Supporting Information for "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 SAR, China

Contents of this files

1. Text S1 to S2
2. Tables S1 to S6
3. Figures S1 to S29

Introduction

Figures S1-S14 and Table S1 show the properties of the data used in this study. Tables S2-S4 and Figures S15-S16 show the process of hyperparameter tuning. Figure S17 shows the distribution of signal to noise ratio for different subsets of the test data from Alaska, Japan and Hawaii. Figure S18 shows the model performance versus frequency index for the testing waveforms from northern California. Figure S19 and S20 show the histograms of picking residuals for VTs and LPs, respectively. Tables S6-S5 and Figures S21-S23 show the evaluation results for the 3 tasks defined in (Münchmeyer et al., 2022). Figures S24 and S25 show the recalls and precisions of different models, respectively. Figures S26

and S27 show the F1 scores calculated using the definition of positive and negative based on waveform traces (Zhu & Beroza, 2019; Mousavi et al., 2020). Figures S28 and S29 show the performances of different models on the INSTANCE data set (Michelini et al., 2021).

Text S1. Frequency index

Frequency index (FI) is a metric used to quantify the dominant frequency content of an earthquake from seismic waveforms (Buurman & West, 2010),

$$FI = \log_{10} \frac{\overline{A}_{upper}}{\overline{A}_{lower}}, \tag{1}$$

where $\bar{A}_{lower}$ and $\bar{A}_{upper}$ are the mean spectral amplitudes in a predefined high-frequency band and a low-frequency band, respectively. Following Song, Tan, and Roman (2023), we choose 1-5 Hz and 10-15 Hz as the low and high frequency bands, respectively. Time windows starting 1s prior to and ending 6s after P arrivals are extracted to calculate FIs. If there are multiple components at a station, the average of FI values of available components is used as the FI value for this station. The frequency index of a seismic event is defined as the average of FIs at all stations that have recorded this event (Matoza et al., 2014).

Text S2. Performance metrics

Since phase picking is not a binary classification task, we need to redefine positive and negative to calculate precision, recall and F1-score. There are discrepancies in performance reporting among different researchers. Some studies consider a waveform trace as a true positive as long as there is a predicted pick sufficiently close to the labeled pick on this waveform (Zhu & Beroza, 2019; Mousavi et al., 2020). However, false predictions may be

underestimated when the model predicts incorrect picks at the same time, leading to a higher reported precision.

Here, we base the definition of positives and negatives on sampling points (points sampled from a continuous analog signal) instead of entire waveform traces. The model output is time series of "probability" of P and S. To get predicted picks from the probability time output by the models, we first extract segments of probability curves above a given decision threshold and the peak positions of these extracted segments are considered as predicted pick times. If a predicted pick occurs within a threshold around a true pick, it is counted as a true positive prediction (TP). Following (Mousavi et al., 2020), the threshold is chosen as 0.5s. Note that this threshold for distinguishing true picks from false picks is different from the probability threshold which is used to extract picks from a "probability" time series output by the model. In the case where there are multiple predicted picks near the true pick, they are counted as only one true positive. If there are no predictions within 0.5s around a true pick, it is counted as a false negative (FN). If there are no true picks around a predicted pick, it is counted as a false positive (FP). Precision is the fraction of predicted picks that are actually correct, calculated as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. Recall is the fraction of testing manual picks that have been correctly identified by the model, calculated as $\mathrm{TP}/(\mathrm{TP}+FN)$. F1 score is calculated as $2\times(\mathrm{Precision}+\mathrm{Recall})/(\mathrm{Precision}+\mathrm{Recall})$, which is the harmonic mean of the precision and recall. Those samples that are not labeled as true phase arrivals and also not picked by the model are considered as true negatives (TN). For example, considering a 30s waveform with a sampling rate of 100Hz which contains 3001 samples, if there is one manual

P pick and the model gives 10 predicted picks one of which is close to the manual P pick, there are 1 TP, 9 FPs and 2991 TNs. Considering that true negatives are not involved in precision and recall and they heavily outnumber TP, FP and FN, we do not count the number of true negatives when calculating precision, recall and F1 score.

Münchmeyer et al. (2022) evaluated the performance of a model in terms of 3 tasks: (1) event detection, (2) phase identification and (3) onset time picking. For event detection, they used 1 minus noise probability as the score for detection. If the peak detection score for a waveform is above a given threshold, the waveform is considered as a positive detection. ROC (receiver operating characteristic curve) and its AUC (area under the curve) value are used to evaluate the detection performance. In phase identification, they used the ratio of the maximum value of P probability to the maximum value of S probability as the decision score. The Matthews correlation coefficient was used to evaluate the phase identification. The fraction of outliers, root mean square error and mean absolute error were used to evaluate onset time picking. To generate a proper testing set for phase identification and onset time picking, they randomly selected a 10s window around P or S arrivals for each testing waveform, and make sure only one phase is located in the selected window. However, this way of evaluation use the maximum probability value within the tested window as the prediction result, which is different from practical applications of a deep-learning picker where a trigger algorithm is used to retrieve picks from an output probability curve.

References

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.

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.

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.

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., 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.

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.

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival-

time picking method. *Geophysical Journal International*, *216*(1), 261–273.

Table S1. The number of waveform traces in our dataset, including volcano-tectonic earthquakes (VTs), long-period earthquakes (LPs) and noise. The number of corresponding events is given in brackets. Since different waveforms may originate from the same source, the sum of events in the training, validation and testing sets does not necessarily equal the total events. Note that splitting waveforms from the same event to different data sets does not result in data leakage, because waveforms recorded at different stations have been influenced by different path effects and different background noise, thus representing unique examples. The 4,841 LP waveforms and 4,841 VT waveforms in northern California as well as the 6,224 LFE waveforms in Japan that are used as extra test sets are not shown in this table.

	Total traces	Training set	Validation set	Test set
Whole dataset	323,088 (70,352)	270,224 (68,996)	17,744 (13,346)	35,120 (23,700)
Earthquake	303,088 (70,352)	257,763 (68,996)	15,190 (13,346)	30,135 (23,700)
Noise	20,000 (0)	12,461 (0)	2,554 (0)	4,985 (0)
LP earthquakes	151,431 (33,886)	128,802 (33,364)	7,551 (6,609)	15,078 (11,798)
VT earthquakes	151,657 (36,466)	128,961 (35,632)	7,639 (6,737)	15,057 (11,902)
Alaska LPs	51,942 (15,701)	44,263 (15,511)	2,544 (2,370)	5,135 (4,497)
Alaska VTs	50,899 (15,519)	43,198 (15,151)	2,598 (2,377)	5,103 (4,354)
Hawaii LPs	16,906 (2,351)	14,404 (2,323)	811 (666)	1,691 (1,132)
Hawaii VTs	16,814 (2,766)	14,346 (2,702)	806 (653)	1,662 (1,119)
Japan LPs	82,583 (15,834)	70,135 (15,530)	4,196 (3,573)	8,252 (6,169)
Japan VTs	83,944 (18,181)	71,417 (17,779)	4,235 (3,707)	8,292 (6,429)

Table S2. Performance metrics on the validation set for 12 PhaseNet networks trained with different hyperparameters: learning rate, batch size and σ_{label} which is the standard deviation of the Gaussian function used for labeling training data. The networks were randomly initialized. Each model is trained up to 400 epochs, and the epoch at which the loss on the validation set is the lowest is saved as the final result. MAE is the mean absolute error of picks. The picking residuals outside the interval $(-1,\ 1)s$ are considered as outliers and not involved in the calculation of MAE. The row with the higest F1 score is highlighted in bold face. For each network, we have tried various decision thresholds and choose the one with the highest F1-score as the optimal threshold. Figure S16 presents the threshold tuning for preferred models.

Network	Hyperparameters			Decision threshold		F1 score		MAE (s)	
	Batch size	Learning rate	σ_{label}	P picking	S picking	P picking	S picking	P picking	S picking
PhaseNet	1024	0.0010	20	0.31	0.34	0.9169	0.8842	0.0767	0.1148
PhaseNet	1024	0.0010	10	0.29	0.23	0.9110	0.8762	0.0750	0.1186
PhaseNet	1024	0.0005	20	0.32	0.31	0.9158	0.8844	0.0779	0.1162
PhaseNet	1024	0.0005	10	0.29	0.25	0.9124	0.8787	0.0762	0.1162
PhaseNet	1024	0.0001	20	0.32	0.31	0.9090	0.8773	0.0810	0.1182
PhaseNet	1024	0.0001	10	0.30	0.25	0.9005	0.8643	0.0766	0.1200
PhaseNet	512	0.0010	20	0.31	0.31	0.9157	0.8843	0.0778	0.1173
PhaseNet	512	0.0010	10	0.29	0.24	0.9115	0.8756	0.0745	0.1187
PhaseNet	**512**	**0.0005**	**20**	**0.39**	**0.34**	**0.9181**	**0.8866**	**0.0755**	**0.1146**
PhaseNet	512	0.0005	10	0.28	0.24	0.9134	0.8782	0.0758	0.1184
PhaseNet	512	0.0001	20	0.37	0.34	0.9106	0.8805	0.0788	0.1171
PhaseNet	512	0.0001	10	0.27	0.24	0.9001	0.8644	0.0809	0.1230

Table S3. Performance metrics on the validation set for 12 EQTransformer networks trained with different hyperparameters: learning rate, batch size and σ_{label} which is the standard deviation of the Gaussian function used for labeling training data. The networks were randomly initialized. Each model is trained up to 400 epochs, and the epoch at which the loss on the validation set is the lowest is saved as the final result. MAE is the mean absolute error of picks. The picking residuals outside the interval $(-1,\ 1)s$ are considered as outliers and not involved in the calculation of MAE. The row with the highest F1 score is highlighted in bold face. For each network, we have tried various decision thresholds and choose the one with the highest F1-score as the optimal threshold. Figure S16 presents the threshold tuning for preferred models.

Network	Hyperparameters			Decision threshold		F1 score		MAE (s)	
	Batch size	Learning rate	σ_{label}	P picking	S picking	P picking	S picking	P picking	S picking
EQTransformer	**1024**	**0.0010**	**20**	**0.22**	**0.25**	**0.9245**	**0.8919**	**0.0877**	**0.1242**
EQTransformer	1024	0.0010	10	0.15	0.16	0.9212	0.8878	0.0856	0.1182
EQTransformer	1024	0.0005	20	0.23	0.24	0.9216	0.8905	0.0911	0.1271
EQTransformer	1024	0.0005	10	0.16	0.15	0.9176	0.8861	0.0860	0.1241
EQTransformer	1024	0.0001	20	0.23	0.27	0.9149	0.8842	0.0956	0.1307
EQTransformer	1024	0.0001	10	0.15	0.16	0.9148	0.8814	0.0924	0.1283
EQTransformer	512	0.0010	20	0.19	0.27	0.9232	0.8887	0.0887	0.1238
EQTransformer	512	0.0010	10	0.17	0.13	0.9213	0.8869	0.0855	0.1230
EQTransformer	512	0.0005	20	0.22	0.23	0.9216	0.8916	0.0895	0.1243
EQTransformer	512	0.0005	10	0.13	0.15	0.9191	0.8855	0.0868	0.1215
EQTransformer	512	0.0001	20	0.20	0.18	0.9166	0.8817	0.0957	0.1350
EQTransformer	512	0.0001	10	0.15	0.14	0.9133	0.8798	0.0900	0.1264

Table S4. Performance metrics of the models trained with random initial weights and those first initialized with pre-trained weights. The performance is evaluated on the validation set. For pre-training, we use the network weights pre-trained on INSTANCE dataset (Münchmeyer et al., 2022) as the starting point before training. The hyperparameters batch size, learning rate and σ_{label} are the same as the preferred ones highlighted in bold face in Table S2-S3.

Network	Initialized with weights pre-trained on	Decision threshold		F1 score		MAE (s)	
		P picking	S picking	P picking	S picking	P picking	S picking
EQTransformer	None	0.22	0.25	0.9245	0.8919	0.0877	0.1242
EQTransformer	INSTANCE	0.22	0.22	0.9250	0.8916	0.0876	0.1256
PhaseNet	None	0.39	0.34	0.9181	0.8866	0.0755	0.1146
PhaseNet	INSTANCE	0.39	0.34	0.9175	0.8833	0.0750	0.1165

Table S5. AUC scores for the event detection task defined by (Münchmeyer et al., 2022, section 2.1.1), which is much simpler than picking. If the peak of the output probability curve for a test example is larger than the threshold, it is considered as a positive prediction.

Model	LP test set	VT test set
EQTransformer retrained in this study	0.9993	0.9994
PhaseNet retrained in this study	0.9992	0.9994
Original EQTransformer(Mousavi et al., 2020)	0.9776	0.9839
Original PhaseNet (Zhu & Beroza, 2019)	0.9932	0.9937
EQTransformer trained on INSTANCE (Münchmeyer et al., 2022)	0.9934	0.9907
PhaseNet trained on INSTANCE (Münchmeyer et al., 2022)	0.9695	0.9784

Table S6. Matthews correlation coefficients for the phase discrimination task (Münchmeyer et al., 2022).

