IEEE Geoscience and Remote Sensing Letters and Machine Learning,Volume:19,Issue Date:12.January.2022

Machine learning for fast and reliable source-location estimation in earthquake early warning

Omar M. Saad, Yunfeng Chen, Daniel Trugman, M. Sami Soliman, Lotfy Samy, Alexandros Savvaidis, Mohamed A. Khamis, Ali G. Hafez, Sergey Fomel, and Yangkang Chen

***Abstract*—We develop a random forest (RF) model for rapid earthquake location with an aim to assist earthquake early warning (EEW) systems in fast decision making. This system exploits P-wave arrival times at the first five stations recording an earthquake and computes their respective arrival time differences relative to a reference station (i.e., the first recording station). These differential P-wave arrival times and station locations are classified in the RF model to estimate the epicentral location. We train and test the proposed algorithm with an earthquake catalog from Japan. The RF model predicts the earthquake locations with a high accuracy, achieving a Mean Absolute Error (MAE) of 2.88 km. As importantly, the proposed RF model can learn from a limited amount of data (i.e., 10% of the dataset) and much fewer (i.e., three) recording stations and still achieve satisfactory results (MAE$<$5 km). The algorithm is accurate, generalizable, and rapidly responding, thereby offering a powerful new tool for fast and reliable source-location prediction in EEW.**

***Index Terms*—Earthquake Early Warning (EEW) system; Machine learning; Earthquake Location.**

I. Introduction

EARTHQUAKE hypocenter localization is essential in the field of seismology and plays a critical role in a variety of seismological applications such as tomography, source characterization, and hazard assessment. This underscores the importance of developing robust earthquake monitoring systems for accurately determining the event origin times and hypocenter locations. In addition, the rapid and reliable characterization of ongoing earthquakes is a crucial, yet challenging, task for developing seismic hazard mitigation tools like earthquake early warning (EEW) systems [1]. While classical methods have been widely adopted to design EEW systems, challenges remain to pinpoint hypocenter locations in real-time largely due to limited information in the early stage of earthquakes. Among various key aspects of EEW, timeliness is a crucial consideration and additional efforts are required to further improve the hypocenter location estimates with minimum data from 1) the first few seconds after the P-wave arrival and 2) the first few seismograph stations that are triggered by the ground shaking.

The localization problem can be resolved using a sequence of detected waves (arrival times) and locations of seismograph stations that are triggered by ground shaking. Among various network architectures, the recurrent neural network (RNN) is capable of precisely extracting information from a sequence of input data, which is ideal for handling a group of seismic stations that are triggered sequentially following the propagation paths of seismic waves. This method has been investigated to improve the performance of real-time earthquake detection [2] and classification of source characteristics. Other machine learning based strategies have also been proposed for earthquake monitoring. Comparisons between traditional machine learning methods, including the nearest neighbor, decision tree, and the support vector machine, have also been made for the earthquake detection problem [3]. However, a common issue in the aforementioned machine learning based frameworks is that the selection of input features often requires expert knowledge, which may affect the accuracy of these methods. Convolution neural networks-based clustering methods have been used to regionalize earthquake epicenters [4] or predict their precise hypocenter locations [5]. In the latter case, three-component waveforms from multiple stations are exploited to train the model for swarm event localization.

In this study, we propose a RF-based method to locate earthquakes using the differential P-wave arrival times and station locations (Figure 1). The proposed algorithm only relies on P-wave arrival times detected at the first few stations. Its prompt response to earthquake first arrivals is critical for rapidly disseminating EEW alerts. Our strategy implicitly considers the influence of the velocity structures by incorporating the source-station locations into the RF model. We evaluate the proposed algorithm using an extensive seismic catalog from Japan. Our test results show that the RF model is capable of determining the locations of earthquakes accurately with minimal information, which sheds new light on developing efficient machine learning.

II. Method

A. Differential traveltime based epicenter prediction

To estimate the epicenters of earthquakes, the RF model is trained via a supervised learning scheme. Two sets of properties, including the differential P-wave arrival times and station locations, are utilized as the model input, which can be expressed as

$$X = [T_i, Y_i, Z_i], \tag{1}$$

where T_i represents the P-wave travel time of the i^{th} station relative to that of a reference station, and Y_i and Z_i are the corresponding latitude and longitude of the target station. In this study, we set the P-wave arrival time at the first recording

O.M.Saad, M.S. Soliman, L.Samy, and A.G. Hafez are with National Research Institute of Astronomy and Geophysics (NRIAG), Egypt.

