Geophysical Research Letters

RESEARCH LETTER
10.1029/2024GL108438

Deep-Learning-Based Phase Picking for Volcano-Tectonic and Long-Period Earthquakes

Yiyuan Zhong[1] and **Yen Joe Tan**[1]

[1]Earth and Environmental Sciences Programme, Faculty of Science, The Chinese University of Hong Kong, Hong Kong S.A.R., China

Key Points:

- We compile a data set of seismic waveforms from various volcanic regions globally
- We show that existing deep-learning phase pickers' performances deteriorate with decreasing earthquake frequency content
- Our retrained models perform better and are more generalizable for monitoring volcano seismicity, especially long-period earthquakes

Supporting Information:

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

Correspondence to:

Y. J. Tan,
yjtan@cuhk.edu.hk

Citation:

Zhong, Y., & Tan, Y. J. (2024). Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes. *Geophysical Research Letters*, *51*, e2024GL108438. https://doi.org/10.1029/2024GL108438

Received 24 JAN 2024
Accepted 31 MAY 2024

Abstract The application of deep-learning-based seismic phase pickers has surged in recent years. However, the efficacy of these models when applied to monitoring volcano seismicity has yet to be fully evaluated. Here, we first compile a data set of seismic waveforms from various volcanoes globally. We then show that the performances of two widely used deep-learning pickers deteriorate systematically as the earthquakes' frequency content decreases. Therefore, the performances are especially poor for long-period earthquakes often associated with fluid/magma movement. Subsequently, we train new models which perform significantly better, including when tested on two data sets where no training data were used: volcanic earthquakes along the Cascadia subduction zone and tectonic low-frequency earthquakes along the Nankai Trough. Our model/workflow can be applied to improve monitoring of volcano seismicity globally while our compiled data set can be used to benchmark future methods for characterizing volcano seismicity, especially long-period earthquakes which are difficult to monitor.

Plain Language Summary Earthquake activity at volcanic regions is often monitored to indicate volcanic activity. Identifying the time when the energy radiated from an earthquake source arrives at a seismometer is essential for locating the earthquake, which can be difficult for volcanic earthquakes because of high noise levels, high event rates, and obscured onsets. Previous studies have demonstrated that deep learning can excel in picking the arrival times of regular earthquakes. However, it is unclear how sensitive these detectors are to earthquakes in volcanic regions. Here, we first compile a data set of earthquakes from various volcanoes globally. We then show that existing deep-learning-based detectors can miss a large fraction of these earthquakes, especially those without an abrupt change in signal amplitude. We then provide two new models which can better detect volcanic earthquakes than existing models. Our model/workflow can be applied to improve monitoring of volcanic earthquakes globally.

1. Introduction

Detecting and identifying onsets of seismic phases is fundamental to locating seismicity. Manual inspection by experienced analysts is viewed as the gold standard but is extremely laborious and time-consuming. This makes it difficult to handle the ever-increasing volumes of seismic data and periods with extremely high seismicity rate such as during volcanic unrests. On the other hand, early automatic methods, such as the short-term average over long-term average method (STA/LTA) (Allen, 1978), suffer from low accuracy and require a number of parameters to be tuned carefully. Over the past two decades, the matched-filter technique has been shown to be an effective method (Chamberlain et al., 2017; Gibbons & Ringdal, 2006) to search for repeating or near-repeating earthquakes based on waveform similarity. However, this method is only capable of detecting earthquakes in the vicinity of known template events. In recent years, deep-learning-based phase pickers/event detectors (e.g., Kriegerowski et al., 2019; Mousavi et al., 2020; Ross et al., 2018; Soto & Schurr, 2021; Zhu & Beroza, 2019) have been gaining increasing attention due to their picking accuracy being comparable to human analysts (Chai et al., 2020) and high efficiency. Their application has surged in recent years, including for delineating seismicity in fault zones, subduction zones, oceanic transform faults, and volcanoes (e.g., Chen et al., 2022; Garza-Girón et al., 2023; Gong et al., 2023; Jiang et al., 2022; Liu et al., 2023; Liu et al., 2024; Tan et al., 2021; Wilding et al., 2023; Zhang et al., 2024). However, it can be difficult to predict deep-learning models' performance for out-of-distribution data that are not well represented by training data (Teney et al., 2022; Wenzel et al., 2022).

Seismicity which often correlates with magmatic/volcanic processes and sometimes represents eruption precursors (Acocella et al., 2023; White & McCausland, 2019) is an important monitoring observable at volcanoes.

Two types of earthquakes are commonly observed in volcanic regions: volcano-tectonic earthquakes (VTs) and long-period earthquakes (LPs), which are classified mainly based on their waveform frequency content but may imply different source processes (e.g., Chouet & Matoza, 2013; Matoza & Roman, 2022; Saccorotti & Lokmer, 2021, and references therein). VTs share common spectral characteristics with regular tectonic earthquakes and have impulsive onsets. They mostly originate from shear fractures in the solid part of an edifice or the underlying crust, hence only indirectly indicate magmatic activity. In comparison, most conceptual source models of LPs involve fluids, for example, resonating fluid-filled cracks (Chouet & Matoza, 2013), thermal stresses in cooling magmas (Aso & Tsai, 2014), pressurization of exsolved volatiles from stalled magmas (Wech et al., 2020), and rapidly growing bubbles in ascending magmas (Melnik et al., 2020). Therefore, LPs are often interpreted as a more direct evidence of fluid movement (e.g., Song et al., 2023). However, compared to VTs, LPs are more difficult to detect and catalog because they are depleted of high frequency content and have emergent phase onsets (Pitt et al., 2002; Shapiro et al., 2017), which also hinders the study of their source mechanisms.

Some recent studies have applied existing deep-learning phase pickers, which were trained using regular tectonic earthquake waveforms, to monitor volcano seismicity (e.g., Bannister et al., 2022; Garza-Girón et al., 2023; Li et al., 2023; Mittal et al., 2022; Retailleau et al., 2022; Suarez et al., 2023; Wilding et al., 2023). However, there is currently no large-scale, systematic evaluation of the efficacy of these existing models for volcano monitoring. For instance, their performances for volcanic earthquakes may be impaired by different waveform characteristics, emergent onsets of long-period events, and high/different background noise in volcanic regions (Lapins et al., 2021). While there have been a few deep-learning studies using seismic data near volcanoes (Armstrong et al., 2023; Kim et al., 2023; Lapins et al., 2021; Manley et al., 2022; Titos et al., 2020), limited data distributions (individual volcano) make these models less generalizable to other volcanic regions. In addition, most deep-learning studies that involved long-period earthquakes focused on event classification (e.g., Manley et al., 2022; Titos et al., 2020), but none of the existing phase picking models explicitly included long-period earthquakes in their training and tests (e.g., Armstrong et al., 2023; Kim et al., 2023; Lapins et al., 2021).

