Local Magnitude Estimation via an Attention-Based Machine Learning Model

Ji Zhang[1], Aitaro Kato[*1], Huiyu Zhu[2,3], and Wei Wang[4,5]

Abstract

Rapid and reliable earthquake magnitude estimation is crucial for disaster management, scientific research, and resource conservation across multiple fields, especially during the initial stages of event detection. The most reliable traditional methods rely on complete waveform records, including earthquake epicenter distance and waveform amplitude, which can delay magnitude assessment. Machine learning techniques offer a promising avenue for capturing nonlinear relationships within seismic data, enhancing both information extraction and timeliness in magnitude estimation. In this study, we introduce an Attention-based machine learning model for MAGnitude estimation (AMAG) tailored for real-time earthquake monitoring. Using two independent datasets for training and testing, the results demonstrate the efficacy of our method in accurately predicting earthquake magnitudes. The magnitude prediction errors on the two test sets are −0.2 and −0.1, respectively, and the picking errors for both are 0.02 s. Our approach can be used directly for different time windows and signal lengths (at least 1 s) without retraining. We investigate the influence of signal-to-noise ratio, distances, and the integration of attention mechanisms. The attention mechanism facilitates the identification of the first motion and provides insights into the network's focus areas, thereby establishing a relationship between waveform characteristics and earthquake magnitude. In addition, we systematically explore the impact of network architectures, loss functions, and signal lengths on prediction performance. Our findings reveal that a network with a depth of four layers and a convolution kernel size of five yields optimal prediction accuracy, with mean square error identified as the most effective loss function. When the input waveform is six seconds long, with equal durations of noise and signal, the model's prediction accuracy is optimized. Our study underscores the potential of machine learning-based magnitude estimation for real-time earthquake monitoring, offering novel opportunities to mitigate natural disaster impacts, minimize casualties, and safeguard lives and property.

Cite this article as Zhang, J., A. Kato, H. Zhu, and W. Wang (2025). Local Magnitude Estimation via an Attention-Based Machine Learning Model, *Seismol. Res. Lett.* **96**, 2187–2200, doi: 10.1785/0220240289.

Supplemental Material

Introduction

Quantifying the size (magnitude) of earthquakes is crucial for scientific understanding and assessing their societal impact (Stein and Wysession, 2009). The amplitude of seismic waveforms reflects earthquake size once corrections are made for distance-related geometric spreading and attenuation. The Richter scale, introduced by Richter (1935) for earthquakes in southern California, measures local magnitude (M_L). However, local magnitudes are inadequate for global studies. Alternatives include body-wave magnitude (m_b), surface-wave magnitude (M_s), and moment tensor (M_w). Magnitudes offer two key advantages: they are directly measured from seismograms without complex signal processing, and they provide intuitively understandable units.

Until recently, utilizing just a few seconds of waveform data to estimate earthquake magnitude posed significant challenges. However, with advancements in artificial intelligence and machine learning techniques (LeCun *et al.*, 2015; Goodfellow *et al.*, 2016; He *et al.*, 2016), seismic data processing and analysis have seen widespread adoption. Techniques such as seismic denoising (Perol *et al.*, 2018; Zhu *et al.*, 2019; Wang and Zhang, 2023), earthquake detection, and phase picking (Li *et al.*, 2018; Ross *et al.*, 2018; Zhu and Beroza, 2018; Zhou *et al.*, 2019;

1. Earthquake Research Institute, University of Tokyo, Tokyo, Japan; 2. Institute of Engineering Mechanics, China Earthquake Administration, Harbin, China, https://orcid.org/0000-0003-4448-7351 (HZ); 3. Key Laboratory of Earthquake Engineering and Engineering Vibration, China Earthquake Administration, Harbin, China; 4. Key Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, https://orcid.org/0000-0002-5934-4268 (WW); 5. College of Earth and Planetary Sciences, University of Chinese Academy of Sciences; Beijing, China

*Corresponding author: akato@eri.u-tokyo.ac.jp

Mousavi *et al.*, 2020; Yang *et al.*, 2020; Xiao *et al.*, 2021), first-motion polarity determination (Ross *et al.*, 2018; Zhang *et al.*, 2023), and location (Zhang *et al.*, 2020; Wang *et al.*, 2024) have become common.

Researchers are exploring various approaches to assess earthquake magnitude using limited waveform data. Some approaches involve employing classification methods to categorize earthquake magnitudes, whereas another employs regression methods to estimate magnitudes. For instance, Lomax *et al.* (2019) utilized a convolutional neural network (CNN) for rapid earthquake parameter assessment, classifying earthquake magnitudes into 20 categories. Mousavi and Beroza (2020) utilized CNN and recurrent neural network (RNN) to predict local magnitudes with high accuracy based on single-station waveforms, even without instrument response correction. Some aim to evaluate earthquake magnitude through full-waveform analysis (e.g., Lomax *et al.*, 2019; Mousavi and Beroza, 2020; van den Ende and Ampuero, 2020; Saad *et al.*, 2022). Saad *et al.* (2022) employed a vision transformer network to assess earthquake magnitude using 30 s waveforms, with the real-time assessment based on 4 s waveforms necessitating network reset and retraining. Similarly, van den Ende and Ampuero (2020) utilized graph neural networks for magnitude assessment across multiple stations, leveraging station information but still relying on full-waveform data. Although full-waveform analysis ensures comprehensive evaluation, it often sacrifices timeliness in real-time earthquake monitoring.

In addition, attempts have been made to achieve real-time earthquake magnitude assessment using multistation triggered methods (e.g., Kuang *et al.*, 2021; Münchmeyer *et al.*, 2021). Kuang *et al.* (2021) presented a magnitude neural network (MagNet) based on full-waveform recordings from a network of seismic stations. This method can effectively minimize the influence of abnormal noise in the data. Münchmeyer *et al.* (2021) introduced a novel model for real-time magnitude and location estimation using attention-based transformer networks, incorporating waveforms from dynamically varying station sets and outperforming deep learning baselines in both magnitude and location estimation performance. This approach holds significant potential for a range of applications, from routine earthquake monitoring to early warning systems.

The magnitude is predicted within the first few seconds after the *P*-wave arrival. Chakraborty *et al.* (2022) used CNN and RNN to evaluate earthquake magnitude based on 512-sample-point waveforms containing 1–2 s *P*-wave signals, offering promise for earthquake early warning (EEW) systems. If the input size is not 512 or the effective *P*-wave window length is not within 1–2 s, this method will not work, which greatly reduces the applicability of this method. Wang *et al.* (2022) propose the EEWNet that can predict magnitude between 4.0 and 5.9 as early as the first 0.5 s *P* wave arrives. Hou *et al.* (2024) designed a deep-learning, multiple-seismometer-based magnitude estimation method using three heterogeneous multimodalities with a specific transformer architecture to achieve the magnitude estimation. This model performs less error level than the *Pd* approach on magnitude estimation, especially on the high magnitude.

In this study, we introduce an attention-based neural network (Vaswani *et al.*, 2017; Mousavi *et al.*, 2020; Niu *et al.*, 2021; Zhang *et al.*, 2023) for real-time earthquake local magnitude (M_L) estimation (AMAG). The utilization of CNNs for feature extraction and attention mechanisms for feature focus and visualization has been widely documented across various seismic processing applications (Mousavi *et al.*, 2020; Zhang *et al.*, 2023). We adopt the U-net (Ronneberger *et al.*, 2015) architecture combined with long short-term memory (LSTM) networks to assess the magnitude of waveforms at each moment. Our focus lies in achieving real-time magnitude estimation of a single station, with the capability to evaluate magnitude using *P*-wave waveforms after just over 1 s. We use single-station data for magnitude analysis for two main reasons. First, in most cases where dense seismic arrays are unavailable, single-station analysis ensures real-time processing and efficient handling of real-time data streams. Second, incorporating station locations for multistation data streams remains a significant challenge. Our well-trained model is adaptable to various time windows without requiring transfer learning or retraining. We investigate model performance across different architectures and loss functions. We emphasize that network depth should be carefully balanced. Excessive depth can lead to learning unnecessary details, whereas insufficient depth may hinder the model's ability to capture the data-label relationship effectively. A suitable loss function aids in fast model convergence, while proper regularization terms help constrain the model. We determine optimal parameters for our model, including a convolution kernel size of 5, a network depth of 4, and the mean square error (MSE) loss function. Many existing studies utilize fixed time windows, offering advantages in model prediction improvement and ease of training. However, the model with a fixed time window is less compatible because other users who want to use it need to set the same parameters. We explore the impact of different signal and noise lengths on model predictions under fixed and variable time-window conditions. The attention mechanism facilitates the identification of data features in regions receiving more focus, thereby enhancing model performance. Furthermore, we demonstrate that attention mechanism results have implications for deeper cognitive models. Our method outperforms other approaches in terms of phase picking, event classification, and magnitude estimation, exhibiting superior pickup and magnitude prediction effects, alongside unique adaptability across different regions and data sizes.

