A multi-task encoder-decoder to separating earthquake and ambient 1 noise signal in seismograms

Jiuxun Yin[1], Marine Denolle[2], and Bing He[3]

[1]Harvard University
[2]University of Washington
[3]University of Rhode Island

November 26, 2022

Abstract

Seismograms contain multiple sources of seismic waves, from distinct transient signals such as earthquakes to ambient seismic vibrations such as microseism. Ambient vibrations contaminate the earthquake signals, while the earthquake signals pollute the ambient noise's statistical properties necessary for ambient-noise seismology analysis. Separating ambient noise from earthquake signals would thus benefit multiple seismological analyses. This work develops a multi-task encoder-decoder network to separate transient signals from ambient signals directly in the time domain for 3-component seismograms. We choose the active-volcanic Big Island in Hawai'i as a natural laboratory given its richness in transients (tectonic and volcanic earthquakes) and diffuse ambient noise (strong microseism). The approach takes a noisy seismogram as input and independently predicts the earthquake and noise waveforms. The model is trained on earthquake and noise waveforms from the STandford EArthquake Dataset (STEAD) and on the local noise of a seismic station. We estimate the network's performance using the Explained Variance (EV) metric on both earthquake and noise waveforms. We explore different network architectures and find that the long-short-term-memory bottleneck performs best over other structures, which we refer to as the WaveDecompNet. Overall we find that WaveDecompNet provides satisfactory performance down to signal-to noise-ratio (SNR) of 0.1. The potential of the method is 1) to improve broadband SNR of transient (earthquake) waveforms and 2) to improve local ambient noise to monitor the Earth structure using ambient noise signals. To test this, we apply a short-time-average to a long-time-average (STA/LTA) filter and improve the detection 27 times. We also measure single-station cross-correlation and autocorrelations of the recovered ambient noise and establish their improved coherence through time and over different frequency bands. We conclude that WaveDecompNet is a promising tool for a range of seismological research.

A multi-task encoder-decoder to separating earthquake and ambient noise signal in seismograms

Jiuxun Yin[1], Marine A. Denolle[2], Bing He[3]

[1]Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA

[2]Department of Earth and Space Sciences, University of Washington, Seattle, WA, USA

[3]Graduate School of Oceanography, University of Rhode Island, Narragansett, RI, USA

*Corresponding author: Jiuxun Yin (jiuxun_yin@g.harvard.edu)

Abstract

Seismograms contain multiple sources of seismic waves, from distinct transient signals such as earthquakes to ambient seismic vibrations such as microseism. Ambient vibrations contaminate the earthquake signals, while the earthquake signals pollute the ambient noise's statistical properties necessary for ambient-noise seismology analysis. Separating ambient noise from earthquake signals would thus benefit multiple seismological analyses. This work develops a multi-task encoder-decoder network to separate transient signals from ambient signals directly in the time domain for 3-component seismograms. We choose the active-volcanic Big Island in Hawai'i as a natural laboratory given its richness in transients (tectonic and volcanic earthquakes) and diffuse ambient noise (strong microseism). The approach takes a noisy seismogram as input and independently predicts the earthquake and noise waveforms.

The model is trained on earthquake and noise waveforms from the STandford EArthquake Dataset (STEAD) and on the local noise of a seismic station. We estimate the network's performance using the Explained Variance (EV) metric on both earthquake and noise waveforms. We explore different network architectures and find that the long-short-term-memory bottleneck performs best over other structures, which we refer to as the WaveDecompNet. Overall we find that WaveDecompNet provides satisfactory performance down to signal-to noise-ratio (SNR) of 0.1.

The potential of the method is 1) to improve broadband SNR of transient (earthquake) waveforms and 2) to improve local ambient noise to monitor the Earth structure using ambient noise signals. To test this, we apply a short-time-average to a long-time-average (STA/LTA) filter and improve the detection 27 times. We also measure single-station cross-correlation and autocorrelations of the recovered ambient noise and establish their improved coherence through time and over different frequency bands. We conclude that WaveDecompNet is a promising tool for a range of seismological research.

1 Introduction

A seismogram is a record of how the ground moves and usually contains a rich mix of different seismic signals. They may be transient such as the ground motions caused by earthquakes, surface processes (glacier sliding (Weaver and Malone, 1979; Lipovsky et al., 2019), landslides (Keefer, 1984; Weichert et al., 1994)), human activities (cars, trains, ships, machinery from factories, (Schippkus et al., 2020)). They may be more diffuse such as the microseism (Cessaro, 1994), the seismic hum (Rhie and Romanowicz, 2004), river noise (Burtin et al., 2008), urban life (Lecocq et al., 2020). The transient motions capture the seismic signature of their source (earthquakes, landslides, glacial sliding) and thus are essential information to understand these processes (event detection, location, discrimination, source properties). The diffuse ambient seismic field, on the other hand, found its use in correlation seismology to extract spatial and temporal variations in the Earth structure (Aki, 1957; Claerbout, 1968; Shapiro et al., 2005; Sens-Schönfelder and Wegler, 2006). Therefore, separating the earthquake and ambient noise signals can significantly improve seismological studies from different perspectives: robust event source characterization and robust imaging and monitoring of the Earth's interior.

Many studies across the sciences have focused on removing the diffuse ambient data from the transient signals. In seismology, the diffuse field is often considered as "noise". The task of denoising earthquake signals aims to improve the signal quality, and the most commonly used method is Fourier-based spectral filtering. This approach assumes that the earthquake and ambient noise signals are well separated in the frequency domain. Although this technique has been proven to be effective in numerous cases, it can distort the signals and/or cause artifacts around the impulsive signals (Douglas, 1997; Mousavi and Langston, 2017).

Earthquake and ambient noise signals often overlap in the frequency domain, and direct filtering may be challenging to separate them (Peterson et al., 1993; McNamara et al., 2019). Denoising using time-frequency representations of the signal is another widely applied and effective technique to separate the earthquake and ambient noise signals when they overlap. Many innovative algorithms and methods have been developed, for example, using time-frequency transforms such as the Stockwell S-transform (Stockwell et al., 1996), the Radon transform, the wave-packet transform, the continuous wavelet transform, or others using f-x or f-k filtering, singular spectrum analysis, sparse transform-based denoising, which are extensively reviewed and discussed in Mousavi and Langston (2017). Most of these transform-based denoising methods achieve noise suppression through thresholding methods, that is, determining some hard- (Donoho and Johnstone, 1994) or soft- (Chang et al., 2000) thresholds to separate seismic and noise signals. Although those transform-based methods are shown to be very practical and possible to be automated (Mousavi et al., 2016;

Mousavi and Langston, 2017); they still require manual intervention, and the parameter tuning is often performed using trial-and-error approaches.

With the recent leap of computational power, memory and data storage, machine learning (ML) has provided a diverse set of powerful tools in the geosciences (Bergen et al., 2019). Many ML algorithms are built to (1) automatically perform complex prediction task; (2) create a representation that approximates numerical simulations or captures relationships; (3) reveal new patterns, structures, or relationships from data (Bergen et al., 2019). ML algorithms are powerful in many different seismological tasks, including but not limited to waveform classification and earthquake detection (Li et al., 2018; Perol et al., 2018; Kong et al., 2019; Mousavi et al., 2019b, 2020; Beroza et al., 2021; Johnson et al., 2021), phase picking and association (Li et al., 2018; Meier et al., 2019; Zhu et al., 2019a; Liu et al., 2020; Mousavi et al., 2020; Walter et al., 2021), source location and characterization (Perol et al., 2018; Mousavi and Beroza, 2019; Ren et al., 2020; van den Ende and Ampuero, 2020; Kuang et al., 2021; Münchmeyer et al., 2021a; Zhou et al., 2021), earthquake early warning (Li et al., 2018; Münchmeyer et al., 2021b), and many others.

Machine-learning methods have also been developed to denoise and decompose the seismic data (Chen et al., 2016; Siahsar et al., 2017; Saad et al., 2018; Zhu et al., 2019b; Tibi et al., 2021). Saad et al. (2018) develop a stacked denoising auto-encoder to smooth the ambient noise and output a time series for better arrival time detection, in a similar fashion to an impulsive filter (Allen, 1978). Zhu et al. (2019b) develop a DeepDenoiser network, which applies deep learning to the short-time Fourier transform and focuses on classification on pixels of the spectrogram of seismograms (Zhu et al., 2019b; Tibi et al., 2021).

This study aims to separate the waveforms of both earthquake and ambient noise signals through the ML network, thereby benefiting both earthquake and ambient-noise seismology. We develop a new multi-task encoder-decoder ML network, which we name WaveDecompNet, to separate both types of signal simultaneously. We treat the problem as a time series extrinsic regression (Tan et al., 2021) and directly work on the seismic data in the time domain. This setting can save human intervention by not tuning parameters for the time-frequency representation and provides the excellent potential to be applied in a near real-time framework given its lower computation cost. The performance of the WaveDecompNet is noticeable even at a low signal-to-noise ratio.

We take the Big Island of Hawai'i as a natural laboratory of the seismically complex environment. Oceanic islands record strong microseismic signals, which are the basis for ambient noise seismology (Longuet-Higgins, 1950). The volcanic region also exhibits dynamic tectonics and volcanic activities. We choose the broadband seismic station IU.POHA at Pohakuloa, Hawai'i. The richness of this dynamical system presents particular challenges in monitoring the volcano-tectonic activities and the temporal evolution of structural changes from ambient-noise seismology.

We further test the applicability of our technique by performing two standard single-station measurements. We first test the improvements on detecting transients by applying the standard short-time-average through long-time-average (STA/LTA) method (Allen, 1978) and find increased trigger rates and potentially improved picking accuracy. We also measure the coherence of single-station ambient-noise cross-correlation functions and find increased stability in the coda of these functions. Our results highlight the potential for WaveDecompNet to improve seismograms used in both event-based seismology and noise-based seismology.

2 Network design of WaveDecompNet

The multi-task encoder-decoder network handles the time series seismic data directly. The input to the network is 1-minute long, 3-component (East-West, North-South, and up-down) raw seismograms. The output of the network is two 1-minute long, 3-component seismograms (earthquake and noise, Fig.1). In order to seek computational efficiency, no pre-processing of the data is applied except the amplitude normalization of the waveforms.

2.1 Network architecture

The encoder-decoder network is a popular network design in ML problems, such as generating dialogues (Serban et al., 2017), semantic image labelling (Badrinarayanan et al., 2015, 2017), detection of image forgeries (Bappy et al., 2019), and prediction of vehicle trajectory (Park et al., 2018). In this study, our encoder-decoder network consists of 3 major parts: the encoder branch, the two decoder branches, and two bottleneck blocks in between:

1. The role of the encoder is to extract useful, high-level features from the seismic time series. Through training with sufficient data and updating its parameters, the encoder aims to learn features of the input data that can help characterize the earthquake and ambient noise signals. We use one-dimensional (1D) convolutional layers with an increasing number of kernels to extract high-level features with a minimal number of parameters (Fig.1). The stride of the convolution is adjusted to down-sample the time series along the time axis. We have tested the use of MaxPooling instead of convolutional strides but found poorer network performance. After each 1D convolutional layer, batch normalization is applied to normalize the output to zero-mean and unit variance. Finally, a rectified linear unit function (ReLU) is used as the activation function to add non-linearity to the network for better regression from time series to time series.