In this study, we first compile a data set of seismic waveforms from various volcanic regions. We then show that the performances of two widely used deep-learning pickers, PhaseNet (Zhu & Beroza, 2019) and EQTransformer (Mousavi et al., 2020), deteriorate when applied off-the-shelf to volcanic seismic data, especially for long-period earthquakes. We then train new models that achieve significantly better performances for monitoring volcano seismicity.

2. Data Set of Seismic Waveforms From Volcanic Regions

We assemble a data set of 157,082 LP waveforms (35,181 events), 157,308 VT waveforms (38,335 events), and 20,000 noise waveforms recorded by 261 seismic stations around 34 volcanoes in Alaska from 1997 to 2017 (Power et al., 2019), 77 stations around six volcanoes in Hawaii from 2012 to 2021 (Hawaiian Volcano Observatory/USGS, 1956), 193 stations around eight volcanoes in Northern California from 1985 to 2019 (NCEDC, 2014), 225 stations around 10 volcanoes along the Cascade Range from 1981 to 2024 (University of Washington, 1963) and 973 stations around 88 volcanoes in Japan from 2004 to 2023 (National Research Institute for Earth Science and Disaster Resilience, 2019). The geographical distribution of the volcanoes colored by event numbers is shown in Figure 1. In Alaska, about 75% of the LPs are from Shishaldin, Spurr, Gareloi, and Iliamna, and about 50% of the VTs are from Spurr, Martin, Makushin, Augustine, and Redoubt. More than 80% of the LPs and VTs in Hawaii are from Kīlauea. In Japan, about 30% of the LPs are from Mount Meakan, Mount Fuji and Mount Yakedake and about 10% of the VTs are from Mount Kurikoma. In Northern California, ~70% of the LPs are from Mammoth Mountain while ~45% of the VTs are from the Clear Lake Volcanic Field. The LPs in the Cascade Range are dominated by Mount St. Helens, Mount Baker and Newberry (~70%). See Table S1 in Supporting Information S1 for data set splitting, Figure S1 in Supporting Information S1 for the event locations, Figure S2 in Supporting Information S1 for the recording stations, Figure S3 in Supporting Information S1 for the distribution of all volcanoes, Figure S4 in Supporting Information S1 for the time span of the data set, and Figures S5–S9 in Supporting Information S1 for the distributions of signal-to-noise ratios, source depths, epicentral distances and earthquake frequency content. All the event waveforms have both manually picked P and S phase arrivals. The LPs had already been labeled by analysts. We choose the noise waveforms from the same stations as the event waveforms by visual inspection. Most waveforms contain three components (77%) (Figure S10 in Supporting Information S1) and are from earthquakes located within 50 km of an active volcano (95%) (Figure S11 in Supporting Information S1). Since there are far more available VTs than LPs, we only include a similar

10.1029/2024GL108438

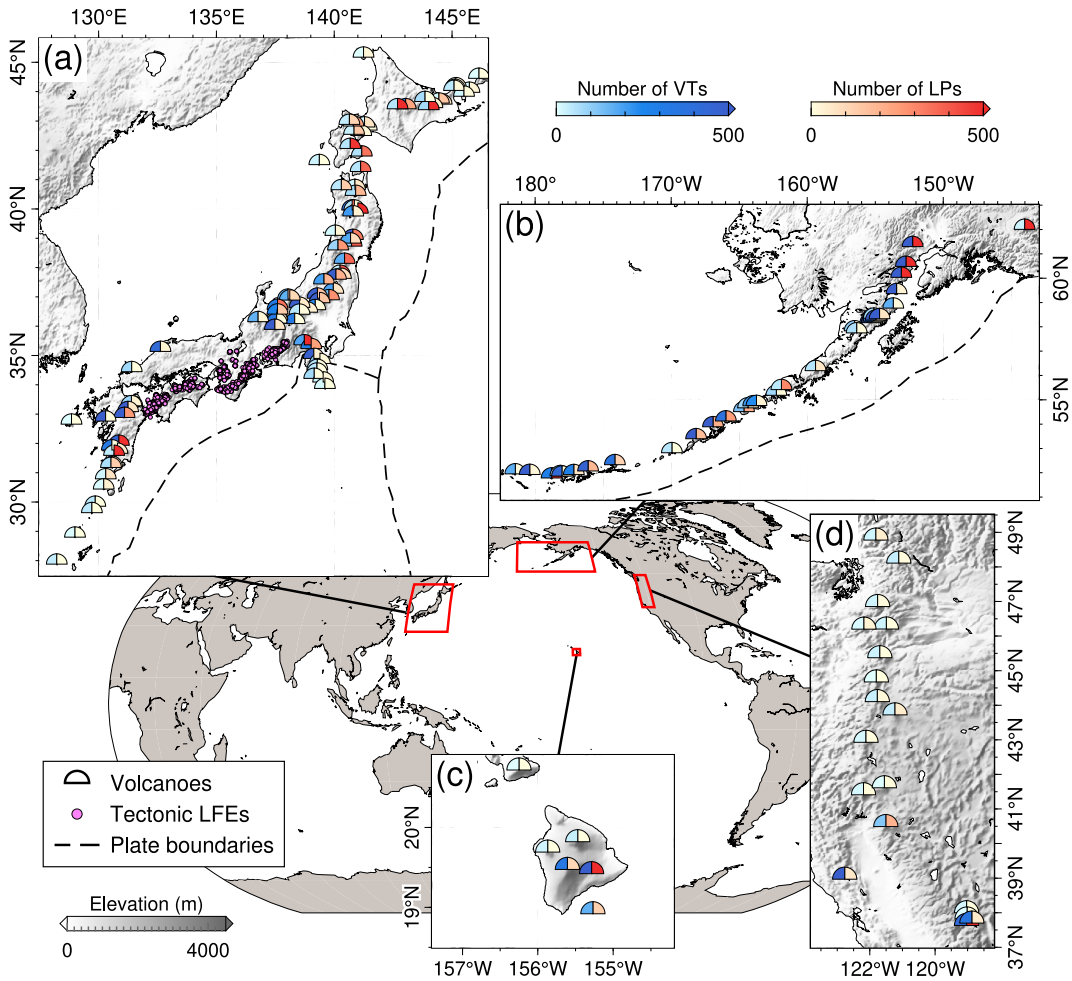

Figure 1. Geographical distribution of active volcanoes (filled semicircles) with seismic events used in our data set. The volcanoes are colored by the number of VTs (left halves, in blue) and LPs (right halves, in red) within 50 km. The seismic waveforms of volcano-tectonic earthquakes and volcanic long-period earthquakes from Japan (a), Alaska (b) and Hawaii (c) are split into a training set, a validation set and a test set, while the data from volcanoes along the Cascadia subduction zone in the Western United States (d) and the tectonic low-frequency earthquakes (LFEs) (purple circles in a) from Japan are only used for testing.

