

FULL PAPER

Open Access



Real-time earthquake magnitude estimation via a deep learning network based on waveform and text mixed modal

Baorui Hou^{1,2}, Yueyong Zhou³, Shanyou Li^{1,2}, Yongxiang Wei³ and Jindong Song^{1,2*}

Abstract

Rapid and accurate earthquake magnitude estimations are essential for earthquake early warning (EEW) systems. The distance information between the seismometers and the earthquake hypocenter can be important to the magnitude estimation. We designed a deep-learning, multiple-seismometer-based magnitude estimation method using three heterogeneous multimodalities: three-component acceleration seismograms, differential P-arrivals, and differential seismometer locations, with a specific transformer architecture to introduce the implicit distance information. Using a data-augmentation strategy, we trained and selected the model using 5365 and 728 earthquakes. To evaluate the magnitude estimation performance, we use the root mean square error (RMSE), mean absolute error (MAE), and standard deviation error (σ) between the catalog and the predicted magnitude using the 2051 earthquakes. The model could achieve RMSE, MAE, and σ less than 0.38, 0.29, and 0.38 when the passing time of the earliest P-arrival is 3 s and stabilize to the final values of 0.20, 0.15, and 0.20 after 14 s. The comparison between the proposed model and model ii, which is retrained without the specific architecture, indicates that the architecture contributes to the magnitude estimation. The P-arrivals picking error testing indicates the model could provide robust magnitude estimation on EEW with an absolute error of less than 0.2 s.

Keywords Magnitude estimation, Earthquake early warning, Deep learning, Real time, Multiple seismometers

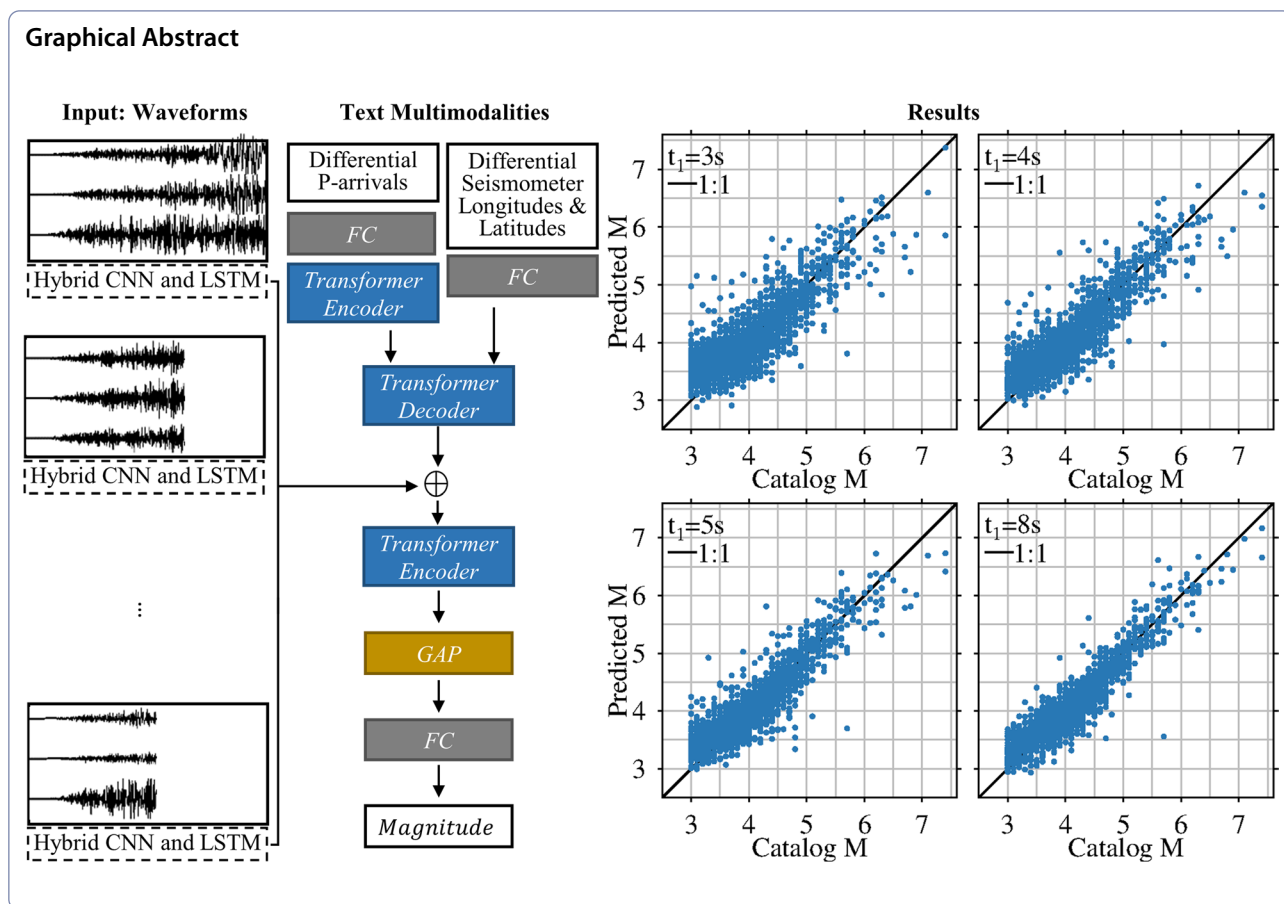
*Correspondence:

Jindong Song
jdsong@iem.ac.cn

Full list of author information is available at the end of the article



©The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



Introduction

Rapid and accurate earthquake magnitude estimation during quaking is essential for earthquake early warning (EEW) systems, especially in providing quick information about the earthquake to the public or specific users (Yamada et al. 2021). The earthquake magnitude measures the size of the earthquake empirically using the peak amplitude after the distance is corrected carefully (Funasaki & Earthquake Prediction Information Division 2004; Katsumata 2004; Moriwaki 2017; Richter 1935; Tsuboi 1954), such as the Japan Meteorological Agency Magnitude (M_{JMA}) is defined based on multiple magnitudes (the local meteorological office magnitude, displacement magnitude, and velocity magnitude). The classical methods applied to EEW systems would be roughly classified into the predominant period (Kanamori 2005; Heidari 2018), peak amplitude (Wu & Zhao 2006; Kuyuk and Allen 2013; Colombelli et al. 2020), and energy (Festa et al. 2008), which still play a vital and foundational role in EEW. However, the accuracy of the magnitude estimation is associated with the determination of the hypocenter location which the challenges still exist to pinpoint the hypocenter location in real-time primarily due to

limited information in the early stage of the earthquake (Saad et al. 2022b) not yet considering the peak integrated amplitude influenced by the low-frequency noise (Yamada and Mori 2009). There are two representative strategies to extract information from the waveforms for magnitude estimation: the classical methods (Colombelli et al. 2020; Kanamori 2005; Kuyuk and Allen 2013; Wu and Zhao 2006) and deep-learning methods (Kuang et al. 2021; Mousavi & Beroza 2020a; Münchmeyer et al. 2021; Saad et al. 2022a). The classical methods could be considered the seismologist expertise-based physics features extraction strategy, and the deep-learning methods could be considered the automatic and direct features extraction strategy assisted by the seismologist expertise. Some seismologists are trying to utilize deep-learning methods to obtain low error levels of magnitude estimation performance and make some progress, especially in the early stage of the earthquake (Münchmeyer et al. 2021).

In recent years, researchers have adopted deep-learning approaches to solve problems in various seismic fields from waveforms, such as earthquake detection (Meier et al. 2019; Perol et al. 2018; Reynen and Audet 2017), phase picking (Mousavi et al. 2020; Zhu and Beroza

2019), estimation of earthquake location (Lomax et al. 2019; Mousavi and Beroza 2020b; Münchmeyer et al. 2021), and magnitude estimation (Kuang et al. 2021; Lomax et al. 2019; Mousavi & Beroza 2020a; Münchmeyer et al. 2021; Saad et al. 2022a). The architectures of these magnitude estimation algorithms are mainly the convolutional neural networks (CNN), the long-short memory networks (LSTM), and the transformer networks. The CNN could extract significant features due to its architecture with a parameter-sharing design. However, it requires scale/normalization input or other neural networks to avoid the domination of features with higher amplitudes on the prediction results (Saad et al. 2022a). As a result of the importance of the waveform's amplitude, it is challenging to use only CNN on magnitude estimation. Compared with CNN, the LSTM is insensitive to the non-normalization waveforms due to the gate mechanism with the Tanh and Sigmoid activation function (Hochreiter and Schmidhuber 1997; Mousavi et al. 2019). It is suitable to combine with CNN on magnitude estimation, such as the algorithms (Lomax et al. 2019; Mousavi and Beroza 2020a). Although the methods are not suitable for real-time in the current structure, their design concepts are still worth learning to design a real-time deep-learning magnitude estimation model; that is, CNN extracts significant features, and LSTM avoids the domination of features with higher amplitudes on the prediction results. The LSTM is specially designed for time series, which could process unfixed time using the different lengths of waveforms recorded on different seismometers due to the wave propagation and seismometer distribution during quaking. The transformer networks (Vaswani et al. 2017) could weigh the features according to their relationship using the attention mechanism, which could be suitable for processing the features from multiple seismometers with different lengths of waveforms in real-time. The typical method (Münchmeyer et al. 2021) utilized the CNN to extract the onsite features and combine them with six transformer encoders. To our knowledge, it belongs to the earliest use of transformer encoder for multiple seismometers. In addition, they proposed a set of practices to build a model for fast earthquake source characterization, which is significant to help more seismologists establish a suitable deep-learning model in their fields. Moreover, the transformer networks for extracting features corresponding to different times using a single waveform for magnitude estimation have made the process, and the seismologists are developing it to suit real-time situations (Saad et al. 2022a). Several real-time methods currently use CNN (such as Van Den Ende and Ampuero 2020), seemingly introducing largely noise or invalid zeros, which may still be debated (Saad et al. 2022a).

Most magnitude estimation methods are not real-time or network approaches. These models provide evidence that distance information is essential for magnitude estimation, directly introducing distance or automatically extracting from P/S phases in waveform (Kuang et al. 2021; Mousavi and Beroza 2020a). For the magnitude estimation approaches in EEW, it is ideal to introduce the location information based on the final earthquake location. Unfortunately, the estimation of the earthquake location might vary with the increase in earthquake information during quaking. A recent study provides a random forest-based approach to accurately estimate the earthquake location using differential P-wave arrivals and seismometer locations recorded on the five earliest seismometers (Saad et al. 2022b). The method inspires us to introduce location information on magnitude estimation for processing the variety of earthquake location estimations at the early stage of an earthquake. However, making direct model fusion on the above random forest model and a deep-learning magnitude estimation model using waveforms is challenging. On the other hand, it should be carefully considered to utilize the two types of heterogeneous multimodality data of an earthquake on magnitude estimation based on the deep-learning methods: the time series sequence data of the waveforms and the text data of the P-wave arrivals and seismometer locations, which the two type multimodalities might be related to different physics meanings. Recently, multimodal machine learning (MMML) could construct models that can process and relate information from multiple heterogeneous modalities (Baltrušaitis et al. 2018; Lahat et al. 2015), having been paying attention to a variant of multi-disciplinary fields, e.g., computer vision (Luo et al. 2022; Radford et al. 2021; Wang et al. 2021a, b), natural language processing (Gong et al. 2021; Liu et al. 2021). Some research indicates that transformer architectures perform well on multimodality fusion tasks of MMML (Gong et al. 2021; Nagrani et al. 2021; Zhu et al. 2020) and provide a way to process multimodality data of an earthquake. Although the real-time magnitude estimation methods (Münchmeyer et al. 2021) could generally extract distance information from the waveforms recorded on each seismometer, it may not always work at the initial of the earthquake, especially when the seismometers are far from the earthquake location, which leads to the larger travel time between the P/S phases. Thus, to avoid the potential situation, we expect to introduce the additional location information by the multiple triggered seismometers. Considering that the scale/normalization influence could decrease by incorporating the maximum amplitude as additional input still be debated (Kuang et al. 2021; Lomax et al. 2019; Münchmeyer et al. 2021; Mousavi and Beroza 2020a) when only using CNN, we adapt the CNN–LSTM model as the onsite extraction as the models (Lomax et al. 2019; Mousavi and

Beroza 2020a). In this paper, we propose deep-learning-based methods for real-time magnitude estimation using three heterogeneous multimodalities: multiple seismometer waveforms, differential P-wave arrivals, and differential seismometer locations, with a specific architecture to introduce the distance information by automatically processing the variety of hypocenter location estimation. We evaluate the magnitude estimation performance of our model and provide evidence for the effectiveness of the specific architecture using the root mean square error (RMSE), mean absolute error (MAE), and standard deviation error (6).