2. The role of the decoder is to translate the learned features from the encoder branch and reconstruct the separated

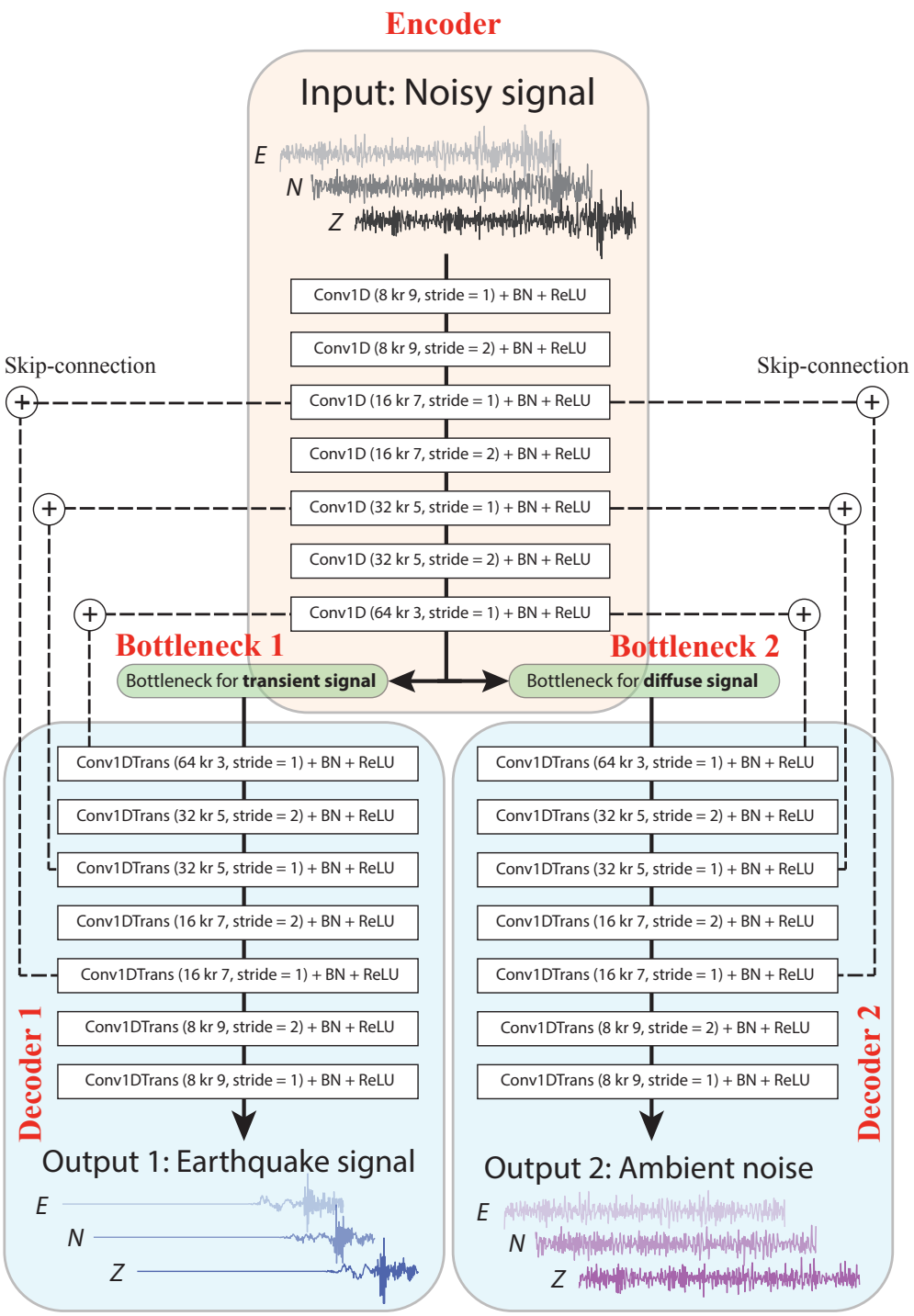

Figure 1: The multi-task encoder-decoder separates earthquake and ambient noise signals. The network consists of 5 main blocks: the encoder branch, two bottlenecks, and two decoder branches. The encoder and decoder branches contain 7 one-dimensional convolutional and transpose convolutional layers. The layer parameters "x kr y, stride=z" refer to x kernels with y features and stride of z. Each convolutional or transpose convolutional layer is followed by a batch normalization (BN) layer and a ReLU activation layer. The structure of the bottleneck block is tested and discussed in the main text for details. 6 residual connections (skip-connection layer by summation) directly connect the encoder to the decoder to improve the convergence of training and prediction performance (He et al., 2016; Ronneberger et al., 2015; Zhu et al., 2019a).

earthquake and noise time series. A two-branch decoder block handles both the separated earthquake and noise waveforms individually and performs better than a single branch network that only outputs the earthquake signal. The branches are composed of 1D transpose convolutional layers. In symmetry with the encoder block, the number of kernels gradually decreases, and we manually select the stride to incorporate the high-level features back into the time domain. Like the encoder block, we also apply batch-normalization and ReLU activation following each 1D transpose convolutional layer. The parameters of the two branches are learned independently.

3. The bottleneck blocks link the encoder and decoder blocks. Their purpose is to learn the mapping relation between the encoder-extracted features of the composite waveform (earthquakes and noise) and the features of the separated earthquake and noise time series, respectively. The design of the bottleneck block greatly impacts the performance of the algorithm and is subject to investigation in this study.

2.2 Data

We use the earthquake waveform data from the STEAD (Mousavi et al., 2019a) (STandford EArthquake Dataset, available at `https://github.com/smousavi05/STEAD`) because of its broad coverage of global earthquakes. This data set is curated to provide many high signal-to-noise ratio waveforms of local (source-receiver distance less than 350 km) earthquakes and a set of "noise" (non-earthquake) signals recorded globally. There are 234,526 samples of 3-component seismograms of ambient noise and 1,030,231 samples of 1-minute 3-component seismograms associated with 450,000 earthquakes located at various regions globally (Mousavi et al., 2019a). We randomly pick 144,000 STEAD earthquake waveforms based on their SNRs, which is defined as:

$$SNR = \frac{||S||^2}{||N||^2}, \tag{1}$$

where $||S||^2$ and $||N||^2$ are the power of signal and noise, respectively. We only keep the highest quality earthquake waveforms with SNR $> 10^4$ in the STEAD dataset to approximate a noise-free signal, then lowpass-filter the waveform below 5 Hz and down-sample them from 100 Hz to 10 Hz.

The data set for ambient noise windows combines a "global noise" from STEAD noise waveforms with a "local noise" from IU.POHA station. We randomly pick 100,000 3-component STEAD samples, low-pass filtering first and then down-sample them to 10 Hz. For the "local noise", we select noise waveforms from 1-month-long continuous seismic

data recorded by IU.POHA (from July 31, 2021, to September 1, 2021), and down-sample the data to 10 Hz. The continuous data may contain known and unknown earthquakes. To keep the spectral features of the noise and reduce the effects from these transient signals, we shuffle the phases as follows. First, we transform the seismic data into the Fourier domain, which gives the amplitude and phase spectra. We keep the amplitude spectrum but assign a random phase using a uniform distribution $-\pi$ to π to each frequency value, then we apply the inverse Fourier transform. We obtain 44,000 samples of 1-minute ambient noise time series.

To represent more time series for better generalization, we apply two strategies of data augmentation (Zhu et al., 2020). First, we shift the arrival time of earthquake signals randomly in a uniform distribution of -30 to 60 s to handle the uncertain arrival time of earthquake signals in the real application, and allow for a redistribution of the weights in the encoder 1D convolutional layers. Second, we randomly scale the amplitude of earthquake waveforms related to the noise signals to increase the range of SNR from 10^{-2} to 10^{4}. We linearly stack the shifted and scaled earthquake and ambient noise waveforms to generate the composite waveforms. We report that the resulting SNR of our training data is uniformly distributed in the logspace.

Normalization of the input data is necessary to stabilize the optimization. We use a standard scaler to normalize the time series to zero-mean and unit-variance. We rescale the earthquake and noise signals using that factor after the two decoder branches of WaveDecompNet. The loss error is the mean square error (MSE) function, and the training is performed using the Adam optimizer (Kingma and Ba, 2014).

The training, validation, and testing data sets are split using 60%-20%-20%. Only the training data are used to train the network, update the model parameters and minimize the loss function. We use a batch size of 128 during training. The validation data is used to track over-fitting during the training. Over-fitting is also mitigated with an early stopping strategy (`https://github.com/Bjarten/early-stopping-pytorch`) with the patience of 10; that is, the training automatically terminates if the value of validation loss remains unimproved for 10 epochs.

After training, we use the test data set to evaluate the model performance, especially for choosing the best one from models using different bottlenecks (more details follow in the next section).

After training, we evaluate the model performance on the test data set. Figures 2 - 3 show two examples with different types of noise signals: Figure 2 contains the regional noise from the IU.POHA station, which is energetic mostly low-frequency below 1 Hz and characteristic of microseism at a dominant period of 7 s, Fig.3 contain the STEAD noise signal, which is broadband and rich in high frequencies in a band that overlaps with the earthquake signal. Visually, the earthquake waveforms are relatively well recovered over a broad range of frequencies. In particular, it is

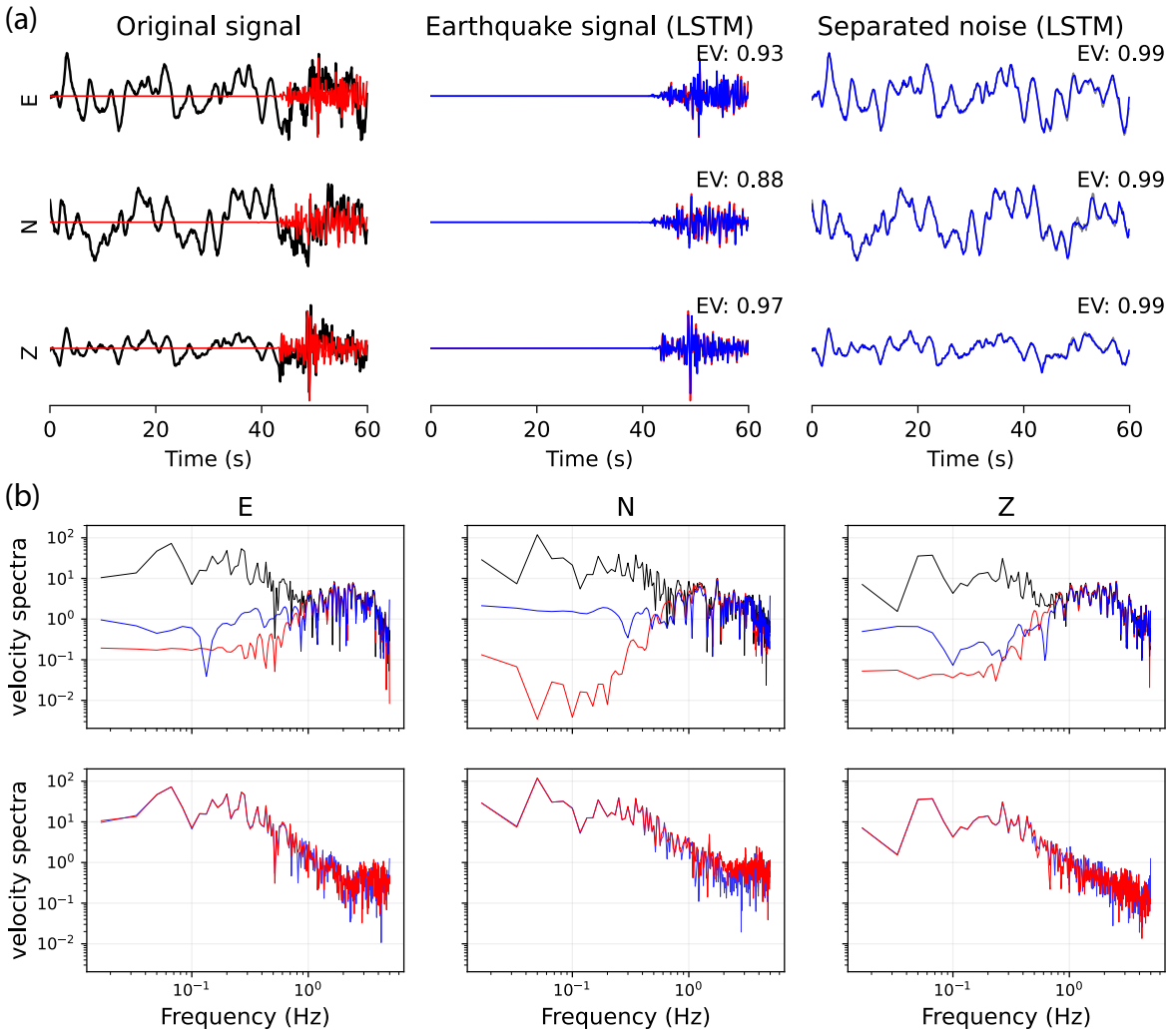

Figure 2: Example 1 of waveform separation using one of the bottleneck architecture (LSTM, see section 2.3). (a) 3-component (E-East, N-North, Z-vertical from top to bottom) velocity seismograms normalized with the same scaling factor of maximum amplitude in each component. (left panels) One STEAD earthquake waveform (in red) and IU.POHA local noise is stacked to get the noise input waveform (in black). (Middle panels) Comparison between the separated earthquake waveforms (blue) with the ground truth earthquake waveform (red). (right panels) Comparison of the separated noise waveform (blue) and ground truth noise waveform (red). (b) 3-component waveform Fourier amplitude spectra. (Top panels) The spectrum of the input waveform is shown in black, with the ground truth earthquake spectrum (in red), the separated earthquake spectrum (in blue). (Bottom panels) The ground truth noise spectrum is shown in red, and the separated noise spectrum is shown in blue.

