase associator. *arXiv preprint arXiv:2310.11157*. https://doi.org/10.26443/seismica.v3i1.1130

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., et al. (2022). Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, *127*(1), e2021JB023499. https://doi.org/10.1029/2021jb023499

Murray, K. E., Brooks, C., Walter, J. I., & Ogwari, P. O. (2023). Oklahoma's coordinated response to more than a decade of elevated seismicity, GSA special paper 559 in recent seismicity in the southern Midcontinent, USA. *Scientific, Regulatory, and Industry Responses*. https://doi.org/10.1130/2023.2559

Nakata, N. (2016). Acquisition of aftershock sequence of the 2016 M5.6 sooner Lake earthquake [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/Y7_2016

NCEDC. (2014). Northern California earthquake data center. UC Berkeley seismological laboratory [Dataset]. https://doi.org/10.7932/NCEDC

Niksejel, A., & Zhang, M. (2024). OBSTransformer: A deep-learning seismic phase picker for OBS data using automated labelling and transfer learning. *Geophysical Journal International*, *237*(1), 485–505. https://doi.org/10.1093/gji/ggae049

Ogwari, P., & Walter, J. (2023). Harnessing the geothermal potential of Oklahoma sedimentary basin [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/7rm8-5370

Oklahoma Geological Survey. (1978). Oklahoma seismic network [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/OK

Oklahoma Geological Survey. (2018). Oklahoma consolidated temporary seismic networks [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/O2

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578. https://doi.org/10.1126/sciadv.1700578

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18* (pp. 234–241). Springer international publishing.

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901. https://doi.org/10.1785/0120180080

Rubinstein, J. L., & Mahani, A. B. (2015). Myths and facts on wastewater injection, hydraulic fracturing, enhanced oil recovery, and induced seismicity. *Seismological Research Letters*, *86*(4), 1060–1067. https://doi.org/10.1785/0220150067

Saad, O. M., Chen, Y., Siervo, D., Zhang, F., Savvaidis, A., Huang, G. C. D., et al. (2023). EQCCT: A production-ready earthquake detection and phase-picking method using the compact convolutional transformer. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/tgrs.2023.3319440

SCEDC. (2013). Southern California earthquake center [Dataset]. *Caltech*. https://doi.org/10.7909/C3WD3xH1

Shapiro, S. A., Dinske, C., & Kummerow, J. (2007). Probability of a given-magnitude earthquake induced by a fluid injection. *Geophysical Research Letters*, *34*(22). https://doi.org/10.1029/2007gl031615

Skoumal, R. J., Barbour, A. J., Rubinstein, J. L., & Glasgow, M. E. (2024). Reduced injection rates and shallower depths mitigated induced seismicity in Oklahoma. *The Seismic Record*, *4*(4), 279–287. https://doi.org/10.1785/0320240030

U.S. Geological Survey. (1989). NetQuakes [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/NQ

U.S. Geological Survey. (2016). U.S. Geological survey networks [Dataset]. *International Federation of Digital Seismograph Networks*. https://doi.org/10.7914/SN/GM

Van der Baan, M., & Calixto, F. J. (2017). Human-induced seismicity and large-scale hydrocarbon production in the USA and Canada. *Geochemistry, Geophysics, Geosystems*, *18*(7), 2467–2485. https://doi.org/10.1002/2017gc006915

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Walter, J. I., Chang, J. C., & Dotray, P. J. (2017). Foreshock seismicity suggests gradual differential stress increase in the months prior to the 3 September 2016 M w 5.8 Pawnee earthquake. *Seismological Research Letters*, *88*(4), 1032–1039. https://doi.org/10.1785/0220170007

Walter, J. I., Frohlich, C., & Borgfeldt, T. (2018). Natural and induced earthquakes in the Texas and Oklahoma Panhandles. *Seismological Research Letters*, *89*(6), 2437–2446. https://doi.org/10.1785/0220180105

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., & Woelfel, I. (2021). easyQuake: Putting machine learning to work for your regional seismic network or local earthquake study. *Seismological Research Letters*, *92*(1), 555–563. https://doi.org/10.1785/0220200226

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., Woelfel, I., Chang, J. C., et al. (2020). The Oklahoma geological survey statewide seismic network. *Seismological Research Letters*, *91*(2A), 611–621. https://doi.org/10.1785/0220190211

Wessel, P., Smith, W. H., Scharroo, R., Luis, J., & Wobbe, F. (2013). Generic mapping tools: Improved version released. *EOS Transactions of the American Geophysical Union*, *94*(45), 409–410. https://doi.org/10.1002/2013eo450001

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., et al. (2022). SeisBench—A toolbox for machine learning in seismology. *Seismological Society of America*, *93*(3), 1695–1709. https://doi.org/10.1785/0220210324

Xiao, H. (2026). Transfer learning seisBench tutorial associated with OKLAD (Oklahoma labeled AI dataset) (Version 202603). *Zenodo*. https://doi.org/10.5281/zenodo.19244808

Xiao, H., Walter, J., Ogwari, P., Thiel, A., Woelfel, I., Gregg, N., & Mace, B. (2026). Oklahoma labeled AI dataset for seismology (1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.18991761

Yeck, W. L., Hayes, G. P., McNamara, D. E., Rubinstein, J. L., Barnhart, W. D., Earle, P. S., & Benz, H. M. (2017). Oklahoma experiences largest earthquake during ongoing regional wastewater injection hazard mitigation efforts. *Geophysical Research Letters*, *44*(2), 711–717. https://doi.org/10.1002/2016gl071685

Zhao, M., & Chen, S. (2021). The generalization ability research of deep learning algorithm in seismic phase detection of regional seismic network. *Earthquake*, *41*(1), 166–179.

Zhao, M., Xiao, Z., Chen, S., & Fang, L. (2023). DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. *Earthquake Science*, *36*(2), 84–94. https://doi.org/10.1016/j.eqs.2022.01.022

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