Model	LP test set	VT test set
EQTransformer retrained in this study	0.9621	0.9787
PhaseNet retrained in this study	0.9570	0.9764
Original EQTransformer (Mousavi et al., 2020)	0.7899	0.9086
Original PhaseNet (Zhu & Beroza, 2019)	0.7333	0.9354
EQTransformer trained on INSTANCE (Münchmeyer et al., 2022)	0.7717	0.9422
PhaseNet trained on INSTANCE (Münchmeyer et al., 2022)	0.8330	0.9463

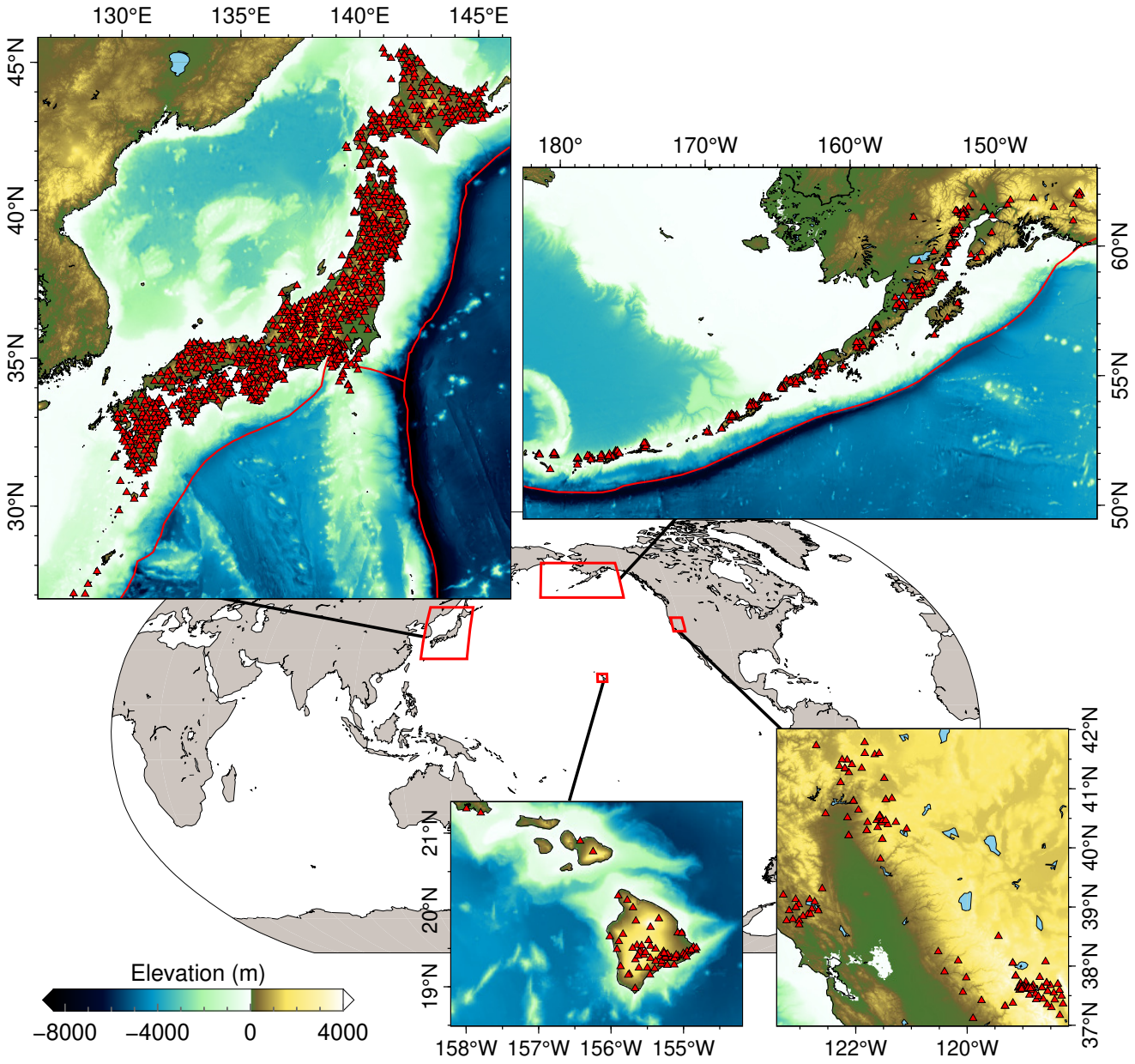

Figure S1. The geographical distribution of seismic stations (red triangles) with waveforms included in our data set, including the data set in Table S1, the northern California test set and the test set of Japan tectonic LPs.

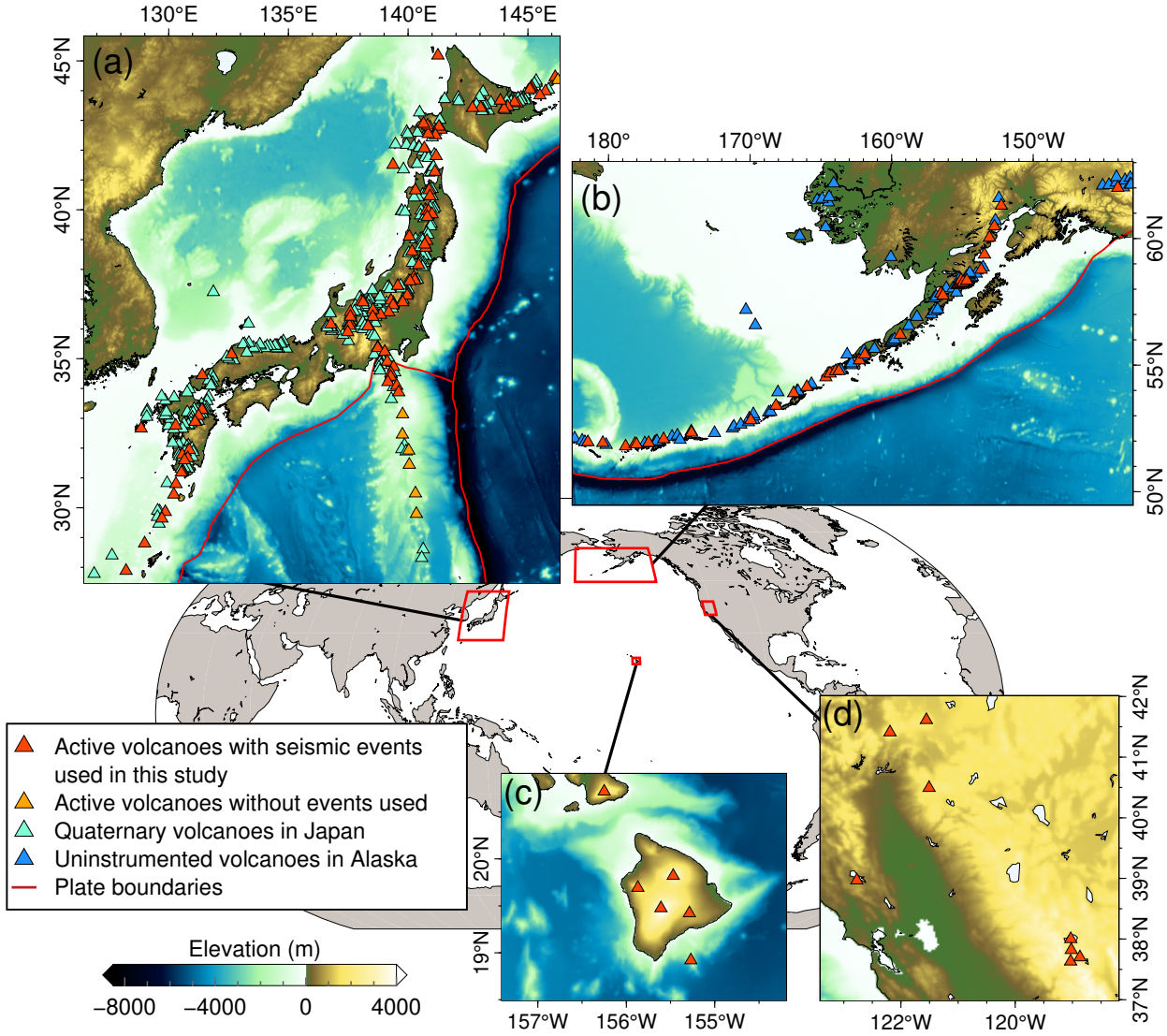

Figure S2. The geographical distribution of the active volcanoes with seismic events included in our data set (red triangles). The active volcanoes in Japan without seismic events use (orange triangles), quaternary volcanoes in Japan (green triangles) and uninstumented volcanoes in Alaska (blue triangles) are also shown.

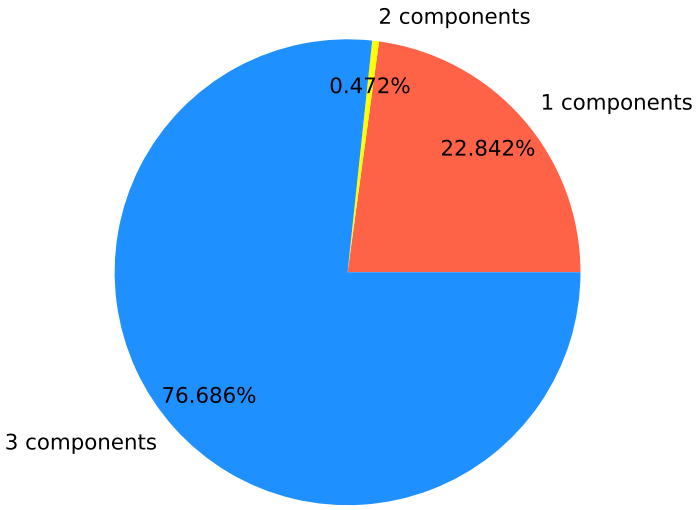

Figure S3. The proportion of seismograms with different numbers of components in Table S1. For one-component or two-component records, we fill in zeros for the remaining components before training.

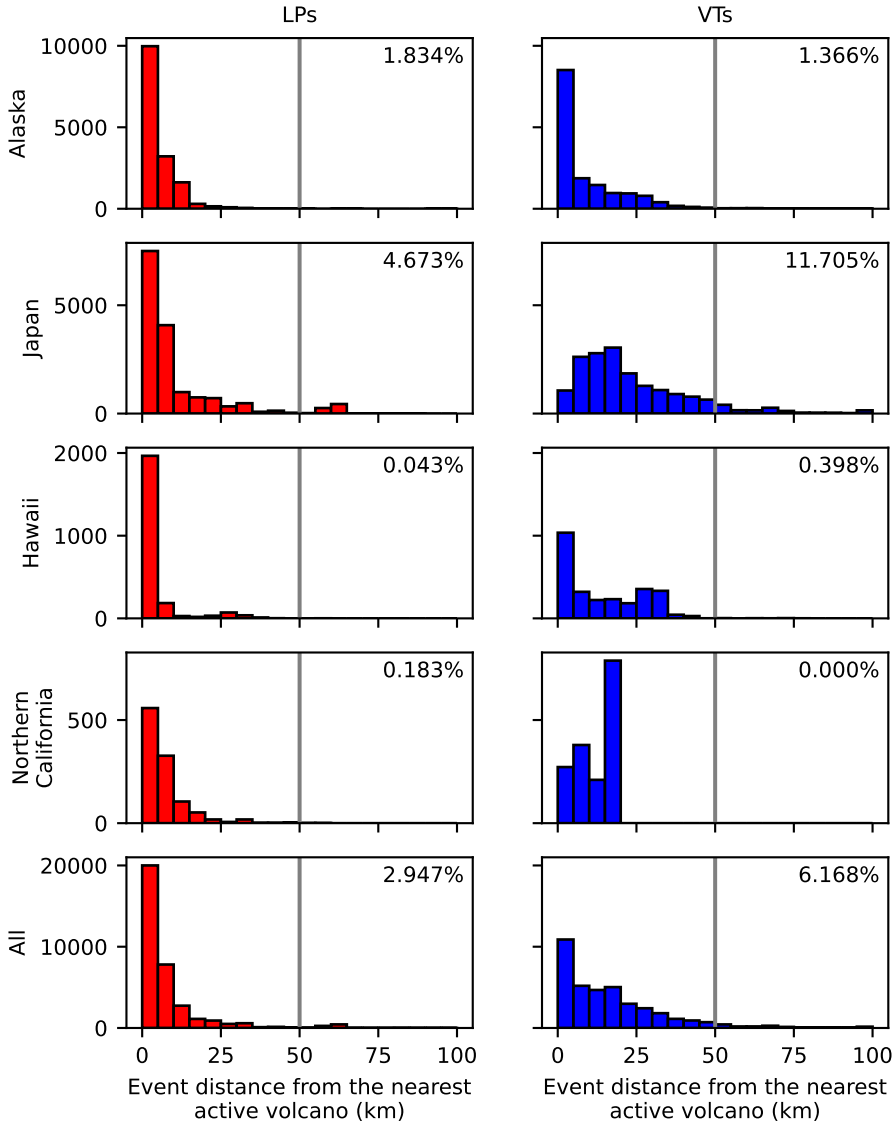

Figure S4. Histogram of the distances of events in our data set to the nearest active volcano. The numbers in the top right corner indicate the fractions of events that are more than 50 km away from the neareat volcano.

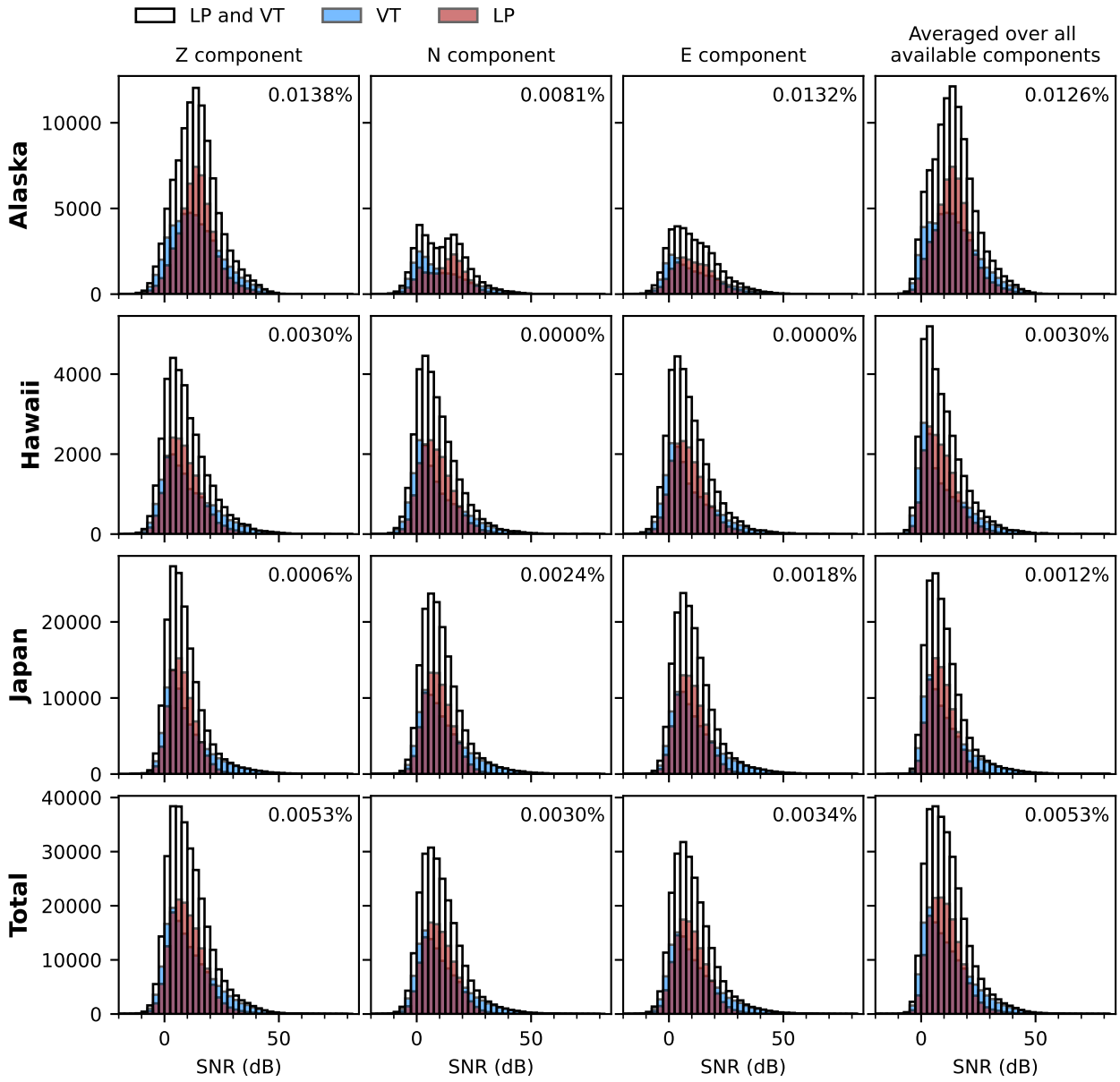

Figure S5. The distribution of signal-to-noise ratios (SNR) of the waveform traces in Table S1. Each column shows SNRs of waveforms from different regions for the same component, while each row represents SNRs of waveforms from the same region but for different components. The numbers in the top right corner indicate the fraction of samples outside the range of the x-axis. SNR is calculated as $\text{SNR} = 20 \log_{10} \frac{|S|_{95}}{|N|_{95}}$, where $|S|_{95}$ is the 95 percentile of absolute amplitudes in a 5s window right after the S arrival and $|N|_{95}$ is that in a 5s window before the P arrival.