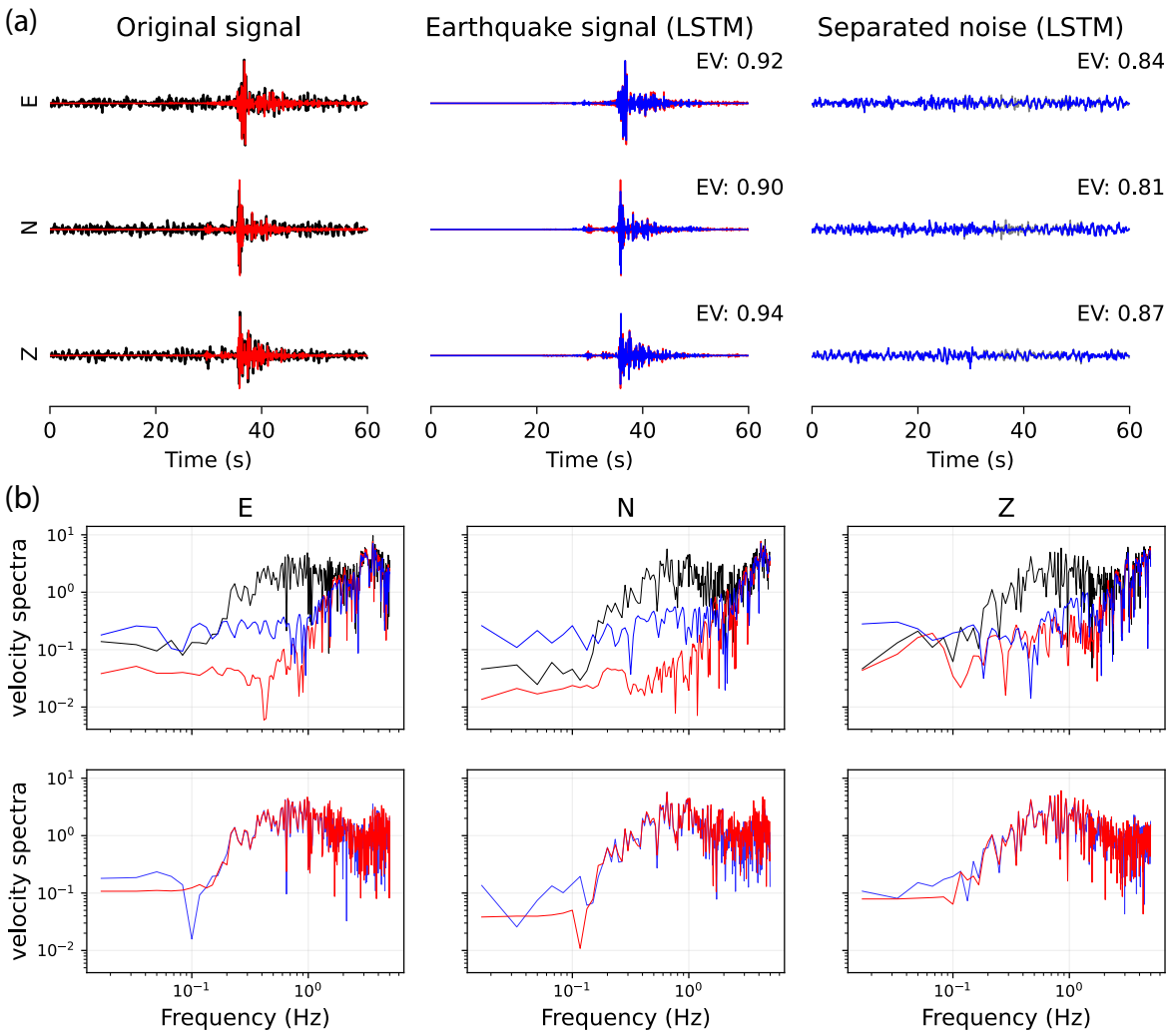

Figure 3: Example 2 of waveform separation. Same as in Figure 2 except that the noise waveform is from the STEAD dataset.

able to decompose the signals with overlapping frequency content (Figures 2 - 3 (b)), which is often a challenge for filtering-based denoising methods (Douglas, 1997; Mousavi and Langston, 2017).

2.3 Choice of the network bottleneck

The bottleneck block aims at learning the mapping relation between features in the encoder and decoder, and those features are necessary to reconstruct the separated signals in the decoder blocks. There are multiple choices for the bottleneck structure in time series analysis. We explore five of them and evaluate their impacts on model performance:

1. None: no specified bottleneck. The encoder and decoder are directly connected. The total number of trainable parameters in the network is 78,090.

2. Linear: a linear regression layer between the encoder and the decoder. The total number of trainable parameters in the network is 86,410.

3. LSTM: a bidirectional long-short-term-memory (LSTM) layer between the encoder and the decoder (Hochreiter and Schmidhuber, 1997). The total number of trainable parameters in the network is 178,442.

4. Attention mechanism: a multi-head attention layer between the encoder and the decoder (Vaswani et al., 2017). We use a 4-head dot-product self-attention layer with a dropout probability of 0.2. Other numbers of heads were tested but did not significantly affect the results. The total number of trainable parameters in the network is 110,858.

5. Transformer: the standard transformer encoder layer made up of self-attention and feed-forward network (Vaswani et al., 2017). The transformer model has been shown to be a powerful tool in different seismological applications such as earthquake detection and phase picking (Mousavi et al., 2020), earthquake source characterization (Münchmeyer et al., 2021a) and early warning (Münchmeyer et al., 2021b). We only use one layer but find adding more layers can greatly downgrade the model performance, which we attribute to insufficient training. The total number of trainable parameters in the network is 640,394.

2.4 Model Training

We use the identical training and validation data set to train all these networks. We train and validate over the batch size of 128. We require each network to be trained for at least 30 epochs. After 30 epochs, we apply the same early

stopping patience of 10 epochs. These steps can help guarantee the convergence of each model's optimization while avoiding over-fitting (Fig.4).

The overall training behaves properly for all models. During model training, the training loss curve keeps decreasing. The validation loss curve decreases and approaches the training loss curve as the training goes, and finally stays almost constant at some epoch, which indicates the convergence of optimization and the model is no longer improved. LSTM and attention models achieve the minimum final loss value for both training and validation data sets (Figs. 4 (c)-(d)). The transformer model, although with more model parameters than any other models, shows a higher validation loss value than that of the LSTM and attention bottlenecks (Fig.4 (e)). The minimal None and Linear bottlenecks exhibit the highest final loss values for both training and validation loss (Figs. 4 (a)-(b)). We also show the partial loss curves from individual branches of the earthquake and the noise waveforms, approximately half of the total loss. The validation loss for the earthquake waveforms is slightly higher than that of the earthquake loss probably due to the complexity of transient earthquake signals.

Next, we test the trained models with the 28,800 samples of the test data set. The test data set is not included during the training process, so it can be used to evaluate the model performance. The LSTM and Attention models have achieved the minimum mean test loss value of 0.0503 and 0.0557, respectively. The transformer model has a mean test loss value of 0.0836. The None and Linear models have mean test loss values of 0.0831 and 0.0893, respectively. The distribution of test loss values are also shown in Figure A.1.

Furthermore, we inspect the waveform fitting for different models. For the same input $\mathbf{X}$ (composite waveform), we obtain the predicted output/waveform $\tilde{\mathbf{y}}$ and compare it with the ground truth waveform y. Other seismic denoising studies have reported improved SNR values as performance metrics (Zhu et al., 2019a; Tibi et al., 2021). Tibi et al. (2021) also use the signal-to-distortion ratio (SDR) metric (Nakajima et al., 2018), but the SDR metric is unbounded. Here, we calculate the EV score for both separated earthquake waveforms and ambient noise waveforms: $EV = 1 - \frac{Var(\mathbf{y}-\tilde{\mathbf{y}})}{Var(\mathbf{y})}$, Var means variance of the time series. The best possible EV score is 1.0, corresponding to perfect waveform reconstruction. An EV score of 0.0 means that no waveform has been reconstructed ($\tilde{\mathbf{y}} = 0$). A negative EV score means a false waveform reconstruction, for example, in the time window where there is no earthquake waveform ($\mathbf{y} \approx 0$) but the network reconstructs a spurious waveform ($\tilde{\mathbf{y}} \neq 0$).

The comparative results are shown in Fig.5. All models can reconstruct both earthquake and noise waveforms with over half of the tested samples that achieve a high EV score around 1 (Fig.5). The network with the LSTM bottleneck recovers most test samples with a high EV score of around 1 for the earthquake waveforms (Fig.5 (a)). There is a bimodal distribution in the EV scores for earthquake waveforms. All models show two peaks around EV scores of 1

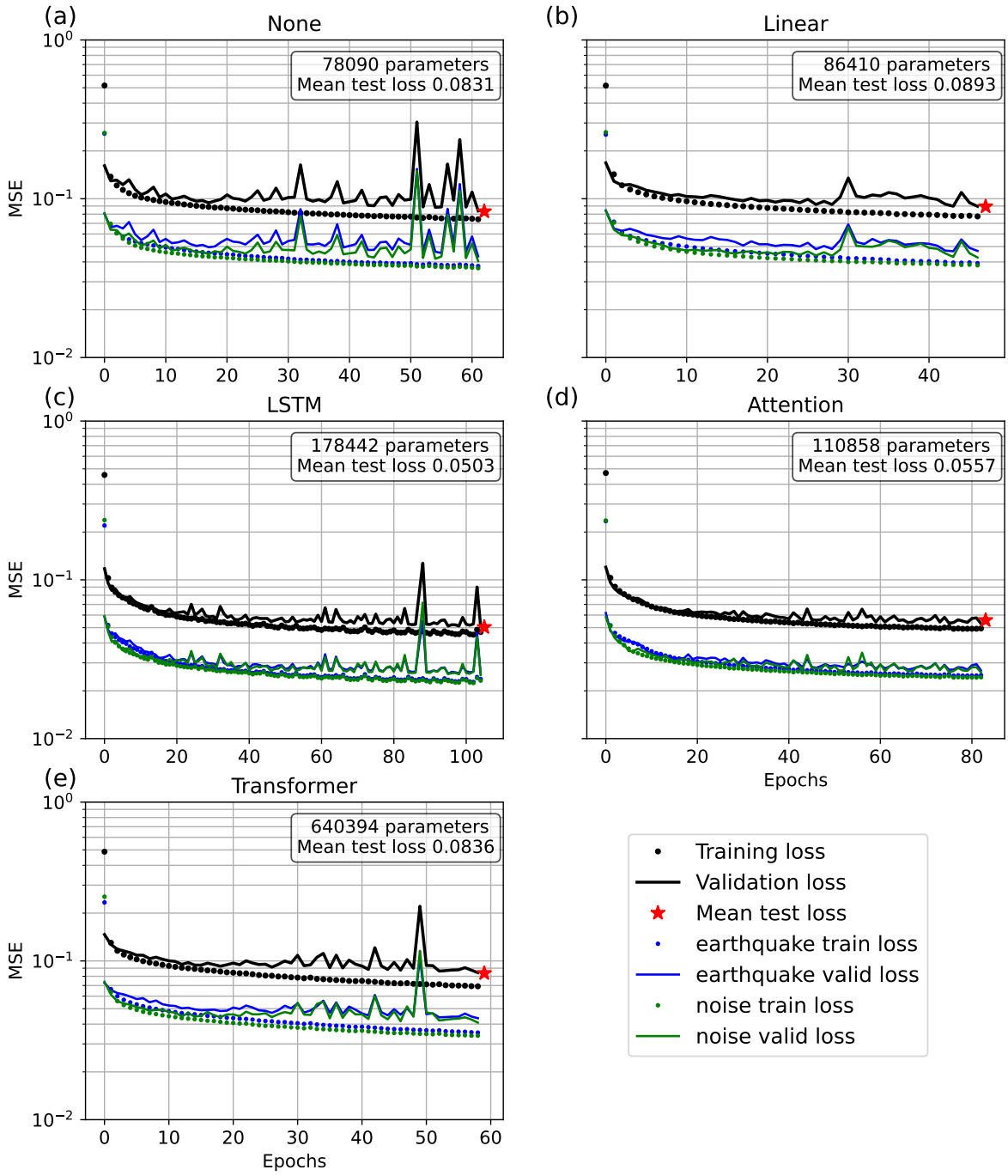

Figure 4: Training curves of networks with different bottlenecks: (a) None, (b) Linear; (c) LSTM; (d) Attention; (e) Transformer. Dots and solid lines indicate the loss from training data set and validation data set, respectively. Colors indicate different part of loss: total loss in black, earthquake waveform (decoder 1) in blue and noise waveform (decoder 2) in green. The red star indicates the loss from test data set.

