Deep-learning-based phase picking for volcano seismicity

Yiyuan Zhong[1], **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 the first data set of seismic waveforms from various volcanic regions globally.
- We show that existing deep-learning phase pickers' performances deteriorate with decreasing volcanic earthquake frequency content.
- Our retrained models perform better and are more generalizable for monitoring volcano seismicity, especially long-period earthquakes.

Corresponding author: Yen Joe Tan, `yjtan@cuhk.edu.hk`

Abstract

The application of deep-learning-based seismic phase pickers for earthquake monitoring has surged in recent years. However, the efficacy of these models when applied to monitoring volcano seismicity has yet to be evaluated. Here, we first compile a dataset 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 volcanic earthquake waveforms from northern California where no training data are used 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 dataset 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, a type of artificial intelligence, can excel in picking the arrival times of regular earthquakes. However, the efficacy of these models when applied to monitoring volcanic earthquakes has yet to be evaluated. Here, we first compile a dataset of earthquakes from various volcanoes globally. We then show that existing deep-learning-based models do not work well for these events, especially those with predominantly low-frequency energy. We then train two new models which perform better than existing models for volcanic earthquakes. 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 (Gibbons & Ringdal, 2006; Chamberlain et al., 2017) 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 pickers (e.g. Ross et al., 2018; Zhu & Beroza, 2019; Mousavi et al., 2020; Soto & Schurr, 2021) 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. Tan et al., 2021; Jiang et al., 2022; Chen et al., 2022; Gong et al., 2023; Liu et al., 2023; Wilding et al., 2023; Garza-Girón et al., 2023). However, it can be difficult to predict deep-learning models' performance for out-of-distribution data that are not well represented by training data (Wenzel et al., 2022; Teney et al., 2022).

Seismicity which often correlate with magmatic/volcanic processes and sometimes represent eruption precursors (White & McCausland, 2019; Acocella et al., 2023) 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; Saccorotti & Lokmer, 2021; Matoza & Roman, 2022, 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, e.g. 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 bubble 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 because they are depleted of high frequency content and have emergent phase onsets (Pitt et al., 2002; Shapiro et al., 2017).

Some recent studies have applied existing deep-learning phase pickers, which were trained using regular tectonic earthquake waveforms, to monitor volcano seismicity (Mittal et al., 2022; Bannister et al., 2022; Suarez et al., 2023; Li et al., 2023; Garza-Girón 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 models trained with seismic data near volcanoes (Lapins et al., 2021; Kim et al., 2023; Armstrong et al., 2023), limited data distribution (individual volcano) make these models less generalizable to other volcanic regions. In addition, none of these studies explicitly included long-period earthquakes in their analyses (Lapins et al., 2021; Kim et al., 2023; Armstrong et al., 2023).

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 Dataset of seismic waveforms from volcanic regions

We assemble a data set of 156,272 LP waveforms (34,980 events), 156,498 VT waveforms (38,115 events), and 20,000 noise waveforms recorded by seismic stations deployed around 34 volcanoes in Alaska (Power et al., 2019), 6 volcanoes in Hawaii (Hawaiian Volcano Observatory/USGS, 1956), 8 volcanoes in northern California (NCEDC, 2014) and 88 volcanoes in Japan (National Research Institute for Earth Science and Disaster Resilience, 2019). The geographical distribution of the events is shown in Figure 1. See Table S1 in the supporting information for more details about data set splitting, Figure S1 for the distribution of recording stations, Figure S2 for the distribution of volcanoes and Figures S3-S14 for other properties of the data. All the event waveforms have both manually picked P and S phase arrivals. Most waveforms contain 3 components (77%) (Figure S3) and are from earthquakes located within 50 km of an active volcano (95%) (Figure S4). Since there are far more available VTs than LPs, we only include a similar 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 60s before the P pick and ending 60s 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 datasets, e.g. STEAD (Mousavi et al., 2019) and INSTANCE (Michelini et al., 2021), our data set has a wider distribution of frequency index (Figures S7-S10)