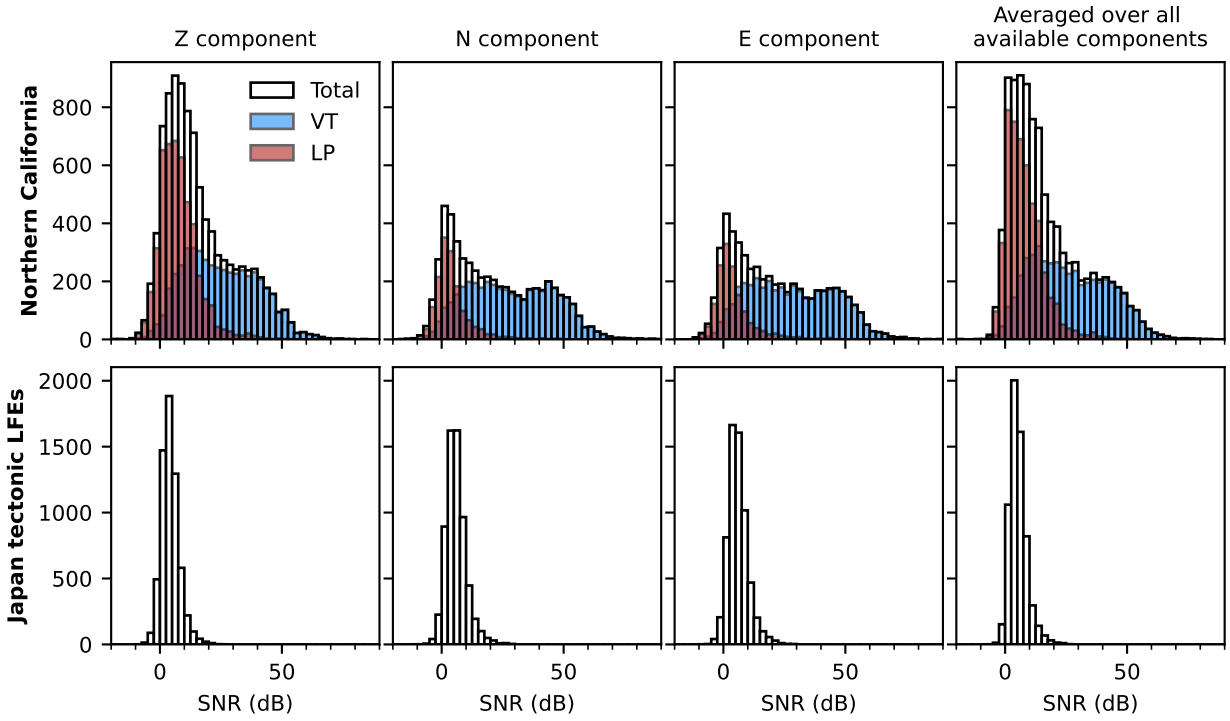

Figure S6. The distribution of signal-to-noise ratios (SNR) of the waveform traces in the northern California test set and the test set of Japan tectonic LFEs.

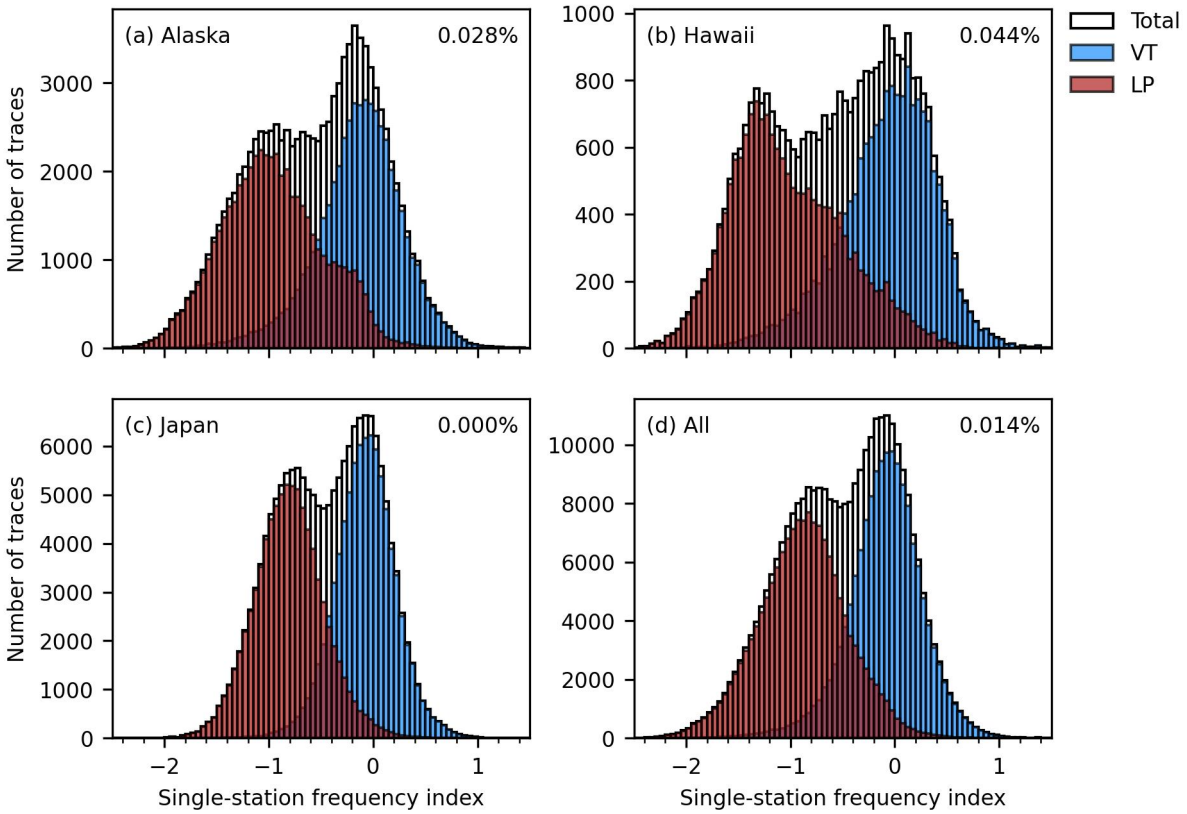

Figure S7. The distribution of single-station frequency index (FI) values of the earthquake waveforms in Table S1. The numbers in the top right corner indicate the fraction of samples outside the range of the plotted x-axis.

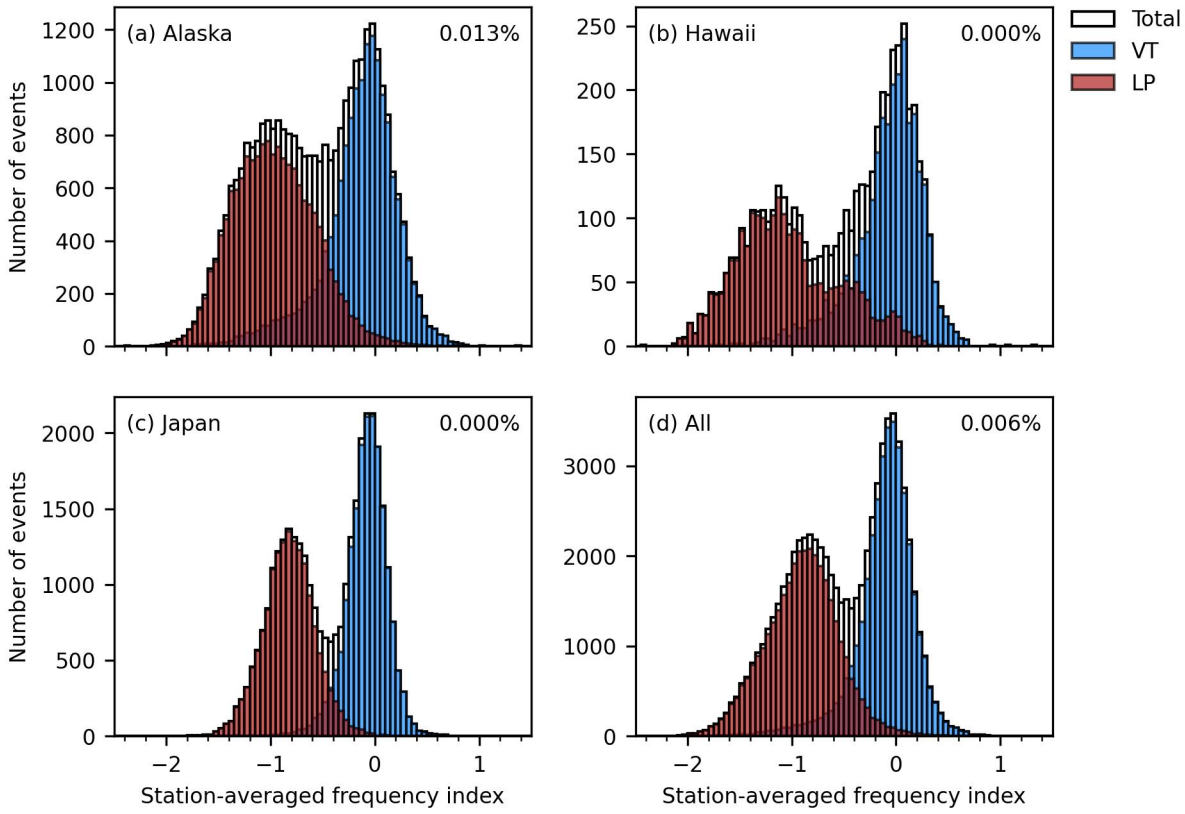

Figure S8. The distribution of event-based frequency index (FI) values of the earthquakes in Table S1, which are calculated by averaging FI values over all recording stations. The numbers in the top right corner indicate the fraction of samples outside the range of the plotted x-axis.

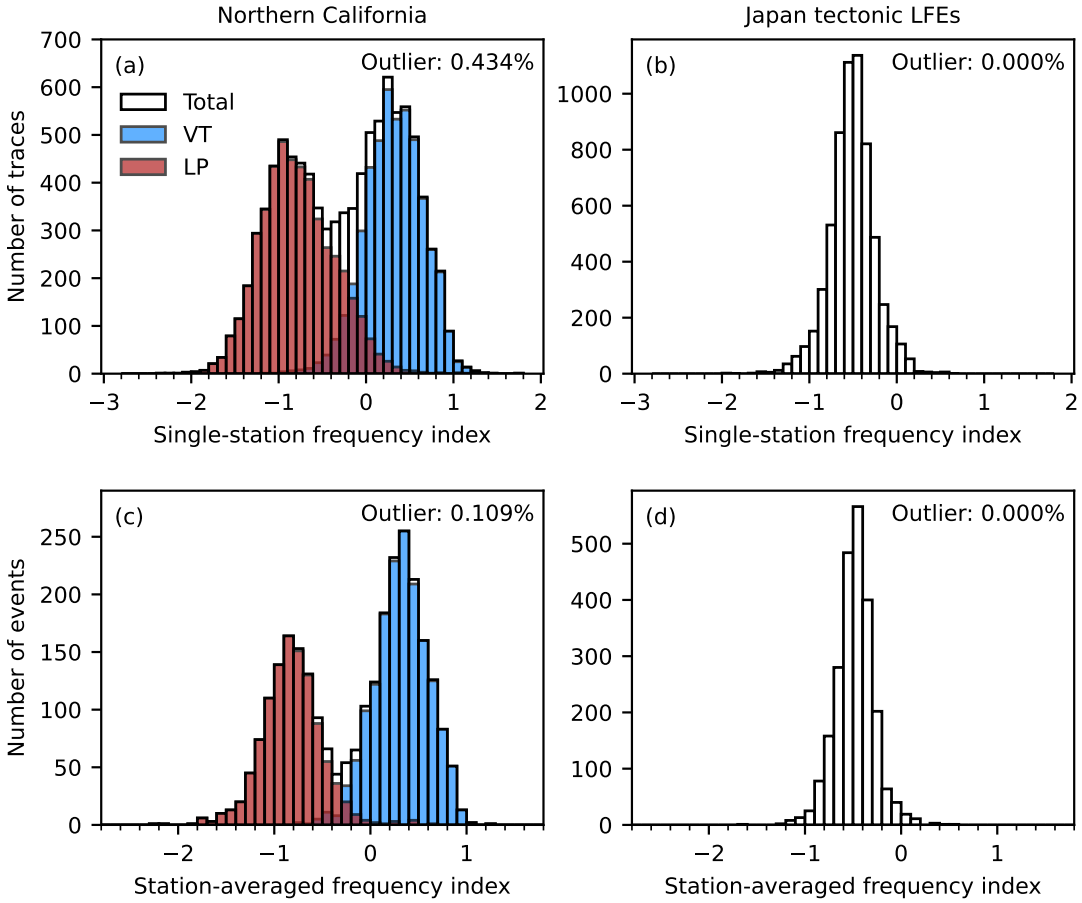

Figure S9. The distribution of frequency index (FI) values for the northern California test set (a, c) and the test set of Japan tectonic LFEs (b, d). The top row (a, b) and the bottom row (c, d) show the single-station FI and the event-based FI, respectively. The numbers in the top right corner indicate the fraction of samples outside the range of the plotted x-axis.

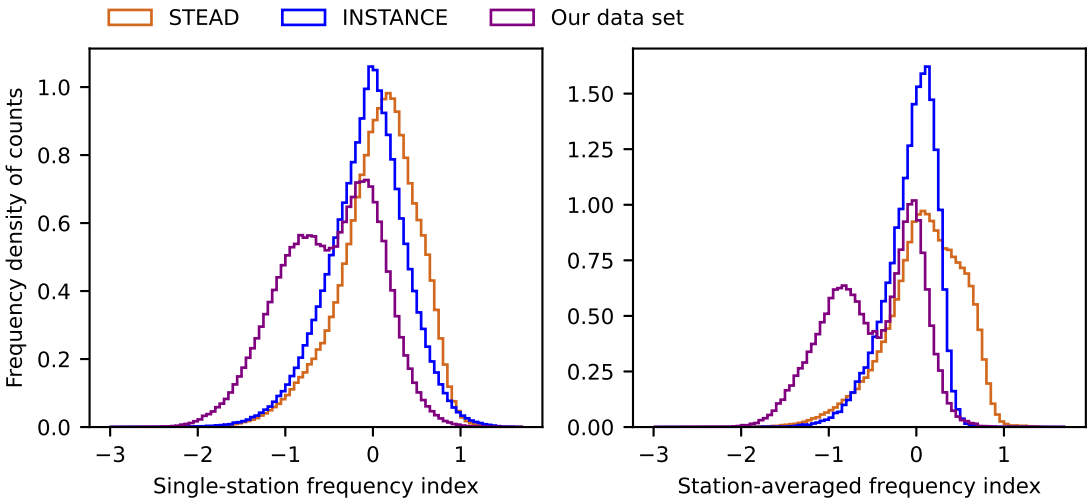

Figure S10. Comparison of frequency index distributions of INSTANCE (Michelini et al., 2021), STEAD (Mousavi et al., 2019) and our data set. The y axis, frequency density, is defined as $\frac{\text{Counts in a bin/Total counts}}{\text{Bin width}}$.