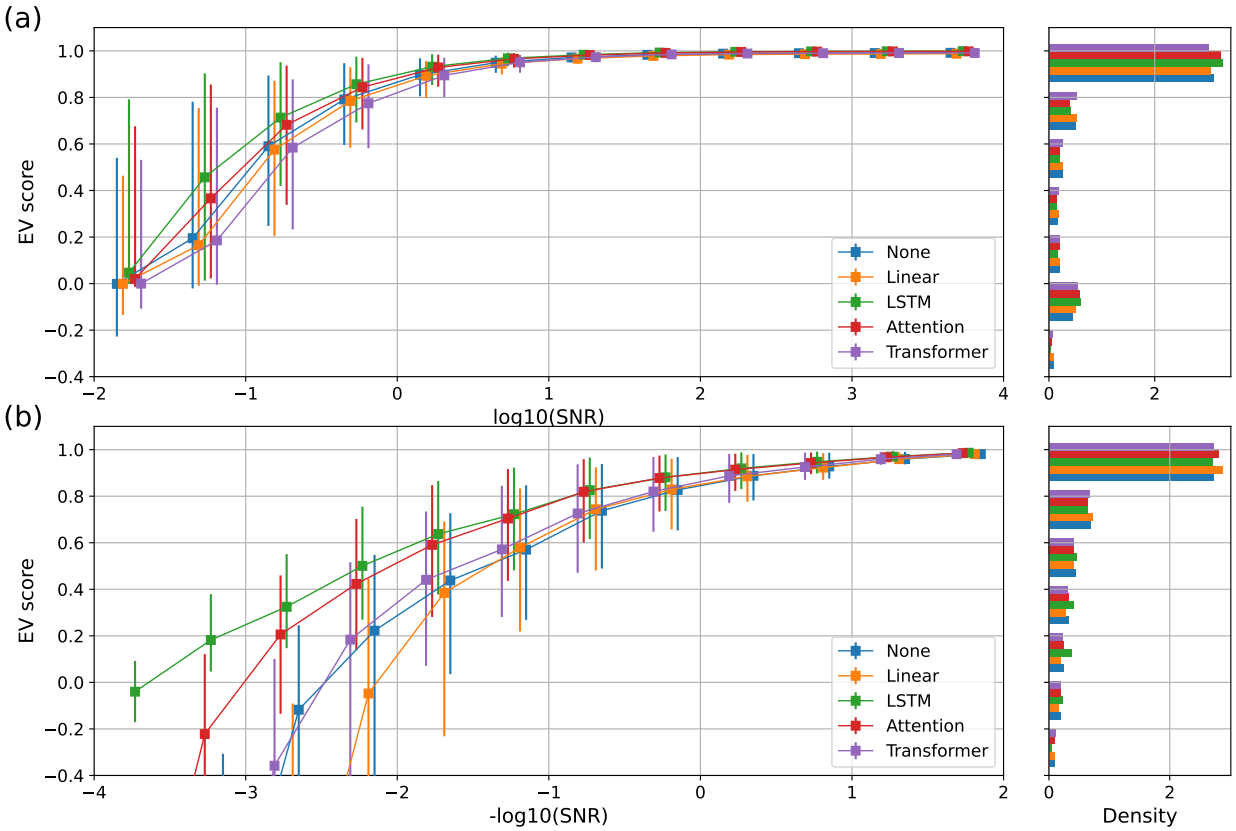

Figure 5: Comparison between networks with different bottlenecks. (a) Variation of earthquake EV score with SNR of the noisy input signal. (b) Variation of noise EV score with SNR of the noisy input signal. Colored lines show the median EV score from networks with different bottlenecks. The error bars are calculated from the median values of EV score for test samples above and below the median EV score. Right panels show the histograms of the EV score of each network with the same color scheme.

and 0, especially when the earthquake-to-noise ratio is less than 1. The bimodal pattern is suggestive of the behaviors of this encoder-decoder network. In cases that the network can recognize, the network can reconstruct the waveforms accurately. In cases that the network can hardly recognize, the network tends to output a time series of zeroes, which leads to almost 0 EV score. All networks show similar performance in EV score for the noise waveforms (Fig.5 (b)). We find no obvious bimodal pattern in the ambient noise reconstruction, which indicates a lower likelihood of outputting strictly zero noise. There are, however, spurious reconstructions of the noise waveforms for lower noise-to-earthquake amplitude ratios, or when transient signals dominate the time series (Fig.5 (b)).

Moreover, we explore how the EV score varies with SNR for both earthquake (Fig.5 (c)) and ambient noise waveforms (Fig.5 (d)). First, all models present the same pattern that the EV score monotonically increases with the corresponding amplitude of signals, quantified by SNR for earthquake waveforms or 1/SNR for ambient noise waveforms, respectively. This is well expected because it is easier for the ML network to learn the intrinsic features of the signal waveforms and reconstruct the earthquake and ambient noise signals for higher amplitude. All models perform similarly. Take the case of earthquake waveform as an example (Fig.5 (a)). For SNR $> 10^1$, all models can correctly separate the earthquake waveform almost perfectly (EV score ~ 1). The performance of any model drops as the SNR decreases. For example, at SNR = 1, the median EV score of models is about 0.8 to 0.9, and LSTM has the best performance. The discrepancy between model performance is exacerbated at low SNR. For instance, when the SNR is small, $= 10^{-1} = 0.1$, it is visually difficult to extract the earthquake signal. However, the LSTM model can still achieve a median EV score of about 0.6, the Attention model has a median EV score of 0.5, the None model has a median EV score of 0.48, the Linear model has a median EV score of about 0.4, and the Transformer model has a median EV score of 0.3. For even smaller SNR values, $= 10^{-1.8} = 0.02$, all models tend to fail with most EV score being 0. The variance of the EV scores also increases with decreasing SNRs, indicating that there are more uncertainties in the reconstructed waveforms. Similar behaviors can be observed for the EV score of ambient noise part, with larger amplitudes of noise yields to better model performance (Fig.5 (b)).

One possible explanation for the different performances between bottlenecks architectures is the difference in model complexity. None and Linear models have fewer parameters than other models, so they may not be enough to understand the internal features of the seismograms properly. The None and Linear models have higher training loss (about 0.08) than other models,suggesting a not good model. The Linear model presents a larger loss value (Fig.4 (b)) and poorer waveform fitting (Fig.5) than the None model, implying the inability of linear regression as the bottleneck layer for this waveform decomposition problem. On the other hand, the Transformer model is more complex than the other models. Its mean test loss (0.0836) is slightly lower than that of the Linear model (0.0893), and the earthquake

waveform fitting is almost the same as the Linear model (Fig.5 (a)) but the noise waveform fitting is similar to that of the None model (Fig.5 (b)). The LSTM and attention models share a similar overall complexity and achieve the lowest test loss and the most stable training. The similarity and systematically low values of the training and validation losses for the LSTM and the attention bottleneck may also indicate that those two models have already well "learned" the features in the training data sets (Fig.4 (c) and (d)).

To summarize, we evaluate the performance of models with different types of bottleneck models by testing the same test data set. We find that the model performance can differ due to model complexity. In general, the order of performance of our model is LSTM $>$ attention $>$ None $>$ Linear $\approx$ Transformer based on the variation of EV score with SNR. The LSTM bottleneck outperforms other bottlenecks in reconstructing both earthquake and noise waveform, especially for a situation with low SNR (Fig.5 (c)). It is interesting to note that LSTM performs better than the attention and Transformer models, which implies that the sequential information is essential for reconstructing the waveforms. We speculate that the feature extraction of the encoder branch suffices at representing the temporal sequencing in the bottleneck layer. The conventional limitations of LSTM that long memory is not long enough are no longer important.

3 Application to continuous seismic data

We now apply WaveDecompNet to continuous time series. It is straightforward to apply the model to any continuous data, provided that it has the same sampling rate. We select continuous recordings at IU.POHA from July 31, 2021, to September 1, 2021. We first down-sample the three-component 1-month-long waveforms to 10 Hz. Next, the typical pre-processing steps to apply machine-learning models is a) windowing to 1-minute long time series (600 samples) without overlap and b) applying the data normalization using the standard scaler. The most intuitive order to apply these processing steps are a), then b). We found that ordering a) then b) leads to spurious effects when concatenating back the 1-minute waveforms into a 1-month long waveform due to offset (means) and trends that rendered the application to continuous time series unpractical. Instead, we experimented with the order of b) then a) and found much better performance without artifacts when stitching back the waveforms.

We normalize the entire month-long time series by removing its mean and scaling with its standard deviation (STD) to have the zero-mean, unit-variance time series. We slide through the data with 1-minute long windows (600 samples) without overlap. We apply the WaveDecompNet to all 1-minute long windows, concatenate all ML-filtered windows, and scale back the 1-month long time series with the standard deviation and mean for both the earthquake/transient

time series and the noise time series.

We show the results of separating transients and noise waveforms in Figure 6. Most of the transient signals have been well separated, with a significantly suppressed noise level (Fig.6 (b)). The residuals between the original waveform and the reconstructed waveforms are obtained from subtracting the earthquake and noise recovered waveform from the input waveform (Fig.6 (d)). Overall, the residuals are low. However, they are large between August 10, 2021, and August 20, 2021, and these are due to the teleseismic (30$\circ$ to 100$\circ$ angular distance) earthquakes. We mark the P arrival, calculated from TauP with IASP91 Earth model, of these large M5.5+ teleseismic earthquakes to illustrate that in Figure 6. No teleseismic waveform was used during the training, in part because the input data length is restricted to one minute. Therefore, our model does not handle longer seismic periods at this stage, and the coda reconstruction of these long waveforms is imperfect. Nevertheless, the general envelop pattern of those earthquake waveforms can still be recovered. In the following section, we test the validity and usefulness of these transient waveforms by applying a standard impulsivity filter most commonly used detection method in seismology.

The separated noise waveforms exhibit more leveled, constant amplitudes throughout the month (Fig.6 (c)). Some transient signals remain, especially in the coda of teleseismic earthquakes. For additional evaluation of the usefulness of this network, we apply the single-station correlation functions used in ambient-seismic noise monitoring in a later section.

Most of the previous denoising networks, such as the DeepDenoiser (Zhu et al., 2019b), construct the noise time series from direct subtraction of the "denoised" earthquake waveforms from the raw data. Unlike the DeepDenoiser, the WaveDecompNet has two branches that learn features of the earthquake and noise waveforms, somewhat independently since their only connection is through a residual connection to the encoder branch. Because the noise window is not the linear difference between the original and the transient/earthquake signal, we also investigate the waveform residuals and show them in Figure 6 (d). In general, the amplitudes of the residual waveforms are small (about 10 times lower than the ambient noise in the standard deviation of waveform amplitude, Fig.6 (c)) except for some large residuals during large teleseismic events. Including more teleseismic earthquake waveforms in the model training can potentially help to mitigate those large residuals, and we leave as a future direction to explore.