Data

We employ the STanford EArthquake Dataset (STEAD) by Mousavi *et al.* (2019) for both training and testing our model. In addition, we extend the model evaluation by testing it on the

High Sensitivity Seismograph Network Japan (Hi-net) data, is a nationwide seismic observation network operated by the National Research Institute for Earth Science and Disaster Resilience (National Research Institute for Earth Science and Disaster Resilience [NIED], 2019). This approach allows for comprehensive validation of the model's performance across different datasets, displaying robustness and generalizability.

STEAD

The STEAD (Mousavi *et al.*, 2019) is a comprehensive repository comprising 1.0 million local seismic records and 0.2 million noise records, making it a valuable resource for global-scale seismic research. It encompasses essential waveform data captured by seismic instruments, along with detailed instrument information. For earthquake data, in addition to station details, STEAD provides comprehensive information such as origin time, epicentral location, depth, magnitude, magnitude type (including M_L, m_b, M_s, and so on), focal mechanism, and arrival times of P and S phases. Moreover, recorded signal attributes such as signal-to-noise ratio (SNR) for each component, coda-end time, and epicentral distance are also included. Each seismic record consists of three waveforms, with each waveform comprising 6000 samples representing 60 s of ground motion (counts) recorded by high-gain velocity seismometers in the east–west, north–south, and vertical directions, respectively. Among the magnitude types provided in the STEAD dataset, M_L magnitude accounts for 69.8% of the data and is primarily selected for analysis. Most seismograms in the dataset were recorded within 110 km of earthquakes, with SNRs typically ranging between 10 and 40 decibels (dB). For analysis purposes, seismograms with SNRs exceeding 10 dB are selected. Preprocessing steps, including detrending, tapering, demeaning, and band-pass filtering (1–20 Hz), are applied to enhance data quality and consistency. After applying the selection criteria, a total of 64,811 traces are retained. These traces are further divided into training, validation, and testing datasets in an 8:1:1 ratio, facilitating robust model development and evaluation for seismic analysis tasks. Figure S1a, available in the supplemental material to this article, shows the magnitude distribution of STEAD data used for the model.

Hi-net

We download the new testing data for the year 2020 from the Hi-net website. Following the approach outlined by the STEAD method, this dataset comprises ~128,000 three-component waveform traces, covering over 13,500 earthquakes as shown in Figure S1b. The magnitude of events in this dataset ranges from 0 to 5.7 (Fig. S1c), providing a diverse range of seismic activity for study. The data have a time sampling rate of 100 Hz and include comprehensive attributes such as earthquake catalog information and temporal evolution of data attributes, including channel types, SNRs, and phase picks. These attributes document the progression of the network earthquake monitoring system over time. For analysis purposes, only traces with magnitudes provided in the M_L scale are selected, and traces with SNR lower than 10 dB are excluded. ~10,000 waveforms are randomly selected from the dataset for testing purposes. Preprocessing steps for this dataset closely resemble those applied to the STEAD data, including detrending, tapering, demeaning, and band-pass filtering, ensuring data consistency and quality across both datasets. These preprocessing steps lay the foundation for robust model testing and evaluation on the Hi-net dataset. In addition, we utilize 10,000 noise data to test the classification of our model.

Method

Figure 1 illustrates the network architecture of the AMAG model, composed of four key components: encoder, decoder, LSTM, and attention mechanism module.

1. Encoder: The encoder of the AMAG network is responsible for capturing essential features from the input waveform data. It preprocesses the input data and extracts relevant information that will be used for subsequent processing. The encoder consists of convolutional blocks (EBlock) that include convolution (conv2d), batch-normalization (BN, Ioffe and Szegedy, 2015), and LeakyReLu (Maas *et al.*, 2013) layers,

$$\mathrm{EBlock}(x) = \mathrm{CB2}(\mathrm{CB1}(x)). \tag{1}$$

Sub-block (CB1) is utilized to extract waveform features, with the same padding method in convolution layers ensuring consistency between input and output sequences, as shown in function (2),

$$\mathrm{CB1}(x) = \mathrm{LeakyReLU}(\mathrm{BN}(\mathrm{conv2d}(x,\mathrm{stride} = 1))). \tag{2}$$

Unlike traditional CNN models that employ max-pooling for downsampling, the AMAG decoder achieves downsampling (CB2) by adjusting the step size of the convolutional layers, as shown in function (3). This approach, referred to as special step size convolution, enables feature extraction and dimensionality reduction,

$$\mathrm{CB2}(x) = \mathrm{LeakyReLU}(\mathrm{BN}(\mathrm{conv2d}(x,\mathrm{stride} = 2))). \tag{3}$$

The depth of the encoder (d) is set to 4, determined through testing to achieve optimal performance. In addition, a convolution kernel size of 5 is chosen to capture sufficient spatial information. The number of channels in the encoder increases exponentially with depth, following the formula:

$$n\mathrm{channel} = 8 \times 2^{(d-1)}. \tag{4}$$

The encoder obtains a special code representing the waveform.

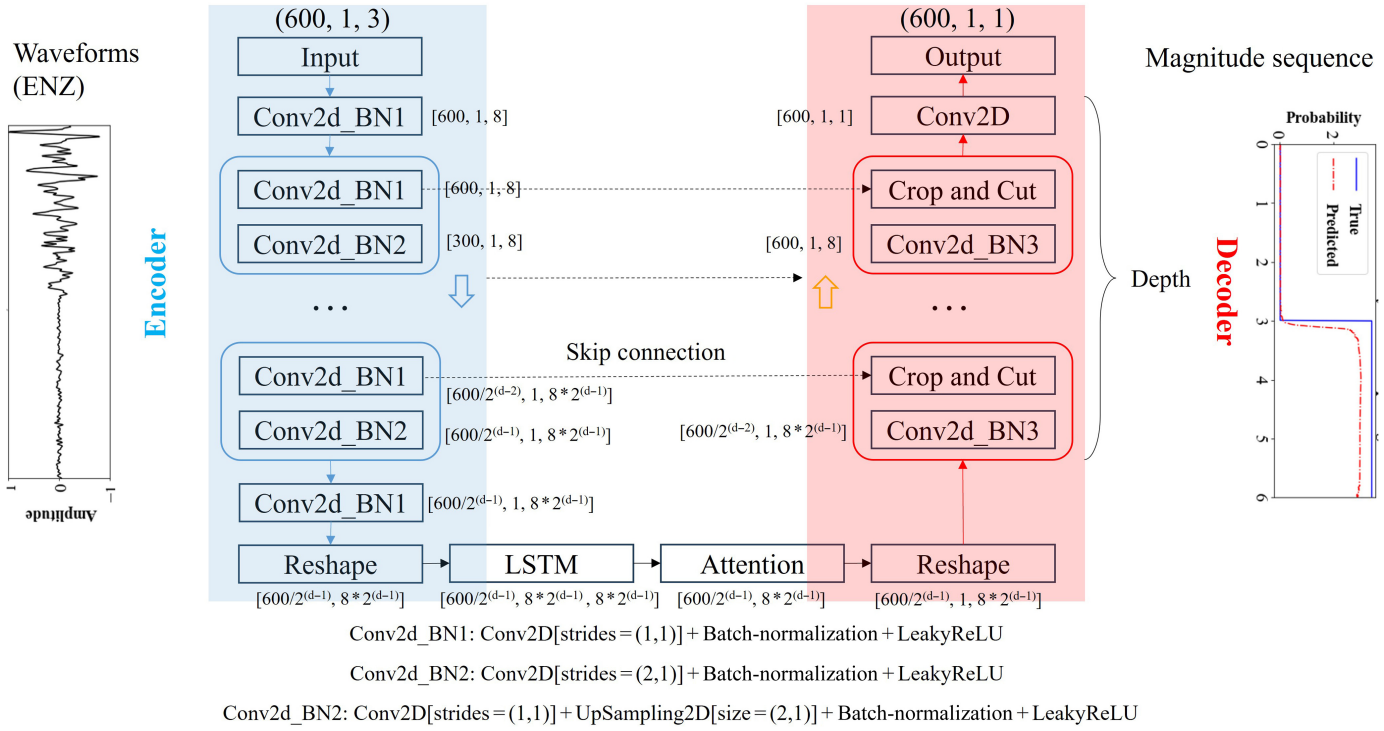

Figure 1. The AMAG model architecture. The AMAG model is based on U-net, which consists of the encoder, long short-term memory (LSTM), attention, and decoder. The input is three-component waveforms (6 s) and the output is the magnitude sequence. The color version of this figure is available only in the electronic edition.