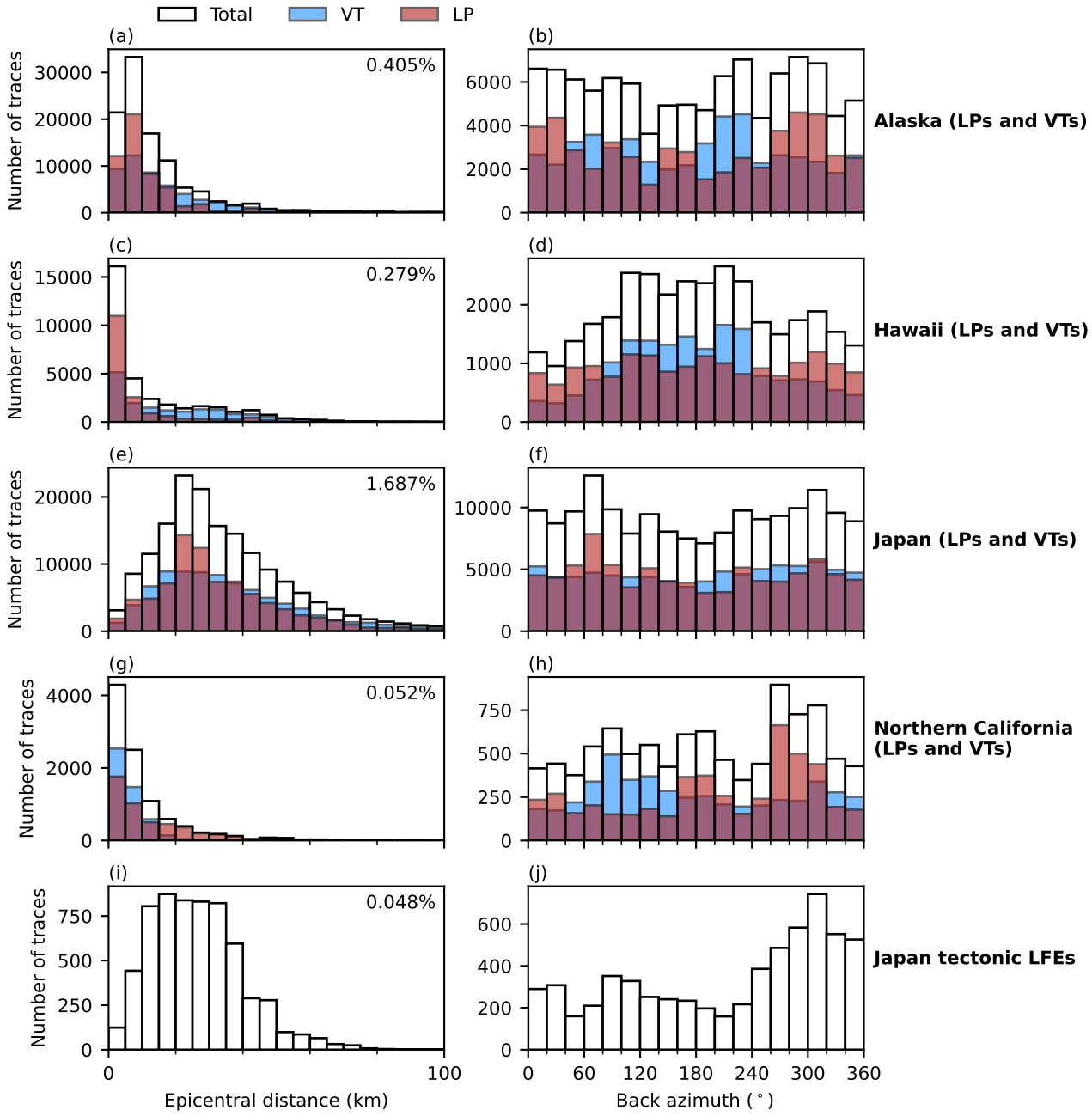

Figure S11. The distribution of epicentral distances (the first column) and back azimuths (the second column) of LP and VT waveforms from Alaska (a, b), Hawaii (c, d), Japan (e, f), northern California (g, h) and tectonic LFE waveforms from Japan (i, j). We adopt the logarithmic scale in the first column to make the number of traces with large epicentral distances visible. The fraction of traces recorded at an epicentral distance greater than 100 km is shown in the top right corner of each panel in the first colum.

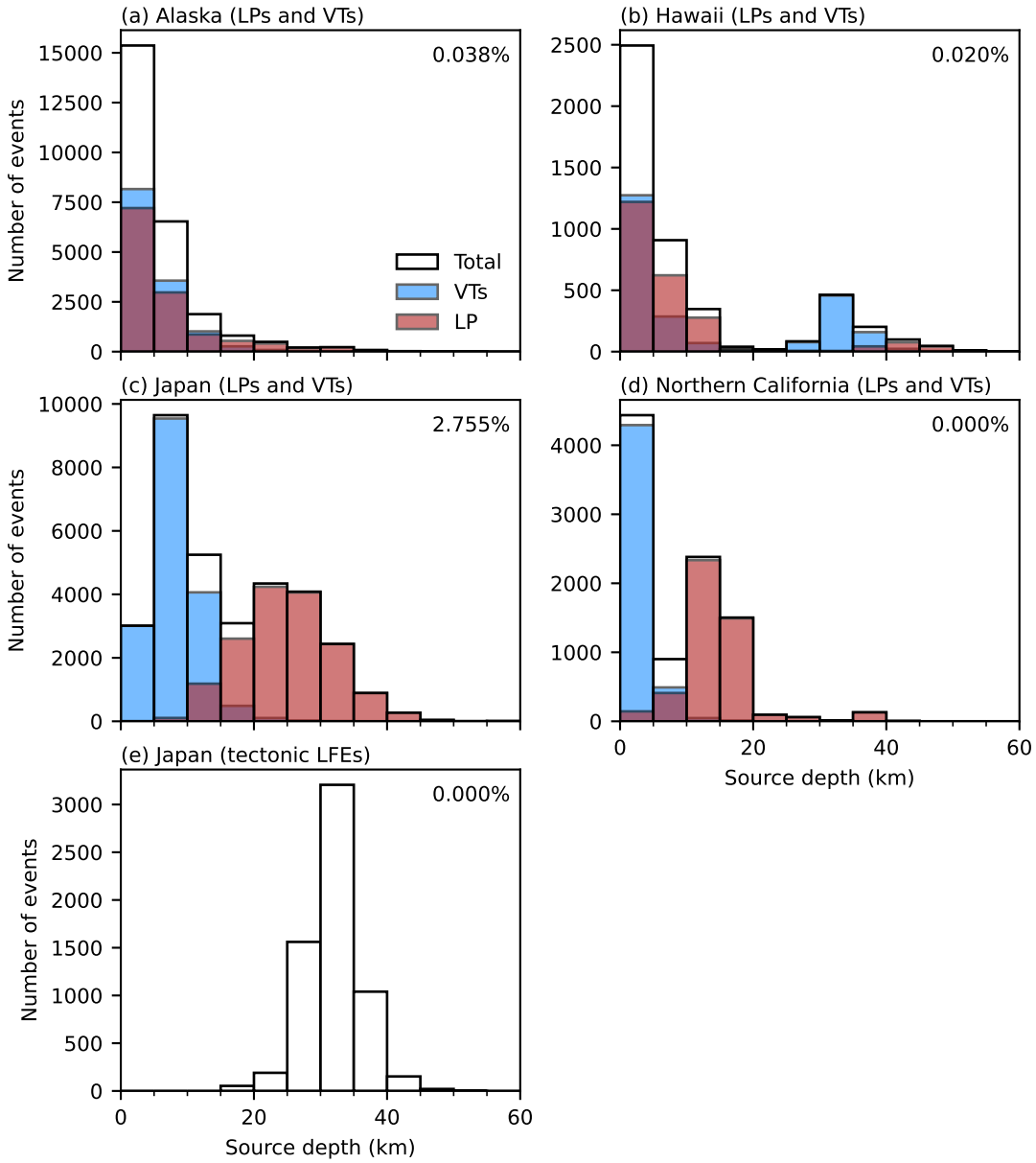

Figure S12. The distribution of source depths of the VTs and LPs in our data set from Alaska (a), Hawaii (b), Japan (c) northern California (d) as well as tectonic LFEs near the Nankai trough from Japan (e). The fraction of events deeper than 60 km is shown in the top right corner of each panel.

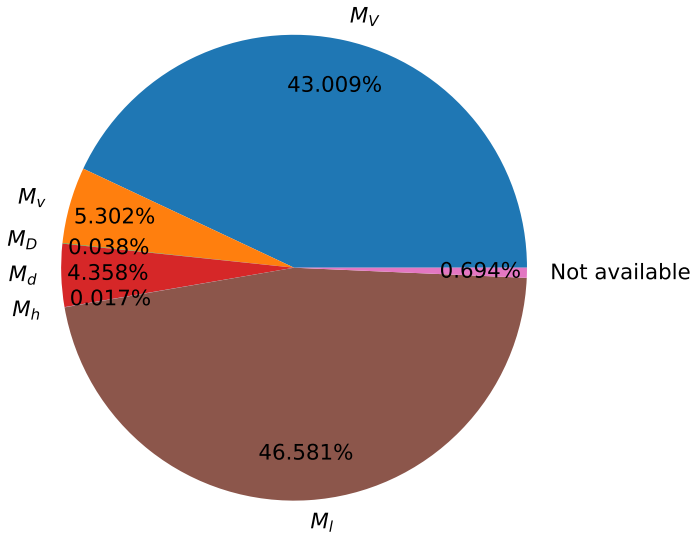

Figure S13. The proportion of different magnitude types. M_l is the local magnitude. M_d is the duration magnitude. M_h is nonstandard magnitudes used by USGS (`https://www.usgs.gov/programs/earthquake-hazards/magnitude-types`). M_V, M_v and M_D are magnitudes used by JMA (Japan Meteorological Agency), where M_V is the velocity magnitude, M_v is similar to M_V but for only 2 or 3 stations, M_D is the displacement magnitude (`https://www.data.jma.go.jp/svd/eqev/data/bulletin/catalog/notes_e.html`).

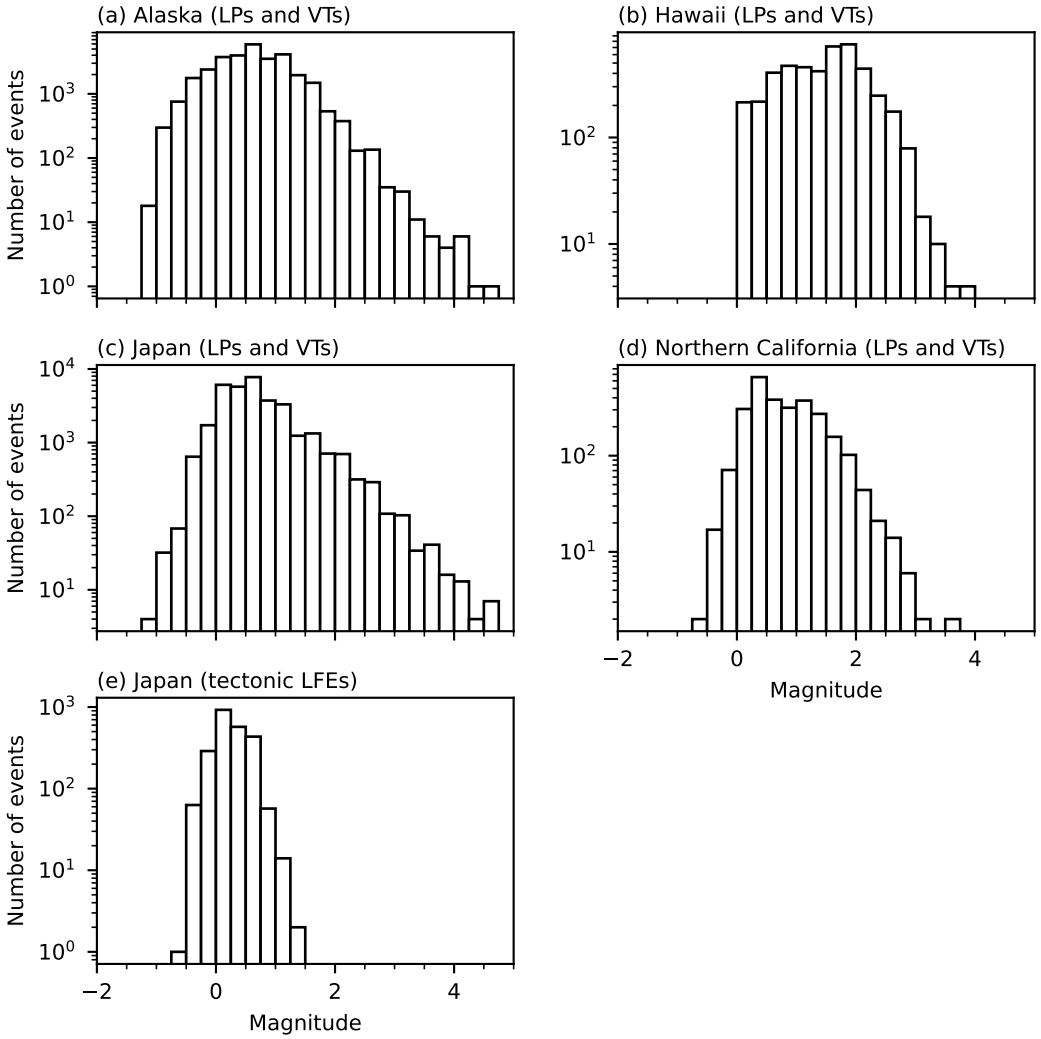

Figure S14. Histogram of magnitudes of the VTs and LPs in our data set from Alaska (a), Hawaii (b), Japan (c) northern California (d) as well as tectonic LFEs near the Nankai trough from Japan (e).