3.1 Application to detecting earthquakes using STA/LTA trigger

We apply a recursive short-term-average (STA) to long-term-average (LTA) trigger method (STA/LTA) to the continuous data (Vanderkulk et al., 1965; Allen, 1978; Withers et al., 1998). This particular STA/LTA algorithm produces

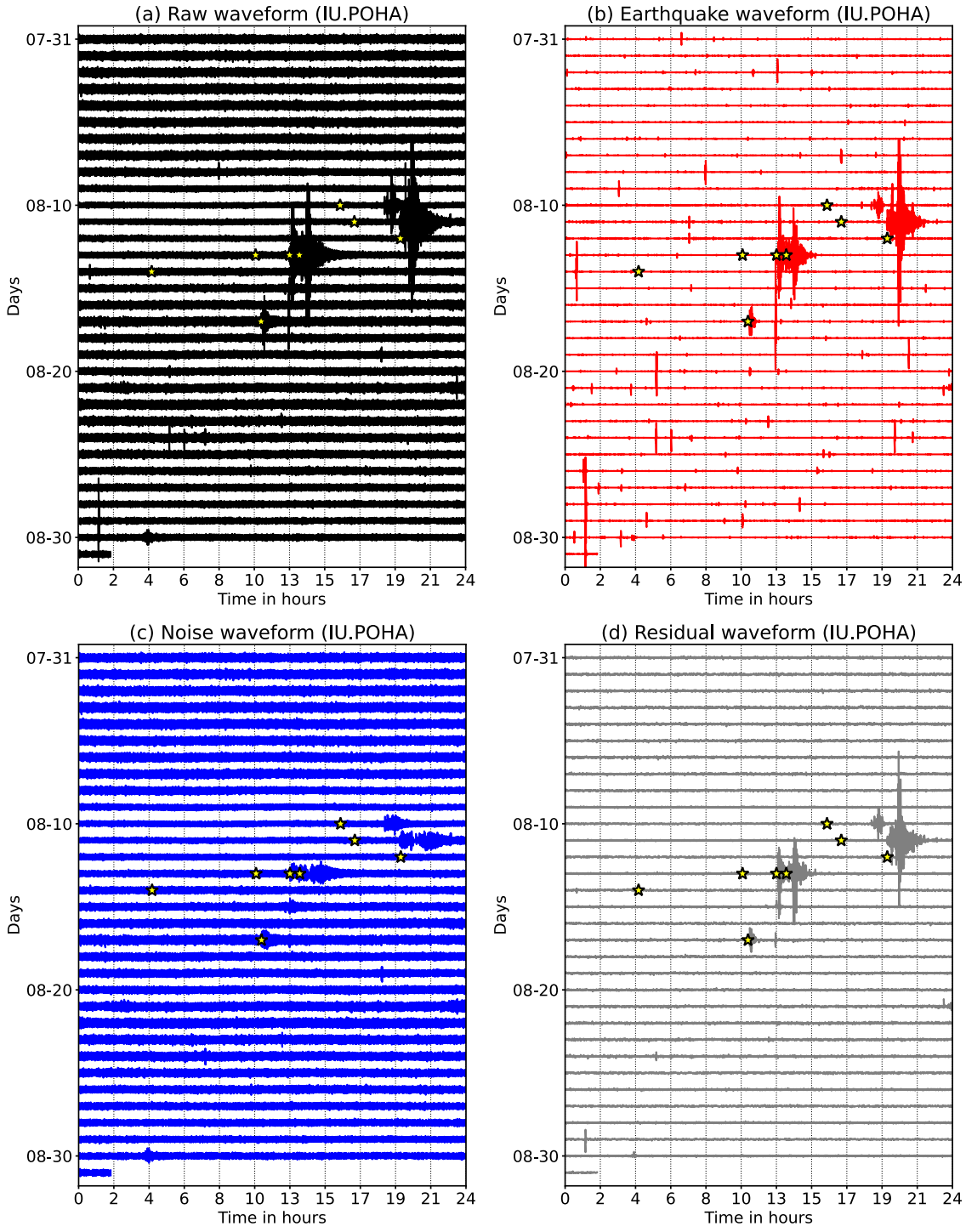

Figure 6: Application to continuous seismic data from an island station IU.POHA. (a) One-month raw waveform from IU.POHA; (b) Separated earthquake waveform; (c) Separated noise waveform; (d) Waveform residuals from subtracting the separated earthquake and noise waveforms from the raw waveform. The yellow stars label the P wave arrivals, TauP calculates with IASP91 velocity model(Kennett and Engdahl, 1991), of large earthquakes (M5.5+) between August 10, 2021, to August 20, 2021, from the International Seismological Centre catalog provided by default by Obspy and IRIS FDSN event server.

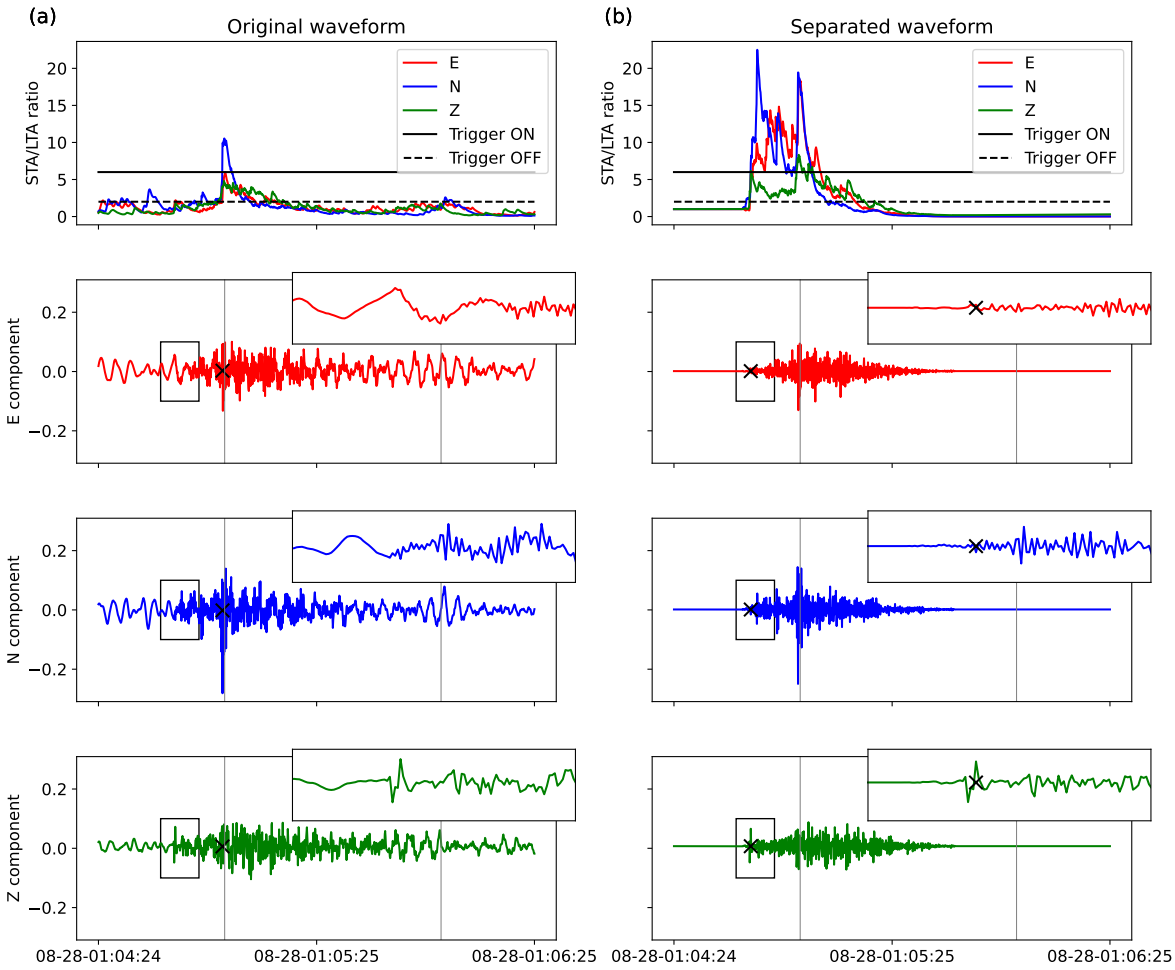

Figure 7: Example of STA/LTA detection algorithm: (a) original waveform, (b) separated earthquake waveform. (Top panels) Recursive STA/LTA ratio from the waveform in the chosen window. Black solid and dashed lines indicate the trigger thresholds on and off, respectively. (bottom panels) Red, blue and green lines indicate E, N, Z components. Black crosses show the picks from STA/LTA. Gray vertical bars indicate the edges of the 1-minute time windows when applying WaveDecompNet. The inset figures show the zoom-in waveforms within the boxes.

a decaying exponential impulse response, and that is sharper impulse than the original STA/LTA algorithm (Withers et al., 1998). The settings of the STA/LTA parameters are chosen from Trnkoczy (2009) and also from trial-and-error tests. The short time window length is set at 2.0 s, the long time window length is set to 60.0 s, the on-threshold is 6.0, and the off-threshold is 2.0. Because we simultaneously run STA/LTA detection on 3-component waveforms, we perform a coincidence trigger with a threshold of 2, which means that a detection trigger occurs when the STA/LTA ratios of any of the two components exceed the on-threshold.

The STA/LTA time series are a lot cleaner in the separated earthquake waveforms than in the original seismograms (Fig.7), which is manifested in two aspects. First, the increased signal-to-noise ratio of the separated earthquake waveforms improves the accuracy of the detection time automated by STA/LTA triggers. For the example shown in

Fig.7, the arrival time cannot be correctly picked in the raw data using automated STA/LTA thresholding detector (Fig.7 (a)). Nevertheless, with the noise separated by WaveDecompNet, the event arrival can be easily detected, and the accuracy of picking the first arrivals can be improved by about 15 seconds (Fig.7 (b)). Second, we can detect many smaller signals (either smaller magnitude or more distant events) buried in the noise, which is suggested by the increased number of coincidence triggers in the separated waveform, from 38 in the original time series to 1031 in the separated waveforms. The wavefield separation increases the detection by a factor of 68 in the current STA/LTA settings. This ratio varies with the choice of threshold, from 4 (1899 events vs. 478 events) for a coincidence threshold of 1 and 7 (147 events vs. 19 events) for a coincidence threshold of 3. Tuning the parameters of this detector is not the scope of this study but would be necessary in the deployment of this algorithm in specific cases.

3.2 Application to ambient noise monitoring using single-station cross-correlations

Single-station correlations are related to the zero-offset Green's function (Claerbout, 1968; Draganov et al., 2009; Saygin et al., 2017; Clayton, 2020). Monitoring phase changes in the single-station measurements have enabled the monitoring of changes in the near-surface environment that occur during earthquakes (Wegler and Sens-Schönfelder, 2007; Viens et al., 2018), volcanic unrest (De Plaen et al., 2016), and to monitor shallow hydrology (Illien et al., 2021). Here, we do not attempt to verify that the single-station correlation is proportional to the Green's function. Instead, we evaluate the temporal stability of the single-station cross-correlations.

We calculate all 9 components of the correlation tensor. We select 1-minute long windows, pad them with zeroes from 600 samples to 2048 samples (204.8 s). We then follow the spectral method from Viens et al. (2020) to calculate the ambient noise single-station correlation function (ACF):

$$ACF_{ij}(t) = F^{-1}\left(\frac{\hat{a}_i \hat{a}_j^*}{|\hat{a}_i||\hat{a}_j|}\right), \tag{2}$$

where i, j corresponds to components (E, N, and Z), $\hat{a}_i$ is the Fourier transform of the i-component waveform, * represents the complex conjugate, $F^{-1}(\cdot)$ is the inverse Fourier transform. We whiten the amplitude spectrum using a running mean as in conventional processing (Bensen et al., 2007) $|\cdot|$ of 32 samples in the frequency domain.

We sub-stack the correlations functions every 6 hours to evaluate their stability through time. We show the causal part (positive lag) of correlation functions in different frequency bands: Low Frequency (LF) 0.1 - 1.0 Hz (Fig.8); Medium Frequency (MF) 1.0 - 2.0 Hz (Fig.9); High Frequency (HF) 2.0 - 4.0 Hz (Fig.10). Each figure shows the single-station

correlations from the original raw waveforms and ones obtained from the separated waveform.

We also stack all of the correlations to form a reference stack, from July 31, 2021, to September 1, 2021. We calculate the correlation coefficient between each 6-hour stack and the month-stack reference waveform and show them in Figure 11. We use the stability of cross-correlation as a success metric of ambient seismic noise recovery.

Because of transient earthquake signals in the raw waveforms, we can see large fluctuations (mostly reduced amplitudes) in the correlation functions at all frequency bands, especially for the E-Z, N-Z, Z-E, Z-N functions (Figs.8 - 10). These fluctuations in the correlation functions and drops in their coherence arise from transient signals in the original time series. However, many of these fluctuations disappear when using the separated noise signals to calculate the correlation functions. Furthermore, some of the coda phases that are weak in the original correlation functions appear clearly in the correlation functions built from the separated noise. These coda phases potentially correspond to seismic wavespeed interfaces or discontinuity beneath the seismic station. With the transient earthquake signals removed, WaveDecompNet can help constrain the velocity structure underneath the seismic station. Additional work remains to be done to verify the nature of these coda phases and whether they can be related to Earth structure.

As expected, the improvement on the correlation functions coherence is substantial (Fig.11). As shown in Figures 8 - 10, the transient earthquake signals can break the coherence among correlation functions, and lead to low correlation coefficients between each function and the reference (see Fig.11). On the other hand, the correlation coefficients from separated noise mostly have stable values closer to 1, confirming the enhanced coherence of the cross-correlation functions from continuous ambient noise data. We find that the coherence from separated noise drops in some time windows (for example, day 1 - day 7 in Fig.11 (a) and day 24 - day 30 in Fig.11 (b)). This can be possibly attributed to the poor reconstruction of the ambient noise signals. We also notice that for some components, the separated noise does not improve the coherence at all, for example, the N-Z component in MF (Fig.11 (b)) and Z-E component in HF (Fig.11 (c)). Further study is needed to understand these less-dominant issues in ambient noise applications.