Methods

Inputs

The input to the model consists of three heterogeneous multimodalities, including the three-component acceleration seismograms from multiple seismometers, differential P-wave arrivals (T), and differential seismometer locations (L). We construct the acceleration seismograms based on the P-wave arrivals as the following process: We set $t_1 + t_{noise}$ and $t_i + t_{noise}$ as the lengths of the acceleration waveforms recorded on the earliest and later P-wave arrival seismometers. t_i is the waveform length after P-wave arrival recorded on the i^{th} seismometer and t_{noise} is the noise length. To simulate the real situation, t_i should be $t_1 - \Delta t_i$ which Δt_i is the travel time between the i^{th} P-wave arrival seismometer and the earliest P-wave arrival seismometer ($i=1, 2, \dots$). We clipped the waveforms after the P-wave arrival recorded on the i^{th} seismometer when t_i is greater or equal to 1.0 s, reference t_1 from 1 to 30 s with an interval of 1 s. To make the model learn the noise condition, we set the length of the noise (t_{noise}) as 1 s before the P-wave arrival at each seismometer. Based on the previous study (Mousavi & Beroza 2020a) about the effectiveness of the amplitude on magnitude estimation, we do not adopt any normalization process. We assumed the P-wave arrival could be identified when the length of the P-wave is greater or equal to 1 s. For the other inputs of the differential P-wave arrivals and differential seismometer locations could be expressed as

$$T = [\Delta t_1, \Delta t_2, \dots, \Delta t_i] \quad (1)$$

$$L = [(\Delta lat_1, \Delta lon_1), (\Delta lat_2, \Delta lon_2), \dots, (\Delta lat_i, \Delta lon_i)] \quad (2)$$

where Δlat_i represents the numerical latitude difference between the i^{th} and the earliest P-arrival seismometer, Δlon_i represents the numerical longitude difference between the i^{th} and the earliest P-arrival seismometer. Considering the computational cost, we set the maximum number of seismometers to 20. In addition, to simulate the real-time condition, the i^{th} P-arrival seismometer should be triggered at the same time.

Model

We build the real-time earthquake magnitude estimation model based on the CNN, LSTM, and transformer architecture. The proposed model consists of four parts: single waveform feature extraction, implicit distance information extraction, feature fusion, and magnitude output. We adopt CNN and LSTM layers to extract each three-component acceleration waveform. We mainly utilize the CNN layer without any activation unit to downsample the three-component dimensions of the waveform. The kernel size and step size are 3×1 . To the i^{th} P-arrival seismometer with waveform dimension $3 \times 100 \times (t_i + 1)$ (the sampling frequency is 100 Hz), the dimension tunes to $100 \times (t_i + 1)$ by the CNN layer which keeps the time series of the waveform. Then, we adopt the LSTM with 32 units to extract the features in advance, in which the activation and recurrent activation are Tanh and Sigmoid activation functions (Hochreiter and Schmidhuber 1997), respectively. The LSTM could have the same weights on the time series, which could be more suitable for processing the unfixed time length of the waveforms. We set the input time window to the LSTM as 0.5 s (50 points) and slip the time window with an interval of 0.5 s. We selected the final time window output by the LSTM units as the single waveform features. To avoid possible information leakage, we selected the penultimate time window output by the LSTM as the final single waveform features when the length of the last input time window is less than 0.5 s. The advantage of the LSTM design is that it could make the dimension of the features the same as the features extracted from each three-component acceleration waveform with unfixed time length.

For the implicit distance information extraction architecture, we mainly utilized the transformer encoder and decoder architectures to extract and fuse features. We tune the dimension and extract features of the differential P-wave arrivals or the differential seismometer locations by a fully connected layer (FC) without any activation units, respectively. The input dimension is $i \times 1$ (or $i \times 2$) and the output dimension is $i \times 32$. The dimension process could make the following calculation easily by the transformer encoder and decoder architectures. We first extract the time feature vectors by the transformer encoder with the input of the differential P-arrivals after the FC layer. The time feature vectors might indicate the travel time relationships between the seismometers and the location of an earthquake. Then, we obtain the location feature vectors using the transformer decoder by fusing the input of the differential seismometer locations after the FC layer (as the query vector) and the time feature vectors (as the value and key vectors). The location feature vectors might indicate the distance between the seismometers and the location of an earthquake. The

transformer encoder contains a self-attention layer, a layer normalization layer, two FC layers, and one dropout rate layer, as shown in Fig. 1. The transformer decoder has one attention layer, three layer normalization layers, two FC layers, and one dropout rate layer (Fig. 1). The two FC layers contain 64 and 32 neurons, respectively, the first FC layer followed by a Relu (rectified linear unit) activation (Nair & Hinton 2010) unit and a dropout layer (Srivastava et al. 2014). The above architecture differs from the traditional architecture (Vaswani et al. 2017). We adopt the above architecture as the previous study (Xiong et al. 2020), which could be easily trained without a warm-up stage.

We fuse the features from multiple seismometers through a transformer encoder and a global average pooling (GAP) layer. We first add the location feature vectors on the waveform features from multiple seismometers. Then, we adopt the transformer encoder on the adding features to obtain the weight features. We use a global

average pooling (GAP) layer to downsample the weight features. The GAP could be flexible to make the model process multiple seismometer features. Then, we use an FC layer with one neuron to estimate the magnitude.

Results

To ensure the data quality, we selected the earthquakes recorded on the K-NET and KiK-net between January 2008 and June 2020 as the following criteria (National Research Institute for Earth Science and Disaster Resilience, 2019). (1) The P-wave arrivals could be identified; (2) epicenter distances are less than 2 degrees; (3) magnitudes of the earthquakes are greater or equal to M_{JMA} 3.0; (4) depths of the earthquakes are less than 300 km; (5) acceleration waveforms in three components are recorded on the seismometers equipped at the ground surface or upbore (Okada et al. 2004); (6) signal-to-noise ratios (SNR) are greater than 10 which the SNR is defined based on the previous study (Wang et al. 2021a, b); and

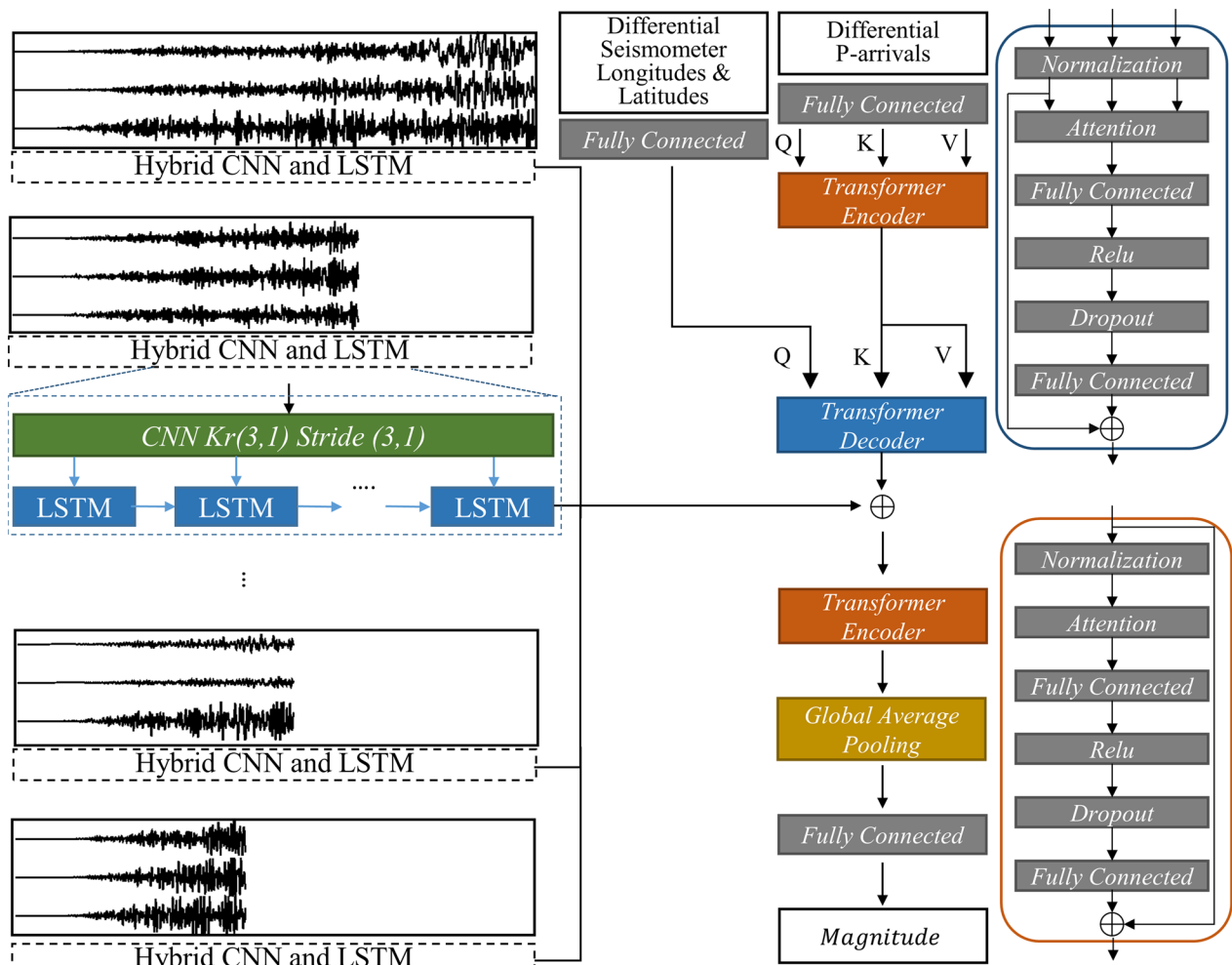


Fig. 1 Architecture of the proposed model for magnitude estimation

(7) seismometers are greater than 5. We selected 297,099 three-component waveforms from 8144 earthquakes in Japan’s inland and offshore areas. We obtained the above acceleration waveforms from the online data sets. The information concerning the event location, seismometer location, and magnitude are provided by the JMA (Japan Meteorological Agency). After picking the P-arrivals manually and removing the baseline offset by subtracting the mean before P-arrival, we filtered these waveforms with corner frequencies of 0.075 Hz using a high-pass digital infinite impulse response (IIR) filter. Then, we extracted the earthquakes after the origin time of 11 November 2016 as the testing data set, which is used to evaluate the model. We randomly split the earlier earthquakes into the training and validation data sets. The training and validation data sets are used to train and select the model. The training, validation, and testing data sets consist of 5,365, 728, and 2,051 earthquakes, with magnitude ranges from 3.0 to 9.0, 3.0 to 7.3, and 3.0 to 7.4, as shown in Fig. 2b. As there are only two earthquakes whose magnitudes exceed 7.0 on the testing data set, we add the earthquake with a magnitude of 7.4, and the origin time is 2022/03/16 23:36. For each data set, most earthquakes’ magnitudes and depths range from 3.0 to 5.0 (Fig. 2b) and 0 to 100 km (Fig. 2c). Figure 2d shows the distribution between the

magnitudes and epicenter distances (Δ) of these earthquakes’ waveforms in three data sets, and the testing data set has a few points with $\Delta \leq 50$ km when the magnitude is ≥ 5.0 . Most epicenter distances are distributed from 20 to 100 km in each data set (Fig. 2e). We set the catalog magnitude provided by the Japan Meteorological Agency as a label, and we expected the model to estimate the catalog magnitude, which could be considered as JMA magnitude. We trained the model for 150 epochs using the mean square error (MSE) loss function with an Adam optimizer (the initial and final learning rate is 0.001 and 0.0001) and randomly selected the 129th model when validation loss was stable during training processing. To avoid overfitting, we adopted a data augmentation strategy and randomly selected 20 seismometers per earthquake during training, if possible. During the model evaluation, we evaluated the model using the 20 earliest P-arrival seismometers per earthquake. We achieved the data augmentation strategy by resampling the earthquakes with the unfixed ratios to make the sample number of different magnitude bins approximately the same. The ranges of the magnitude bins are 3.0~3.5, 3.5~4.0, 4.0~4.5, 4.5~5.0, 5.0~5.5, 5.5~6.0, 6.0~6.5, 6.5~, respectively.