number of VT waveforms as the number of available LP waveforms. We remove data with large spikes and errors (e.g., events with S pick prior to P pick). For waveforms from Japan, we download event waveforms whose length may vary for different events and different stations. For waveforms from the US, we download event waveforms starting from 60 s before the P pick and ending 60 s after the S pick. Hence waveforms in our data set have different lengths, which will be trimmed in the subsequent processing stages. Compared with previous data sets, for example, STEAD (Mousavi et al., 2019) and INSTANCE (Michelini et al., 2021), our data set has a wider distribution of frequency index by design (Figure 2b and Figures S6 and S7 in Supporting Information S1) which is a measure of the dominant frequency content of an earthquake (Buurman & West, 2010; Matoza et al., 2014) (Text S1 in Supporting Information S1), suggesting it includes a greater variety of seismic events, although the waveform number of our data set (334,390) is one fourth that of STEAD (1,265,657) or INSTANCE (1,291,537). While a small amount of data from a single volcano may be enough to tune methods that lean on expert knowledge to study that volcano, a large and global data set is necessary for a deep neural network to learn generalizable knowledge from analyst picks. To the best of our knowledge, this is the first data set of seismic waveforms compiled from various volcanic regions globally for machine learning.

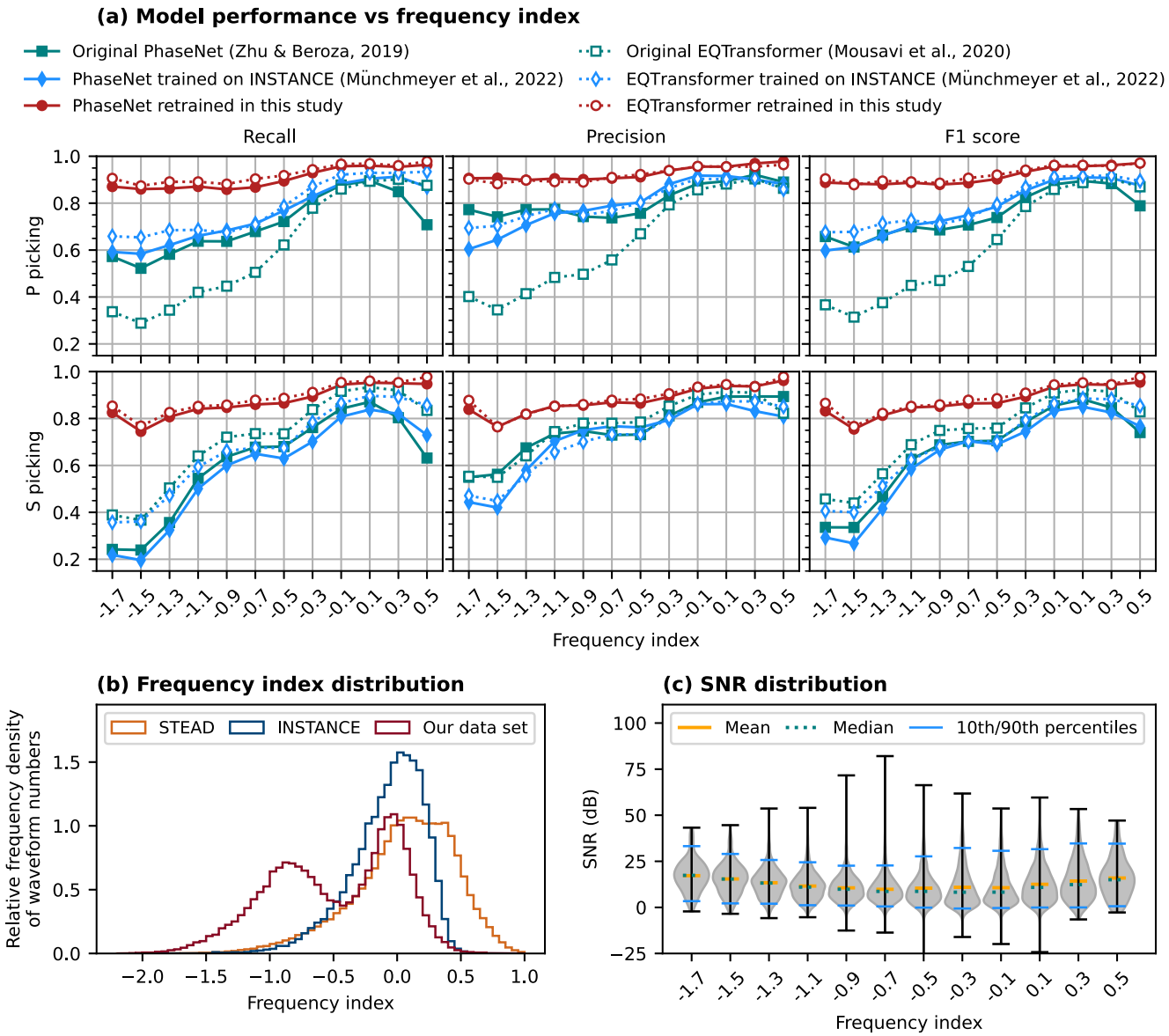

Figure 2. (a) Performances of various models on subsets of testing waveforms with different frequency index values. The tick labels on the *x* axis represent the centers of frequency index bins with a width of 0.2. The F1 scores here are slightly higher than those in Figure 3a because noise waveforms, to which frequency index is not applicable, are not included in this test. (b) Comparison of frequency index distributions of INSTANCE (Michelini et al., 2021) and STEAD (Mousavi et al., 2020) and our data set. (c) The distributions of signal to noise ratio for testing waveforms in different frequency index bins.