2. LSTM: The seismic data are a time series, and we use the LSTM (Hochreiter and Schmidhuber, 1997) to further acquire its time-series features. First, we need to transform the extracted features into the time dimension using the Reshape function. The LSTM module is integrated into the network to model temporal dependencies and capture long-range dependencies in the input waveform data. LSTM units enable the network to retain and utilize information over extended time periods, facilitating accurate prediction of earthquake magnitude. The encoder generates a special code representing the waveform, and we further analyze its time-series features using LSTM. By setting "return_sequences" for LSTM to true, we ensure that the output dimension aligns with the number of convolution cores at the last layer of the encoder.
3. Attention mechanism module: the attention mechanism module enhances the network's capability to focus on relevant features and regions within the input waveform data. By dynamically allocating attention weights to different parts of the input sequence, the attention mechanism improves the model's performance in capturing critical information for earthquake magnitude estimation. We use the attention mechanism to analyze the temporal features of acquired time. Through cross-correlation operations within the data, the attention mechanism can obtain a weight matrix that represents which parts of the data have the most influence on the outcome and are given greater weight (Zheng *et al.*, 2018),

$$a^{t,t'} = \mathrm{softmax}(\mathrm{sig}(W^a(\tan h(W^t h^t + W^{t'} h^{t'} + b^h)) + b^a)), \tag{5}$$

$$O^t = \sum_{t'=1}^{n} a^{t,t'} \cdot h^{t'}. \tag{6}$$

Formula (5) details the calculation of the attention matrix $a^{t,t'}$, in which h^t and $h^{t'}$ are hidden state representations at timesteps t and t', respectively; W and b are weight matrices and bias vectors, respectively; "sig" denotes the elementwise sigmoid function, softmax function (Goodfellow *et al.*, 2016) ensures that the scores are converted into a probability distribution, where the sum of all attention scores equals 1. This step highlights the most relevant parts of the input data by assigning higher weights to more critical sections. The output of the attention layer, O^t, at timestep t is given by the summation of hidden states at all other steps, $h^{t'}$, weighted by their similarities the current hidden state, $a^{t,t'}$, by dot product as shown in function (6). We can see the impact on the results by looking at the attention map ($a^{t,t'}$). After the output of the attention mechanism, it needs to be reshaped into the encoded dimension.

4. Decoder: The decoder block (DBlock) plays a crucial role in reconstructing the output sequence based on the features learned by the encoder, LSTM, and attention module, as shown in function (7). The operation (CCut) of crop and

cut connects to the encoder part. Sub-blocks (CB3) involve building blocks using convolution layers, upsampling2d, batch-normalization layers, and leakyReLu,

$$\mathrm{DBlock}(x) = \mathrm{CCut}(\mathrm{CB3}(x)), \tag{7}$$

$$\mathrm{CB3}(x) = \mathrm{LeakyReLU}(\mathrm{Upsampling2D}(\mathrm{conv2d}(x,\mathrm{stride} = 1),\mathrm{size} = (2,1))). \tag{8}$$

The number of convolution kernels in each layer of the decoder is the same as the number of convolution kernels in the corresponding encoder. Finally, we integrate the decoder results using a convolutional layer with one channel.

Overall, the AMAG network leverages a carefully designed architecture comprising an encoder, decoder, LSTM, and attention mechanism modules to effectively process seismic waveform data and accurately predict earthquake magnitudes as shown in function (9),

$$\mathrm{AMAG}(x) = \mathrm{Conv2D}([\mathrm{DBlock}(\mathrm{Re}(\mathrm{Atten}(\mathrm{LSTM}(\mathrm{Re}([\mathrm{Eblock}(x)]_d)))))]_d). \tag{9}$$

MSE is utilized as the loss function, and training involves iterating through 100 generations with early stopping (Prechelt, 1998; Raskutti *et al.*, 2014) implemented to prevent overfitting. Specifically, if the validation loss does not decrease in 10 consecutive generations, the iteration is halted. We use the Adam optimizer with an initial learning rate set to 0.001 (Kingma and Ba, 2014),

$$\mathrm{MSE} = \sum_{i=0}^{n} (y^{\mathrm{true}} - y^{\mathrm{pred}})^2. \tag{10}$$

The input for the model is 6 s of three-component seismic data (600, 1, 3), which contains the *P*-wave signal longer than 1 s in length. (600, 1, 3) represents 600 sample points in time, with data from one station and three components. The output is a magnitude sequence (600, 1, 1). Unlike the labeling approach in the CREIME method by Chakraborty *et al.* (2022), where noise is assigned a label of −4, we adopt a different strategy to ensure accurate magnitude prediction. We set the noise label to 0 to prevent underestimation of magnitude predictions. To distinguish between signal and noise labels, we add 1 to the magnitude of the signal, effectively designating it as the signal label. This adjustment helps maintain clarity and accuracy in identifying signal instances within our magnitude prediction framework,

$$\mathrm{label}(t) = \begin{cases} 0, & t < t_p \\ \mathrm{mag} + 1, & t \geq t_p \end{cases}, \tag{11}$$

in which t_p is the *P* arrival time.

By structuring labels in this way, we can not only predict the earthquake magnitude but also assess the seismic phase. In the magnitude sequence label, noise is designated as 0, whereas the signal is labeled with the corresponding magnitude. This labeling strategy results in a distinct transition at the boundary between noise and signal, which can be leveraged for seismic phase picking.

Results

We present some results predicted by the model on the STEAD testing set (Fig. 2). From top to bottom, the Z-component waveform, the magnitude prediction results with true labels (blue solid lines) and prediction labels (red dashed lines), and attention maps in Figure 2. We display the noise length, signal length, and window (input) length in Figure 2b. Figure 2a–c illustrates that, regardless of the length of the signal input, the model can make reasonable predictions. We also display examples of underestimation (Fig. 2d,e) and overestimation (Fig. 2f) of magnitude estimation. Furthermore, these results demonstrate that the attention map has high weights around *P* arrival to effectively identify it, regardless of its position within the signal.

Figure 3 shows the results of the well-trained model on the STEAD testing set. We display the relationship between true magnitudes and predicted magnitudes in Figure 3a. The *x*-axis represents the actual earthquake labels, whereas the *y*-axis represents the predicted magnitudes. The color indicates relative density, with warmer colors representing higher density areas. Ideally, data points should align along the diagonal, indicating accurate predictions. Deviations from the diagonal suggest bias in the predictions; points above the diagonal indicate overestimation, whereas points below indicate underestimation. From the relative density distribution, it is evident that the overall distribution follows a diagonal pattern, although it consistently falls near the central diagonal. It is noteworthy that if the predicted magnitude is less than −0.5, we classify the seismic waveform as noise. This approach allows the model to serve as a classifier for distinguishing between earthquakes and noise. We calculate the distribution of magnitude prediction errors, as shown in Figure 3b. The overall distribution is centered around the negative axis, with a mean error of −0.1. The mean absolute error (MAE) is calculated as 0.4, and the variance is 0.5. These results are based on predictions from seismic data recorded by a single seismic station within just a few seconds. We also fit a linear relationship between the peak displacement amplitude of the *P* wave in the 3 s (*Pd*) and the magnitude values. The fitted empirical relationship yields an MSE of 0.5 and a variance of 0.7. The network's prediction performance surpasses the empirical method based on *Pd*.

As described in the Method section, we leverage the magnitude prediction sequence to facilitate seismic phase picking. Utilizing waveforms amenable to magnitude assessment, we can estimate phase picking by identifying the sharp step in