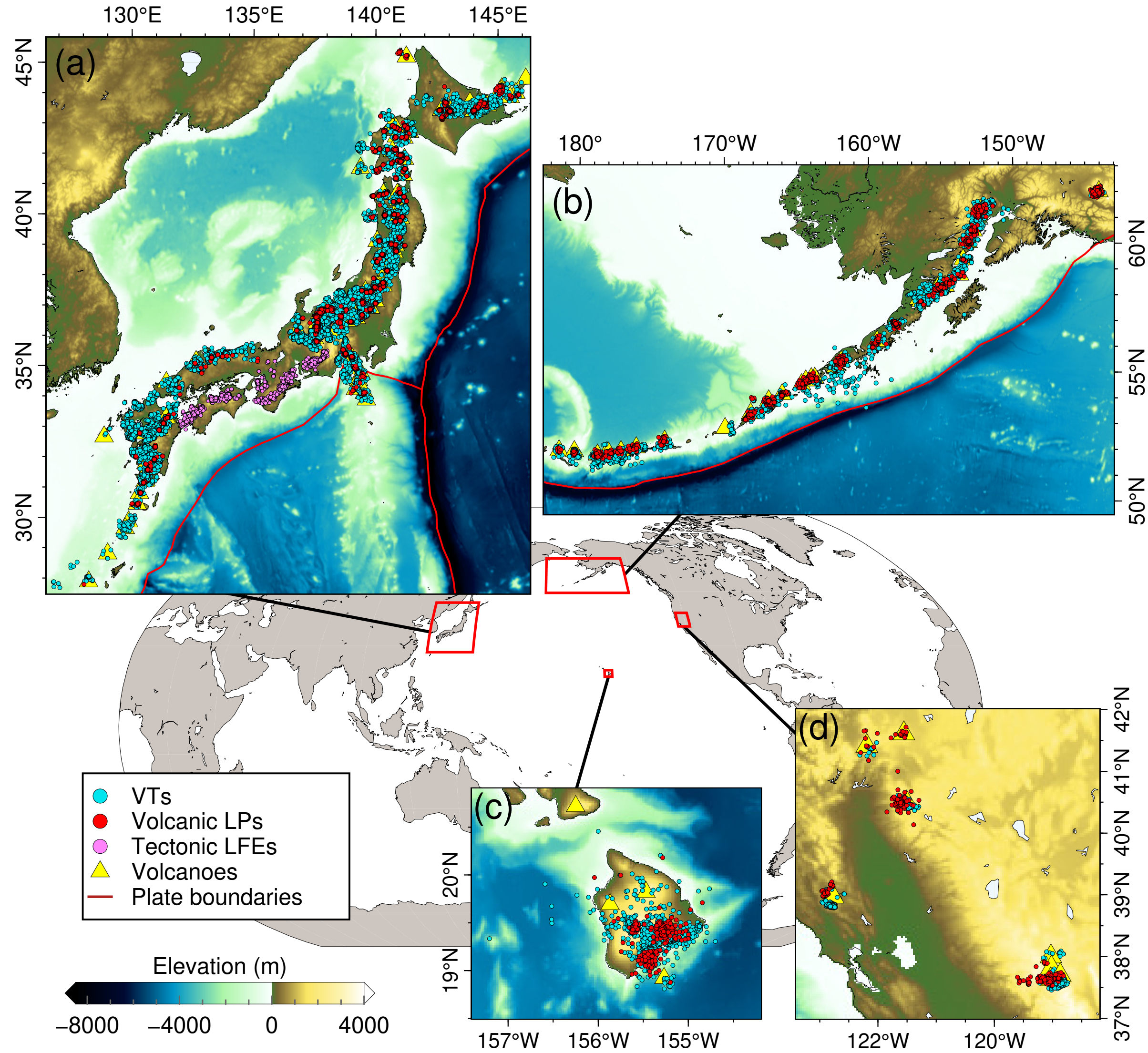

Figure 1. Geographical distribution of the earthquakes used in this study. The seismic data of volcano-tectonic earthquakes (cyan circles) and volcanic long-period earthquakes (red circles) from Japan (a), Alaska (b) and Hawaii (c) are split into a training set, a validation set and a test set, while the data from northern California (d) and the tectonic low-frequency earthquakes (LFEs) (purple circles) from Japan are only used for testing. Yellow triangles mark active volcanoes with seismic events used in this study.