3. Evaluation of Existing Deep-Learning Phase Pickers

We use 15,078 LP waveforms and 15,057 VT waveforms from Alaska, Hawaii and Japan to evaluate two most widely used models: PhaseNet (Zhu & Beroza, 2019) and EQTransformer (Mousavi et al., 2020), which are the best performing architectures in a recent benchmark study (Münchmeyer et al., 2022). PhaseNet is a U-net with 1D convolutional layers originally trained on earthquakes from Northern California. EQTransformer is a stack of convolutional layers, long short-term memory (LSTM) units, and self-attentive layers originally trained on the global data set STEAD (Mousavi et al., 2019). We divide the testing waveforms into subsets according to frequency index values to evaluate how the model performance varies with the dominant frequency content. We randomly extract 30 s windows around the manual picks of the testing waveforms. For each waveform, the same window is used to test different models. Since EQTransformer operates on a 60 s window, we will only focus on the 30 s target window of the output (Münchmeyer et al., 2022). We use precision, recall and F1-score to evaluate

the results. Precision is the fraction of output picks that are actually correct. Recall is the fraction of manual picks that are correctly identified by the model. F1 score is the harmonic mean of precision and recall (Text S2 in Supporting Information S1). The original versions of PhaseNet and EQTransformer are tested here since they are most widely used. Considering that the original EQTransformer and PhaseNet were trained under the TensorFlow framework (Abadi et al., 2015) that is different from the platform we use (pyTorch) and that they were not trained on the same data set, we also include the variants of EQTransformer and PhaseNet trained on the INSTANCE data set (Michelini et al., 2021) for comparison, which were trained by Münchmeyer et al. (2022) and available in the SeisBench package (Woollam et al., 2022). The model output is time series of "probability" of P and S. To get predicted picks from the probability time series output by the models, we first extract segments of probability curves above a given threshold and the peak positions of these extracted segments are considered as pick times. The model-specific threshold is tuned (Figure S12 in Supporting Information S1) on the validation set (Table S1 in Supporting Information S1).

The recalls, precisions and F1 scores of the original models decrease systematically with decreasing frequency index (Figure 2). For example, the F1 score of PhaseNet decreases from ∼0.9 to ∼0.6 for P picking and from ∼0.85 to ∼0.35 for S picking as the frequency index decreases from ∼0.5 to ∼−1.7. Compared with precision, the recall exhibits a greater deterioration, which can be as low as 0.3 for P picking and 0.2 for S picking, indicating that most LPs in the test set have been overlooked. We observe a similar trend for the models trained on INSTANCE (Münchmeyer et al., 2022). This is unlikely to be related to changes in signal-to-noise ratio distribution since we do not observe significant systematic changes in signal-to-noise ratio with frequency index (Figure 2c and Figure S14 in Supporting Information S1). In addition, both the performances on LPs and VTs decrease with signal-to-noise ratio, and these models show lower performances for LPs than for VTs at the same signal-to-noise ratio (Figure S15 in Supporting Information S1). Our results suggest that these existing models will likely underreport LPs compared to VTs when directly applied to monitoring volcano seismicity (Bannister et al., 2022; Garza-Girón et al., 2023; Li et al., 2023; Mittal et al., 2022; Suarez et al., 2023; Wilding et al., 2023), which is not ideal since LPs often indicate fluid/magma movements (Chouet & Matoza, 2013; Matoza & Roman, 2022; Song et al., 2023). Therefore, we decided it would be valuable to train a new phase picker specifically for volcano seismicity.

4. Training Deep-Learning Phase Pickers for Volcano Seismicity

Among our data set, 151,431 LP waveforms, 151,657 VT waveforms and 20,000 noise waveforms from Alaska, Hawaii and Japan corresponding to 70,352 events are grouped into a training set (83.64%), a validation set (5.49%) and a test set (10.87%) (Table S1 in Supporting Information S1). Here, the earthquake waveforms in the test set are the same as those presented in the previous section. An extra test set comprising 5,651 waveforms from 1,295 LP events and 5,651 waveforms from 1,869 VT events near 18 volcanoes along the Cascadia subduction zone in the Western United States (Figure 1d) is used to test how our model generalizes to a region where no training data have been used. In addition, 6,224 waveforms of 2,356 tectonic low-frequency earthquakes (LFEs), which also have lower frequency indices, along the Nankai trough in Japan are used as another test set to investigate whether our model works for tectonic LFEs associated with shear slip on the subduction zone plate interface (Obara & Kato, 2016).

We use our data set to train two new models based on the PhaseNet and EQTransformer architectures implemented in the SeisBench package (Woollam et al., 2022). All the waveforms are resampled to 100 Hz. We normalize each component of a waveform by removing the mean and dividing it by the maximum value. We perform data augmentation by randomly modifying the waveforms at each step of training. The modifications include randomly shifting waveforms, adding gaps to waveforms, adding Gaussian noise and superimposing a training example on the shifted and normalized version of another training example that is randomly multiplied by a factor. Each type of augmentation is performed with a given probability. Normalization is performed before and after data augmentation. The labels for phase arrivals are Gaussian functions with peaks aligning with manual picks. Event types are not included as labels because the models are only aimed at phase picking while event classification can be conducted in a postprocessing stage using frequency index. At each step of training, a batch of waveform examples are randomly selected, normalized, randomly augmented, labeled, and input into the Adam optimization algorithm (Kingma & Ba, 2015) to adjust the model weights.

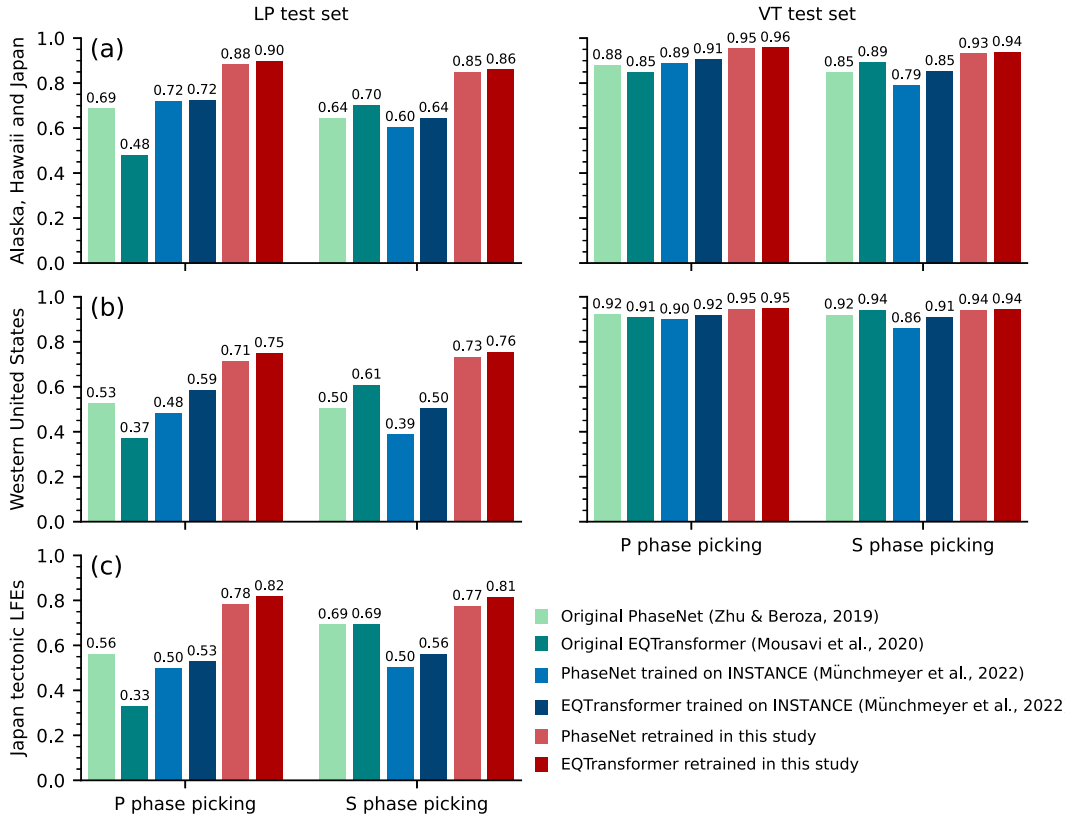

Figure 3. F1 scores of different models evaluated on the testing waveforms from (a) the same regions as the training data, (b) volcanoes along the Cascadia subduction zone in the western United States where no data have been used in the training and (c) tectonic LFEs in Japan. The precision and recall are given in Figures S23 and S24 in Supporting Information S1.