4 Conclusion and Discussion

We develop a machine-learning-based model, WaveDecompNet, to separate earthquake and ambient noise signals from raw seismic data. We combine the STEAD and local ambient noise to form a sufficient overall data set to train and test WaveDecompNet. Our network consists of three parts: one encoder branch, two decoder branches, and two bottlenecks. We systematically explore the performance of models using different types of bottlenecks, and we find

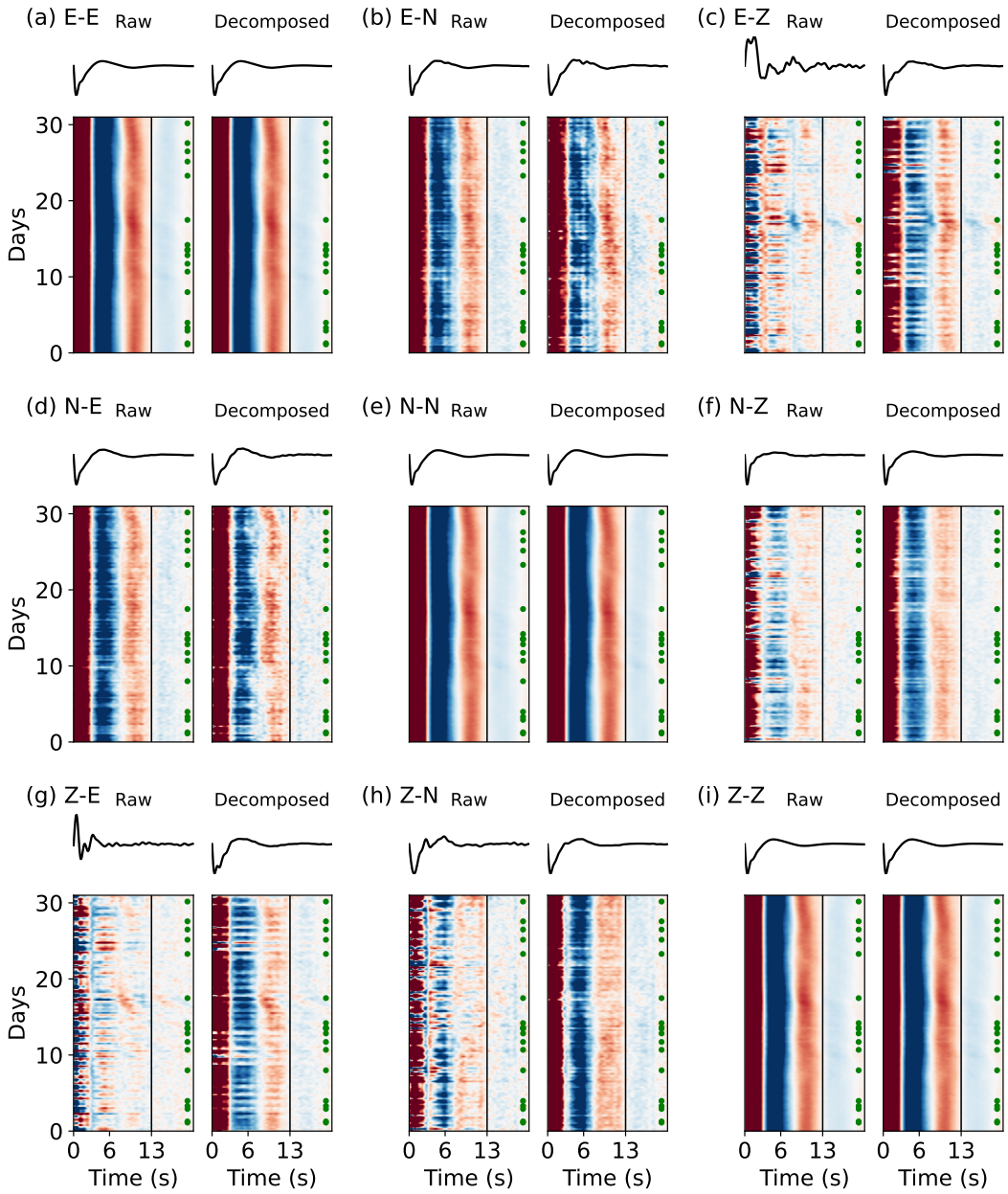

Figure 8: Single-station cross-correlation and auto-correlation functions filtered in the LF low frequency band (0.1 - 1.0 Hz) for the original raw waveforms (a) and the separated noise waveforms (b). Green dots show the P wave arrivals of M5.5+ teleseismic earthquakes.

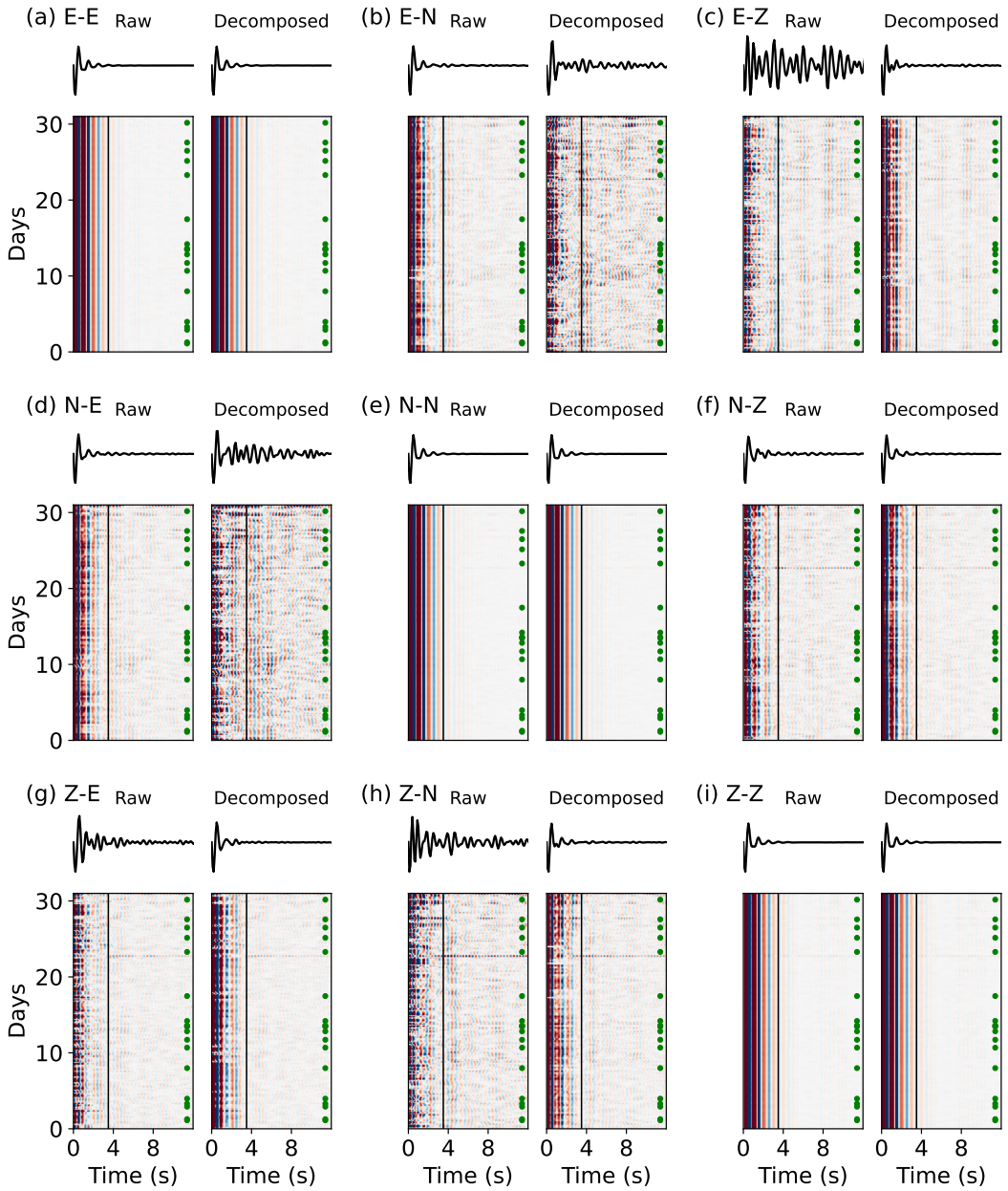

Figure 9: Same as Figure 8 for the MF medium frequency band 1.0-2.0 Hz.

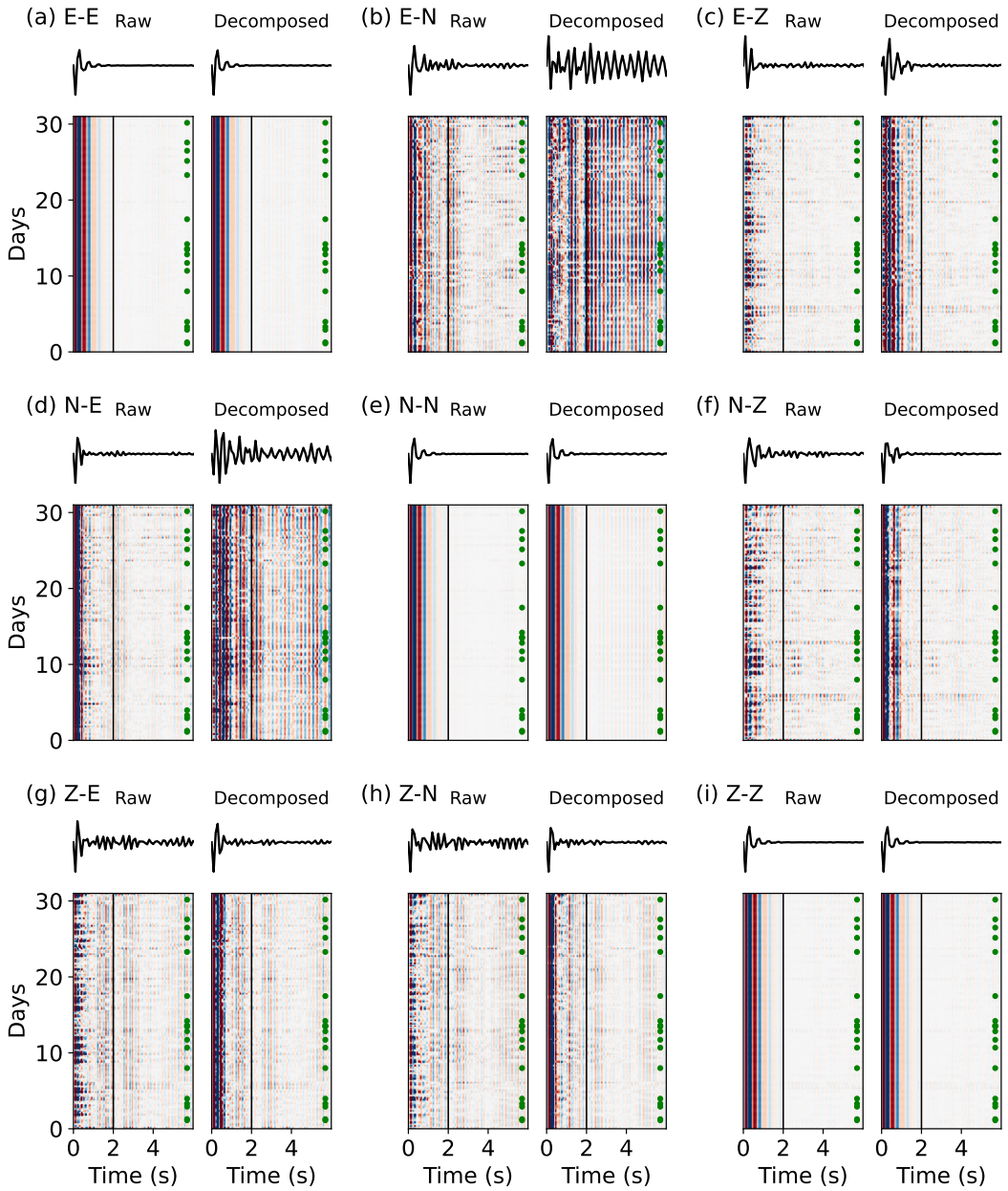

Figure 10: Same as Figure 8 for the HF high frequency band 2.0-4.0 Hz.