We evaluated the proposed model by the root mean square error (RMSE), mean absolute error (MAE), and

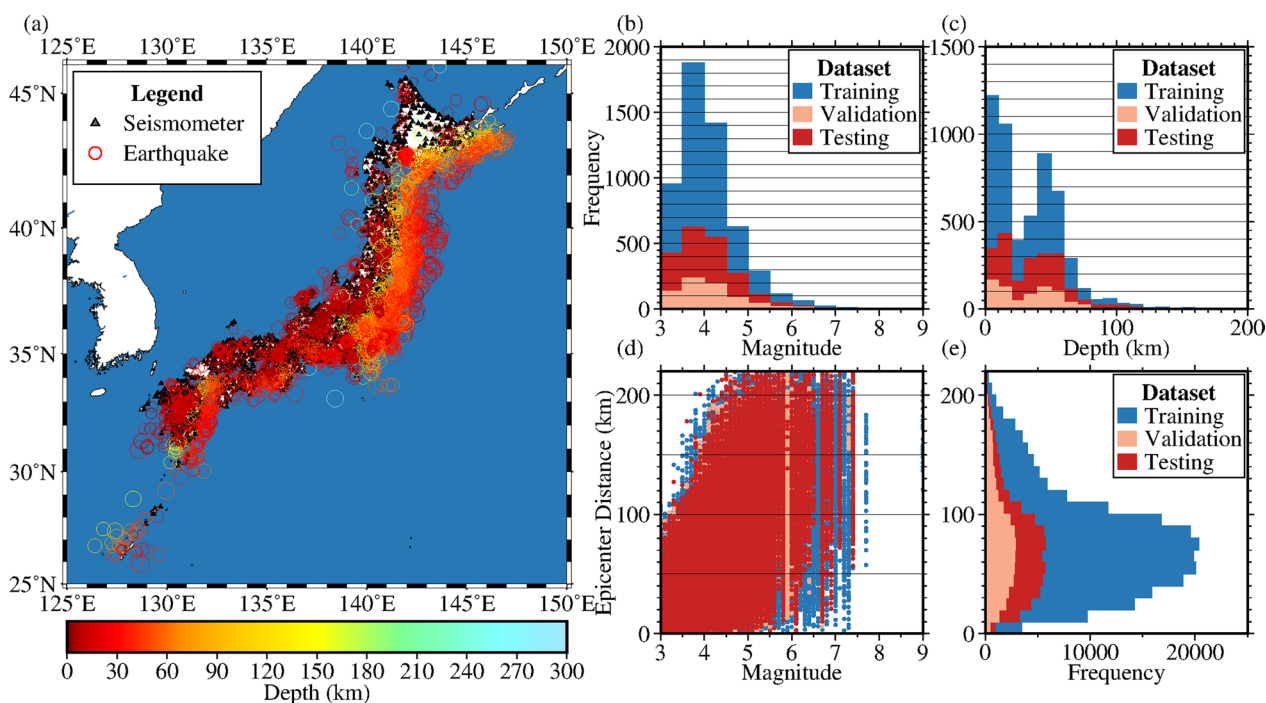


Fig. 2 Distribution of earthquake and seismometer locations. **a** indicates the earthquake epicenter and seismometer locations. The size of the circles is proportional to the magnitude, and the different colors indicate the focal depth. **b, c** show the frequency distribution of the earthquake magnitudes and focal depths. **d** indicates the distribution between epicenter distances and magnitudes. **e** indicates the frequency distribution of the epicenter distances. Different colors indicate different data sets in **b–e**

standard deviation error (σ) between the catalog and predicted magnitude with time (t_1) increases from 1 to 30 s after the first P-wave arrival per earthquake (Fig. 3). The MAE could measure the average absolute difference between the statistical model's predicted and catalog magnitude. Compared with the MAE, the RMSE is more sensitive to the predicted magnitudes with an absolute error > 1.0 (Chai and Draxler 2014). For the EEW application, the prediction error may not be expected when the absolute value exceeds 1.0, especially when the error exceeds 1.0. We considered the σ as a supplement metric that measures the stability of the prediction errors by the model. To the three evaluation metrics, the less values mean the model has lower error levels and is more stable on magnitude estimation. The error curves of the three evaluation metrics could be considered as the prior knowledge to guide the EEW systems to alert the public or specific users with the different tolerations. To show the magnitude performance of the proposed model, we chose the classical model (Kuyuk and Allen 2013) as a

baseline comparison and the real-time deep-learning method from Münchmeyer et al. (2021) as a CNN-based comparison. The classical model utilized the linear relationships between the magnitude, epicenter distance, and the vertical peak displacement of P-wave (P_d) for the magnitude estimation, which its basis is that P_d should be proportional to the rate of moment release (Aki and Richards 2002) in the far field (Trugman et al. 2019). To avoid the inclusion of the S-wave, we adopted the theoretical S-arrivals based on the previous study (Colombelli et al. 2014). Assuming the epicenter location as a known parameter for the classical model, we directly averaged the magnitudes calculated from the triggered seismometers, using the coefficients in the previous study (Kuyuk and Allen 2013). The real-time deep-learning model (Münchmeyer et al. 2021) utilized the normalized time series and logarithmic peak absolute value of the acceleration waveforms and geographical location (latitude and longitude) from multiple seismometers to estimate the magnitude Gaussian mixture, which contains the

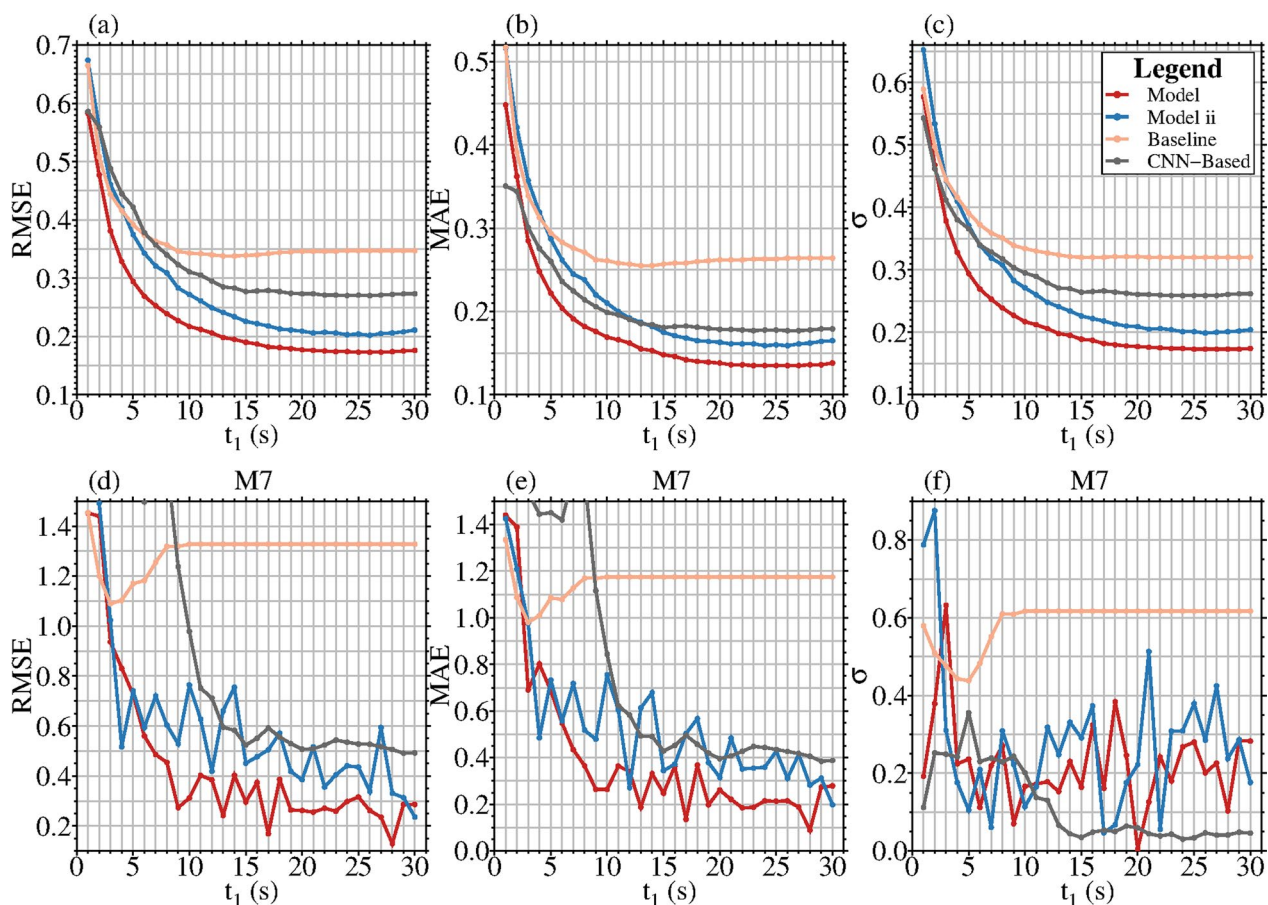


Fig. 3 RMSE (a and d), MAE (b and e), and σ (c and f) curves with time (t_1) increasing on the testing data set. The top figures indicate the curves with magnitudes ranging from 3.0 to 7.4. The bottom figures indicate the curves with magnitudes greater than 7.0. Different colors indicate different models

onsite feature extraction part (the CNNs and the multi-layer perceptron) and the multiple seismometers' feature combination part (the six transformer architectures and the trained weights). The time window of the input corresponding to the onsite time series is fixed to 30 s, and pad zeros to 30 s if the time series length is less than 30 s. Considering the Gaussian mixture, it could not be assumed that the predicted uncertainties are indeed well calibrated, as they mentioned (Münchmeyer et al. 2021); we made the model output the magnitude by the minor change using an FC layer instead of the multi-layer perceptron used to predicted mixture Gaussian. The minor change cannot influence the comparison with our proposed model. Then, we retrained the deep-learning model using the same data sets and loss function as the proposed model, according to the technical details Münchmeyer et al. mentioned (2021), mainly the pre-training model trick.