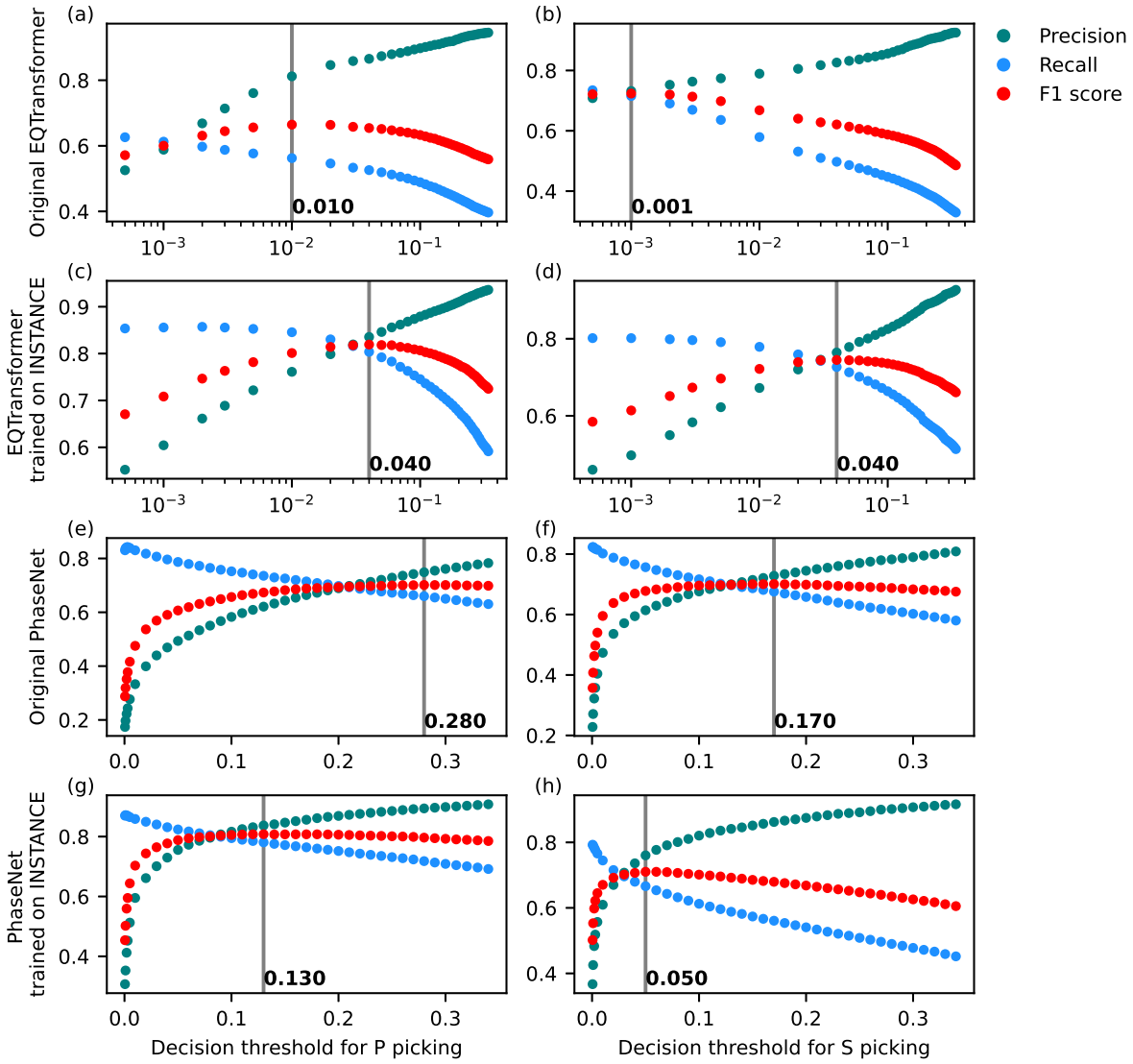

Figure S15. Threshold tuning for the original PhaseNet (Zhu & Beroza, 2019), EQTransformer (Mousavi et al., 2020) and their variants trained on the INSTANCE data set (Münchmeyer et al., 2022) The performance is evaluated on the validation set in Table S1. The left and right columns show the performance metrics for P picking and S picking, respectively. The first (a, b) and second (c, d) rows show the performance metrics of the original EQTransformer network trained on STEAD (Mousavi et al., 2019, 2020) and the EQTransformer network trained on INSTANCE (Michelini et al., 2021; Münchmeyer et al., 2022), respectively. The third (e, f) and fourth (g, h) rows show the performance metrics of the original PhaseNet network trained on California earthquakes (Zhu & Beroza, 2019) and the PhaseNet network trained on INSTANCE (Michelini et al., 2021; Münchmeyer et al., 2022), respectively. The gray lines and the numbers show the optimal thresholds found at the highest F1 scores.

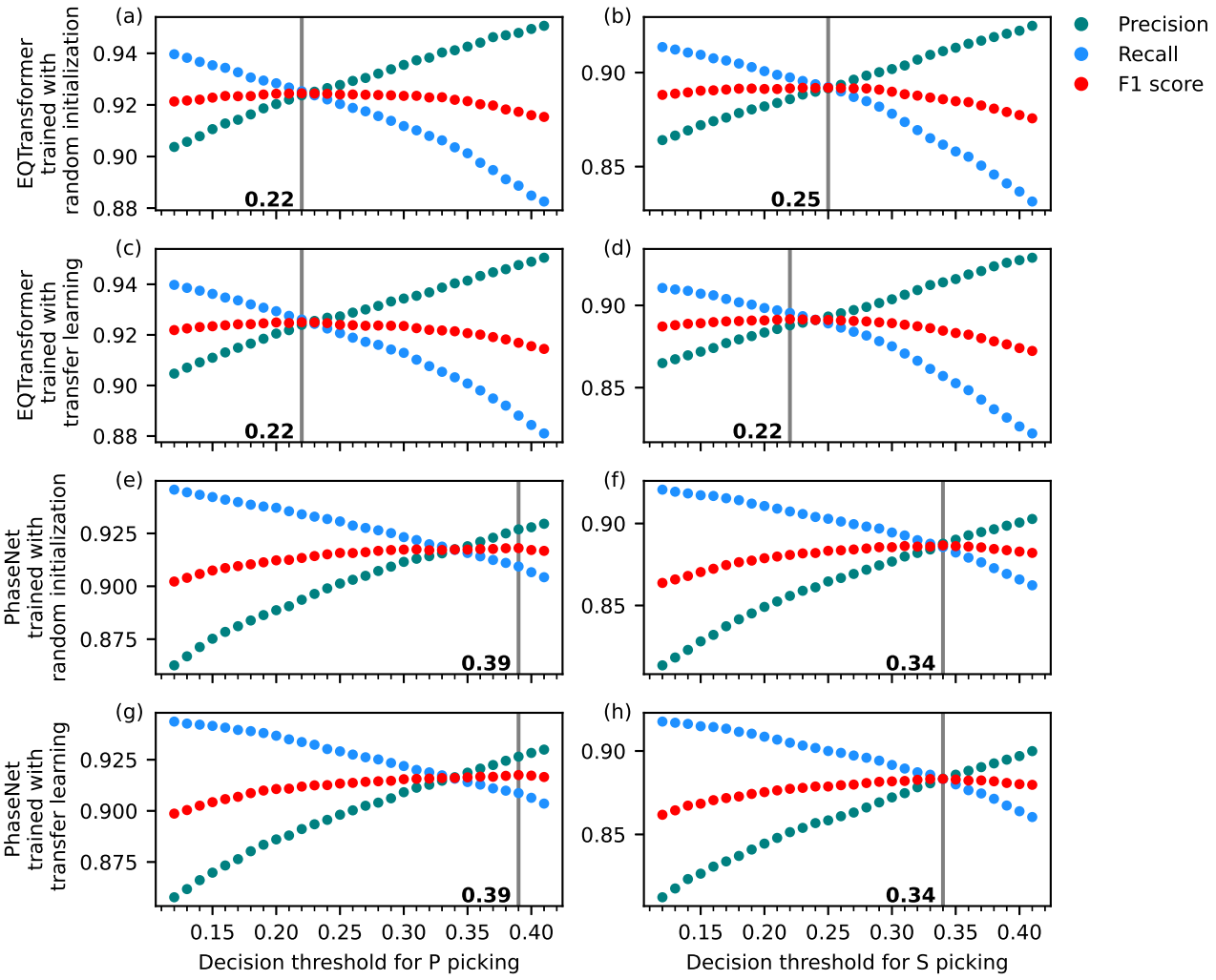

Figure S16. Threshold tuning for EQTransformer and PhaseNet networks trained for volcano seismicity in this study (Table S4). The performance is evaluated on the validation set (development set). The left and right columns show the precision, recall and F1 score for P picking and S picking, respectively. The first (a, b) and second (c, d) rows show the performance metrics of the EQTransformer networks trained with randomly initialized weights and initial weights pre-trained on INSTANCE , respectively. The third (e, f) and fourth (g, h) rows show the performance metrics of the PhaseNet networks trained with randomly initialized weights and initial weights pre-trained on INSTANCE, respectively. The gray lines and the numbers show the optimal thresholds found at the highest F1 scores.

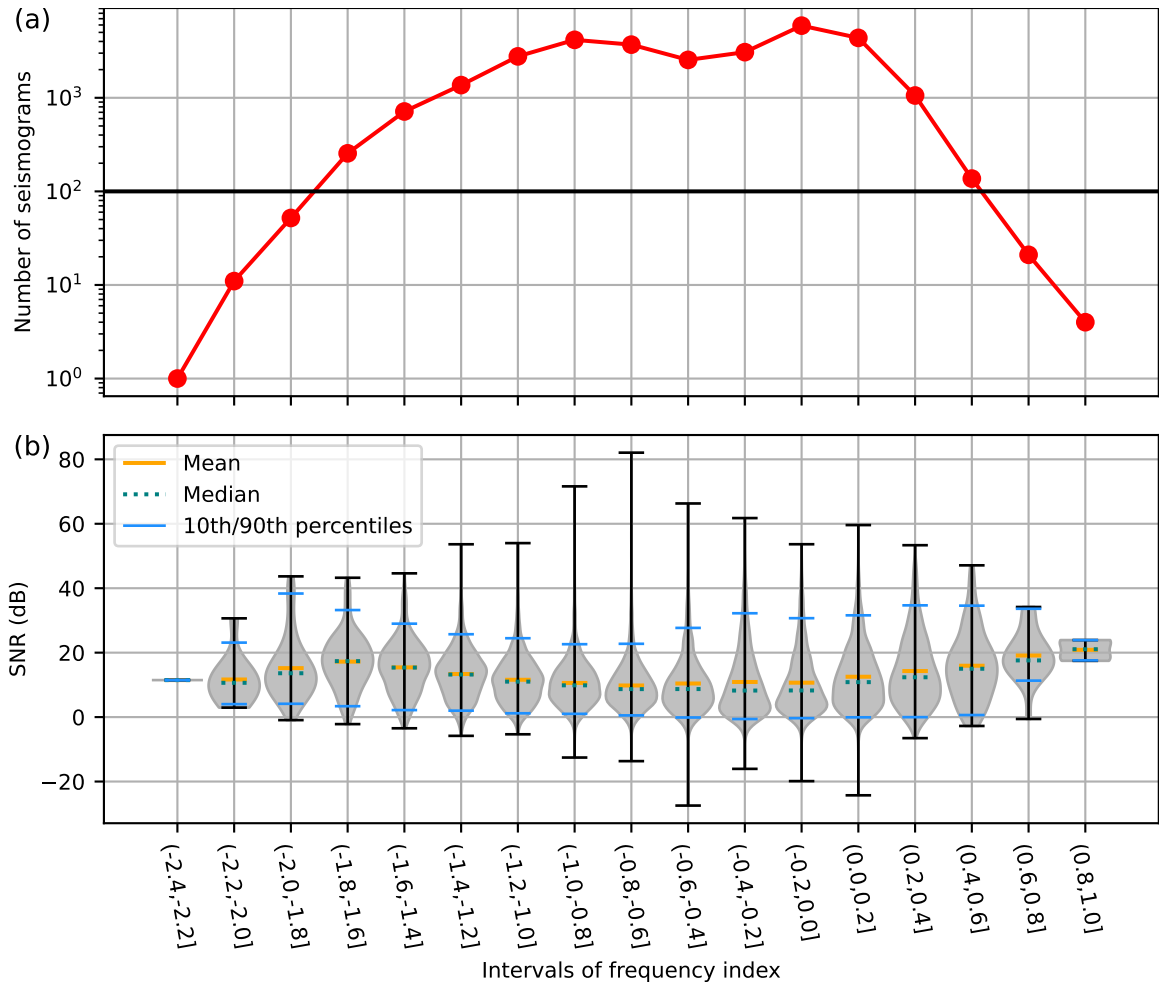

Figure S17. (a) The numbers of waveform traces in the test set for different frequency index bins. We only use the subsets with more than 100 traces for testing, i.e. those above the horizontal black line. (b) The distribution of signal to noise ratio for each subset. The vertical lines show the SNR ranges. The gray area is the estimated probability density for the SNR distribution.

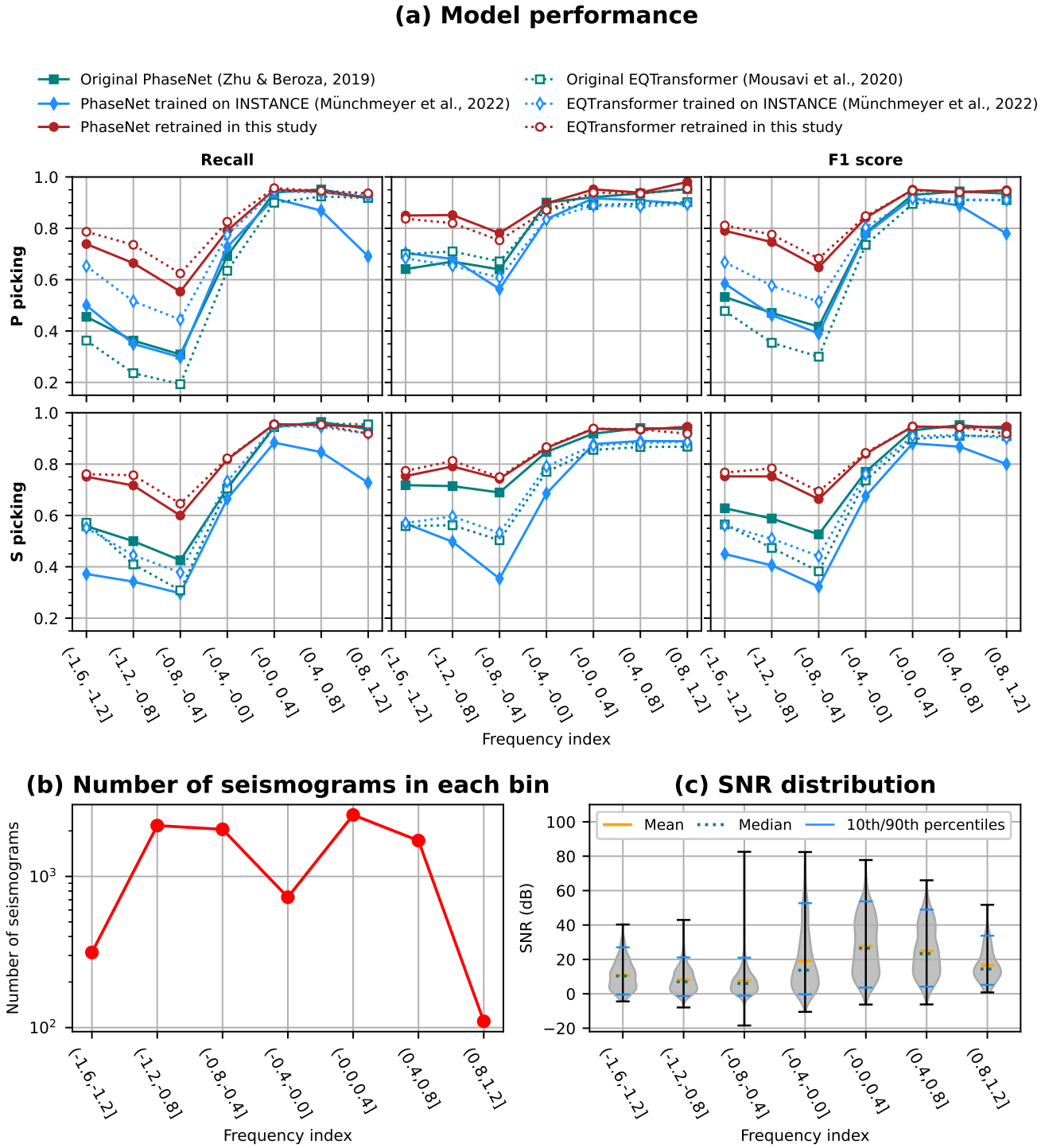

Figure S18. (a) Model performance on subsets of the testing waveforms from northern California. (b) Number of waveforms in each subset. (c) SNR distribution in each subset.