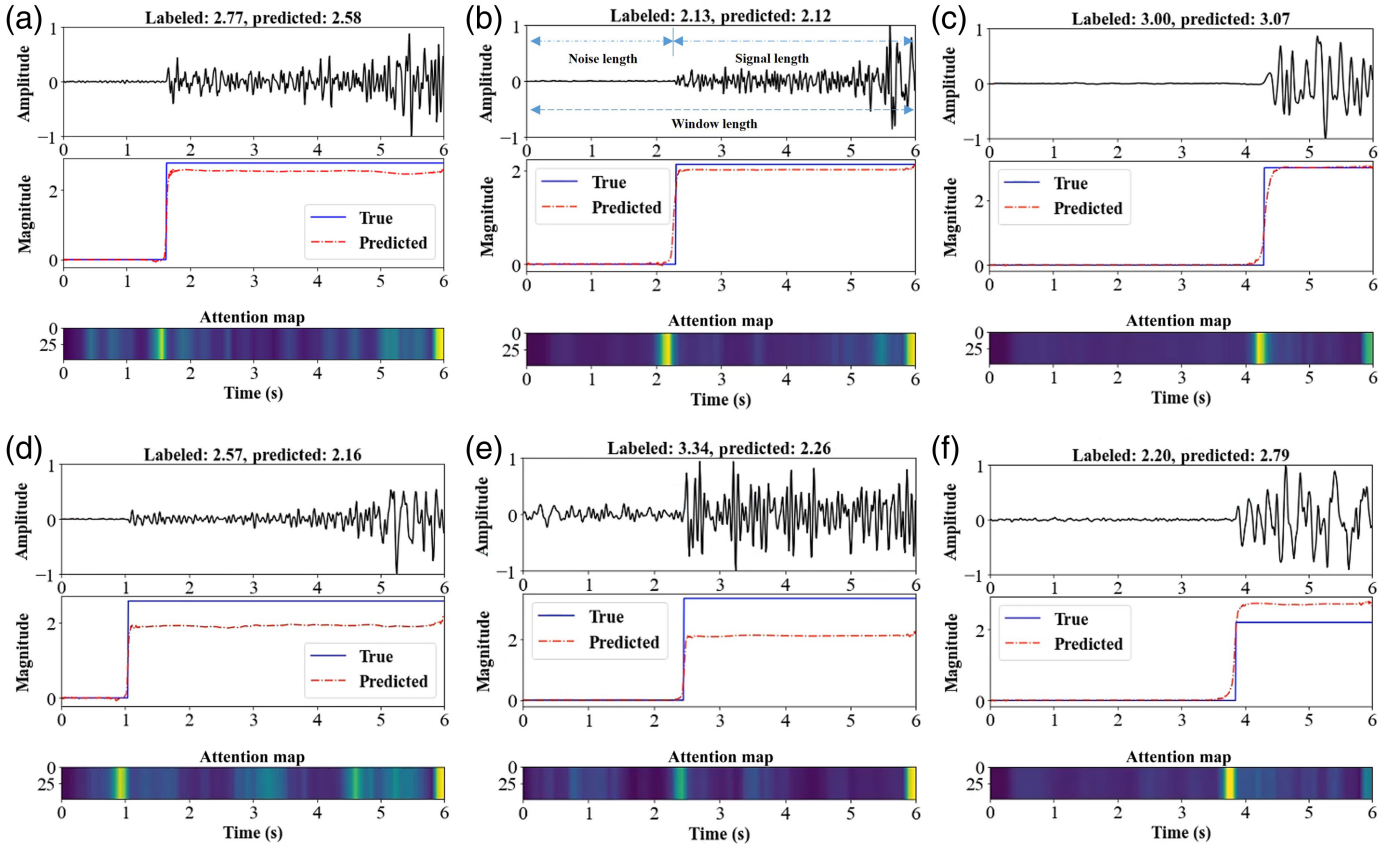

Figure 2. Some predicted results. From top to bottom, the Z-component waveform, the magnitude prediction results with true labels (blue solid lines) and prediction labels (red dashed lines), and attention maps, where warmer colors indicate higher weights. We show the noise length, signal length, and window length in panel (b). The title presents the true and predicted magnitude. (a–c) Some good instances. (d–f) Some not good instances of underestimation or overestimation. The color version of this figure is available only in the electronic edition.

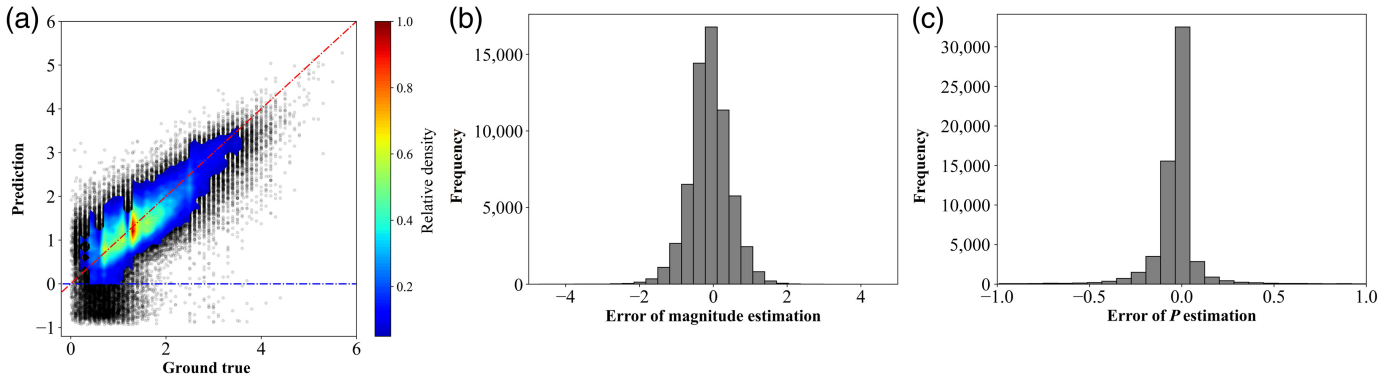

Figure 3. Model performance in Stanford Earthquake Dataset (STEAD). (a) The true magnitudes and predicted magnitudes. The color indicates relative density, with warmer colors representing higher density areas. (b) The error of magnitude estimation. (c) The error of phase picking. The color version of this figure is available only in the electronic edition.

the predicted labels, as shown in Figure 3c. We observe that the phase picking error is primarily concentrated around 0, with a mean error of 0.06 s, an MAE of 0.18 s, and a standard deviation (STD) of 0.59 s. These findings demonstrate that our method also effectively and accurately picks up onsets of seismic phases. In addition, the model predicts 10,000 waveforms in 19 s using one central processing unit, equating to less than 2 ms per waveform, making it well-suited for real-time applications.

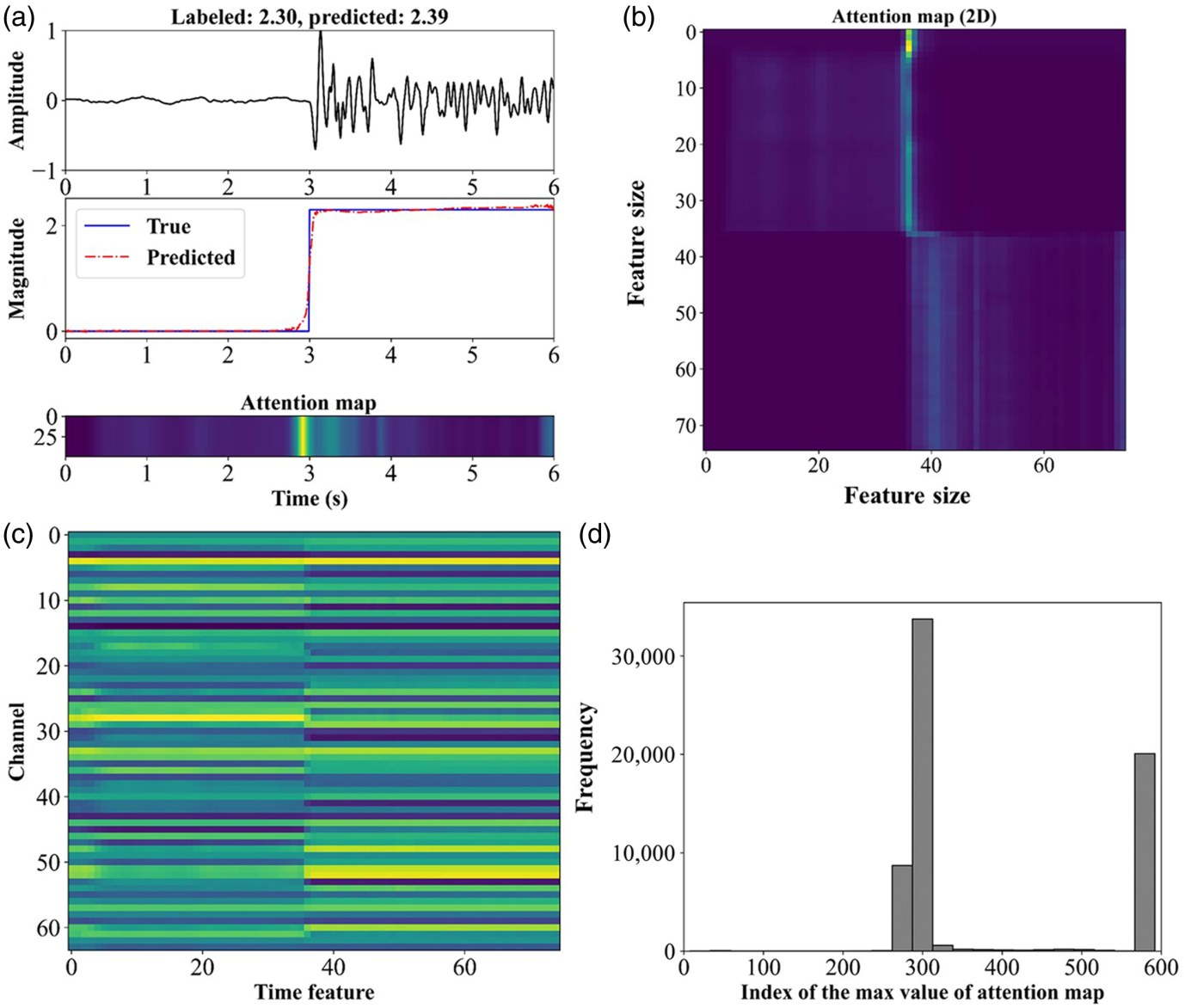

Figure 4. Attention mechanism. (a) One predicted example. From top to bottom, the Z-component waveform, the magnitude prediction results with true labels (blue solid lines) and prediction labels (red dashed lines), and attention maps, where warmer colors indicate higher weights. (b) 2D attention map, weight matrix of the attention layer. (c) The output of the attention layer. The *x* and *y* axes represent the time feature and the number of channels, respectively. This result is obtained by multiplying the features extracted by the LSTM with the attention weight matrix shown in panel (b). (d) The distribution of the indexes of the max values of attention maps. The *x* and *y* axes represent the index and frequency of the maximum values of the attention matrix, respectively. The color version of this figure is available only in the electronic edition.

Discussion

Attention maps

The weight of the attention mechanism can help us explore the location of the information that the network pays attention to and help us understand the network. In Figure 2, we can also observe that the attention mechanism effectively focuses on *P* arrivals. Here, we display the attention matrix and the output of the attention layer to explore the function of attention mechanisms, as shown in Figure 4.