For the proposed model, the RMSE, MAE, and σ are 0.38, 0.29, and 0.38 at 3 s and are stable with the final values of 0.20, 0.15, and 0.20 after 14 s, respectively. The RMSE, MAE, and σ of the classical model are 0.44, 0.34, and 0.44 at 3 s and stabilize to 0.35, 0.26, and 0.32 after 10 s, respectively. To the CNN-based model, the RMSE, MAE, and σ are 0.49, 0.39, and 0.44 at 3 s and stabilize to 0.28, 0.21, and 0.27 after 15 s. Compared with the baseline and CNN-based models, the errors of our proposed magnitude estimation model decrease more rapidly and achieve a smaller final value. The proposed model could achieve the final value of the baseline and the CNN-based models when the time (t_1) is approximate 4 and 8 s, as shown in Fig. 3a–c. We further analyzed the three earthquakes from the testing data set to evaluate the performance of the M7 earthquakes with the catalog magnitudes exceeding 7.0, which the EEW is more concerned with. The M7 RMSE of the proposed model is 0.94 and less than the 1.09 and 1.61 corresponding to the classical and CNN-based models when the time (t_1) is 3 s. Compared with the two models, the M7 RMSEs of the proposed model decrease more rapidly with a lower final value of 0.29, which means the proposed model has lower error levels on the three earthquakes with a magnitude ≥ 7.0 . Moreover, each model's M7 MAE and σ errors at different times show the same trend as the M7 RMSEs (Fig. 3). For each model, the M7 RMSE, MAE, and σ are larger than the errors of the whole magnitude bins. In addition to considering the smaller quantity of M7 may influence the results, the larger error levels of M7 may also be related to the rupture progress of earthquakes with the magnitude ≥ 7.0 , which is more complex and has a longer duration (Trugman et al. 2019). The proposed model has less errors (lower error levels) on magnitude estimation using the testing data set, which the following

reasons may cause: (1) the proposed model extracts more information from the three-component waveforms, the differential P-arrivals, and the differential seismometer locations than just utilizing the peak displacement of P-wave; (2) the proposed model is trained on simulating a complex situation by random selecting the triggered seismometers which lead the model could be more suitable to the real-time situation without random seismometers selected; and (3) the difference between the data sets in the study and the previous study (Münchmeyer et al. 2021) may lead the CNN-based model having larger errors than their results. Except for containing the different earthquakes, the data sets of this study contain the waveforms recorded on the seismometers equipped at the ground surface from the K-NET and uphole from the KiK-net, without the waveforms recorded on seismometers equipped at the downbore from the KiK-net. However, as the earliest deep-learning-based magnitude estimation model (Münchmeyer et al. 2021) using multiple seismometers in real time, the concepts still have great value, such as combining the multiple seismometer features by utilizing the transformer encoder architectures, the positional embedding (seismometer location), the event token, and pre-training model.

Figure 4 indicates the predicted results by the proposed model when time (t_1) is at 1, 2, 3, 4, 5, 6, 7, and 8 s after the earliest P-wave arrival. The figure could show the magnitude estimation performance corresponding to errors in detail. Overall, with time (t_1) increases, the top sub-figures from a to f directly indicate that the points distribute more convergence around the 1:1 relationship line as the errors decrease. Similarly, as the bottom subfigures show (the blue color bars), the better performance trend is also reflected in the increased numbers of earthquakes in which the prediction errors are less than 0.5. When t_1 is equal to 3 s, the results of the proposed model indicate that most of the prediction errors are within ± 0.5 range. However, for the earthquakes of magnitudes ≥ 6.5 , the model still underestimates magnitude with an absolute prediction error exceeding 1.0 when t_1 is 3 s and the underestimation is mitigated with t_1 increases, as the top subfigures show. Based on the physical model from Trugman et al. (2019), the physical model suggests a weak rupture predictability based on the peak displacement after 50% of the rupture duration. Münchmeyer et al. (2021) analyzed the magnitude saturation based on the physical model; that is, the value of the saturation magnitude could be expected to be 5.7 after 1 s, 6.4 after 2 s, 7.0 after 4 s, and 7.4 after 8 s. Considering the complexity of rupture progress, triggered seismometer distribution, and data distribution of the training data set in real-time situations, the model could only achieve the

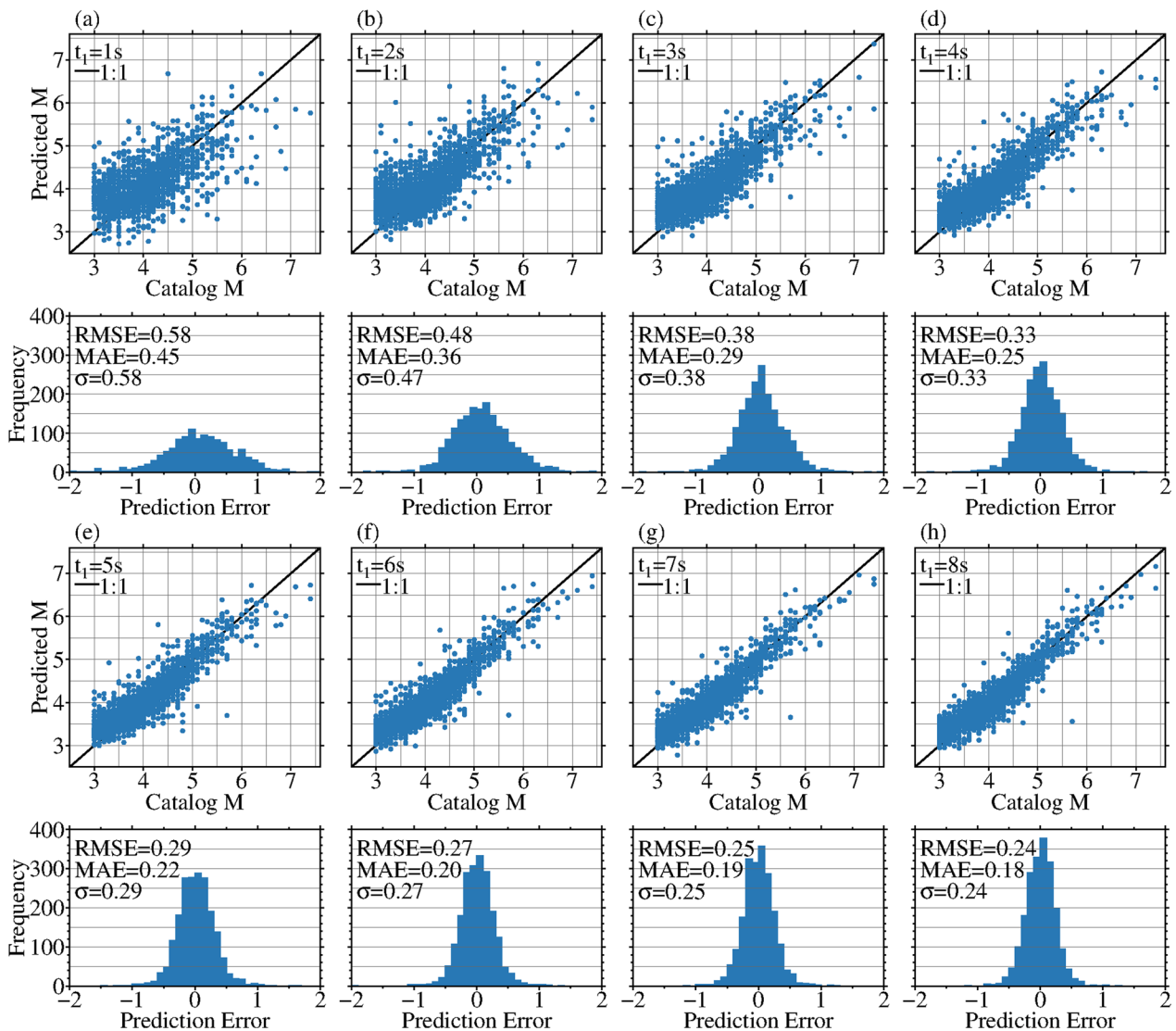


Fig. 4 Results predicted by the proposed model on the testing data set when time (t_1) is at 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g) and 8 (h) s

threshold somewhat. Moreover, we observed that there are several points with the error > 1 when t_1 is 1 or 2 s. The more significant error points occurred inland with only 1 or 2 triggered seismometers in which the epicenter distance is less than 10 km and the earthquake’s depth is less than 24 km, as shown in Fig. 4a ($t_1=1$ s), b ($t_1=2$ s). To the data-driven model, the training data set lacks seismometers with an epicenter distance of less than 10 km, as shown in Fig. 2e, which may be one reason for the overestimated magnitude only using 1 or 2 seismometers with the $\Delta < 10$ km. Moreover, we speculated that diversity may be a factor to consider. In this current situation of the distribution of the triggered seismometers, the model also could not contribute to the magnitude estimation by introducing the implicit

location information between the seismometers and earthquake source from the P-arrivals and seismometer locations.

To provide the effectiveness of the encoder–decoder architecture with different P-wave arrivals and seismometer locations, we only removed the above architecture of the proposed model and retrained, called model ii. The RMSE, MAE, and σ of the model ii achieve 0.46, 0.36, and 0.45 at 3 s and stabilize to approximately 0.21, 0.16, and 0.20 after 20 s, respectively. Based on the curves of the errors (Fig. 3a–c), the proposed model has less errors consistently and reaches the final value of model ii with less time. For earthquakes with a magnitude ≥ 7.0 , the curves of the proposed model also perform less errors than those of the model ii. In other words, the results

indicate that the encoder–decoder architectures contribute to the magnitude estimation, especially at the initial earthquake. The differential errors between the proposed model and model ii decrease with time (t_1) increase, indicating that the contribution of the encoder–decoder architectures decreases with time (t_1) increases. Their same single waveform feature extraction process may cause a decreased trend, extracting the information related to distance from longer waveforms, which contain more information, just like the difference between P-arrival and S-arrival could be used to estimate the distance. Based on the previous study (Saad et al. 2022b), the P-arrivals and seismometer location could be used to ensure the earthquake location and be more accurate with more triggered seismometers. Thus, we expected the encoder–decoder architecture to obtain more accurate implicit earthquake location information with more triggered seismometers, which bring more accurate magnitude estimation performance. Figure 5 indicates the error curves with time (t_1) increase when the triggered seismometers exceed 1, 3, 5, 10, and 15. Taking the MAEs as an example, when the time (t_1) is 2 s, the MAEs are 0.36, 0.30, 0.24, 0.18, and 0.19, respectively. Similarly, the other two errors also decreased with the increased

number of triggered seismometers. The prediction error curve for each earthquake at different triggered seismometer conditions indicates that the error decreased trend with the triggered seismometers increased, as shown in Fig. 5d. There are almost no earthquakes with an absolute error ≥ 1.0 when the triggered seismometers ≥ 3 . The error analysis indicates that the proposed model obtains better magnitude estimation performance with more triggered numbers and could provide evidence that encoder–decoder architecture obtains implicit location information.

Discussion

Hyperparameter optimization of the model

To obtain the optimal network architecture of the proposed model, we tuned the network architecture and selected the final architecture based on the MSE using the validation data set. Based on our design criteria in the Method Section, the LSTM units are a crucial and fundamental parameter that influences all the layers. In other words, we could change the units to tune the network architecture. First, we set the several numbers of the LSTM units as 2, 4, 8, 16, 32, and 64. We found that when the units exceed 8, the proposed model begins

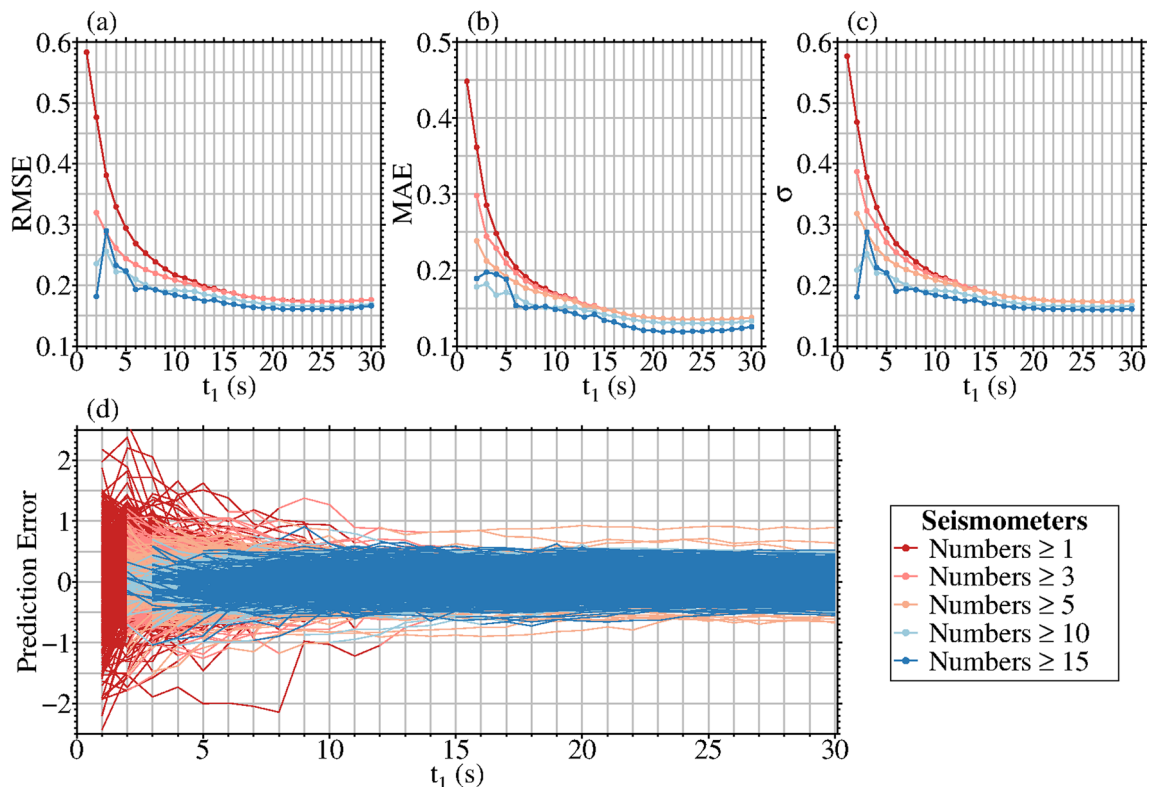


Fig. 5 Error curves predicted by the proposed model with time (t_1) increasing on the testing data set when the triggered seismometers per earthquake are greater or equal to 1, 3, 5, 10, and 15. **a–c** indicate the MSE, MAE, and b curves, respectively. **d** indicates the prediction error curve of each earthquake. Different colors indicate different triggered seismometer numbers