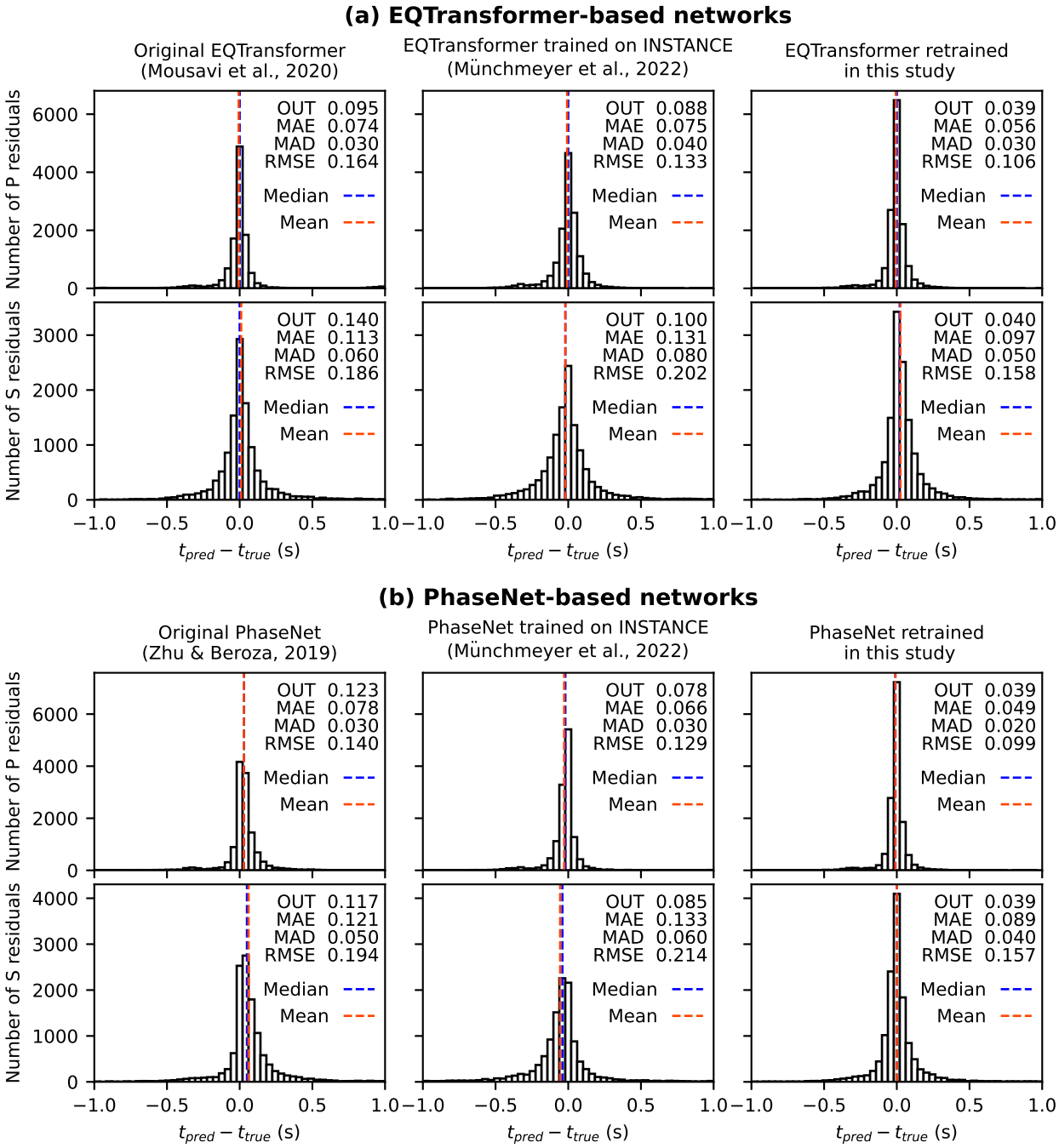

Figure S19. Histogram of residuals between the manual picks in the VT test set and the picks predicted by the EQTransformer-based networks (a) and the PhaseNet-based networks (b). The numbers in the upper right corner show the fraction of residual outside the (-1, 1)s interval (OUT), the mean absolute error (MAD), the median absolute deviation (MAD) and the root mean square error (RMSE). The MAE, RMSE and MAD are calculated only for the residuals within (-1 ,1)s to avoid strong influence of outliers.

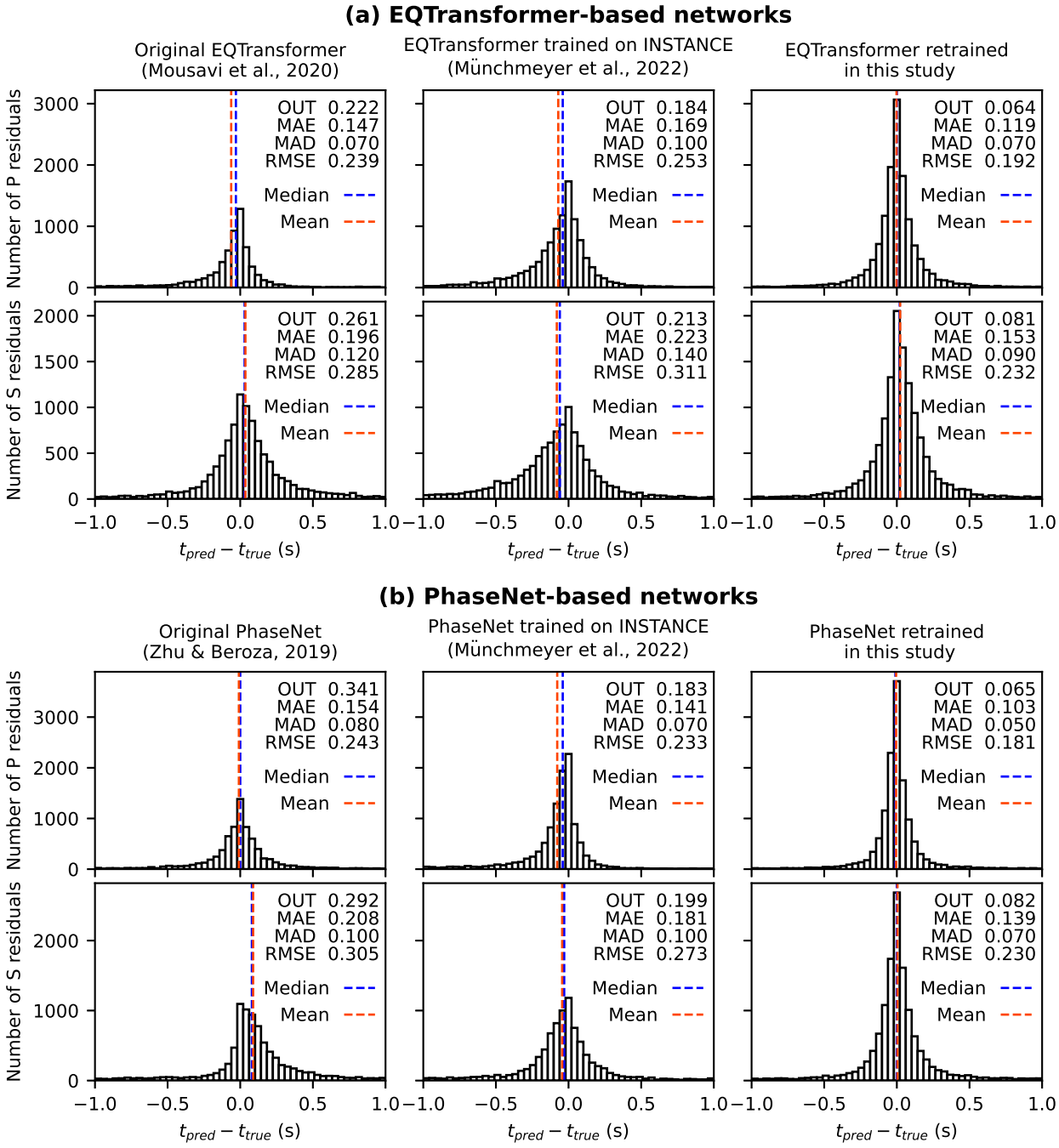

Figure S20. Similar to Figure S19 but for the LP test set.

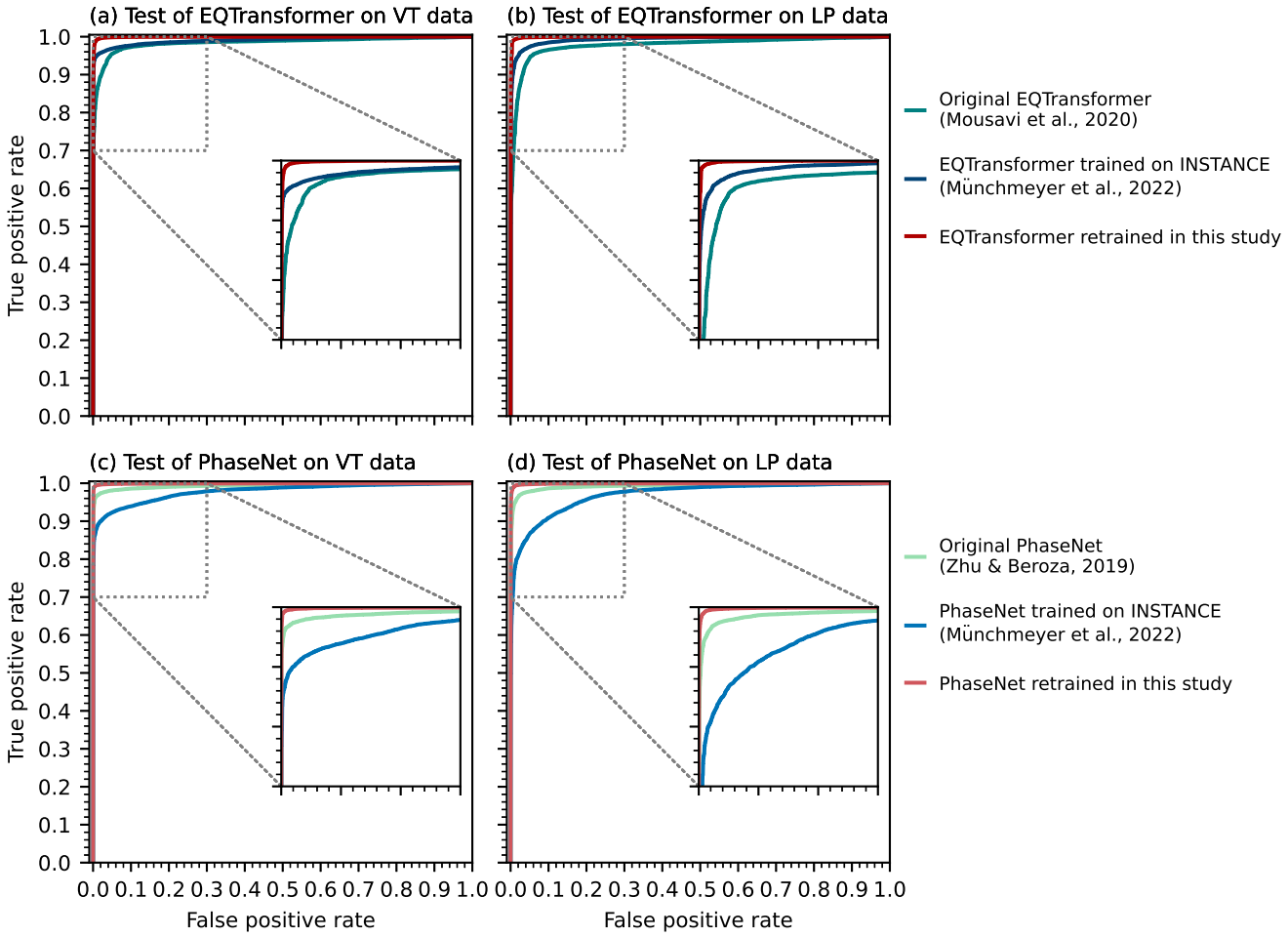

Figure S21. Receiver operating characteristics (ROC) for event detection. The first row shows the ROC curves for the EQTransformer-based networks while the second row is for the PhaseNet-based networks. If the output probability curve for a test example is larger than the decision threshold, it is considered as a positive prediction. The test data are from Alaska, Hawaii and Japan. The LP test set and VT test set (Table S1) are evaluated separately.