Figure 4a displays the seismic waveform and prediction results alongside the attention mechanism diagram. Figure 4b illustrates the original 2D diagram of the attention mechanism, and Figure 4c demonstrates the result of 64 channels output through the attention mechanism. The attention weight concentrates near abrupt changes in the data, typically where the *P* wave begins. The output of attention reveals a clear interface line that separates noise and signal components, facilitating the extraction of distinct features from these two parts. To better confirm that the attention weights are concentrated near the *P* wave, we fix the input signal and noise durations to 3 s each (300 samplings). We calculate the position of the maximum value in the attention weight matrix, as shown in Figure 4d. The results show that

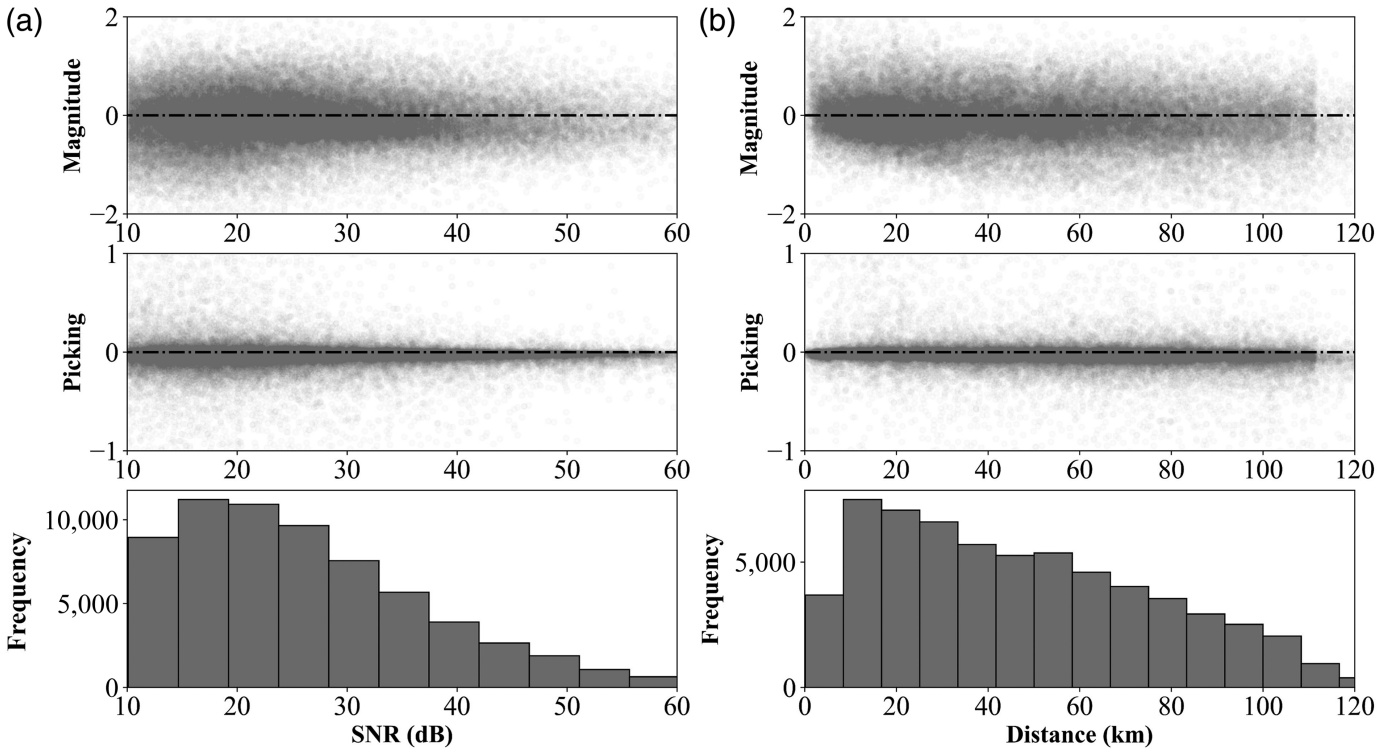

Figure 5. Prediction errors of magnitude and phase picking as a function of (a) signal-to-noise ratios (SNRs) and (b) distances.

the largest weights are mainly concentrated around the index 300, further indicating that the attention mechanism focuses on *P*-wave information. For weak *P*-wave arrivals (Fig. S2), the attention mechanism still performs effectively. Notably, the maximum attention weight often appears at the front position of the *P* wave, indicating the region where the signal experiences drastic changes. However, generating a large weight of attention requires a signal change. Without it, this phenomenon cannot occur. It is worth considering whether anomalies before the signal can indeed be mined from the data itself. In addition to concentrating the energy of the attention mechanism near the *P* wave, it also concentrates at the end of the signal (e.g., Fig. 2a,b,d, e) because there are also large signal changes here. Although the attention map shows a significant weight at the end of the signal, it has minimal impact on the picking process when combined with the *P*-wave picking results (Fig. 3c). The weight remains focused around the vicinity of the *P* wave, ensuring accurate picking.

SNR and distance

Figure 5 illustrates the errors in magnitude prediction and phase picking across different SNRs and epicentral distances. Because the SNR increases, the magnitude prediction becomes more stable, although a certain degree of underestimation is observed. Generally, higher SNRs would be expected to yield more accurate predictions; however, the scarcity of high SNR data contributes to this underestimation phenomenon. Furthermore, high SNR data are often associated with major earthquakes, exacerbating the impact of underestimation. The scarcity of major earthquake data in the training set leads to an underestimation in their predictions. For machine learning regression models, there is a tendency to minimize errors by focusing on the dominant patterns in the training data that include many small-magnitude events. Because the SNR increases, *P*-wave picking becomes more stable, aligning with our expectations. Regarding epicentral distance, the prediction error of magnitude increases slowly with greater distances. This is partly due to the insufficient data at large epicentral distances and partly because the event duration increases with distance, making prediction more challenging. Similarly, the error in *P*-wave picking does not significantly change with increasing epicentral distance, indicating that the SNR distribution is uniform across different epicentral distances. Through analysis, the effects of SNR and epicentral distance on the network model are not significant, as we initially selected data with an SNR greater than 10 dB. The large-magnitude prediction errors (Fig. 3b) are mainly concentrated in low SNR data (Fig. 5a) and large epicentral distances (Fig. 5b). The primary reason is the lack of prominent features of *P*-wave arrivals, such as significant signal attenuation due to large epicentral distances or waveforms from stations located in nodal regions.

Time windows and signal length

In many network models, there is a fixed time window including a fixed signal length. For example, the ViT network (Saad *et al.*, 2022) uses a 30 s time window and a 20 s signal, whereas

TABLE 1
Model Performance on Different Time Windows

	Magnitude			*P* Arrival		
Time Windows	**Mean**	**MAE**	**STD**	**Mean**	**MAE**	**STD**
5 s	−0.24	0.49	0.60	−0.01	0.23	0.28
4 s	−0.25	0.51	0.61	−0.02	0.11	0.24
3 s	−0.32	0.54	0.62	0.05	0.11	0.23

MAE, mean absolute error; STD, standard deviation.

TABLE 2
Model Performance Under Different Architectures

	Depth					
Error	**7**	**6**	**5**	**4**	**3**	**Mean**
Kernel size						
7	1.156	1.174	1.142	1.184	1.216	1.174
5	1.164	1.513	1.365	**1.099**	1.312	1.290
3	1.236	1.555	1.125	1.527	1.307	1.350
Mean	1.185	1.414	1.210	1.270	1.278	/

The bold value represents the optimal solution among all structure tests.

the CREIME network (Chakraborty *et al.*, 2022) employs a 5.12 s time window and 1–2 s signal length. This fixed data configuration offers advantages such as easier training and more focused learning. However, it suffers from poor adaptability and requires additional preprocessing methods.

Our test results (Fig. 2) demonstrate that our model can handle different input signal lengths with a 6 s fixed time window. We tested our waveforms under various time windows without retraining the model. Specifically, we applied the model directly to time windows of 5, 4, and 3 s, each with a fixed 1 s signal input. Figure S3 displays some examples of these situations and Figure S4 displays the predicted results (Table 1) of magnitude and phase picking. In all three cases, the prediction results for magnitude and phase are quite similar, demonstrating that accurate predictions for both can be achieved. For the prediction of earthquake magnitude, MAE is ∼0.5, with a variance of around 0.6. The MAE for phase picking is 0.11 s, and the variance is 0.23 s. When the window is set to 3 s, the prediction error increases slightly, indicating that less data input leads to larger errors. More data are needed to build relationships, and noise is an important part of that. The results indicate that our method is adaptable to different time windows with 1 s signal input and can produce satisfactory results. Many models rely on fixed input sizes, and if the input does not match this size, the model will fail. In real-time earthquake detection, seismic data are a continuous stream of real-time data. We hope to leverage as much useful information as possible. As real-time seismic data arrive, our model can also provide results in real time. Such flexible input allows for better compatibility with other programs, making it more convenient for researchers to use. It also offers the potential for dynamically detecting changes in magnitude.

Architectures and loss function

Selecting appropriate network parameters is crucial yet challenging. In this study, we investigate the impact of network parameter selection within a certain range and determine the optimal model parameters based on this analysis. The choice of convolution kernel size typically ranges from 3 to 7, reflecting common practice in feature extraction. Larger convolution kernels capture broader information but require more parameters, whereas smaller kernels offer a narrower receptive field. Increasing the number of convolution layers can expand the receptive field and introduce more nonlinear relations to fit the data. Deeper networks accommodate more parameters and can handle complex problems, but they also escalate computation and training time. Conversely, shallow networks may fail to capture intricate data relationships. Given that downsampling occurs during the encoder process, excessively deep networks must be avoided to preserve timing characteristics. Hence, we consider a range of 3–7 for both encoder and decoder depth.