to have the ability to estimate the magnitude. Since the LSTM units are greater than 32, the MSE of the proposed model is similar. Thus, we chose the LSTM units as 32. Second, we added a dropout layer between the two FC layers in each transformer encoder or the decoder to avoid overfitting. We tested the dropout rate as 0.1, 0.2, 0.3, 0.5, and 0.7. We found that the magnitude estimation performance is not significantly decreased when the dropout rate is less or equal to 0.5. Considering the advice value in the previous study (Srivastava et al. 2014) and our testing results, we finally chose 0.3 as the dropout rate. We mainly utilized the attention mechanism to obtain the location information and weigh the onsite features. We did not test the variation of the magnitude estimation performance with the number of the transformer decoder or encoder increases. While tuning the LSTM units and dropout rate, the transformer decoder or encoder number with different inputs is 1. The tuning results seemly indicate there is no need to tune the number of the transformer decoder or encoder. Thus, we set 1 as the transformer decoder or encoder number with different inputs.

Magnitude estimation performance using only P-wave

The proposed model estimates the magnitude using the unfixed length waveforms from the triggered seismometers, the waveform contains only P-wave or both P/S-wave in real-time situations. However, for the proposed model, whether the P-wave contributes to the magnitude estimation and the error level on magnitude estimation only using the P-wave information might be unclear. On the other hand, it might be challenging to analyze the contribution directly from the P-wave based on the proposed model using the LSTM architecture. To indicate the P-wave contribution to the magnitude estimation of the proposed model, we retrained the proposed model and model ii only using P-wave based on theoretical S-arrivals (Colombelli et al. 2014), called model-P and model ii-P. Then, we utilized the same data sets to train and select the models (only using P-wave) using the same criteria as the other models above. Considering the P-arrival times could contribute to the magnitude estimation when the triggered seismometers are greater than 3 (as shown in Fig. 5), we evaluated the models using the earthquakes from the testing data set when the triggered seismometers exceed 3. As shown in Fig. 6a–c, the RMSE, MAE, and δ of the proposed model-P are 0.41, 0.31, and 0.40 at 3 s and stabilize with the final values of 0.32, 0.24, and 0.31 after 16 s, respectively. To evaluate the performance of the M5.5 earthquakes with catalog magnitude exceeding 5.5, we further analyzed the 76 earthquakes from the testing data set. For the M5.5 earthquakes, the RMSE, MAE, and δ of the proposed model-P are 0.87,

0.69, and 0.71 at 3 s and stabilize with the final values of 0.68, 0.51, and 0.64 after 15 s, respectively. To the model ii-P, the RMSE, MAE, and δ of the model ii-P are 0.49, 0.37, and 0.48 at 3 s and stabilize with the final values of 0.32, 0.24, and 0.32 after 14 s, respectively. For the M5.5 earthquakes, the RMSE, MAE, and δ of the model ii-P are 0.94, 0.74, and 0.71 at 3 s and stabilize with the final values of 0.71, 0.54, and 0.58 after 15 s, respectively. The results of the model-P and model ii-P indicate that the two models only using the P-wave information could be used for magnitude estimation, but the error level is high. In addition, the comparison between the two models only using P-wave and the proposed model suggests that introducing the S-wave or Surface-wave could have lower error levels on the performance of magnitude estimation, especially for the magnitude earthquakes M5.5.

Effect on magnitude estimation from the uncertainty of the P-wave arrival picks

The proposed model for magnitude estimation relies on the P-arrivals, in which the P-arrivals influence the utilization of the length of the acceleration waveforms and P-wave travel times from multiple seismometers. The accuracy of the P-arrival picks should influence the proposed model's magnitude estimation performance. To obtain the effect on the magnitude estimation from the uncertainty of the P-arrival picks, we calculated the magnitudes 50 times by simulating erroneous P-arrival picks on the testing data set. We simulated erroneous P-arrival picks by adding random perturbations on P-arrival times from different absolute picking error bins. The ten absolute picks error bins are 0.0~0.1, 0.1~0.2, ..., 0.8~0.9, and 0.9~1.0 s; for example, the absolute picks error bin is 0.2~0.3 s; we added the random perturbations from -0.2~-0.3 or 0.2~0.3 s. Figure 7 shows the model's simulation results for different absolute picking error bins on the testing data set. Overall, the RMSE, MAE, and δ curves have greater values in a higher absolute picking error bin; the phenomenon means the higher absolute picking error brings a higher error level on magnitude estimation by the proposed model (shown in Fig. 7d–f). The comparison with the RMSE, MAE, and δ curves from different absolute picking error bins shows that the 0.0~0.1 s absolute picking error bin curves coincide with the manual picks curves, and the curves of the 0.1~0.2 s absolute picking error bin are slightly higher than those of the manual picks. However, the curves of the 0.2~1.0 s absolute picking error bin are significantly higher than those of the manual picks, which means a more significant error level on magnitude estimation and the proposed model may not obtain a reliable estimation when the absolute picking error exceeds 0.2 s. To quantitatively calculate the effect of the absolute P-arrival picking

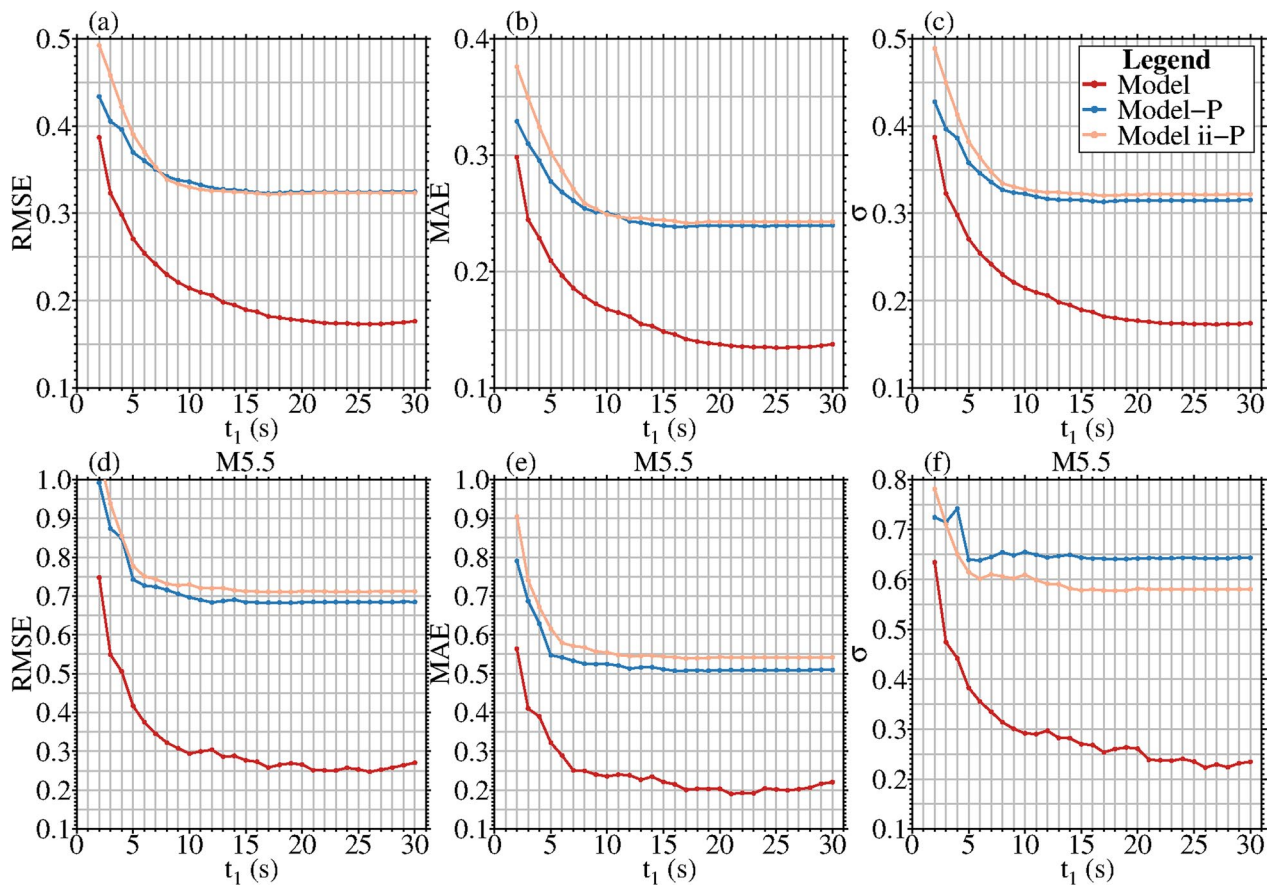


Fig. 6 RMSE (a and d), MAE (b and e), and $\bar{\epsilon}$ (c and f) curves with time (t_1) increasing only using P-wave on the testing data set when the triggered seismometers exceed three. The top figures indicate the curves with magnitudes ranging from 3.0 to 7.4. The bottom figures indicate the curves with magnitudes greater than 5.5. Different colors indicate different models

error bins, we chose the RMSE, MAE, and $\bar{\epsilon}$ errors of different picking error bins when t_1 is at 3 s and evaluated the effect based on the increased ratio of the error (RMSE, MSE and $\bar{\epsilon}$), as shown in Fig. 7a–c. The larger ratio means a higher increase in error levels on magnitude estimation which mean decreased performance. We calculate the ratio by setting the MAE (RMSE or $\bar{\epsilon}$) of the manual picks as the denominator, and the molecule is the error difference between the absolute picks error bin and the manual picks. The effects of the absolute P-arrival picking error bins are reflected similarly in the RMSEs, MAEs, and $\bar{\epsilon}$. Thus, we take the MAEs when t_1 is 3 s as an example (Fig. 7b). The MAE with 0.0~0.1 s absolute picking error bin range is 0.30, approximately the same as the MAE with manual picks. The MAE with 0.1~0.2 s absolute picking error bin range is 0.35 (20.70% increase of the MAE with manual picks), which could be considered a slightly higher increase in error level. To the 0.2~1.0 s absolute picking error bins, the MAEs are greater than 0.40, and the error level significantly increased, at least 37.90%. Thus, by combining the error curves with t_1

(Fig. 7d–f), the model could obtain a robust magnitude estimation when the P-wave arrivals picking errors are less than 0.2 s. Such accuracy is generally achievable with recently proposed picking methods (Mousavi et al. 2020).