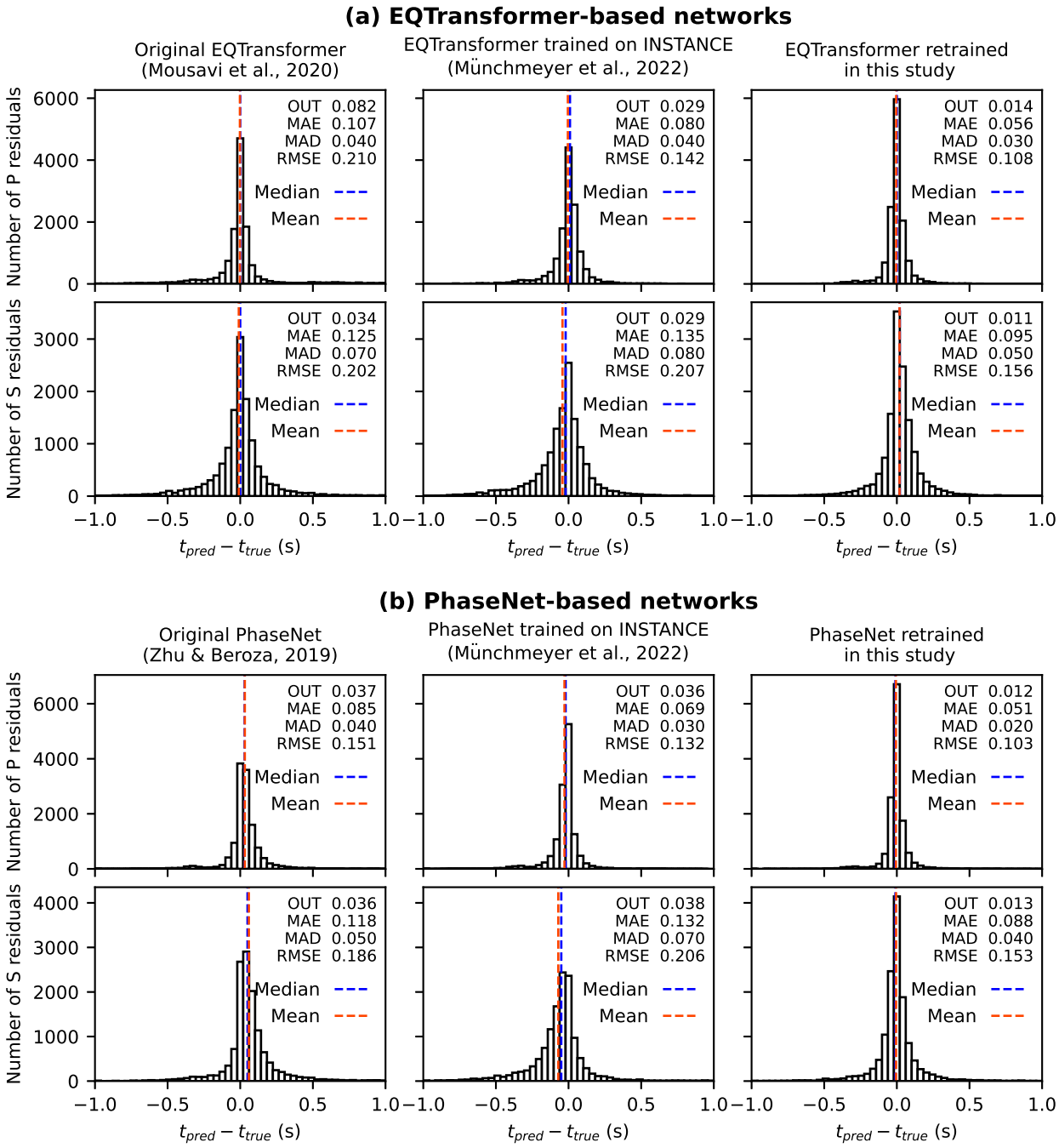

Figure S22. Residuals of phase picks for the VT test set calculated using Münchmeyer et al. (2022)'s evaluation workflow. The difference from Figure S19 is due to the different ways in pre-processing and post-processing. In Münchmeyer et al. (2022)'s workflow, 10s windows containing only P or only S are randomly generated, where the peak position of the output phase probability is taken as the model pick. See (Münchmeyer et al., 2022, Data and Method) for more details. In our evaluation workflow a 30s window containing the P manual pick is randomly generated, which may or may not contains the S pick. We run a trigger algorithm on the output probability curves and find the peaks between trigger on and off times, which may produce more than one model pick for a waveform trace even though there is only one ground truth.

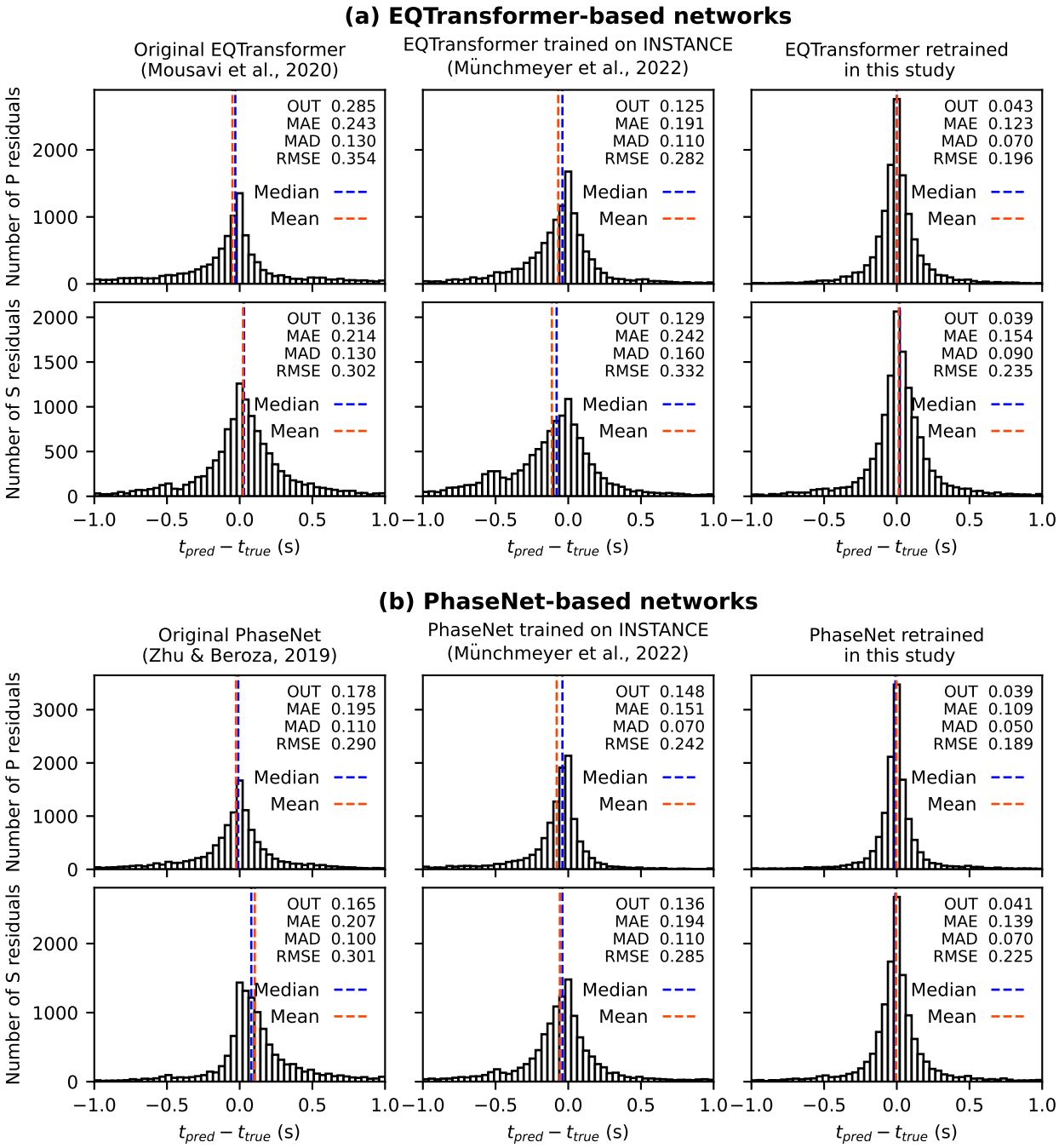

Figure S23. Similar to Figure S22 but for the LP test set. The difference from Figure S20 is due to the different ways of pre-processing and post-processing as explained in the caption of Figure S22.

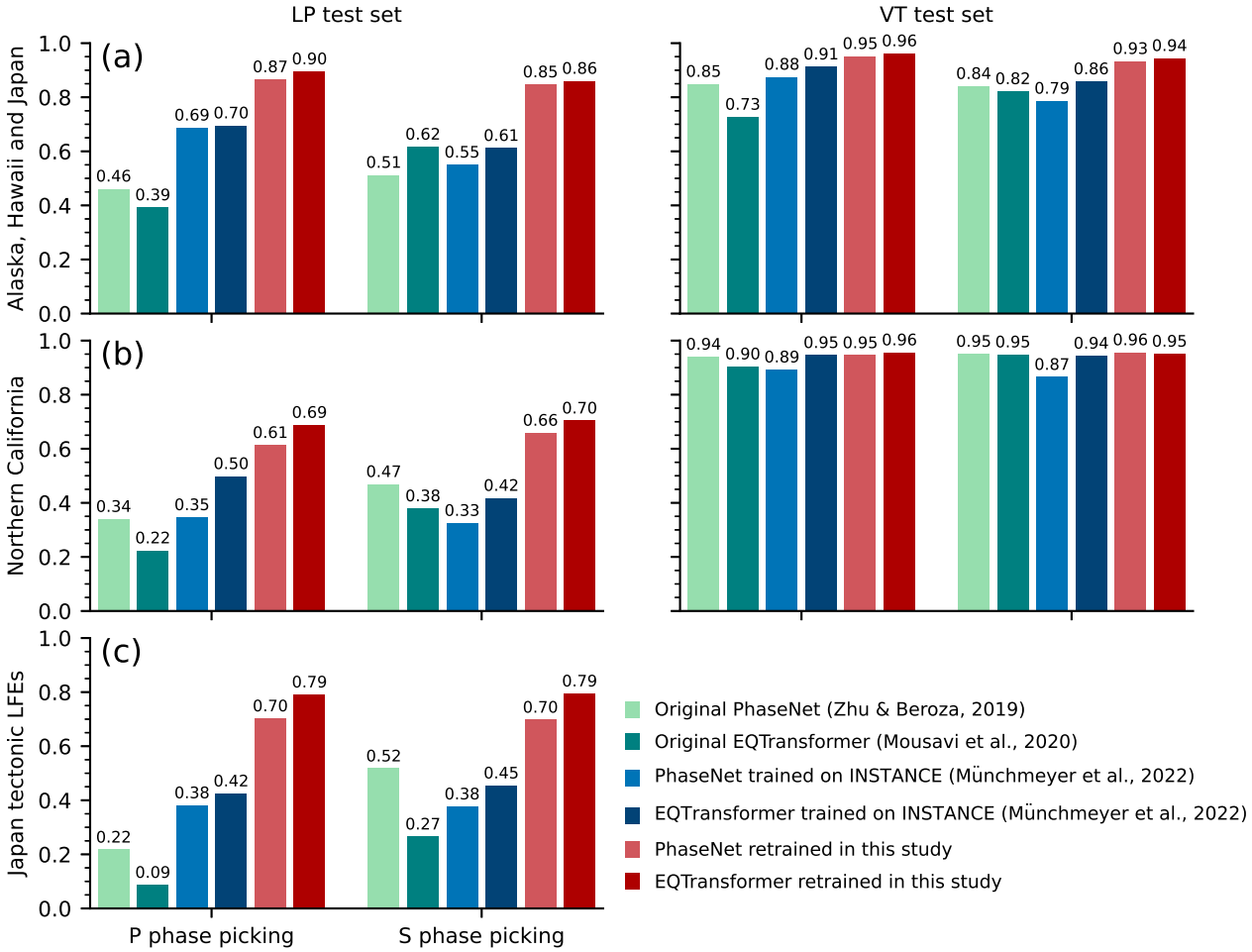

Figure S24. Recalls of different models evaluated on the test waveforms from (a) the same regions as the training data, (b) northern California from where no training data are used and (c) tectonic LP earthquakes in Japan which are generally condiered different from volcanic long-period earthquakes in terms of source processes.

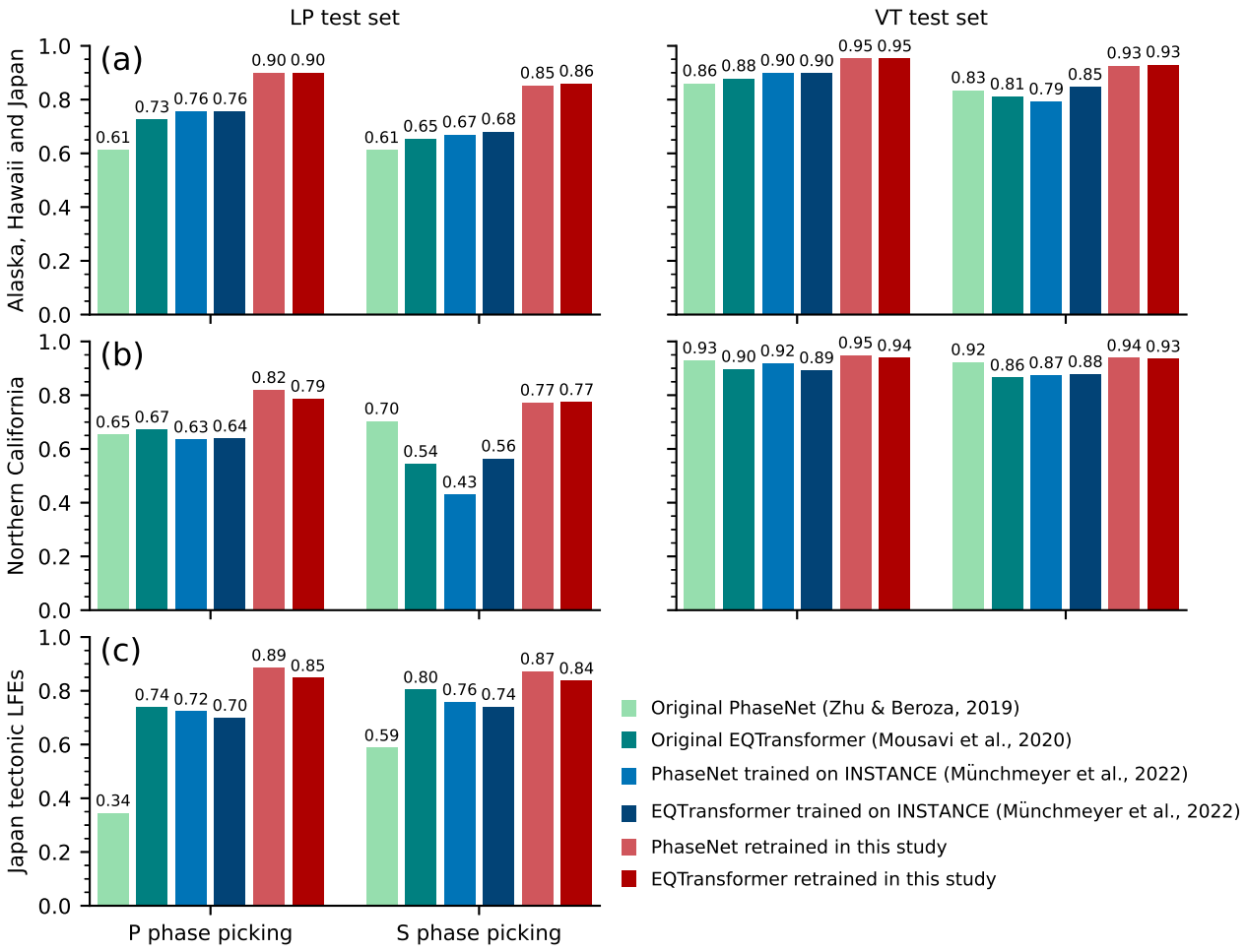

Figure S25. Precisions of different models evaluated on the test waveforms from (a) the same regions as the training data, (b) northern California from where no training data are used and (c) tectonic LP earthquakes in Japan which are generally condiered different from volcanic long-period earthquakes in terms of source processes.