which is a measure of the dominant frequency content of an earthquake (Buurman & West, 2010) (Text S1), suggesting it includes a greater variety of seismic events. To the best of our knowledge, this is the first data set of seismic waveforms compiled from various volcanic regions globally for machine learning.

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 30s 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 60s window, we will only focus on the 30s 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). 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 S15) on the validation set (Table 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 ∼0.5 for P picking and from ∼0.85 to ∼0.25 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 0.4 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 since we do not observe significant systematic changes in signal-to-noise ratio with frequency index (Figure S17). 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; Mittal et al., 2022; Wilding et al., 2023; Garza-Girón et al., 2023; Suarez et al., 2023; Li et al., 2023), which is not ideal since LPs often indicate fluid/magma movements (Chouet & Matoza, 2013; Matoza & Roman, 2022). 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). Here, the earthquake waveforms in the test set are the same as those presented in the previous section. An extra test set comprising 4,841 waveforms from 1,094 LP events and 4,841 waveforms from 1,649 VT events near 8 volcanoes in northern California is used

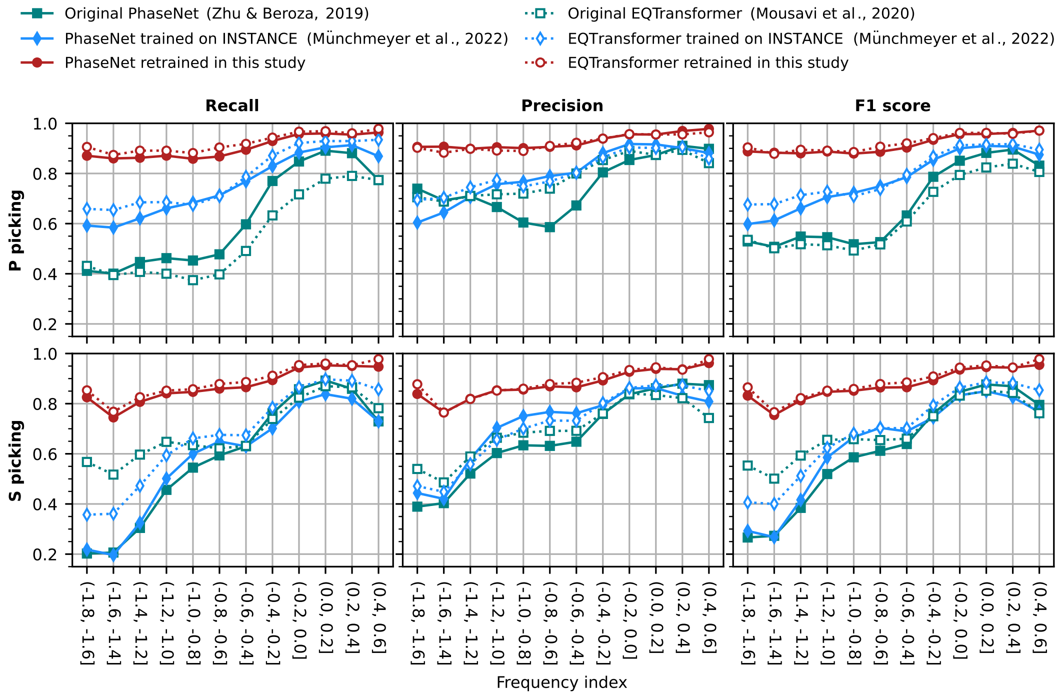

Figure 2. Performances of various models on subsets of testing waveforms with different frequency index values. The F1 scores here is slightly higher than those in Figure 3a because noise waveforms, to which frequency index is not applicable, are not included in this test.