Yunfeng Chen is with Zhejiang University, China.

A. Savvaidis, S. Fomel, D. Trugman, and Yangkang Chen are with The University of Texas at Austin, USA.

M.A. Khamis is with E-JUST, Egypt.

The research is partially supported by Texas Seismological Network Project, the University of Texas at Austin Rising STARs program and Texas Consortium of Computational Seismology (TCCS).

Manuscript received mm dd, yyyy; revised mm dd, yyyy.

station as the reference (i.e., $T_i = t_i - t_1$) and utilize five stations to locate the earthquakes. The input parameters consist of a total of 14 features that are defined as

$$\begin{aligned} T_i &= \{t_2 - t_1, t_3 - t_1, t_4 - t_1, t_5 - t_1\}, \\ Y_i &= \{y_1, y_2, y_3, y_4, y_5\}, \\ Z_i &= \{z_1, z_2, z_3, z_4, z_5\}. \end{aligned} \tag{2}$$

The combination of these features enables the network to determine the relative location (i.e., the latitude/longitude difference) between the earthquake and the reference station.

B. Random Forest (RF)

The final output of a RF model is obtained by averaging the predictions from K trees as

$$\bar{H}(X) = (\frac{1}{K}) \sum_{k=1}^{K} H(X; \theta_k), \tag{3}$$

where $H(X; \theta_k)$ denotes the k^{th} predictor tree, and θ represents the random vector of the RF [6]. The supervised learning is achieved by minimizing the following loss function

$$\sum_{n=1}^{N} (O_n - \bar{H}(X)_n)^2, \tag{4}$$

where O denotes the latitude and longitude difference between the event and the reference station and N represents the number of training earthquakes. We tune two hyperparameters, the maximum number of trees ($mtree$) and the maximum depth of each tree ($mdep$), during the training process. The training of each tree is conducted by randomly drawing M records [6] from the training earthquakes, with a sampling ratio (MS) that varies between 0 and 1. Each node in a decision tree (except for the leaf node) is split into more branches while considering a random subset of features, with the number of features represented by MF. The training process is performed through the following steps:

A) Growing the number of trees to $mtree$.
B) Picking M random records according to the MS factor.
C) Randomly splitting each tree into $mdep$ levels.
D) Randomly selecting the MF at each splitting node.
E) Obtaining the averaging of the $mtree$ trees outputs according to Equation. 3.
F) Obtaining the loss function according to Equation. 4.
G) Repeating steps B-F until the loss function converges.

C. The Architecture of RF

The robust performance of the RF model relies on well-designed network architecture. We tune its parameters by a trial-and-error approach. Firstly, we test the number of trees ($mtree$) from 500 to 10000 with an interval of 500, and the depth of each tree ($mdep$) from 10 to 200 with an interval of 5. The optimal values of $mtree$ and $mdep$ are 1000 and 100, respectively. Secondly, we increase the MS factor from 0.1 to 1 with an interval of 0.1. A MS factor of 1 achieves the optimal result, which indicates that all records contribute positively to the training process. Finally, we test a series of MF, e.g, 2, 3, 5, 7, 8, 9, 11, and 13, and the optimal value is determined to be 8. The architecture of the proposed algorithm is shown in Figure 1.

III. Results

A. Dataset and Model Inputs

We apply the proposed network to an earthquake detection problem in Japan (Figure 2a). We use as ground truth a unified catalog reported by the National Research Institute for Earth Science and Disaster Resilience, the Japan Meteorological Agency, and various institutions. This large catalog includes 2,235,159 regional seismic events recorded by the Hi-net seismic network between January 1^{st}, 2009 and November, 11^{th} 2020. For each event, we extract the source parameters including arrival times, magnitudes, depths, latitudes, and longitudes, as well as the locations of recording stations. We define qualified events for further analysis as those satisfying the following criteria: A) P-wave arrivals are detected at a minimum of five stations, B) Epicenter distances are less than 1° ($\approx$112 km), and C) Magnitudes of the events are greater than 0 M_L. These criteria facilitate a rapid response to earthquakes while ensuring relatively reliable predictions. The final catalog, which contains a total of 1,692,787 qualified events, is characterized by a broad distribution of source parameters and offers an ideal dataset to train and test the proposed algorithm. In this catalog, longitude varies between 121.86° and 146.48° and the latitude between 23.42° and 46.22°. Event magnitude ranges from 0.10 M_L to 7.59 M_L, and depths from 0 to 440.78 km. Note that the intermediate (80-300km) and deep (300+km) events in the training dataset only affect the location accuracy marginally according to some tests we have performed. To train the network, we set the earliest arrival time of a group of P waves as the reference time (t_1), and determine its differential travel times relative to later P phases recorded at the other stations (T_i). The latitudes (Y_i) and the longitudes (Z_i) of the recording stations are also used as input parameters for the RF model (Figure 1). Finally, a total of 1,541 stations from the Hi-net network are included in the training process (Figure 2a).