Prospects

To better illustrate our proposed model’s magnitude estimation of the study region and develop its performance, we analyzed the spatial distribution of the magnitude prediction errors on the testing data set with time (t_1) at 1, 3, 5, 6, 7, and 8 s when the earthquake’s triggered seismometers exceed one (as shown in Fig. 8). For the total study region when t_1 is at 3 s (Fig. 8b), most of the prediction errors of the earthquakes that occurred in the inland region are greater than zero and those of the earthquakes that occurred in the offshore region are less than zero, within ± 0.5 range. With t_1 increases (Fig. 8c–f), the prediction errors of the inland and offshore have a tendency closer to zero, but slight overestimation and underestimation trends still exist. The phenomenon may mainly be related to the distances (epicenter distance and

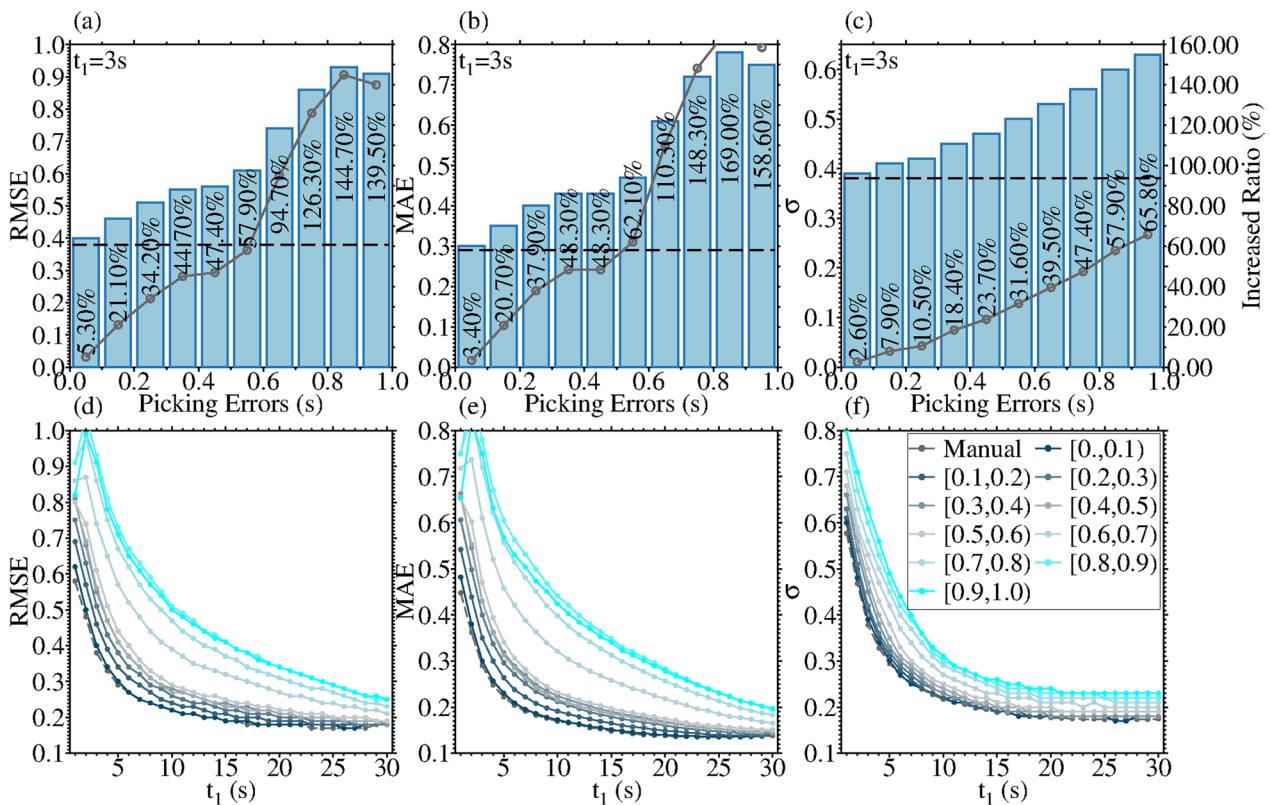


Fig. 7 Simulation results of the proposed model for different picking absolute error bins on the testing data set. **a–c** indicates the results about RMSE, MAE, and σ when time (t_1) is at 3 s. The black dashed lines indicate the results of the model with P-arrivals picked manually. The gray dashed lines indicate the ratio of increased error levels on the magnitude estimation performance from different picking absolute error bins compared with the results with manual P-arrivals. **d–f** indicates the error curves of the model for different picking absolute error bins with time (t_1) increasing. Different colors indicate different picking absolute error bins

depth) and the azimuths of the varying triggered seismometers. The triggered seismometers from the offshore region’s earthquake are generally farther, and the earthquakes have greater depth, in which the distances influence the amplitude of the real-time waveforms to lead to variations in magnitude estimation (Mousavi and Beroza 2020a). The influence of the azimuths is mainly reflected in two aspects: (1) the azimuths from offshore earthquakes are relatively simpler with sparse seismometers may bring uncertainty on P-arrival-based earthquake location (Saad et al. 2022b) and also introduce uncertainty on the implicit distance information provided by our proposed model; (2) the directivity related to the azimuths would also influence the amplitude of the real-time waveforms with varying distances. Figure 8a shows the influence more clearly than Fig. 8b–f; that is, most prediction errors exceeding 1.0 are close to the epicenter with $\Delta \leq 10$ km, and most underestimated prediction errors of the earthquakes occurred offshore when t_1 is at 1 s and only one seismometer is triggered. The directivity and distance influence could generally be decreased with more triggered seismometers, as shown in Fig. 8b–f.

To explore the detailed distance influence on the magnitude estimation of our proposed model, we selected the subset region of the testing data set (latitude from 35°N to 40°N and longitude from 139°E to 144°E), which contains massive earthquakes that occurred in the subduction. There are 399 earthquakes from inland with M_{JMA} ranging from 3.0 to 6.3. The depths and epicenter distances are less than 151 km and 100 km. For the offshore, 588 earthquakes with M_{JMA} from 3.0 to 7.4, the depth ranges from 0 to 90 km, and the epicenter distance is less than 200 km. Considering the potential difference of earthquakes among the crustal, subduction, and upper mantle and highlighting the epicenter distance influence as much as possible, we roughly class the offshore earthquakes into three depth segments based on the earthquake location classification methods (Zhao et al. 2015). The depth segments are 0 to 25 km (the crustal), 25 to 70 km (mainly focus on the subduction), and 70 to 160 km (the upper mantle). We also divided the inland earthquakes into three depth segments. The detailed distribution of the earthquakes can be found in the Additional file 1: Figure S1.

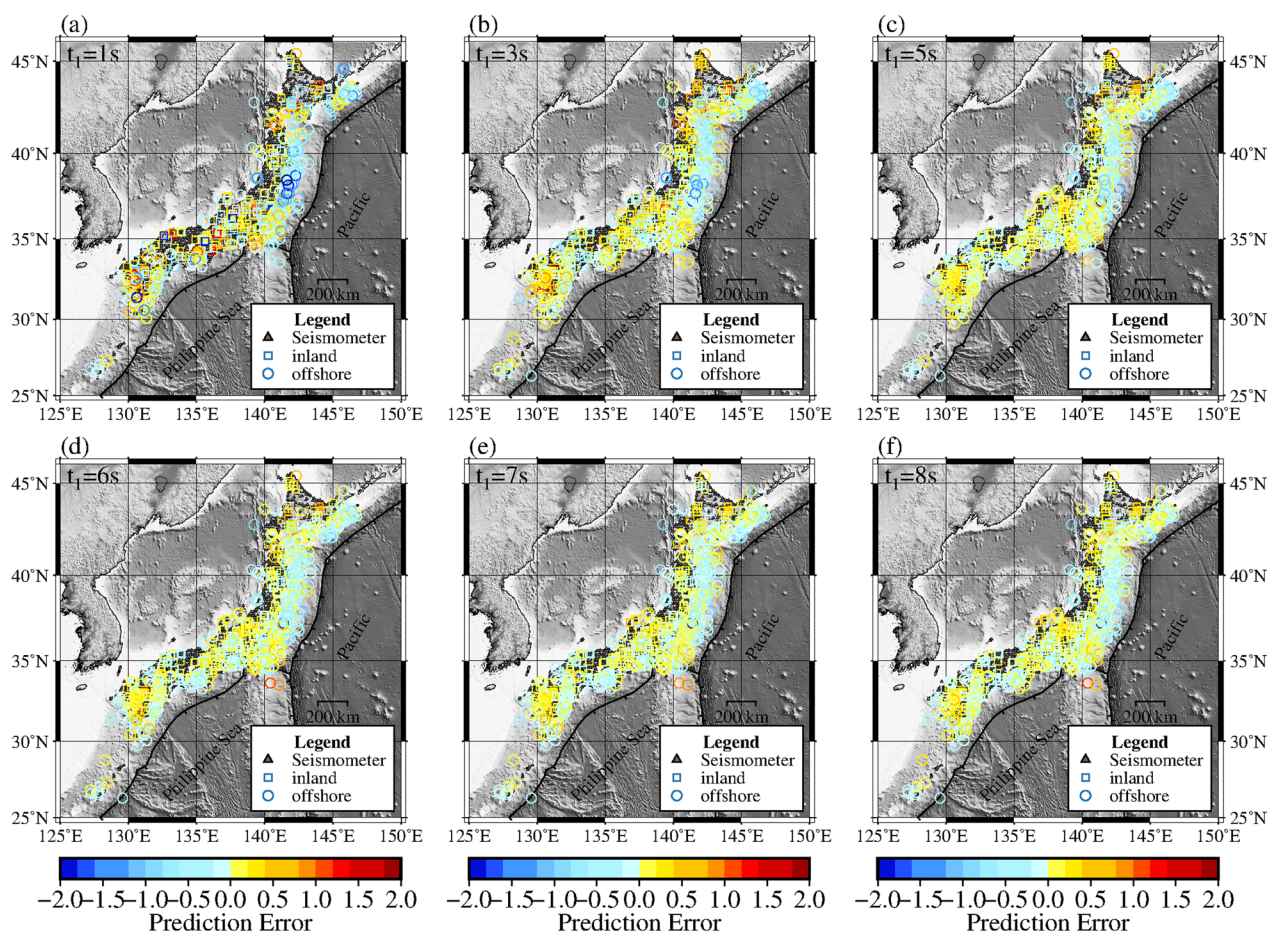


Fig. 8 Spatial distribution of the magnitude prediction errors of the proposed model in the study region with time (t_1) at 1 (a), 3 (b), 5 (c), 6 (d), 7 (e), and 8 (f) s when the earthquake’s triggered seismometers exceed one. The squares and circles represent the earthquakes located in the inland and offshore, respectively. Different colors indicate different error levels

Considering the epicenter distance variation with more triggered seismometers when t_1 increases, it is challenging to quantitatively show the distribution between the prediction errors and the epicenter distances. Thus, we qualitatively analyze the epicenter distance influence on the magnitude prediction errors of the proposed model. Figure 9 shows the spatial distribution of the magnitude prediction errors of the proposed model from three depth segments in the subset region when t_1 is at 1 and 3 s. As shown in Figs. 8, 9 indicates a more significant epicenter distance influence on the magnitude estimation prediction error of the proposed model, such as the prediction error results when t_1 is at 1 s and the depth ranges from 0 to 25 km (Fig. 9a). The prediction errors for earthquakes from 25 to 70 km depth segment are generally less than 0.0 with farther epicenter distance (Fig. 9b, e). However, several points with prediction error greater than 0.0 with farther epicenter distance exist. Apart from the potential magnitude saturation and epicenter distance

influence, the depth distance influences should also be considered. Meanwhile, we noticed that most prediction errors < 0 from inland earthquakes with a depth ≥ 70 km (Fig. 9c, f).