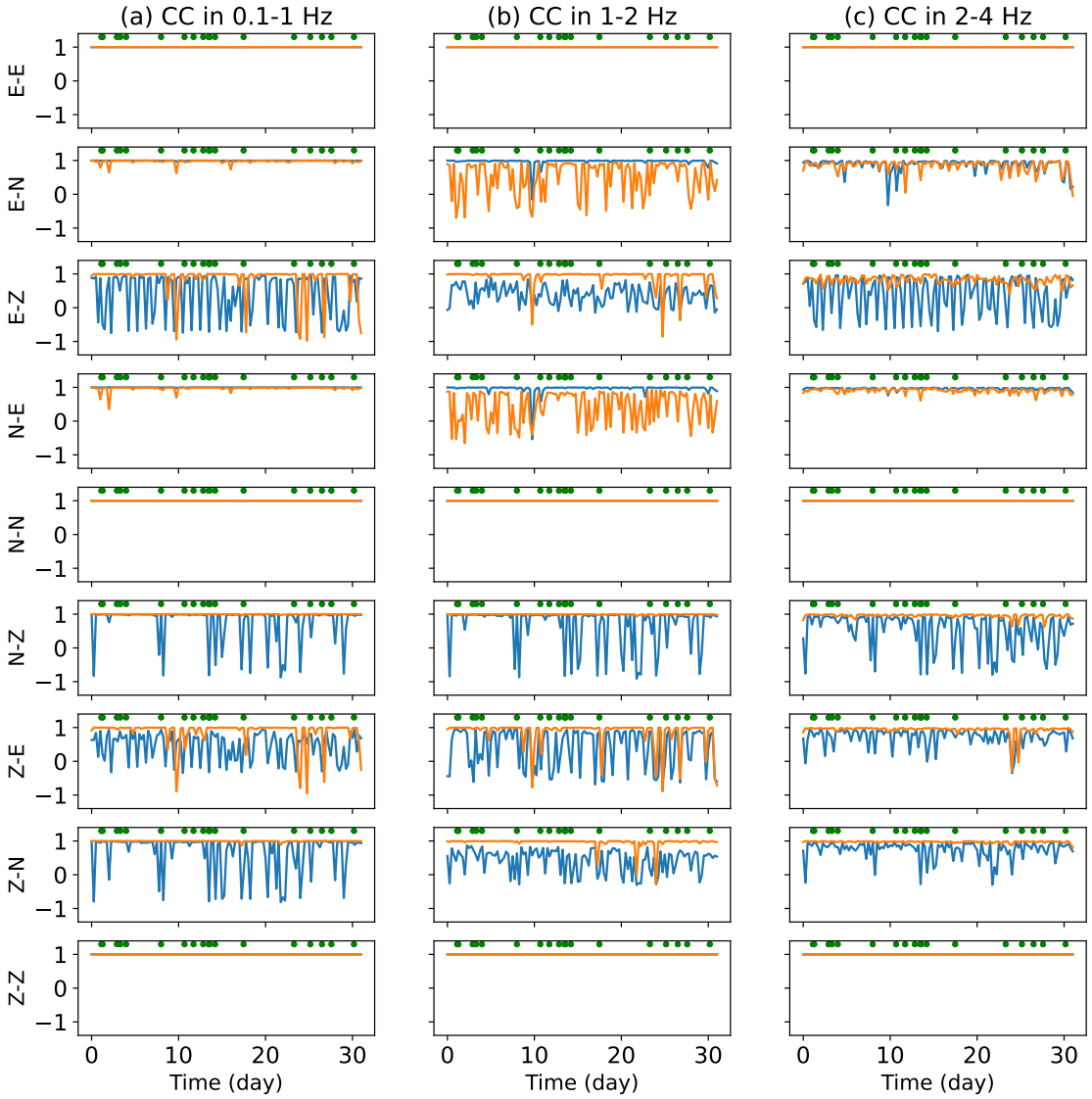

Figure 11: Coherence of the single-station cross-correlation and auto-correlation functions at different frequency bands (a) 0.1 - 1.0 Hz; (b) 1.0 - 2.0 Hz; (c) 2.0 - 4.0 Hz. Blue lines indicate the coherence from the original waveforms, orange lines indicate the coherence from separated ambient noise waveform. The coherence is quantified by the correlation coefficients between each 6-hour averaged correlation function and the 1-month averaged reference correlation function. Green dots show the P wave arrivals of M5.5+ earthquakes in the month calculated using TauP in the IASP91 model.

the network using LSTM bottleneck has the best performance. Next, we test how well our network can be applied to observed continuous data. We apply the trained model directly to a 1-month continuous seismic data at IU.POHA and successfully separate the corresponding earthquake and noise signals, except for the long-duration teleseismic signals. Next, we apply an automated transient detector (STA/LTA) and an established ambient-noise seismology monitoring method to the separated earthquake and noise signals, respectively. Our results show that the quality of both separated earthquake and noise signals has been improved significantly. With the same ML filter, we can obtain more STA/LTA triggers and a highly coherent ambient-noise correlation function.

However, there are some limitations to our current method. First, it only includes waveforms from local earthquakes (< 350 km). The lack of teleseismic waveforms, especially those from large earthquakes, leads to the poor performance of WaveDecompNet when handling the time windows with teleseismic earthquake waveforms. While we extract the general patterns of the teleseismic earthquake waveforms correctly, there remain large residuals in the ambient noise waveforms and residual waveforms. Second, we only include the local noise from a single island station, IU.POHA. We also test with other stations from the Hawaiian Volcano Observatory and find the network trained from IU.POHA can still successfully separate the earthquake, but the coherence of ambient noise worsens, and there are more waveform residuals. Therefore, a good direction to improve the network performance is to include additional and different types of data. For example, we can include teleseismic data for better separation of earthquake waveforms, and we should also include the ambient noise waveforms from other stations and regions for a specific regional or global ambient noise study.

Future developments may involve the integration of multiple stations. The combination of multiple stations to combine the automated triggered events help reduce the false (non-tectonic) detections. It also helps locate the event and build a more complete earthquake catalog. Furthermore, a modification of the network to add more stations may help improve the stability of the inter-station cross-correlations, which in turn can be used for better Earth imaging.

5 Data and resources

The continuous seismic data from IU.POHA (IU: `doi:10.7914/SN/IU`) are downloaded using Obspy (available at `https://github.com/obspy/obspy/wiki`). PyTorch machine learning framework (`https://pytorch.org`) is to build and train the network. The module of self-attention bottleneck is based on Chapter 10.5 of the online book "Dive into Deep Learning" (available at `https://d2l.ai/index.html`). All the codes to reproduce this work are hosted on Github at `https://github.com/yinjiuxun/WaveDecompNet-paper`, WaveDe-

compNet is hosted on https://github.com/yinjiuxun/WaveDecompNet. The authors are grateful for the discussions with Congcong Yuan regarding the model architecture and Yiyu Ni for his comments. This work is supported by the CAREER EAR-1749556 NSF award.

References

Aki, K. (1957). Space and time spectra of stationary stochastic waves, with special reference to microtremors. *Bulletin of the Earthquake Research Institute*, 35:415–456.

Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces. *Bulletin of the seismological society of America*, 68(5):1521–1532.

Badrinarayanan, V., Handa, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Bappy, J. H., Simons, C., Nataraj, L., Manjunath, B., and Roy-Chowdhury, A. K. (2019). Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300.

Bensen, G., Ritzwoller, M., Barmin, M., Levshin, A. L., Lin, F., Moschetti, M., Shapiro, N., and Yang, Y. (2007). Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International*, 169(3):1239–1260.

Bergen, K. J., Johnson, P. A., Hoop, M. V. d., and Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433):eaau0323.

Beroza, G. C., Segou, M., and Mostafa Mousavi, S. (2021). Machine learning and earthquake forecasting—next steps. *Nature Communications*, 12(1):1–3.

Burtin, A., Bollinger, L., Vergne, J., Cattin, R., and Nábělek, J. (2008). Spectral analysis of seismic noise induced by rivers: A new tool to monitor spatiotemporal changes in stream hydrodynamics. *Journal of Geophysical Research: Solid Earth*, 113(B5).

Cessaro, R. K. (1994). Sources of primary and secondary microseisms. *Bulletin of the Seismological Society of America*, 84(1):142–148.

Chang, S. G., Yu, B., and Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. *IEEE transactions on image processing*, 9(9):1532–1546.

Chen, Y., Ma, J., and Fomel, S. (2016). Double-sparsity dictionary for seismic noise attenuation. *Geophysics*, 81(2):V103–V116.

Claerbout, J. F. (1968). Synthesis of a layered medium from its acoustic transmission response. *GEOPHYSICS*, 33(2):264–269. Publisher: Society of Exploration Geophysicists.

Clayton, R. W. (2020). Imaging the Subsurface with Ambient Noise Autocorrelations. *Seismological Research Letters*, 91(2A):930–935.

De Plaen, R. S., Lecocq, T., Caudron, C., Ferrazzini, V., and Francis, O. (2016). Single-station monitoring of volcanoes using seismic ambient noise. *Geophysical Research Letters*, 43(16):8511–8518.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Douglas, A. (1997). Bandpass filtering to reduce noise on seismograms: is there a better way? *Bulletin of the Seismological Society of America*, 87(3):770–777.

Draganov, D., Campman, X., Thorbecke, J., Verdel, A., and Wapenaar, K. (2009). Reflection images from ambient seismic noise. *GEOPHYSICS*, 74(5):A63–A67. Publisher: Society of Exploration Geophysicists.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Illien, L., Andermann, C., Sens-Schönfelder, C., Cook, K., Baidya, K., Adhikari, L., and Hovius, N. (2021). Subsurface moisture regulates himalayan groundwater storage and discharge. *AGU Advances*, 2(2):e2021AV000398.

Johnson, P. A., Rouet-Leduc, B., Pyrak-Nolte, L. J., Beroza, G. C., Marone, C. J., Hulbert, C., Howard, A., Singer, P., Gordeev, D., Karaflos, D., et al. (2021). Laboratory earthquake forecasting: A machine learning competition. *Proceedings of the National Academy of Sciences*, 118(5).

Keefer, D. K. (1984). Landslides caused by earthquakes. *Geological Society of America Bulletin*, 95(4):406–421.

Kennett, B. L. N. and Engdahl, E. R. (1991). Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, 105(2):429–465.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, 90(1):3–14.

Kuang, W., Yuan, C., and Zhang, J. (2021). Real-time determination of earthquake focal mechanism via deep learning. *Nature communications*, 12(1):1–8.

Lecocq, T., Hicks, S. P., Van Noten, K., Van Wijk, K., Koelemeijer, P., De Plaen, R. S., Massin, F., Hillers, G., Anthony, R. E., Apoloner, M.-T., et al. (2020). Global quieting of high-frequency seismic noise due to covid-19 pandemic lockdown measures. *Science*, 369(6509):1338–1343.

Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z., and Andrews, J. (2018). Machine learning seismic wave discrimination: Application to earthquake early warning. *Geophysical Research Letters*, 45(10):4773–4779.

Lipovsky, B. P., Meyer, C. R., Zoet, L. K., McCarthy, C., Hansen, D. D., Rempel, A. W., and Gimbert, F. (2019). Glacier sliding, seismicity and sediment entrainment. *Annals of Glaciology*, 60(79):182–192.

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

Longuet-Higgins, M. S. (1950). A theory of the origin of microseisms. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 243(857):1–35.

McNamara, D. E., Boaz, R. I., Nakata, N., Gualtieri, L., and Fichtner, A. (2019). Visualization of the seismic ambient noise spectrum. *Seismic Ambient Noise*, pages 1–29.

Meier, M.-A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., and Yue, Y. (2019). Reliable real-time seismic signal/noise discrimination with machine learning. *Journal of Geophysical Research: Solid Earth*, 124(1):788–800.

Mousavi, S. M. and Beroza, G. C. (2019). Bayesian-deep-learning estimation of earthquake location from single-station observations. *arXiv preprint arXiv:1912.01144*.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1):1–12.

Mousavi, S. M. and Langston, C. A. (2017). Automatic noise-removal/signal-removal based on general cross-validation thresholding in synchrosqueezed domain and its application on earthquake data. *Geophysics*, 82(4):V211–V227.

Mousavi, S. M., Langston, C. A., and Horton, S. P. (2016). Automatic microseismic denoising and onset detection using the synchrosqueezed continuous wavelet transform. *Geophysics*, 81(4):V341–V355.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. (2019a). Stanford earthquake dataset (stead): A global data set of seismic signals for ai. *IEEE Access*, 7:179464–179476.

Mousavi, S. M., Zhu, W., Sheng, Y., and Beroza, G. C. (2019b). Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Scientific reports*, 9(1):1–14.