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) 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 rescaled version of another training example. 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. At each step of training, a batch of waveform examples are randomly selected, normalized, randomly augmented, labelled, and input into the Adam optimization algorithm (Kingma & Ba, 2015) to adjust the model weights.

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 S2-3). 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 random initialization and initialization from the network weights pre-trained on the INSTANCE data set (Melnik et al., 2020; Münchmeyer et al., 2022), and we choose the one with the highest F1-score on the validation set (Table S4).

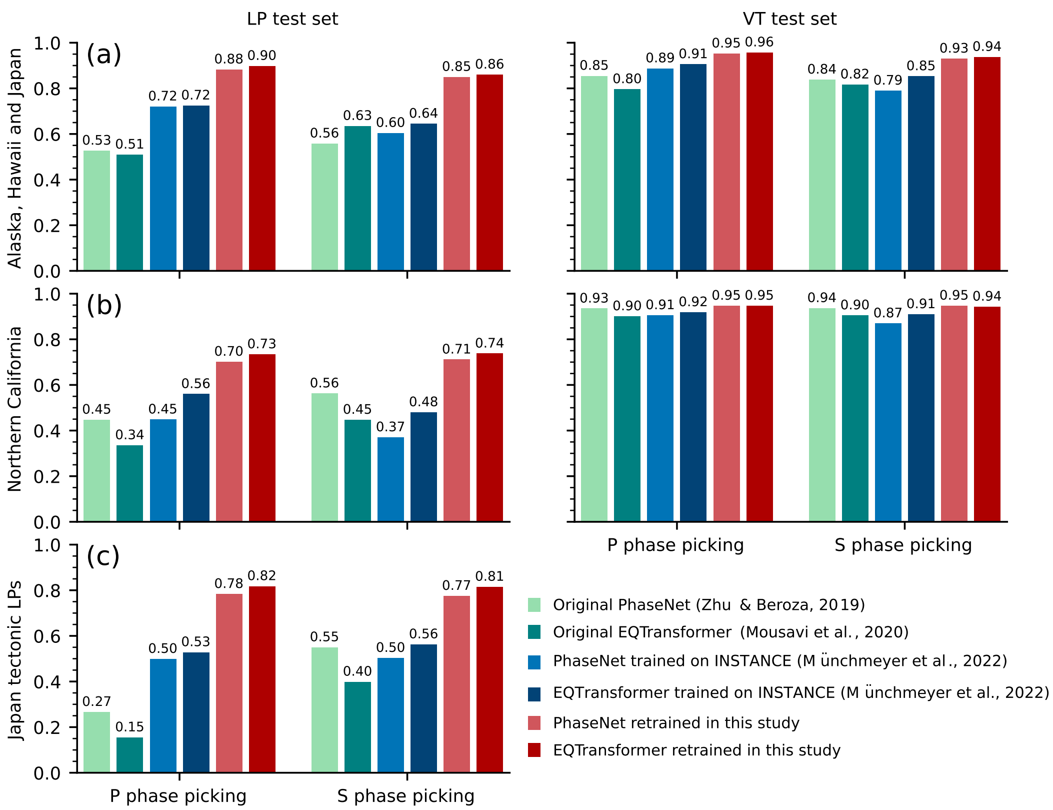

Figure 3. F1 scores of different models evaluated on the testing waveforms from (a) the same regions as the training data, (b) northern California from where no training data are used and (c) tectonic LFEs in Japan. The precision and recall are given in Figures S24-S25 in the supplement.

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 0.39 (P picking) and 0.23 (S picking) higher than those of the original EQTransformer. The performance improvement is smaller for the VTs: the retrained EQTransformer achieves F1 scores 0.16 and 0.12 higher than the original EQTransformer model for P and S picking respectively. The EQTransformer trained on INSTANCE has similar performance to the original EQTransformer except for P picking on the LPs, for which the F1 score of the INSTANCE-based EQTransformer is ∼0.2 higher than that of the original EQTransformer but ∼0.2 lower than that of our retrained 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 (Figure S19-S20). 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 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 0.87, and our models achieve the highest F1 scores (0.95). Notably, the existing pickers perform poorly for LPs, with F1 score ranging from 0.34 to 0.56. 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 0.70 to 0.74. The performance variation with frequency index for this test set (Figure S18) 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 S6 and S18).