To analyze the depth influence on the magnitude estimation of our proposed model, we chose the distribution between the prediction error and depth when t_1 is from 1 to 8 s. Based on the error bars of the prediction errors corresponding to three depth segments, the prediction errors have an overall decreased trend with depth increases, and the trend weakens with t_1 increases, as shown in Fig. 10. The results show that the magnitude estimation of the proposed model varies with depth variation, in which the depth also influences the amplitude of the real-time waveforms. The mean of prediction errors for the offshore region with 25–70 km depth is slightly less than the other two depth segments, especially when $t_1 \leq 3$ s. Two aspects may mainly cause the phenomenon: (1) Compared with the other two

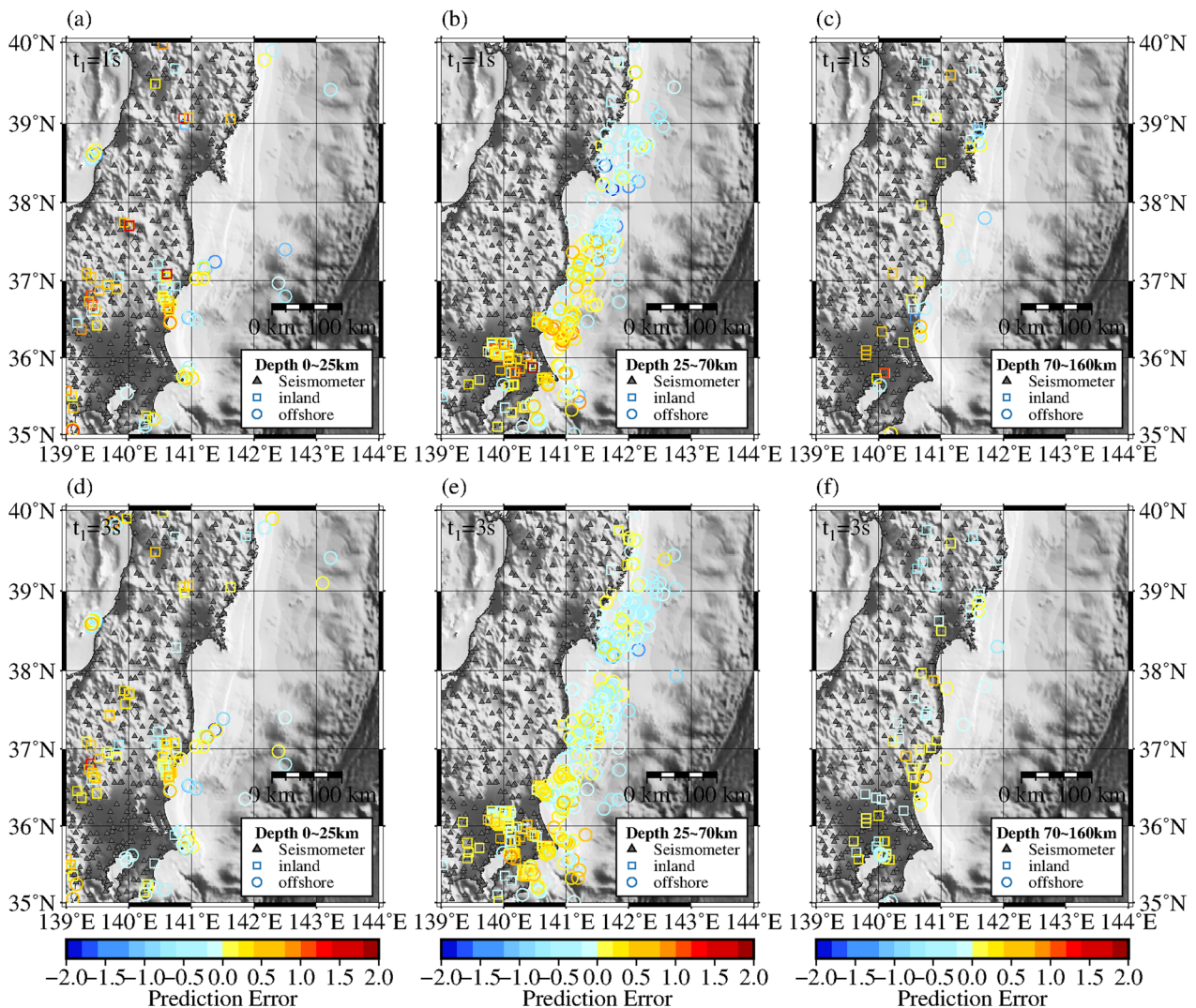


Fig. 9 Spatial distribution of the magnitude prediction errors of the proposed model in the subset region when the t_1 is at 1 and 3 s. The top and bottom subfigures indicate the results when t_1 is at 1 and 3 s, respectively. The left to right subfigures indicate the prediction errors of the earthquakes from different depth segments. The gray triangles represent seismometers. The squares and circles represent the earthquakes located in the inland and offshore, respectively. Different colors indicate the error levels

depth segments, the 25 to 70 km depth segment contains many earthquakes that occurred on the subduction, and the earthquakes have farther seismometers, as shown in Fig. 9 and Additional file 1: Figure S1. Thus, the earthquakes have higher distance levels, and the distance influence leads to a lower mean value of the earthquakes from the depth segment from 25 to 70 km. (2) The 25–70 km depth segment contains more earthquakes with a magnitude ≥ 5.5 than the other two. At the initial of the earthquakes, the earthquake rupture progress and duration may influence the mean prediction errors. Moreover, we will investigate how the azimuthal distribution of the seismometers will affect the magnitude estimation in the future.

To deal with the above issue of the different magnitude estimation performance for earthquakes that occurred inland and offshore regions, there are some potential ways, e.g., introducing more additional information related to distance and azimuth (Iwata et al. 2015; Noda et al. 2012), introducing a simple and rough two-class structure similar to the classifier from Yang et al. (2024) to make the model distinguish the earthquakes occurred region and expect the model automatically correct the difference. Moreover, we also notice that the subset region inland has slightly higher error levels relative to the whole region at the initial of the earthquake, such as the subset region (139°E/144°E/35°N/40°N), which is relative to the earthquake location distribution (occurred

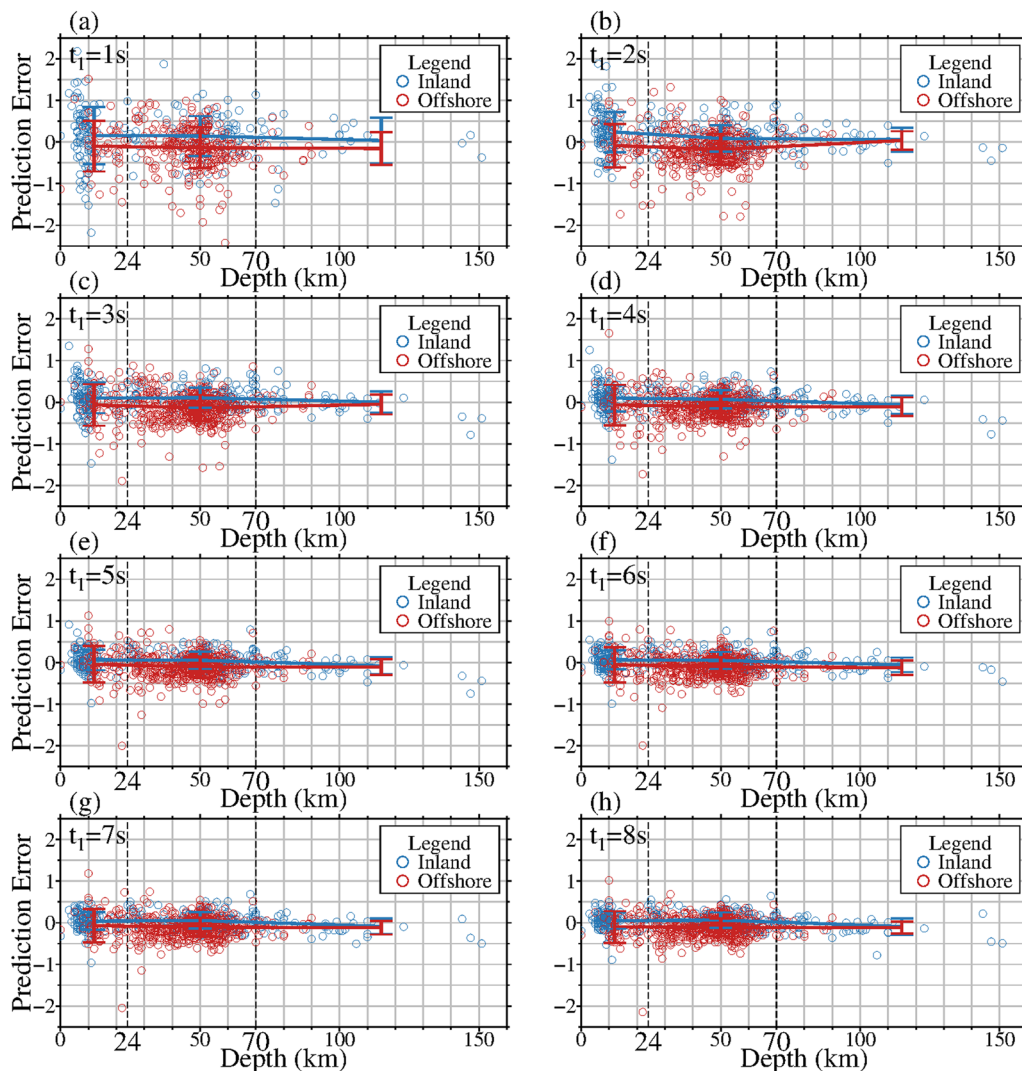


Fig. 10 Distribution of prediction errors with depth when t_1 is 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g), and 8 (h) s. The red color indicates the earthquake from offshore, and the blue color indicates the earthquake from inland. The solid line indicates the error bars from three depth segments (0–25 km, 25–70 km, 70–160 km)

on the crustal or upper mantle). For these subset regions, we will further investigate the potential region difference, and the model could be more suitable for these regions by transferring the current model using the earthquakes from the specific regions. Since the model has a relatively simple structure and a small volume of parameters, it does not seem difficult to transfer.

Conclusions

In this study, we designed a deep-learning network to estimate earthquake magnitude automatically, using three multimodalities: real-time acceleration waveforms, differential P-wave arrivals, and differential locations from multiple seismometers per earthquake. We

adopt a specific architecture to introduce the distance information on the magnitude estimation by automatically processing the variety of hypocenter location estimation. We trained and selected the model based on the training and validation data sets. We utilized the testing data set to evaluate the model on magnitude estimation performance. Compared with the error analysis from the classical and CNN-based methods, the proposed model performs less error level on magnitude estimation, especially on the high magnitude. The comparison of the results between model ii and the proposed model provides evidence that the specific architecture could introduce the distance information on magnitude estimation. In addition, the P-arrivals

picking error testing indicates the model has robustness on EEW with absolute error less than 0.2 s.

Abbreviations

CNN	Convolutional Neural Network
EEW	Earthquake Early Warning
Δ	Epicenter distance
FC	Fully Connected Layer
GAP	Global Average Pooling Layer
L	The differential seismometer locations
LSTM	Long short-term Memory
M_{JMA}	Japan Meteorological Agency Magnitude
MAE	Mean absolute error
MSE	Mean square error
σ	Standard Deviation Error
Relu	Rectified linear unit
RMSE	Root mean square error
SNR	Signal-to-noise ratio
T	The differential P-wave arrivals
t_1	The passing time after the first P-wave arrival per earthquake
t_i	The waveform length after P-wave arrival recorded on the i^{th} seismometer when t_1
t_{noise}	The noise length before P-wave arrival recorded on the i^{th} seismometer
Δt_i	The travel time between the i^{th} P-wave arrival seismometer and the earliest P-wave arrival seismometer
Δlat_i	The numerical latitude difference between the i^{th} and the earliest P-arrival seismometer
Δlon_i	The numerical longitude difference between the i^{th} and the earliest P-arrival seismometer

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40623-024-02005-8>.

Additional file 1: Figure S1. Earthquake map and data distribution of the subset-region. The subfigure a indicates the distribution of the seismometers and earthquake epicenters. Different colors indicate the earthquake depth range and different shapes indicate different geographical regions. Subfigure b indicates the distribution between depths and magnitudes. Subfigure c indicates the distribution between the depths and epicenter distances. The subfigures d to e indicate the frequency distribution of depths and epicenter distances. The different colors indicate different geographical regions.

Acknowledgements

Thanks to the Japanese National Research Institute for Earth Science and Disaster Resilience providing the data. Thanks to Ye Liu and Yuan Wang for providing help with the study. Thanks to Münchmeyer et al. (2021), opening source their codes.