The validation set is used to tune hyperparameters. We try various learning rates 0.0001/0.0005/0.001 and batch sizes 512/1024 to obtain a series of models. Each model is trained for 400 epochs. Loss function on the validation set is monitored for each epoch and the model snapshot at the epoch with the lowest validation loss is used as the final model. For each model, we test different decision thresholds and choose the one with the highest F1-score as the optimal threshold. Then we evaluate each model on the validation set and choose the one with the highest F1-score (Tables S2 and S3 in Supporting Information S1). The preferred learning rate and batch size for PhaseNet are 0.0005 and 512, respectively. They are 0.001 and 1024 for EQTransformer, respectively. We also compare two initialization strategies: using (a) random weights and (b) the network weights pre-trained on the INSTANCE data set (Melnik et al., 2020; Münchmeyer et al., 2022), and we choose the resulting model with the highest F1-score on the validation set (Table S4 in Supporting Information S1).

We first test our models on subsets with different frequency index values as described in the previous section. Our models trained for volcano seismicity show significant performance improvement for waveforms with low frequency index values compared to existing models, with F1 scores for P and S picking of ∼0.9 and ∼0.8, respectively (Figure 2). There is also a slight improvement for waveforms with high frequency index. The overall performances of various models on the whole test set are shown in Figure 3a, where our models show the best performances for both LPs and VTs for both P and S picking. For the LPs, the EQTranformer-based network trained in this study achieves an F1 score of 0.9 for P picking and 0.86 for S picking, which are 0.42 (P picking) and 0.16 (S picking) higher than those of the original EQTransformer. The performance improvement is smaller for the VTs: the retrained EQTransformer achieves F1 scores 0.11 and 0.05 higher than the original EQTransformer model for P and S picking respectively. The EQTransformer trained on INSTANCE has similar performance to the original EQTransformer. A similar amount of improvement is obtained by the PhaseNet-based network trained on our data set. Furthermore, our models give lower picking residuals as indicated by the narrower histograms of residuals (Figures S18 and S19 in Supporting Information S1), with 2σ (i.e., two standard

deviations) of the retrained EQTransformer being 0.38 s for P residuals and 0.46 s for S residuals for LPs, and 0.21 s (P) and 0.32 s (S) for VTs. The retrained EQTransformer shows only a marginally higher F1 score than the retrained PhaseNet, suggesting that the data set plays a more important role than the network architecture in differences in model performances.

Subsequently, we use the test set from Northern California and the Cascade Range in the Western United States (Figure 1d) to investigate how our models generalize to regions where no training data are used (Figure 3b). All the models show great performance for VTs, with F1 scores for P picking larger than 0.9 and F1 scores for S picking larger than 0.86, and our models achieve the highest F1 scores (0.95). Notably, the existing pickers perform poorly for LPs, with F1 score ranging from 0.37 to 0.61. Although all the models experience some performance degradation for LPs compared with the previous test, our retrained models still perform significantly better than the existing models, with F1 scores ranging from 0.71 to 0.76. The performance variation with frequency index for this test set (Figure S16 in Supporting Information S1) also suggests that our models have better generalization abilities when applied to a new region. The poorer performances for LPs could be partly explained by the LP waveforms in this test set having lower signal-to-noise ratios than VT waveforms (Figures S5 and S16 in Supporting Information S1).

Finally, we investigate whether our models also work for tectonic LFEs since both tectonic LFEs and volcanic LPs appear to have similar frequency content, though they are often inferred to reflect different source processes (Aso et al., 2013). Our training set does not explicitly include any tectonic LFE. Here we test the models on LFEs along the Nankai trough from Japan. The result is shown in Figure 3c. Our retrained models outperform the original models and the INSTANCE-based models by 0.22–0.49 for P picking and 0.08–0.31 for S picking, with F1 scores of ∼0.8. We further confirmed that our models also work for regular tectonic earthquakes, since they achieve F1 scores of 0.89 and 0.75 for P and S picking respectively when tested on the INSTANCE data set (Michelini et al., 2021), which is slightly better than the original EQTransformer and PhaseNet but unsurprisingly inferior to the models trained on the INSTANCE data set (Figure S28 in Supporting Information S1).

5. Discussion

5.1. Comparison With Existing Methods

Deep-learning-based pickers have higher accuracy and require less parameters to manually tune than traditional pickers, for example, STA/LTA (Allen, 1978) and the Baer-Kradolfer picker (Baer & Kradolfer, 1987), as demonstrated in previous studies (e.g., Mousavi et al., 2020; Münchmeyer et al., 2022; Zhu & Beroza, 2019). Also, deep-learning-based pickers have greater flexibility than template matching as they are not limited by the availability of suitable template events, and they can contribute to generating more templates. Compared with previous deep-learning models aimed at tectonic earthquakes, our models can better pick volcano seismicity and thus are a step toward improving volcano monitoring. Our compiled data set can also be used to benchmark future methods for monitoring volcanic earthquakes.

Our study is different from a few recent studies that have also trained phase pickers on volcanic earthquakes (Armstrong et al., 2023; Kim et al., 2023; Lapins et al., 2021) in two aspects. First, the previous studies focused exclusively on one volcano and thus it is unclear how well these models can generalize to other volcanoes, while we use data around 146 active volcanoes from different regions. Further efforts in the community might be necessary to cultivate a better data set that covers the wide variety of volcanic signals observed. Second, LPs were not considered in the previous studies despite being an important form of volcano seismicity, while we explicitly included LP earthquakes to build a data set with a more balanced distribution of frequency content. We subsequently demonstrated that our models perform well for both LPs and VTs, and can be generalized to other volcanoes.