Münchmeyer, J., Bindi, D., Leser, U., and Tilmann, F. (2021a). Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. *Geophysical Journal International*, 226(2):1086–1104.

Münchmeyer, J., Bindi, D., Leser, U., and Tilmann, F. (2021b). The transformer earthquake alerting model: a new versatile approach to earthquake early warning. *Geophysical Journal International*, 225(1):646–656.

Nakajima, H., Takahashi, Y., Kondo, K., and Hisaminato, Y. (2018). Monaural source enhancement maximizing source-to-distortion ratio via automatic differentiation. *arXiv preprint arXiv:1806.05791*.

Park, S. H., Kim, B., Kang, C. M., Chung, C. C., and Choi, J. W. (2018). Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1672–1678. IEEE.

Perol, T., Gharbi, M., and Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Science Advances*, 4(2):e1700578.

Peterson, J. et al. (1993). Observations and modeling of seismic background noise.

Ren, C. X., Hulbert, C., Johnson, P. A., and Rouet-Leduc, B. (2020). Machine learning and fault rupture: a review. *Advances in Geophysics*, 61:57–107.

Rhie, J. and Romanowicz, B. (2004). Excitation of earth's continuous free oscillations by atmosphere–ocean–seafloor coupling. *Nature*,

431(7008):552–556.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Saad, O. M., Inoue, K., Shalaby, A., Samy, L., and Sayed, M. S. (2018). Automatic arrival time detection for earthquakes based on stacked denoising autoencoder. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1687–1691.

Saygin, E., Cummins, P. R., and Lumley, D. (2017). Retrieval of the P wave reflectivity response from autocorrelation of seismic noise: Jakarta Basin, Indonesia. *Geophysical Research Letters*, 44(2):792–799. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2016GL071363.

Schippkus, S., Garden, M., and Bokelmann, G. (2020). Characteristics of the ambient seismic field on a large-n seismic array in the vienna basin. *Seismological Society of America*, 91(5):2803–2816.

Sens-Schönfelder, C. and Wegler, U. (2006). Passive image interferometry and seasonal variations of seismic velocities at merapi volcano, indonesia. *Geophysical research letters*, 33(21).

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, 307(5715):1615–1618.

Siahsar, M. A. N., Gholtashi, S., Abolghasemi, V., and Chen, Y. (2017). Simultaneous denoising and interpolation of 2d seismic data using data-driven non-negative dictionary learning. *Signal Processing*, 141:309–321.

Stockwell, R. G., Mansinha, L., and Lowe, R. (1996). Localization of the complex spectrum: the s transform. *IEEE transactions on signal processing*, 44(4):998–1001.

Tan, C. W., Bergmeir, C., Petitjean, F., and Webb, G. I. (2021). Time series extrinsic regression. *Data Mining and Knowledge Discovery*, 35(3):1032–1060.

Tibi, R., Hammond, P., Brogan, R., Young, C. J., and Koper, K. (2021). Deep learning denoising applied to regional distance seismic data in utah. *Bulletin of the Seismological Society of America*, 111(2):775–790.

Trnkoczy, A. (2009). Understanding and parameter setting of sta/lta trigger algorithm. In *New Manual of Seismological Observatory Practice (NMSOP)*, pages 1–20. Deutsches GeoForschungsZentrum GFZ.

van den Ende, M. P. and Ampuero, J.-P. (2020). Automated seismic source characterization using deep graph neural networks. *Geophysical Research Letters*, 47(17):e2020GL088690.

Vanderkulk, W., Rosen, F., and Lorenz, S. (1965). Large aperture seismic array signal processing study. *IBM Final Report, ARPA Contract Number SD-296*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In

Advances in neural information processing systems, pages 5998–6008.

Viens, L., Denolle, M. A., Hirata, N., and Nakagawa, S. (2018). Complex near-surface rheology inferred from the response of greater tokyo to strong ground motions. *Journal of Geophysical Research: Solid Earth*, 123(7):5710–5729.

Viens, L., Jiang, C., and Denolle, M. (2020). Imaging the kanto basin bedrock with noise and earthquake autocorrelations.

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., and Woelfel, I. (2021). Easyquake: Putting machine learning to work for your regional seismic network or local earthquake study. *Seismological Society of America*, 92(1):555–563.

Weaver, C. S. and Malone, S. D. (1979). Seismic evidence for discrete glacier motion at the rock–ice interface. *Journal of Glaciology*, 23(89):171–184.

Wegler, U. and Sens-Schönfelder, C. (2007). Fault zone monitoring with passive image interferometry. *Geophysical Journal International*, 168(3):1029–1033.

Weichert, D., Horner, R. B., and Evans, S. G. (1994). Seismic signatures of landslides: The 1990 brenda mine collapse and the 1965 hope rockslides. *Bulletin of the Seismological Society of America*, 84(5):1523–1532.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., and Trujillo, J. (1998). A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, 88(1):95–106.

Zhou, L., Zhao, C., Zhang, M., Xu, L., Cui, R., Zhao, C., Duan, M., and Luo, J. (2021). Machine-learning-based earthquake locations reveal the seismogenesis of the 2020 mw 5.0 qiaojia, yunnan earthquake. *Geophysical Journal International*.

Zhu, L., Peng, Z., McClellan, J., Li, C., Yao, D., Li, Z., and Fang, L. (2019a). Deep learning for seismic phase detection and picking in the aftershock zone of 2008 mw7. 9 wenchuan earthquake. *Physics of the Earth and Planetary Interiors*, 293:106261.

Zhu, W., Mousavi, S. M., and Beroza, G. C. (2019b). Seismic signal denoising and decomposition using deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9476–9488.

Zhu, W., Mousavi, S. M., and Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks. *Advances in Geophysics*, 61:151–177.

Appendix A

In this appendix, we show the distribution of the test loss values from all 28,800 test samples for models with different bottlenecks.

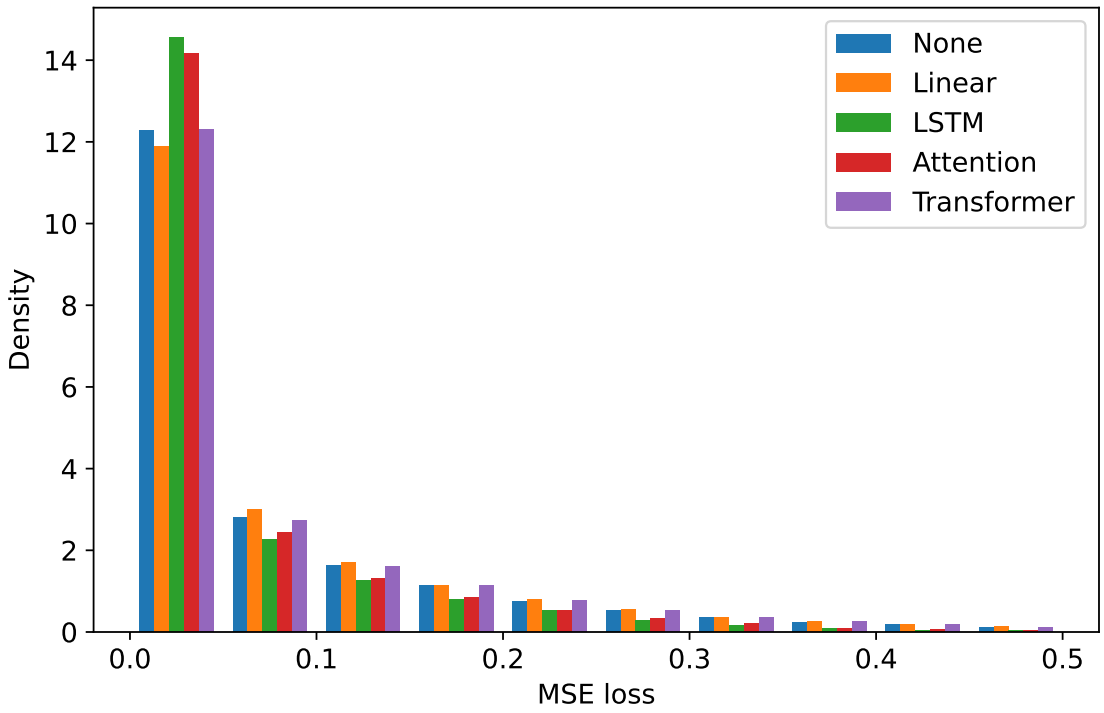

Figure A.1: Distribution of test loss values of the test data set (with 28,800 samples) from different models.

Appendix B

Because of the internal random steps in the training algorithm, the searching path for optimal model parameters may be slightly different, even for the same model initialization (Fig.B.1). Considering that, we fix the model initialization and train each model independently for 11 times, and keep the model with longest epoch as the final model.

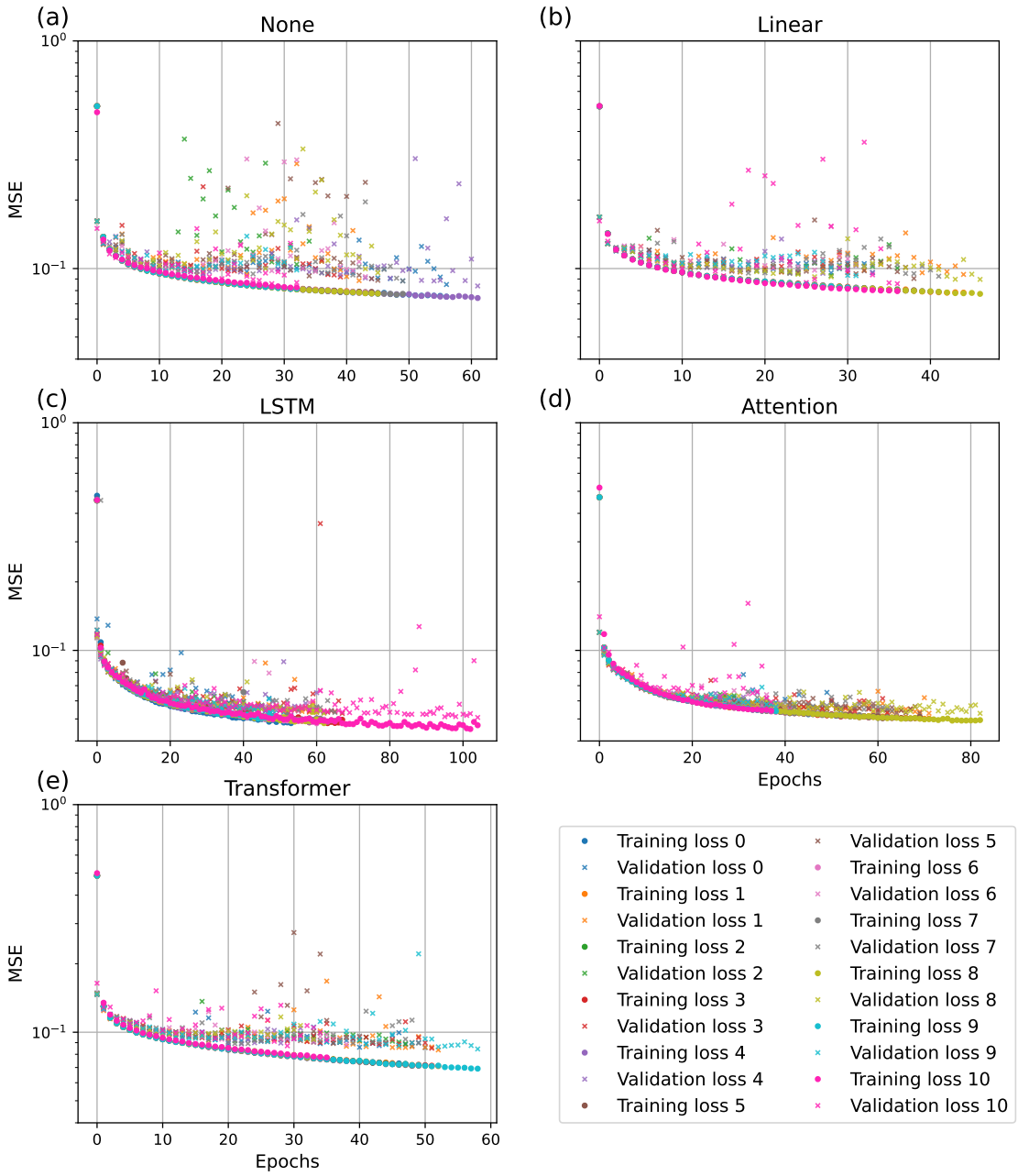

Figure B.1: Each model has been trained for 11 times with the same random seeds for model initialization and data split. The model with longest epoch has been chosen as the final model.