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 a large margin for both P and S picking, with F1 scores of $\sim$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 S29).

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, e.g. STA/LTA (Allen, 1978) and the Baer-Kradolfer picker (Baer & Kradolfer, 1987), as demonstrated in previous studies (e.g. Zhu & Beroza, 2019; Mousavi et al., 2020; Münchmeyer et al., 2022). Also, deep-learning-based pickers have greater flexibility than template matching as they are not limited by the availability of suitable template events. Compared with previous deep-learning models aimed at tectonic earthquakes, our models can better pick volcano seismicity and thus can help to improve volcano monitoring. Our compiled waveform dataset 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 models on volcanic earthquakes (Lapins et al., 2021; Kim et al., 2023; Armstrong et al., 2023) 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

136 active volcanoes from different regions. Second, LPs were not considered in the previous studies despite being an important form of volcano seismicity, while we included LP earthquakes for training. We subsequently demonstrated that our models perform well for both LPs and VTs, and can be generalized to other volcanoes. However, since these studies adopted different data formats, input/output formats, machine-learning frameworks and not all of these models are available, it would be hard to make direct comparisons.

Finally, our study is different from recent studies which focused on tectonic LFEs (Thomas et al., 2021; Lin et al., 2023; Münchmeyer et al., 2023) 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 (1) our models are generalizable to other tectonic environments and (2) 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. Zhu & Beroza, 2019; Mousavi et al., 2020). 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). 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 (Zhu & Beroza, 2019; Mousavi et al., 2020), and the results (Figure S26-S27) show similar trends as those presented in the previous section (Figure 2-3) except that the absolute values are slightly higher.

Alternatively, Münchmeyer et al. (2022) decomposed the evaluation into 3 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 3 tasks following Münchmeyer et al. (2022)'s workflow (Figure S21-S23 and Table S5-S6), 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 dataset of seismic waveforms from various volcanic regions globally, which has a wider distribution of frequency index than previous datasets 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 with 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.

Open Research Section

Our models have been uploaded for peer review, with the archiving at Zenodo currently underway. All seismic data used in this study are publicly available. The seismic waveforms and catalogs in Japan are from the Japan Meteorological Agency (`http://www.jma.go.jp`) and the National Research Institute for Earth Science and Disaster Resilience (`https://www.hinet.bosai.go.jp`) (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, `https://ds.iris.edu/ds/nodes/dmc`). The seismic data and catalogs for northern California are from the Northern California Earthquake Data Center (NCEDC, 2014) (`https://ncedc.org`). We use the plate boundaries by Bird (2003) in Figure 1. The volcano locations are from the Japan Meteorological Agency (`https://www.data.jma.go.jp/vois/data/tokyo/STOCK/souran_eng/menu.htm`), Geological Survey of Japan (`https://gbank.gsj.jp/volcano/Quat_Vol/index_e.html`), Alaska Volcano Observatory (`https://www.avo.alaska.edu/volcano/`), Hawaiian Volcano Observatory (`https://www.usgs.gov/observatories/hvo`) and California Volcano Observatory (`www.usgs.gov/observatories/calvo`). 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 (Falcon & The PyTorch Lightning team, 2019).

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.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015, November). *TensorFlow, Large-scale machine learning on heterogeneous systems.* doi: 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 & Environment*, 1–18.

Alaska Volcano Observatory/USGS. (1988). *Alaska volcano observatory.* International Federation of Digital Seismograph Networks. Retrieved from `https://www.fdsn.org/networks/detail/AV/` doi: 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.

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.

Aso, N., Ohta, K., & Ide, S. (2013). Tectonic, volcanic, and semi-volcanic deep low-frequency earthquakes in western japan. *Tectonophysics*, *600*, 27–40.