Finally, our study is different from recent studies which focused on tectonic LFEs (Lin et al., 2023; Münchmeyer et al., 2024; Thomas et al., 2021) in terms of training data and targets. These studies focused on tectonic LFEs which are a manifestation of creep or slow fault slips (Behr & Bürgmann, 2021), while our target is to pick volcano seismicity including both VTs and LPs. The capability of our models to pick tectonic LFEs is a side benefit and demonstrates that (a) our models are generalizable to other tectonic environments and (b) tectonic LFEs and volcanic LPs have relatively similar waveform characteristics.

5.2. Different Ways of Performance Evaluation

The presented evaluation results for different models depend on the metrics used and how they are calculated, which may vary in different studies. Therefore, it might not be appropriate to directly compare the values reported in different papers. For instance, some studies calculate true positive (TP), false positive (FP), true negative (TN) and false negative (FN) based on waveform traces so that any of the four outcomes TP/FP/TN/FN is assigned to each testing waveform (e.g., Mousavi et al., 2020; Zhu & Beroza, 2019). In this case, a waveform is considered as a true positive as long as there is a predicted pick sufficiently close to the manual pick even if there may also be some falsely predicted picks for the same waveform. Hence, false predictions may be underreported. In contrast, the definition of positive and negative in this paper is based on sampling points, where any of TP/FP/TN/FN is assigned to each sampling point of a waveform rather than the whole waveform (Text S2 in Supporting Information S1). The different definitions of FP and FN lead to different values of recall and precision. We have also calculated the model performances using the definition of positive/negative based on waveform traces (Mousavi et al., 2020; Zhu & Beroza, 2019), and the results (Figure S25 and S26 in Supporting Information S1) show similar trends as those presented in the previous section (Figures 2 and 3) except that the absolute values are slightly higher.

Alternatively, Münchmeyer et al. (2022) decomposed the evaluation into three tasks: event detection, phase identification and onset time picking. This evaluation workflow avoids the ambiguity in the definition of positive/negative for phase picking. However, it uses the maximum probability value within the tested window as the prediction result, which may be inconsistent with the practical application of a deep-learning picker where a trigger algorithm is used to retrieve picks from an output probability curve. Nevertheless, our models also show better performances than existing models when evaluated on the three tasks following Münchmeyer et al. (2022)'s workflow (Figures S20–S22 and Tables S5–S6 in Supporting Information S1), although existing models also perform well on the task of event detection which is easier than phase picking. Therefore, our models show consistently better performances than existing models regardless of the method of performance evaluation.

6. Conclusion

In this study, we first compile a data set of seismic waveforms from various volcanic regions globally, which has a wider distribution of frequency index than previous data sets of tectonic earthquakes. We then show that existing deep-learning-based phase pickers do not generalize well for volcanic earthquakes, with their performances deteriorating as the earthquakes' frequency content decreases, hence direct applications for monitoring volcano seismicity is suboptimal due to potential biases. Finally, we train and test new models using our data set. The test results show that our models can better pick P and S phases of VTs and LPs, and can be generalized to other regions not included in our training data set, including for tectonic LFEs. Therefore, our results can benefit future efforts to improve monitoring of volcano seismicity.

Data Availability Statement

The metadata of the data set in SeiBench format is preserved on Zenodo (Zhong & Tan, 2024a). The final models trained in this paper and the python scripts for data processing and model training/evaluation are available on github and Zenodo (Zhong & Tan, 2024b). All seismic waveforms used in this study are publicly available. The seismic waveforms and catalogs in Japan are from the Japan Meteorological Agency (www.jma.go.jp) and NIED Hi-net (National Research Institute for Earth Science and Disaster Resilience, 2019). The seismic data and catalogs for Hawaii and Alaska are from USGS (Hawaiian Volcano Observatory/USGS, 1956; Alaska Volcano Observatory/USGS, 1988) and Incorporated Research Institutions for Seismology Data Management center (IRIS-DMC, ds.iris.edu/ds/nodes/dmc). The seismic data and catalogs in the Western United States are from the Northern California Earthquake Data Center (NCEDC, 2014) and Pacific Northwest Seismic Network (University of Washington, 1963). We use the plate boundaries by Bird (2003) in Figure 1. The volcano locations are from the Japan Meteorological Agency (www.data.jma.go.jp/vois/data/tokyo/STOCK/souran_eng/menu.htm), Geological Survey of Japan (gbank.gsj.jp/volcano/Quat_Vol/index_e.html), Alaska Volcano Observatory (www.avo.alaska.edu/volcano/), Hawaiian Volcano Observatory (www.usgs.gov/observatories/hvo) and California Volcano Observatory (www.usgs.gov/observatories/calvo) and Cascades Volcano Observatory (www.usgs.gov/cascades-volcano-observatory). We use ObsPy (Krischer et al., 2015) and HinetPy (Tian et al., 2022) to facilitate

waveform downloading. We use the network architectures implemented in the SeisBench package (Woollam et al., 2022). We train the networks under the PyTorch framework (Paszke et al., 2019) using the pytorch-lightning package (github.com/Lightning-AI/pytorch-lightning).

Acknowledgments
This work is supported by the Direct Grant for Research (Grant 4053512) from the Chinese University of Hong Kong, Hong Kong RGC General Research Fund (Grant 14300422), and the Croucher Tak Wah Mak Innovation Award. We thank the editor Christian Huber as well as Alicia Hotovec-Ellis and two other anonymous reviewers for their constructive comments and suggestions. We thank Zhen Zhang, Peifeng Wang, Chengyan Fan and Jinping Zi for their suggestions on improving the figures, and Zhangyu Sun for discussion about evaluation metrics.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow, Large-scale machine learning on heterogeneous systems. https://doi.org/10.5281/zenodo.4724125

Acocella, V., Ripepe, M., Rivalta, E., Peltier, A., Galetto, F., & Joseph, E. (2023). Towards scientific forecasting of magmatic eruptions. *Nature Reviews Earth and Environment*, *5*, 1–18. https://doi.org/10.1038/s43017-023-00492-z

Alaska Volcano Observatory/USGS. (1988). Alaska volcano observatory [Dataset]. https://doi.org/10.7914/SN/AV

Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, *68*(5), 1521–1532. https://doi.org/10.1785/bssa0680051521

Armstrong, A. D., Claerhout, Z., Baker, B., & Koper, K. D. (2023). A deep-learning phase picker with calibrated Bayesian-derived uncertainties for earthquakes in the Yellowstone volcanic region. *Bulletin of the Seismological Society of America*, *113*(6), 2323–2344. https://doi.org/10.1785/0120230068

Aso, N., Ohta, K., & Ide, S. (2013). Tectonic, volcanic, and semi-volcanic deep low-frequency earthquakes in western Japan. *Tectonophysics*, *600*, 27–40. https://doi.org/10.1016/j.tecto.2012.12.015