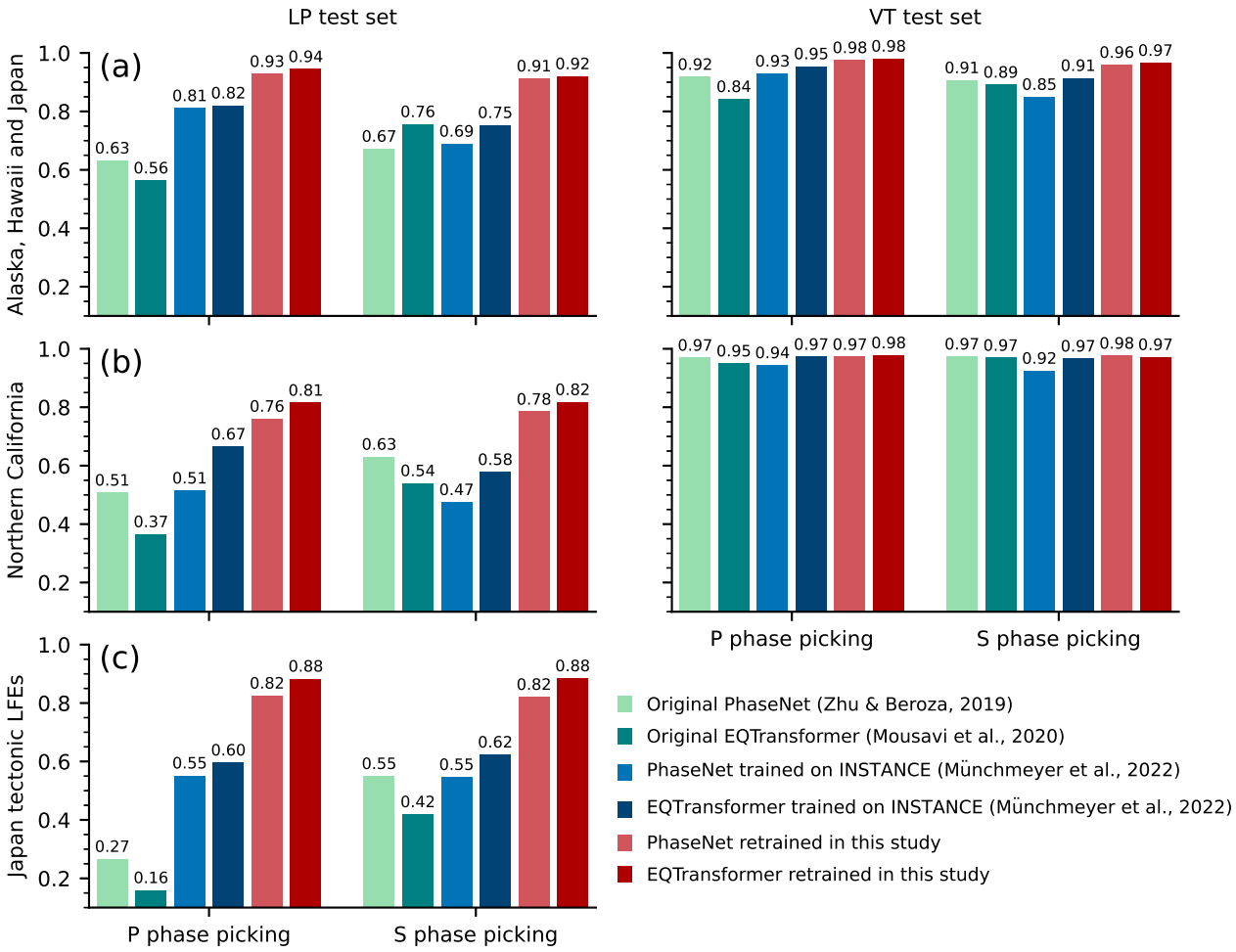

Figure S26. F1 scores of different models calculated using the definition of FP/TN/FN/TN based on waveform traces rather than sampling points. Each row shows the performance for test data from different regions: (a) the same regions as the training data, (b) northern California from where no training data are used, (c) tectonic LP earthquakes in Japan which are generally condiered different from volcanic long-period earthquakes in terms of source processes. The precision and recall are given in Figure S30-S31 in the supplement.

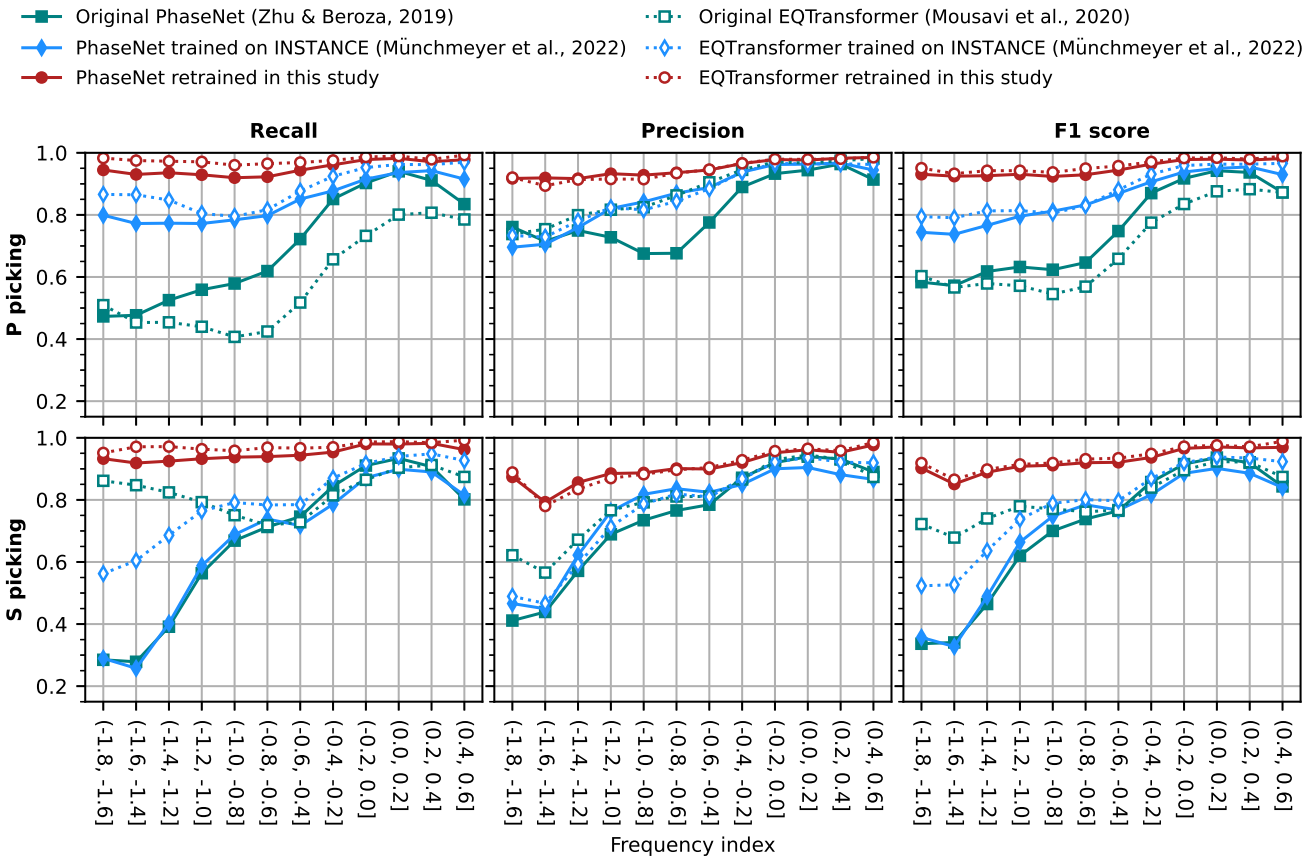

Figure S27. Model performance on subsets of testing waveforms with different frequency index values. Different from Figure 3 in the main paper, the performance in this figure is calculated using the definition of FP/TN/FN/TN based on waveform traces rather than sampling points.

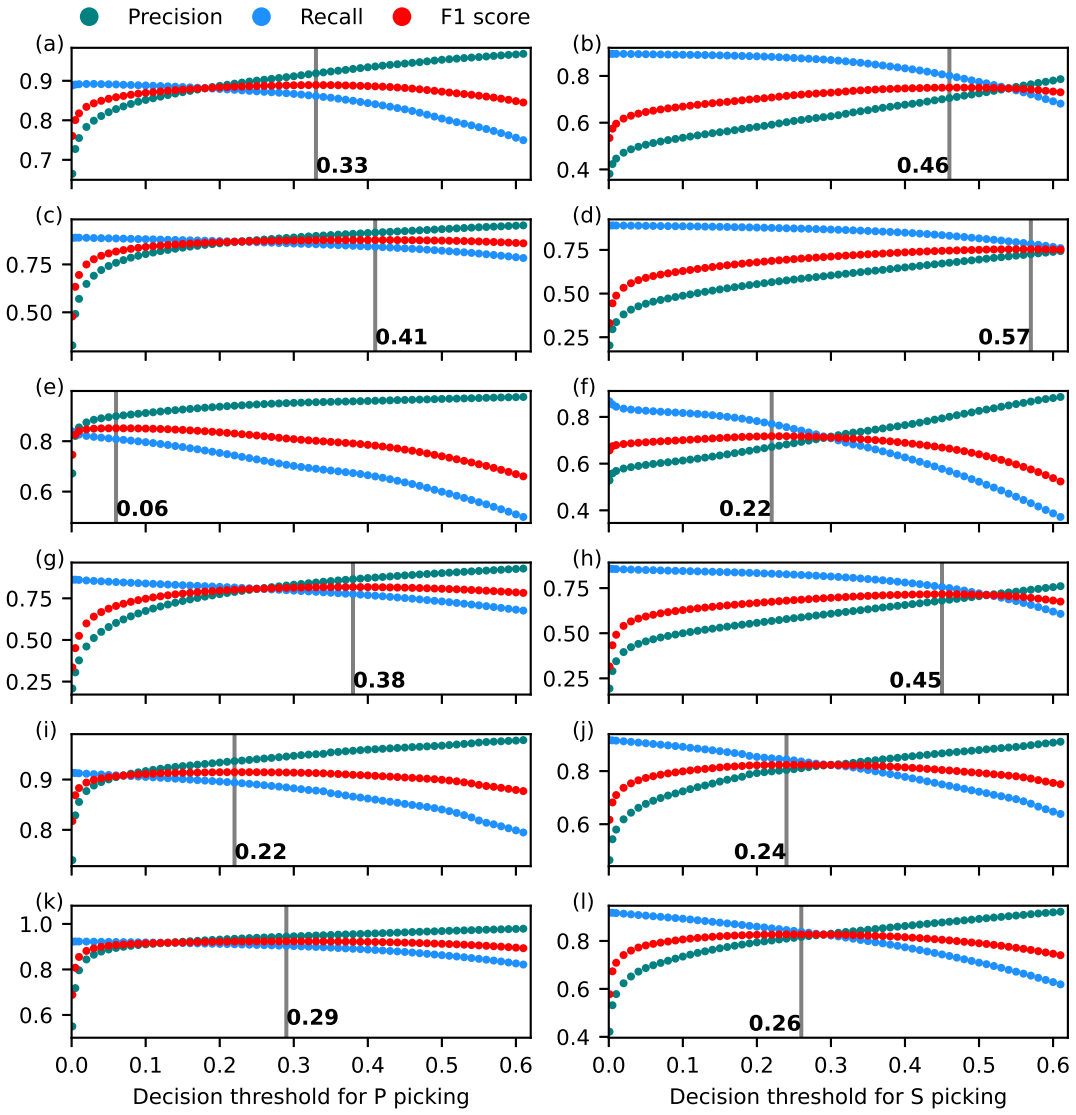

Figure S28. Model performance on the validation set of the INSTANCE data set for the EQTransformer retrained in this study (a-b), PhaseNet retrained in this study (c-d), original EQTransformer (e-f), original PhaseNet (g-h), EQTransformer trained on INSTANCE (i-j), PhaseNet trained on INSTANCE (k-l). The optimal decision thresholds (vertical gray lines) are selected to maximize the F1 score on the validation set.

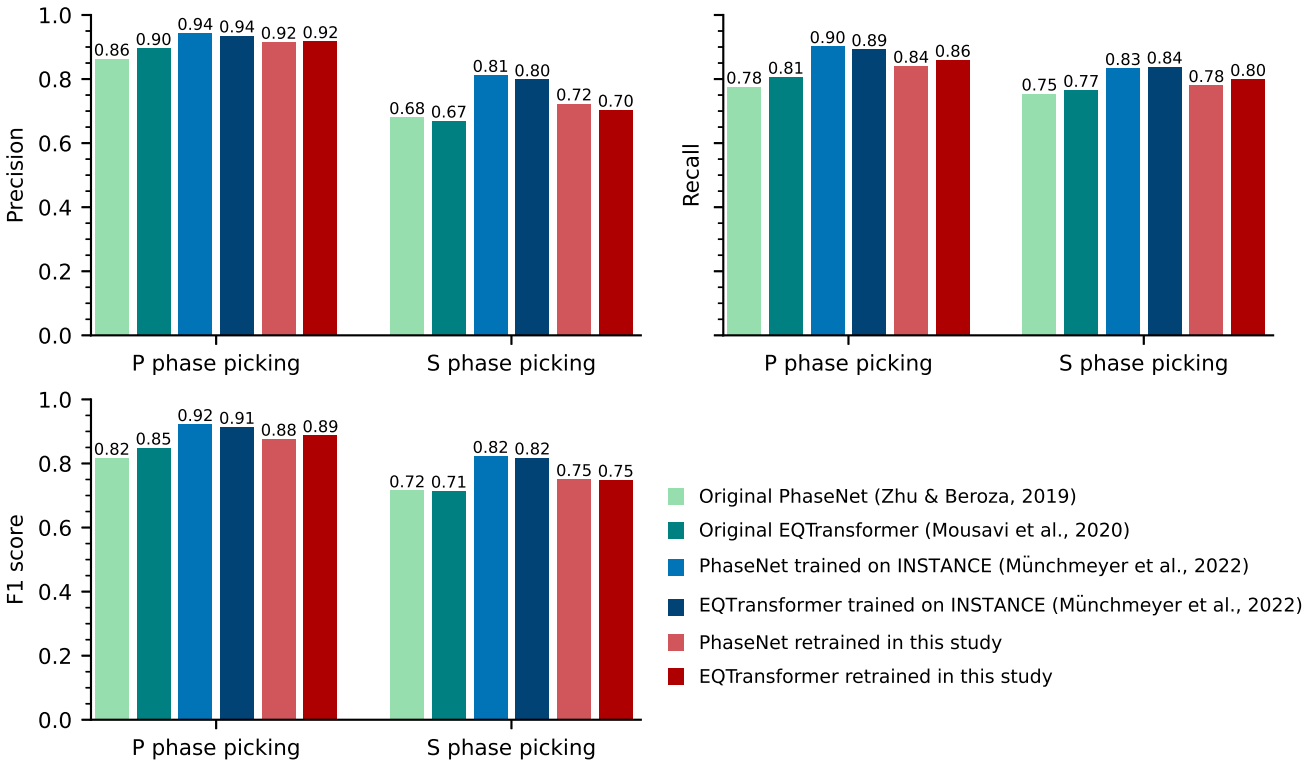

Figure S29. Evaluation of different models on the test set of the INSTANCE data set. The optimal decision thresholds are selected to maximize the F1 score on the validation set of the INSTANCE data set (Figure S28).