To streamline testing efficiency, we randomly select 10,000 data points from the STEAD training dataset for model training and another 10,000 from the test set for evaluation. Table 2 presents model performance under different network parameters. Generally, larger convolution kernels yield better prediction results, whereas deeper networks tend to produce superior outcomes. Optimal model performance is achieved with a convolution kernel size of 5 and a depth of 4. This represents a strategic compromise, leveraging a medium-size convolution kernel and moderate network depth.

The primary function of a loss function is to guide the optimization of the model. It quantifies the disparities between model predictions and actual data, facilitating the evaluation and refinement of the model's learning process. By minimizing the loss value, the machine learning model iteratively adjusts its parameters during training to better capture and understand the underlying patterns and relationships within the data. Two common loss functions used are MSE and MAE as shown in functions (10) and (12). MSE yields a smooth, continuous, and differentiable function, making it convenient for use with gradient descent algorithms. It is particularly effective in promoting convergence, as the gradient decreases with decreasing error, facilitating faster convergence to the minimum. Notably, MSE amplifies errors greater than 1 and reduces errors less than 1, thereby exhibiting sensitivity to outliers and their significant impact on the loss. In contrast, MAE offers resilience against outliers as it computes the absolute value of the error, resulting in a fixed penalty for differences of any size. This stability is advantageous, as it prevents the occurrence of gradient explosion problems and ensures a more robust solution. However,

TABLE 3
Model Performance Under Different Loss Functions

Loss	+/	+W	+STD
MSE	**1.099**	1.246	1.340
MAE	1.632	1.120	1.617
MAE + MSE	1.191	1.206	1.250

Note: +/: no other constraints. +W: add different weights according to the magnitude. +STD: add regularization of standard deviation between label and prediction. MAE, mean absolute error; MSE, mean square error. The bold value represents the optimal solution among all loss function tests.

the discontinuity of MAE at $y - f(x) = 0$ and its uniform gradient distribution across most cases can impede function convergence and model learning (Ng, 2004),

$$\mathrm{MAE} = \sum_{i=0}^{n} |y^{\mathrm{true}} - y^{\mathrm{pred}}|. \tag{12}$$

For earthquake magnitude assessment, we experiment with MSE, MAE, and their combinations (CME = MSE + MAE). Considering that large earthquakes occur less frequently than small earthquakes, we assign higher weights to different earthquake magnitudes to address this imbalance. Different weights are assigned to different magnitudes, as demonstrated in Table S1. Specifically, larger magnitudes correspond to smaller data and larger weights. We propose the weighted MSE (WMSE) and weighted MAE (WMAE), as shown in function (13),

$$\begin{aligned} \mathrm{WMSE} &= \sum_{i=0}^{n} W^i (y^{\mathrm{true}} - y^{\mathrm{pred}})^2 \\ \mathrm{WMAE} &= \sum_{i=0}^{n} W^i |y^{\mathrm{true}} - y^{\mathrm{pred}}|, \end{aligned} \tag{13}$$

$$\begin{aligned} \mathrm{STDE} = &\sum_{i=0}^{n} \mathrm{std}(y_i^{\mathrm{truenoise}} - y_i^{\mathrm{prednoise}}) \\ &+ \mathrm{std}(y_i^{\mathrm{truesignal}} - y_i^{\mathrm{predsignal}}). \end{aligned} \tag{14}$$

Furthermore, we consider the stability of predictions for both noise and signal components, calculating the variance of labels for each part to impose constraints, as shown in function (14).

Ideally, integrating appropriate constraints can enhance the final prediction accuracy. Our tests demonstrate that incorporating various constraints improves prediction accuracy when using MAE as the loss function, as detailed in Table 3. Conversely, for the MSE loss function, which is already well suited to constrain seismic evaluation problems, additional constraints do not notably enhance prediction accuracy.

Model tuning is an important and complex task. In addition, we used the Optuna method (Akiba *et al.*, 2019) to explore the relationship between loss functions, network depths, and kernel sizes, as shown in Figure S5. In conclusion, our optimal model employs network architecture with a convolution kernel size of 5, and a network depth of 4, and utilizes MSE as the loss function. This configuration yields satisfactory performance in earthquake magnitude estimation.

Signal and noise length

We are also curious about the effect of different signal lengths on the training results. To investigate the impact of signal and noise lengths within fixed time windows, we maintain a 6 s time window while adjusting the relative lengths of signal and noise to train the model. To streamline testing efficiency, we randomly select 10,000 data points from the STEAD training dataset for model training and another 10,000 from the test set for evaluation. Specifically, we vary the signal and noise lengths as follows: (5,1), (4,2), (3,3), (2,4), and (1,5). The effects of these variations on final prediction results are summarized in Figure 6a.

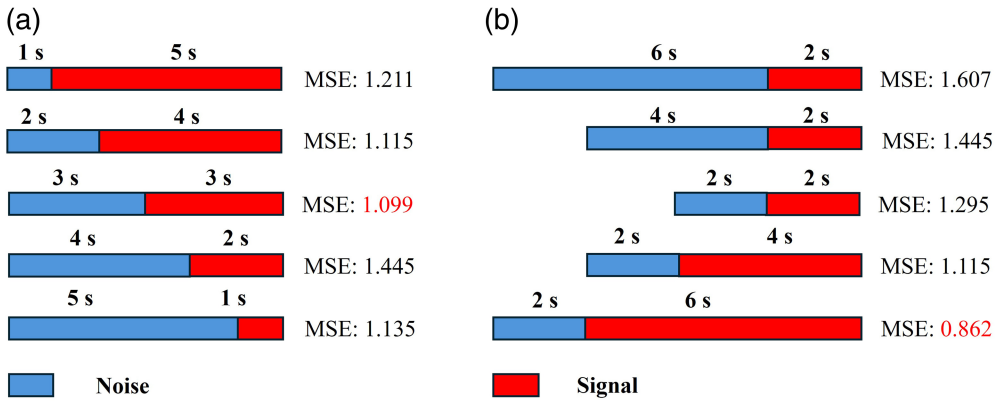

Figure 6. The effect of different signal and noise lengths. (a) Fixed time windows, different signal and noise lengths. (b) Unfixed time windows, with fixed signal or noise lengths. The color version of this figure is available only in the electronic edition.

Given a fixed time window, optimal prediction performance is achieved when the signal and noise lengths are equal. In our network model, besides predicting earthquake magnitudes, we also aim to predict noise. Therefore, when signal and noise lengths are balanced, it results in the most favorable prediction outcome due to an optimal equilibrium between the modeling signal and noise components.

When the time window is not fixed, we can explore the impact of fixing either the

(a)

	AMAG		CREIME	
	Magnitude	*P* arrival	Magnitude	*P* arrival
MEAN	**−0.05**	0.02	0.16	0.02
MAE	**0.46**	**0.04**	0.52	0.05
STD	**0.60**	0.06	0.68	0.06

(b)

	AMAG		CREIME	
	Event	Noise	Event	Noise
Precision	98.59	**99.99**	**99.03**	99.81
Recall	**99.99**	98.77	99.78	**99.16**

Figure 7. Comparison of performance between AMAG model and CREIME model using the High Sensitivity Seismograph Network Japan (Hi-net) stations. (a) Magnitude estimation and phase picking. (b) The performance of the AMAG model and CREIME model as a classifier for events and noise. The bold values highlight the superior results of the two models for different metrics.

signal or the noise length while varying the other. We fix the noise length at 2 s and alter the signal length to 6, 4, and 2 s. Conversely, we fix the signal length at 2 s and adjust the noise length to 2, 4, and 6 s.

As summarized in Figure 6b, it becomes evident that when the noise length is fixed, increasing the effective signal length leads to improved prediction accuracy. This aligns with our expectations, as more effective waveform information involved in the network training process enhances the model's predictive capability. However, this improvement in prediction accuracy comes with the trade-off of reduced time available for EEW. Conversely, when the signal length is fixed, varying the noise length demonstrates a notable effect on prediction performance. As the noise length increases, the model's prediction ability diminishes. This indicates that while effective signal information is crucial, noise interference also significantly impacts the model's judgment.

The results from Figure 6 reveal an interesting phenomenon: when the lengths of signal and noise are 1/5 and 5/1, the prediction results are similar. However, when the lengths change to 2/6 and 6/2, the predictions differ significantly. A reasonable explanation for this discrepancy is that the model struggles to fit the data when the overall signal length increases and is dominated by noise. Conversely, when the length of the effective signal increases, more useful information (such as *S*wave) is available, which helps the model converge more quickly. This indicates that the choice of time window and the lengths of signal and noise have a profound impact on the result. It is not simply that more signals or less noise will always lead to better model predictions.

Taken together, these findings underscore the importance of both effective signal and reasonable noise in seismic prediction models. Optimal prediction results are achieved when the model is trained with enough meaningful signal information while also considering the presence of noise and its potential impact on prediction accuracy.