Aso, N., & Tsai, V. C. (2014). Cooling magma model for deep volcanic long-period earthquakes. *Journal of Geophysical Research: Solid Earth*, *119*(11), 8442–8456. https://doi.org/10.1002/2014jb011180

Baer, M., & Kradolfer, U. (1987). An automatic phase picker for local and teleseismic events. *Bulletin of the Seismological Society of America*, *77*(4), 1437–1445. https://doi.org/10.1785/bssa0770041437

Bannister, S., Bertrand, E. A., Heimann, S., Bourguignon, S., Asher, C., Shanks, J., & Harvison, A. (2022). Imaging sub-caldera structure with local seismicity, Okataina Volcanic Centre, Taupo Volcanic Zone, using double-difference seismic tomography. *Journal of Volcanology and Geothermal Research*, *431*, 107653. https://doi.org/10.1016/j.jvolgeores.2022.107653

Behr, W. M., & Bürgmann, R. (2021). What's down there? The structures, materials and environment of deep-seated slow slip and tremor. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2193), 20200218. https://doi.org/10.1098/rsta.2020.0218

Bird, P. (2003). An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems*, *4*(3). https://doi.org/10.1029/2001gc000252

Buurman, H., & West, M. E. (2010). Seismic precursors to volcanic explosions during the 2006 eruption of Augustine Volcano. In J. A. Power, M. L. Coombs, & J. T. Freymueller (Eds.), *The 2006 eruption of Augustine Volcano, Alaska* (pp. 41–57). U.S. Geological Survey.

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

Chamberlain, C. J., Hopp, C. J., Boese, C. M., Warren-Smith, E., Chambers, D., Chu, S. X., et al. (2017). EQcorrscan: Repeating and near-repeating earthquake detection and analysis in Python. *Seismological Research Letters*, *89*(1), 173–181. https://doi.org/10.1785/0220170151

Chen, H., Yang, H., Zhu, G., Xu, M., Lin, J., & You, Q. (2022). Deep outer-rise faults in the southern Mariana subduction zone indicated by a machine-learning-based high-resolution earthquake catalog. *Geophysical Research Letters*, *49*(12), e2022GL097779. https://doi.org/10.1029/2022gl097779

Chouet, B. A., & Matoza, R. S. (2013). A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *Journal of Volcanology and Geothermal Research*, *252*, 108–175. https://doi.org/10.1016/j.jvolgeores.2012.11.013

Garza-Girón, R., Brodsky, E. E., Spica, Z. J., Haney, M. M., & Webley, P. W. (2023). A specific earthquake processing workflow for studying long-lived, explosive volcanic eruptions with application to the 2008 Okmok Volcano, Alaska, eruption. *Journal of Geophysical Research: Solid Earth*, *128*(5), e2022JB025882. https://doi.org/10.1029/2022jb025882

Gibbons, S. J., & Ringdal, F. (2006). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, *165*(1), 149–166. https://doi.org/10.1111/j.1365-246x.2006.02865.x

Gong, J., Fan, W., & Parnell-Turner, R. (2023). Machine learning-based new earthquake catalog illuminates on-fault and off-fault seismicity patterns at the Discovery Transform Fault, East Pacific Rise. *Geochemistry, Geophysics, Geosystems*, *24*(9), e2023GC011043. https://doi.org/10.1029/2023gc011043

Hawaiian Volcano Observatory/USGS. (1956). Hawaiian volcano observatory network [Dataset]. https://doi.org/10.7914/SN/HV

Jiang, C., Zhang, P., White, M. C. A., Pickle, R., & Miller, M. S. (2022). A detailed earthquake catalog for Banda Arc–Australian Plate collision zone using machine-learning phase picker and an automated workflow. *The Seismic Record*, *2*(1), 1–10. https://doi.org/10.1785/0320210041

Kim, A., Nakamura, Y., Yukutake, Y., Uematsu, H., & Abe, Y. (2023). Development of a high-performance seismic phase picker using deep learning in the Hakone volcanic area. *Earth Planets and Space*, *75*(1), 1–15. https://doi.org/10.1186/s40623-023-01840-5

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kriegerowski, M., Petersen, G. M., Vasyura-Bathke, H., & Ohrnberger, M. (2019). A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms. *Seismological Research Letters*, *90*(2A), 510–516. https://doi.org/10.1785/0220180320

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., & Wassermann, J. (2015). ObsPy: A bridge for seismology into the scientific Python ecosystem. *Computational Science and Discovery*, *8*(1), 014003. https://doi.org/10.1088/1749-4699/8/1/014003

Lapins, S., Goitom, B., Kendall, J.-M., Werner, M. J., Cashman, K. V., & Hammond, J. O. S. (2021). A little data goes a long way: Automating seismic phase arrival picking at Nabro Volcano with transfer learning. *Journal of Geophysical Research: Solid Earth*, *126*(7), e2021JB021910. https://doi.org/10.1029/2021jb021910

Li, J., Tian, Y., Zhao, D., Yan, D., Li, Z., & Li, H. (2023). Magmatic system and seismicity of the Arxan volcanic group in Northeast China. *Geophysical Research Letters*, *50*(6), e2022GL101105. https://doi.org/10.1029/2022gl101105

Lin, J.-T., Thomas, A., Bachelot, L., Toomey, D., Searcy, J., & Melgar, D. (2023). Detection of hidden low-frequency earthquakes in Southern Vancouver Island with deep learning.

Liu, M., Li, L., Zhang, M., Lei, X., Nedimović, M. R., Plourde, A. P., et al. (2023). Complexity of initiation and evolution of the 2013 Yunlong earthquake swarm. *Earth and Planetary Science Letters*, *612*, 118168. https://doi.org/10.1016/j.epsl.2023.118168

Liu, M., Tan, Y. J., Lei, X., Li, H., Zhang, Y., & Wang, W. (2024). Intersection between tectonic faults and magmatic systems promotes swarms with large-magnitude earthquakes around the Tengchong volcanic field, southeastern Tibetan Plateau. *Geology*, *52*(4), 302–307. https://doi.org/10.1130/g51796.1

Manley, G. F., Mather, T. A., Pyle, D. M., Clifton, D. A., Rodgers, M., Thompson, G., & Londoño, J. M. (2022). A deep active learning approach to the automatic classification of volcano-seismic events. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/feart.2022.807926

Matoza, R. S., & Roman, D. C. (2022). One hundred years of advances in volcano seismology and acoustics. *Bulletin of Volcanology*, *84*(9), 86. https://doi.org/10.1007/s00445-022-01586-0

Matoza, R. S., Shearer, P. M., & Okubo, P. G. (2014). High-precision relocation of long-period events beneath the summit region of Kīlauea Volcano, Hawai'i, from 1986 to 2009. *Geophysical Research Letters*, *41*(10), 3413–3421. https://doi.org/10.1002/2014gl059819