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. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JB011180` doi: 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.

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. doi: 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 and Engineering Sciences*, *379*(2193), 20200218. doi: 10.1098/rsta.2020.0218

Bird, P. (2003). An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems*, *4*(3). doi: 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., … Team, E. C. (2020). Using a deep neural network and transfer learning to bridge scales for seismic phase picking. *Geophysical Research Letters*, *47*(16), e2020GL088651.

Chamberlain, C. J., Hopp, C. J., Boese, C. M., Warren-Smith, E., Chambers, D., Chu, S. X., … Townend, J. (2017). EQcorrscan: Repeating and Near-Repeating Earthquake Detection and Analysis in Python. *Seismological Research Letters*, *89*(1), 173-181.

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.

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.

Falcon, W., & The PyTorch Lightning team. (2019, March). *PyTorch Lightning.* Retrieved from `https://github.com/Lightning-AI/lightning` doi: 10.5281/zenodo.3828935

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.

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.

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.

Hawaiian Volcano Observatory/USGS. (1956). *Hawaiian volcano observatory network.* International Federation of Digital Seismograph Networks. Retrieved from `https://www.fdsn.org/networks/detail/HV/` doi: 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.

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.

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.*

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 & Discovery*, *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.

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.

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., … Li, H. (2023). Complexity of initiation and evolution of the 2013 yunlong earthquake swarm. *Earth and Planetary Science Letters*, *612*, 118168. Retrieved from https://www.sciencedirect.com/science/article/pii/S0012821X23001814 doi: https://doi.org/10.1016/j.epsl.2023.118168

Matoza, R. S., & Roman, D. C. (2022). One hundred years of advances in volcano seismology and acoustics. *Bulletin of Volcanology*, *84*(9), 86.

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.

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.

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. doi: 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.

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.

Münchmeyer, J., Giffard-Roisin, S., Malfante, M., Frank, W., Poli, P., Marsan, D., & Socquet, A. (2023). *Deep learning detects uncataloged low-frequency earthquakes across regions.*

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., … others (2022). Which picker fits my data? a quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, *127*(1), e2021JB023499.

National Research Institute for Earth Science and Disaster Resilience. (2019). *NIED Hi-net, National Research Institute for Earth Science and Disaster Resilience.* `https://www.hinet.bosai.go.jp`. doi: 10.17598/NIED.0003

NCEDC. (2014). *Northern california earthquake data center.* `https://ncedc.org/`. doi: 10.7932/NCEDC

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (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. Retrieved from `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

Pitt, A. M., Hill, D. P., Walter, S. W., & Johnson, M. J. S. (2002, 03). Midcrustal, Long-period Earthquakes beneath Northern California Volcanic Areas. *Seismological Research Letters*, *73*(2), 144-152. doi: 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.

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.

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, p. 25-73). Elsevier. Retrieved from `https://www.sciencedirect.com/science/article/pii/B9780128180822000020` doi: 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.

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.

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.

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. doi: https://doi.org/10.1016/j.jvolgeores.2023.107946

Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., ... Segou, M. (2021). Machine-Learning-Based High-Resolution Earthquake Catalog Reveals How Complex Fault Structures Were Activated during the 2016–2017 Central Italy Sequence. *The Seismic Record*, *1*(1), 11-19.

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.

Tian, D., Kriegerowski, M., & Sawaki, Y. (2022). *seisman/hinetpy: 0.7.1.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6810553 doi: 10.5281/zenodo.6810553

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. Retrieved from https://www.science.org/doi/abs/10.1126/science.aba4798 doi: 10.1126/science.aba4798

Wenzel, F., Dittadi, A., Gehler, P. V., Simon-Gabriel, C.-J., Horn, M., Zietlow, D., ... Locatello, F. (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.

Wilding, J. D., Zhu, W., Ross, Z. E., & Jackson, J. M. (2023). The magmatic web beneath hawai'i. *Science*, *379*(6631), 462-468.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., ... Soto, H. (2022). SeisBench–A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, *93*(3), 1695-1709.

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