Compared with CREIME

We retrained the AMAG model using the STEAD training dataset, closely following the data preparation process outlined by CREIME. We compare the AMAG model with the CREIME model, utilizing consistent input parameters to maintain comparability with the CREIME method for effective signal input.

Applying both models separately to the Hi-net dataset, we evaluate their errors in magnitude prediction and initial pickup, as presented in Figure 7a. For AMAG, the mean error in magnitude prediction was −0.05, with an MAE of 0.46 and an STD of 0.60. In comparison, the CREIME method yielded a mean error of magnitude prediction of 0.16, an MAE of 0.52, and an STD of 0.68. Clearly, our method outperformed the CREIME method in the accuracy of earthquake magnitude prediction. Regarding the accuracy of *P*-wave pickup, our model exhibits a mean error of 0.02, an MAE of 0.04, and an STD of 0.06. Conversely, the CREIME method shows a mean error of 0.02, an MAE of 0.05, and an STD of 0.06 for *P*-wave picking. For phase picking, the AMAG model also outperforms the CREIME model in the MAE, the mean error and deviation are the same. The larger picking errors may be attributed to incorrect labeling of *P*-wave arrivals included in the test dataset. We provide some examples of large picking errors to illustrate these discrepancies as shown in Figure 8.

In Figure 8, the predicted magnitude sequence accurately corresponds to the *P* wave, whereas the labeled *P*-wave arrival is incorrect. The attention weights in Figure 8a,b are concentrated at the *P*-wave arrival, whereas in Figure 8c,d, the attention weights are focused on the end of the sequence. Despite some bias in seismic phase prediction, most magnitude estimates remain reliable, as shown in Figure 8b,c.

Furthermore, we assess the ability of both methods to distinguish between events and noise, summarized in Figure 7b. AMAG achieves the precision and recall rate of 98.59% and 99.99% for events, and 99.99% and 98.77% for noise, respectively. Conversely, the CREIME method attained precision and recall rates of 99.03% and 99.78% for events, and 99.81% and 99.16% for noise, respectively. Although our model slightly outperforms the CREIME method in the noise precision and event recall rate, the CREIME model exhibited slightly better performance in the event precision and noise recall rate. Overall, there was no significant difference between the two

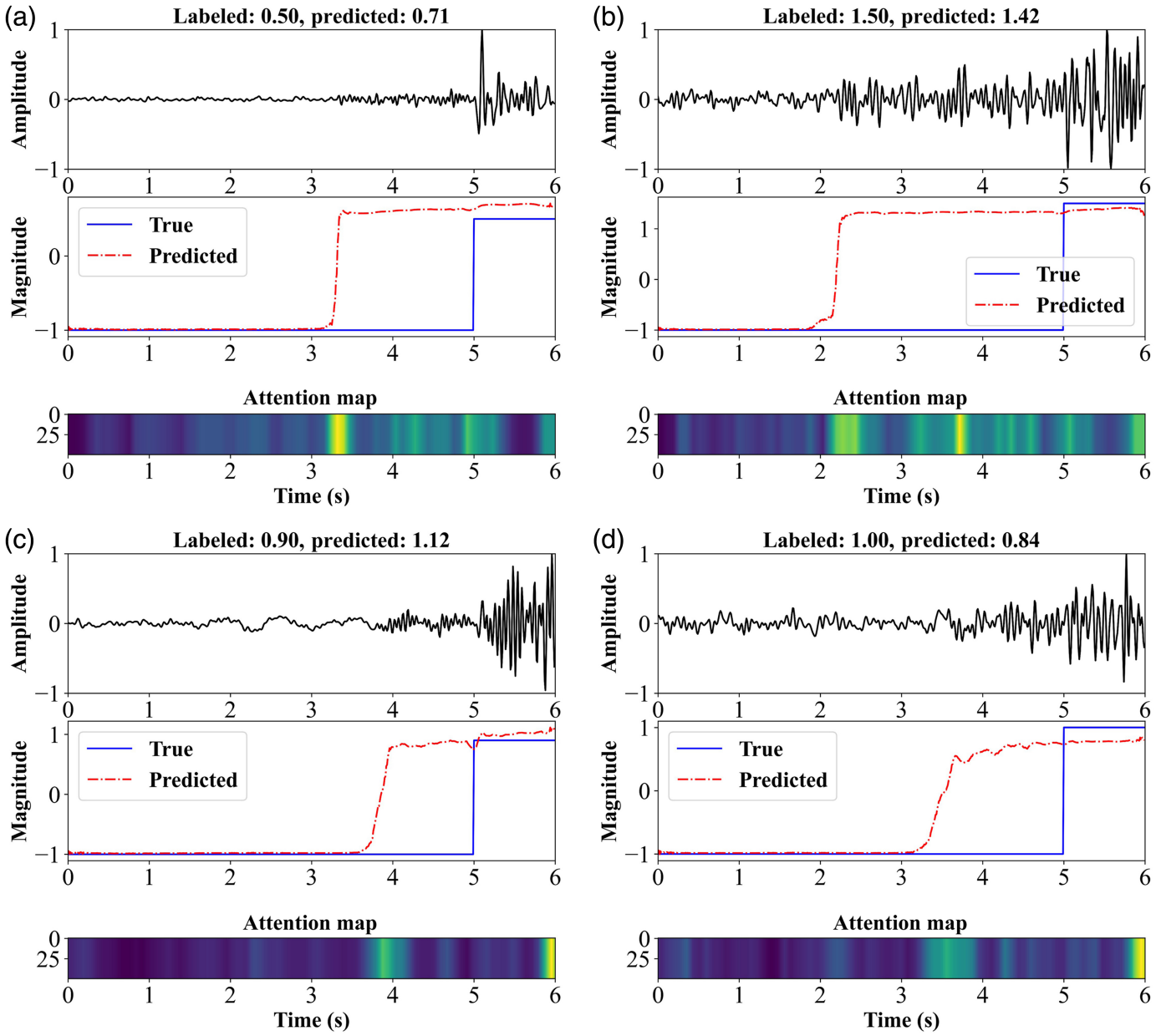

Figure 8. Some predicted results of large picking errors. The labeled *P*-wave arrival is incorrect, while the predicted result is well. (a,b) Attention weights focus on the *P*-wave arrival. (c, d) Attention weights focus on the end of the signal. The color version of this figure is available only in the electronic edition.

models in event and noise classification. Figure S6 illustrates the prediction results and error analysis of both models, providing further insights into their performance characteristics.

In addition, the detection capability was analyzed using continuous waveform data from the KAKH Hi-net station in Miyagi Prefecture over a single day (20 March 2021), during which the **M** 6.9 interplate earthquake (offshore Miyagi Prefecture) and its subsequent aftershocks occurred. A visual inspection of all waveforms was conducted, and the results were compared with the Japan Meteorological Agency (JMA) earthquake catalog and the EQTransformer (EQT) method (Mousavi *et al.*, 2020). The JMA catalog reports 208 events, whereas EQT detects 258 events, 159 of which overlapped with the catalog. Our method identifies 463 events, including 156 additional events not found in either the catalog or EQT results. Visual inspection of these additional events revealed that 10% are noise data. The training data consist primarily of high SNR seismic events (>10 dB) and noise, whereas low SNR seismic events were not included in the training. This implies that our model is not yet fully equipped to function as a standalone earthquake detector. It is highly sensitive to small events, allowing it to detect a greater number of them. However, this sensitivity also introduces a corresponding noise level, leading to false alarms. By leveraging the picks, SNR calculations, and estimated magnitudes, we can enhance the method for earthquake detection,

resulting in a more robust detector with fewer false triggers, though it may miss some small seismic events. Through this comparative analysis, we can better understand the effectiveness and limitations of our model in various seismic scenarios, guiding future improvements and applications in real-time earthquake monitoring.

Conclusions

In this study, we introduce an attention-mechanism-based method for seismic magnitude assessment, providing an effective means to evaluate earthquake magnitudes using short-waveform data. This approach holds promise for applications in real-time earthquake monitoring systems and related fields, such as traffic-light systems for geothermal fields and shale gas regions. Our model can handle inputs with different time windows and varying signal lengths without the need for additional training. This capability allows for seamless integration into various scenarios and enhances compatibility with other models and processing methods.

We analyze the effects of SNR and epicentral distance on magnitude prediction and phase picking. The results demonstrate that our model remains stable under varying conditions. To explore the influence of different signal lengths on the results, we conducted tests with fixed time windows, fixed noise lengths, and fixed signal lengths. We found a balance between signal and noise length, with optimal prediction occurring when both are equal within a 6 s window. Although increasing the effective signal input generally improves prediction accuracy, it can also reduce the timeliness of EEW. We conducted a systematic analysis of the network model's depth and loss function, revealing that optimal model performance often requires a compromise in network parameters. Comparative analysis with the CREIME model, which also addresses earthquake magnitude prediction using partial signal information, highlights the advantages of our method in magnitude assessment and seismic phase picking. Both methods show strengths in event and noise classification, with minor differences between them. In addition, we delve into the impact of the attention mechanism on model performance. The concentration of attention weight near the *P*-wave region and the distinct feature output facilitate noise–signal differentiation. However, the attention mechanism's ability to anticipate seismic events remains limited. Future research directions may involve integrating additional data sources to extract effective pre-earthquake information, thereby enriching real-time earthquake monitoring capabilities.