Author contributions

Conceptualization, B.H., S.L. and J.S.; methodology, B.H.; software, B.H.; validation, B.H., Y.Z. and Y.X.; formal analysis, B.H., S.L. and J.S.; investigation, B.H.; resources, B.H., S.L. and J.S.; data curation, Y.Z., Y.X. and J.S.; writing—original draft preparation, B.H.; writing—review and editing, B.H., S.L. and J.S.; visualization, B.H. and J.S.; supervision, S.L. and J.S.; project administration, S.L. and J.S.; funding acquisition, S.L. and J.S.. All authors have read and agreed to the published version of the manuscript.

Funding

This research was financially supported by the National Natural Science Foundation of China (U2039209, 42304074 and 51408564).

Availability of data and materials

Seismic waveforms were provided by the Japanese National Research Institute for Earth Science and Disaster Resilience (NIED; <http://www.kyoshin.bosai.go.jp/>, last accessed December 2023), Tsukuba, Japan. The figures were made

using Genetic Mapping Tools (Wessel et al. 2013). The magnitude estimation model's codes were written with Python packages TensorFlow (www.tensorflow.org) and Sklearn (<https://scikit-learn.org/stable/>) (Pedregosa et al. 2011).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Laboratory of Earthquake Engineering and Engineering Vibration, Institute of Engineering Mechanics, China Earthquake Administration, Harbin 150080, Heilongjiang, China. ²Key Laboratory of Earthquake Disaster Mitigation, Ministry of Emergency Management, Harbin 150080, Heilongjiang, China. ³Fujian Earthquake Agency, Fuzhou 350003, Fujian, China.

Received: 11 October 2023 Accepted: 7 April 2024

Published online: 17 April 2024

References

- Aki K, Richards PG (2002) Quantitative seismology. University Science Books, Sausalito
- Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 41(2):423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>
- Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7(3):1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Colombelli S, Zollo A, Festa G, Picozzi M (2014) Evidence for a difference in rupture initiation between small and large earthquakes. *Nat Commun*. <https://doi.org/10.1038/ncomms4958>
- Colombelli S, Festa G, Zollo A (2020) Early rupture signals predict the final earthquake size. *Geophys J Int* 223(1):692–706. <https://doi.org/10.1093/gji/ggaa343>
- Festa G, Zollo A, Lancieri M (2008) Earthquake magnitude estimation from early radiated energy. *Geophys Res Lett*. <https://doi.org/10.1029/2008GL035576>
- Funasaki J, Division EPI (2004) Revision of the JMA velocity magnitude (in Japanese). *Quart J Seis* 67:11–20
- Gong Y, Chung YA, Glass J (2021) Ast: audio spectrogram transformer. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2104.01778>
- Heidari R (2018) tps, A new magnitude scaling parameter for earthquake early warning. *Bull Earthquake Eng* 16:1165–1177. <https://doi.org/10.1007/s10518-017-0256-x>
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Iwata N, Yamamoto S, Korenaga M, Noda S (2015) Improved algorithms of seismic parameters estimation and noise discrimination in earthquake early warning. *Q Rep RTRI* 56(4):291–298. https://doi.org/10.2219/rtriqr.56.4_291
- Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Annu Rev Earth Planet Sci* 33:195–214. <https://doi.org/10.1146/annurev.earth.33.092203.122626>
- Katsumata A (2004) Revision of the JMA displacement magnitude (in Japanese). *Quart J Seis* 67:1–10
- Kuang W, Yuan C, Zhang J (2021) Network-based earthquake magnitude determination via deep learning. *Seismol Res Lett* 92(4):2245–2254. <https://doi.org/10.1785/0220200317>
- Kuyuk HS, Allen RM (2013) A global approach to provide magnitude estimates for earthquake early warning alerts. *Geophys Res Lett* 40(24):6329–6333. <https://doi.org/10.1002/2013GL058580>

- Lahat D, Adali T, Jutten C (2015) Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc IEEE* 103(9):1449–1477. <https://doi.org/10.1109/JPROC.2015.2460697>
- Liu D, Wang Z, Wang L, Chen L (2021) Multi-modal fusion emotion recognition method of speech expression based on deep learning. *Front Neurobot*. <https://doi.org/10.3389/fnbot.2021.697634>
- Lomax A, Michelini A, Jozinović D (2019) An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network. *Seismol Res Lett* 90(2A):517–529. <https://doi.org/10.1785/0220180311>
- Luo H, Ji L, Zhong M, Chen Y, Lei W, Duan N, Li T (2022) Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508:293–304. <https://doi.org/10.1016/j.neucom.2022.07.028>
- Meier MA, Ross ZE, Ramachandran A, Balakrishna A, Nair S, Kundzicz P, Li Z, Andrews J, Hauksson E, Yue Y (2019) Reliable real-time seismic signal/noise discrimination with machine learning. *J Geophys Res Solid Earth* 124(1):788–800. <https://doi.org/10.1029/2018JB016661>
- Moriwaki K (2017) Automatic detection of low-frequency earthquakes in southwest Japan using Matched-Filter technique. *Quart J Seis* 81:3 (in Japanese with English abstract). <https://earth-planets-space.springeropen.com/articles/10.1186/s40623-018-0915-4>
- Mousavi SM, Beroza GC (2020a) A machine-learning approach for earthquake magnitude estimation. *Geophys Res Lett*. <https://doi.org/10.1029/2019GL085976>
- Mousavi SM, Beroza GC (2020b) Bayesian-deep-learning estimation of earthquake location from single-station observations. *IEEE Trans Geosci Remote Sens* 58(11):8211–8224. <https://doi.org/10.1109/TGRS.2020.2988770>
- Mousavi SM, Zhu W, Sheng Y, Beroza GC (2019) CRED: a deep residual network of convolutional and recurrent units for earthquake signal detection. *Sci Rep*. <https://doi.org/10.1038/s41598-019-45748-1>
- Mousavi SM, Ellsworth WL, Zhu W, Chuang LY, Beroza GC (2020) Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nat Commun*. <https://doi.org/10.1038/s41467-020-17591-w>
- Münchmeyer J, Bindi D, Leser U, Tilmann F (2021) Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. *Geophys J Int* 226(2):1086–1104. <https://doi.org/10.1093/gji/ggab139>
- Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. *Adv Neural Inf Process Syst* 34:14200–14213
- Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Fürnkranz J, Joachims T (eds) *ICML'10: topics in artificial intelligence*. 27th International Conference on Machine Learning, Haifa, June 2010. Papers in machine learning research, p 807–814. National Research Institute for Earth Science and Disaster Resilience (2019) NIED K-NET, KiK-net. National Research Institute for Earth Science and Disaster Resilience. <https://doi.org/10.17598/nied.0004>
- Noda S, Yamamoto S, Sato S (2012) New method for estimating earthquake parameters for earthquake early warning. *Q Rep RTRI* 53(2):102–106. <https://doi.org/10.2219/rtrirq.53.102>
- Okada Y, Kasahara K, Hori S, Obara K, Sekiguchi S, Fujiwara H, Yamamoto A (2004) Recent progress of seismic observation networks in Japan—Hi-net, F-net, K-NET and KiK-net—. *Earth Planets Space* 56:xxv–xxviii
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
- Perol T, Gharbi M, Denolle M (2018) Convolutional neural network for earthquake detection and location. *Sci Adv*. <https://doi.org/10.1126/sciadv.1700578>
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) *ICML'21: topics in artificial intelligence*. 38th International Conference on Machine Learning, Virtual, July 2021. Papers in machine learning research, vol 139, PMLR, Virtual, p 8748–8763.
- Reynen A, Audet P (2017) Supervised machine learning on a network scale: Application to seismic event classification and detection. *Geophys J Int* 210(3):1394–1409. <https://doi.org/10.1093/gji/ggx238>
- Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seismol Soc Am* 25(1):1–32. <https://doi.org/10.1785/BSSA0250010001>
- Saad OM, Chen Y, Savvaidis A, Fomel S, Chen Y (2022a) Real-time earthquake detection and magnitude estimation using vision transformer. *J Geophys Res Solid Earth*. <https://doi.org/10.1029/2021JB023657>
- Saad OM, Chen Y, Trugman D, Soliman MS, Samy L, Savvaidis A, Khamis MA, Hafez AG, Fomel S, Chen Y (2022b) Machine learning for fast and reliable source-location estimation in earthquake early warning. *IEEE Geosci Remote Sens Lett* 19:1–5. <https://doi.org/10.1109/LGRS.2022.3142714>
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15(1):1929–1958
- Trugman DT, Page MT, Minson SE, Cochran ES (2019) Peak ground displacement saturates exactly when expected: implications for earthquake early warning. *J Geophys Res Solid Earth* 124(5):4642–4653. <https://doi.org/10.1029/2018JB017093>
- Tsubi C (1954) Determination of the Gutenberg-Richter's magnitude of earthquakes occurring in and near Japan. *Zisin*, II 7:185–193
- Van Den Ende MPA, Ampuero JP (2020) Automated seismic source characterization using deep graph neural networks. *Geophys Res Lett*. <https://doi.org/10.1029/2020GL088690>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In Isabelle G, Ulrike VL, Samy B, Hanna MW, Rob F, SVN Vishwanathan, Roman G (eds) *NIPS'17: Advances in Neural Information Processing Systems*. 30th Annual Conference on Neural Information Processing Systems, Long Beach, December 2017. Papers in Advances in Neural Information Processing Systems 30 (NIPS 2017), Virtual, p 5998 – 6008. <https://dblp.org/db/conf/nips/nips2017.html> The paper's BibTeX link is: <https://dblp.org/rec/conf/nips/VaswaniSPUJGK17.html?view=bibtex>
- Wang Y, Li S, Song J (2021a) Magnitude-scaling relationships based on initial P-wave information in the Xinjiang region. *China J Seismol* 25(2):697–710. <https://doi.org/10.1007/s10950-020-09981-w>
- Wang M, Xing J, Liu Y (2021b) Actionclip: a new paradigm for video action recognition. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2109.08472>
- Wessel P, Smith WHF, Scharroo R, Luis J, Wobbe F (2013) Generic mapping tools: improved version released[J]. *Eos Trans Am Geophys Union* 94(45):409–410. <https://doi.org/10.1002/2013EO450001>
- Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning. *Geophys Res Lett*. <https://doi.org/10.1029/2006GL026871>
- Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T (2020) On layer normalization in the transformer architecture. In Hal Daumé III, Singh A (eds) *ICML'20: topics in artificial intelligence*. 37th International Conference on Machine Learning, Virtual, July 2020. Papers in machine learning research, vol 119, PMLR, Virtual, p 10524–10533.
- Yamada M, Mori J (2009) Using tc to estimate magnitude for earthquake early warning and effects of near-field terms. *J Geophys Res Solid Earth*. <https://doi.org/10.1029/2008JB006080>
- Yamada M, Tamaribuchi K, Wu S (2021) The extended integrated particle filter method (IPFx) as a high-performance earthquake early warning system. *Bull Seismol Soc Am* 111(3):1263–1272. <https://doi.org/10.1785/0120210008>
- Yang C, Zhang K, Chen G, Pan Y, Zhang L, Qu L (2024) Application of machine learning to determine earthquake hypocenter location in earthquake early warning. *IEEE Geosci Remote Sens Lett*. <https://doi.org/10.1109/LGRS.2023.3348107>
- Zhao JX, Zhou S, Gao P, Long T, Zhang Y, Thio HK, Lu M, Rhoades DA (2015) An earthquake classification scheme adapted for Japan determined by the goodness of fit for ground-motion prediction equations. *Bull Seismol Soc Am* 105(5):2750–2763. <https://doi.org/10.1785/0120150013>
- Zhu W, Beroza GC (2019) PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophys J Int* 216(1):261–273. <https://doi.org/10.1093/gji/ggy423>
- Zhu H, Wang Z, Shi Y, Hua Y, Xu G, Deng L (2020) Multimodal fusion method based on self-attention mechanism. *Wirel Commun Mob Comput*. <https://doi.org/10.1155/2020/8843186>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.