Melnik, O., Lyakhovsky, V., Shapiro, N. M., Galina, N., & Bergal-Kuvikas, O. (2020). Deep long period volcanic earthquakes generated by degassing of volatile-rich basaltic magmas. *Nature Communications*, *11*(1), 3918. https://doi.org/10.1038/s41467-020-17759-4

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

Mittal, T., Jordan, J. S., Retailleau, L., Beauducel, F., & Peltier, A. (2022). Mayotte 2018 eruption likely sourced from a magmatic mush. *Earth and Planetary Science Letters*, *590*, 117566. https://doi.org/10.1016/j.epsl.2022.117566

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., Giffard-Roisin, S., Malfante, M., Frank, W., Poli, P., Marsan, D., & Socquet, A. (2024). Deep learning detects uncataloged low-frequency earthquakes across regions. *Seismica*, *3*(1). https://doi.org/10.26443/seismica.v3i1.1185

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

National Research Institute for Earth Science and Disaster Resilience. (2019). NIED Hi-net [Dataset]. *National Research Institute for Earth Science and Disaster Resilience*. https://doi.org/10.17598/NIED.0003

NCEDC. (2014). Northern California Earthquake Data Center [Dataset]. https://doi.org/10.7932/NCEDC

Obara, K., & Kato, A. (2016). Connecting slow earthquakes to huge earthquakes. *Science*, *353*(6296), 253–257. https://doi.org/10.1126/science.aaf1512

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pitt, A. M., Hill, D. P., Walter, S. W., & Johnson, M. J. S. (2002). Midcrustal, long-period earthquakes beneath Northern California volcanic areas. *Seismological Research Letters*, *73*(2), 144–152. https://doi.org/10.1785/gssrl.73.2.144

Power, J. A., Friberg, P. A., Haney, M. M., Parker, T., Stihler, S. D., & Dixon, J. P. (2019). *A unified catalog of earthquake hypocenters and magnitudes at volcanoes in Alaska—1989 to 2018* (Tech. Rep.). US Geological Survey. https://doi.org/10.3133/sir20195037

Retailleau, L., Saurel, J.-M., Laporte, M., Lavayssière, A., Ferrazzini, V., Zhu, W., et al. (2022). Automatic detection for a comprehensive view of Mayotte seismicity. *Comptes Rendus. Géoscience*, *354*(S2), 153–170. https://doi.org/10.5802/crgeos.133

Ross, Z. E., Meier, M., 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

Saccorotti, G., & Lokmer, I. (2021). Chapter 2—A review of seismic methods for monitoring and understanding active volcanoes. In P. Papale (Ed.), *Forecasting and planning for volcanic hazards, risks, and disasters* (Vol. 2, pp. 25–73). Elsevier. https://doi.org/10.1016/b978-0-12-818082-2.00002-0

Shapiro, N. M., Droznin, D., Droznina, S. Y., Senyukov, S., Gusev, A., & Gordeev, E. (2017). Deep and shallow long-period volcanic seismicity linked by fluid-pressure transfer. *Nature Geoscience*, *10*(6), 442–445. https://doi.org/10.1038/ngeo2952

Song, Z., Tan, Y. J., & Roman, D. C. (2023). Deep long-period earthquakes at Akutan volcano from 2005 to 2017 better track magma influxes compared to volcano-tectonic earthquakes. *Geophysical Research Letters*, *50*(10), e2022GL101987. https://doi.org/10.1029/2022gl101987

Soto, H., & Schurr, B. (2021). DeepPhasePick: A method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks. *Geophysical Journal International*, *227*(2), 1268–1294. https://doi.org/10.1093/gji/ggab266

Suarez, E., Domínguez-Cerdeña, I., Villaseñor, A., Aparicio, S. S.-M., del Fresno, C., & García-Cañada, L. (2023). Unveiling the pre-eruptive seismic series of the La Palma 2021 eruption: Insights through a fully automated analysis. *Journal of Volcanology and Geothermal Research*, *444*, 107946. https://doi.org/10.1016/j.jvolgeores.2023.107946

Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., et al. (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. https://doi.org/10.1785/0320210001

Teney, D., Lin, Y., Oh, S. J., & Abbasnejad, E. (2022). ID and OOD performance are sometimes inversely correlated on real-world datasets. In *Neural information processing systems*.

Thomas, A. M., Inbal, A., Searcy, J., Shelly, D. R., & Bürgmann, R. (2021). Identification of low-frequency earthquakes on the San Andreas Fault with deep learning. *Geophysical Research Letters*, *48*(13), e2021GL093157. https://doi.org/10.1029/2021gl093157

Tian, D., Kriegerowski, M., & Sawaki, Y. (2022). seisman/HinetPy: 0.7.1 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.6810553

Titos, M., Bueno, A., García, L., Benítez, C., & Segura, J. C. (2020). Classification of isolated volcano-seismic events based on inductive transfer learning. *IEEE Geoscience and Remote Sensing Letters*, *17*(5), 869–873. https://doi.org/10.1109/lgrs.2019.2931063

University of Washington. (1963). Pacific Northwest Seismic Network—University of Washington [Dataset]. https://doi.org/10.7914/SN/UW

Wech, A. G., Thelen, W. A., & Thomas, A. M. (2020). Deep long-period earthquakes generated by second boiling beneath Mauna Kea volcano. *Science*, *368*(6492), 775–779. https://doi.org/10.1126/science.aba4798

Wenzel, F., Dittadi, A., Gehler, P. V., Simon-Gabriel, C.-J., Horn, M., Zietlow, D., et al. (2022). Assaying out-of-distribution generalization in transfer learning. In *Neural information processing systems*.

White, R. A., & McCausland, W. A. (2019). A process-based model of pre-eruption seismicity patterns and its use for eruption forecasting at dormant stratovolcanoes. *Journal of Volcanology and Geothermal Research*, *382*, 267–297. https://doi.org/10.1016/j.jvolgeores.2019.03.004

Wilding, J. D., Zhu, W., Ross, Z. E., & Jackson, J. M. (2023). The magmatic web beneath Hawai'i. *Science*, *379*(6631), 462–468. https://doi.org/10.1126/science.ade5755

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

Zhang, Z., Liu, M., Tan, Y. J., Walter, F., He, S., Chmiel, M., & Su, J. (2024). Landslide hazard cascades can trigger earthquakes. *Nature Communications*, *15*(1), 2878. https://doi.org/10.1038/s41467-024-47130-w

Zhong, Y., & Tan, Y. J. (2024a). The metadata of the dataset of volcanic earthquakes for machine learning compiled in this study [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.11206597

Zhong, Y., & Tan, Y. J. (2024b). The models trained in this study and the scripts for data processing and model training/evaluation [Model]. *Zenodo*. https://doi.org/10.5281/zenodo.11199022

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