Data and Resources

The data used in this study have been downloaded from the web services provided by the Stanford EArthquake Dataset (STEAD: https://github.com/smousavi05/STEAD), Japan Meteorological Agency (JMA: https://www.data.jma.go.jp/svd/eqev/data/bulletin/deck_e.html), and the nationwide high-sensitivity seismograph network (Hi-net: https://www.hinet.bosai.go.jp/). Our source code is available at https://github.com/LolitaZJ/AMAG. All websites were last accessed in July 2024. The supplemental material includes Figures S1–S6 and Table S1. They are supplements to some algorithms and results.

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

The authors sincerely thank the reviewers for their valuable feedback and suggestions, which have significantly enhanced this work. The authors thank the National Key R&D Program of China (2022YFF0503203) for the financial support. This research is supported by SATREPS under the title "Establishment of a Research and Education Complex for Developing Disaster-resilient Societies–MARTEST" promoted by the Japan International Cooperation Agency (JICA) and Japan Science and Technology Agency (JST).

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama (2019). Optuna: A next-generation hyperparameter optimization framework, doi: 10.48550/arxiv.1907.10902.

Chakraborty, M., D. Fenner, W. Li, J. Faber, K. Zhou, G. Rümpker, H. Stoecker, and N. Srivastava (2022). CREIME—A convolutional recurrent model for earthquake identification and magnitude estimation, *J. Geophys. Res.* **127,** no. 7, doi: 10.1029/2022jb024595.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*, The Mit Press, Cambridge, Massachusetts.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, 27–30 June 2016, 770–778.

Hochreiter, S., and J. Schmidhuber (1997). Long short-term memory, *Neural Comput.* **9,** no. 8, 1735–1780, doi: 10.1162/neco.1997.9.8.1735.

Hou, B., Y. Zhou, S. Li, Y. Wei, and J. Song (2024). Real-time earthquake magnitude estimation via a deep learning network based on waveform and text mixed modal, *Earth Planets Space* **76,** no. 1, doi: 10.1186/s40623-024-02005-8.

Ioffe, S., and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift, *Proc. of the 32nd International Conf. on Machine Learning, PMLR*, Vol. 37, 448–456.

Kingma, D. P., and J. Ba (2014). Adam: A method for stochastic optimization, doi: 10.48550/arXiv.1412.6980.

Kuang, W., C. Yuan, and J. Zhang (2021). Network-based earthquake magnitude determination via deep learning, *Seismol. Res. Lett.* **92,** no. 4, 2245–2254, doi: 10.1785/0220200317.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning, *Nature* **521,** no. 7553, 436–444.

Li, Z., M.-A. Meier, E. Hauksson, Z. Zhan, and J. Andrews (2018). Machine learning seismic wave discrimination: Application to earthquake early warning, *Geophys. Res. Lett.* **45,** no. 10, 4773–4779, doi: 10.1029/2018gl077870.

Lomax, A., A. Michelini, and D. Jozinović (2019). An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network, *Seismol. Res. Lett.* **90,** no. 2A, 517–529, doi: 10.1785/0220180311.

Maas, A. L., A. Y. Hannun, and A. Y. Ng (2013). Rectifier nonlinearities improve neural network acoustic models, *Proc. of the 30th International Conf. on Machine Learning*, Atlanta, Georgia.

Mousavi, S. M., and G. C. Beroza (2020). A machine-learning approach for earthquake magnitude estimation, *Geophys. Res. Lett.* **47,** no. 1, doi: 10.1029/2019gl085976.

Mousavi, S. M., W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Commun.* **11,** no. 1, doi: 10.1038/s41467-020-17591-w.

Mousavi, S. M., Y. Sheng, W. Zhu, and G. C. Beroza (2019). STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI, *IEEE Access* **7,** 179,464–17,9476, doi: 10.1109/access.2019.2947848.

Münchmeyer, J., D. Bindi, U. Leser, and F. Tilmann (2021). Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network, *Geophys. J. Int.* **226,** no. 2, 1086–1104, doi: 10.1093/gji/ggab139.

National Research Institute for Earth Science and Disaster Resilience (NIED) (2019). NIED Hi-net, National Research Institute for Earth Science and Disaster Resilience, doi: 10.17598/NIED.0003.

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance, *Twenty-First International Conference on Machine learning - ICML '04*, Banff, June 2004, doi: 10.1145/1015330.1015435.

Niu, Z., G. Zhong, and H. Yu (2021). A review on the attention mechanism of deep learning, *Neurocomputing* **452,** 48–62, doi: 10.1016/j.neucom.2021.03.091.

Perol, T., M. Gharbi, and M. Denolle (2018). Convolutional neural network for earthquake detection and location, *Sci. Adv.* **4,** no. 2, e1700578, doi: 10.1126/sciadv.1700578.

Prechelt, L. (1998). Early stopping-but when? in *Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, 55–69.

Raskutti, G., M. J. Wainwright, and B. Yu (2014). Early stopping and non-parametric regression: an optimal data-dependent stopping rule, *J. Machine Learn. Res.* **15,** no. 1, 335–366.

Richter, C. F. (1935). An instrumental earthquake magnitude scale, *Bull. Seismol. Soc. Am.* **25,** no. 1, 1–32, doi: 10.1785/BSSA0250010001.

Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 5–9 October 2015, 234–241.

Ross, Z. E., M. Meier, E. Hauksson, and T. H. Heaton (2018). Generalized seismic phase detection with deep learning, *Bull. Seismol. Soc. Am.* **108,** no. 5A, 2894–2901, doi: 10.1785/0120180080.

Saad, O. M., Y. Chen, A. Savvaidis, S. Fomel, and Y. Chen (2022). Real-time earthquake detection and magnitude estimation using vision transformer, *J. Geophys. Res.* **127,** no. 5, doi: 10.1029/2021jb023657.

Stein, S., and M. Wysession (2009). *An Introduction to Seismology, Earthquakes, and Earth Structure*, Blackwell Publishing, Malden, Massachusetts.

van den Ende, M. P. A., and J.-P. Ampuero (2020). Automated seismic source characterization using deep graph neural networks, *Geophys. Res. Lett.* **47,** no. 17, doi: 10.1029/2020gl088690.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, Long Beach, California, 4–9 December 2017, doi: 10.48550/arXiv.1706.03762.

Wang, H., and J. Zhang (2023). A deep learning approach for suppressing noise in livestream earthquake data from a large seismic network, *Geophys. J. Int.* **233,** no. 3, 1546–1559, doi: 10.1093/gji/ggad009.

Wang, K., J. Zhang, J. Zhang, Z. Wang, and H. Zhu (2024). Monitoring seismicity in the southern Sichuan Basin using a machine learning workflow, *Earthq. Res. Adv.* **4,** no. 1, 10,0241–10,0241, doi: 10.1016/j.eqrea.2023.100241.

Wang, Y., X. Li, Z. Wang, and J. Liu (2022). Deep learning for magnitude prediction in earthquake early warning, *Gondwana Res.* doi: 10.1016/j.gr.2022.06.009.

Xiao, Z., J. Wang, C. Liu, J. Li, L. Zhao, and Z. Yao (2021). Siamese earthquake transformer: A pair-input deep-learning model for earthquake detection and phase picking on a seismic array, *J. Geophys. Res.* **126,** no. 5, doi: 10.1029/2020jb021444.

Yang, S., J. Hu, H. Zhang, and G. Liu (2020). Simultaneous earthquake detection on multiple stations via a convolutional neural network, *Seismol. Res. Lett.* **91,** no. 1, 246–260, doi: 10.1785/0220200137.

Zhang, J., Z. Li, and J. Zhang (2023). Simultaneous seismic phase picking and polarity determination with an attention-based neural network, *Seismol. Res. Lett.* **94,** no. 2A, 813–828, doi: 10.1785/0220220247.

Zhang, X., J. Zhang, C. Yuan, S. Liu, Z. Chen, and W. Li (2020). Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method, *Sci. Rep.* **10,** no. 1, doi: 10.1038/s41598-020-58908-5.

Zheng, G., S. Mukherjee, X. L. Dong, and F. Li (2018). Opentag: Open attribute value extraction from product profiles, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, United Kingdom, 19–23 August 2018, 1049–1058.

Zhou, Y., H. Yue, Q. Kong, and S. Zhou (2019). Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks, *Seismol. Res. Lett.* **90,** no. 3, 1079–1087, doi: 10.1785/0220180319.

Zhu, W., and G. C. Beroza (2018). PhaseNet: A deep-neural-network-based seismic arrival time picking method, *Geophys. J. Int.* **216,** no. 1, 261–273, doi: 10.1093/gji/ggy423.

Zhu, W., S. M. Mousavi, and G. C. Beroza (2019). Seismic signal denoising and decomposition using deep neural networks, *IEEE Trans. Geosci. Remote Sens.* **57,** no. 11, 9476–9488, doi: 10.1109/tgrs.2019.2926772.

Manuscript received 19 July 2024
Published online 28 February 2025