B. Training and Testing in the Proposed Algorithm

We randomly split the dataset into 90% for training and 10% for testing, which consist of 1,523,508 and 169,279 events, respectively (Figures 2b and 2c). We first train the RF network for the optimal architecture and then predict the event locations of the test dataset. To quantify the accuracy of the predictions, we calculate the Mean Absolute Errors (MAEs) of the latitude and longitude between the reported values from the catalog and those estimated by the RF model. We achieve MAEs of 0.015° ($\approx$1.625 km) and 0.023° ($\approx$2.553 km) for the latitude (Figure 2d) and longitude (Figure 2e), respectively, with corresponding standard deviations of 0.033° and 0.052°. These location errors lead to a distance MAE value of 2.879 km (Figure 2f). The resulting R^2 score reaches 0.9998, which suggests highly consistent values between the predicted and the catalog locations. We define the events with a distance error below 0.1° ($\approx$11.2 km) as true positive (TP), otherwise are false positive (FP), and calculate the accuracy ($\frac{TP}{TP+FP}$). The resulting accuracy rate is 94.39%. We further examine spatial variation in the accuracy across the study region (Figure 3a). To better illustrate our location results, we select a subset of events in central Japan (Figures 3b), where

IEEE Geoscience and Remote Sensing Letters and Machine Learning,Volume:19,Issue Date:12.January.2022

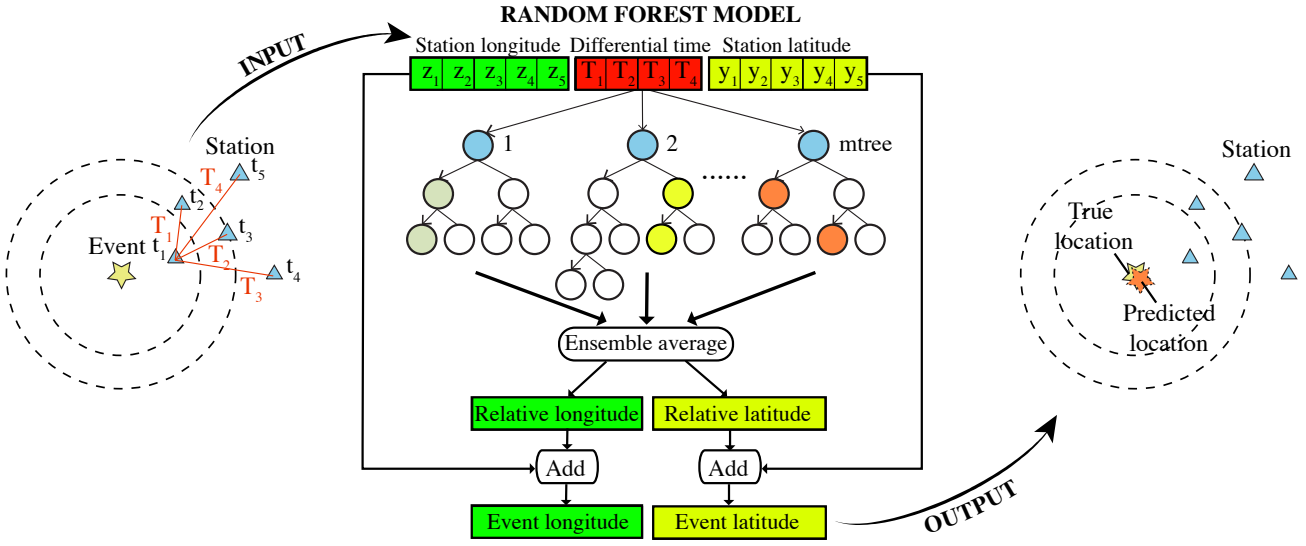

Fig. 1: The architecture of the proposed source-location prediction framework.

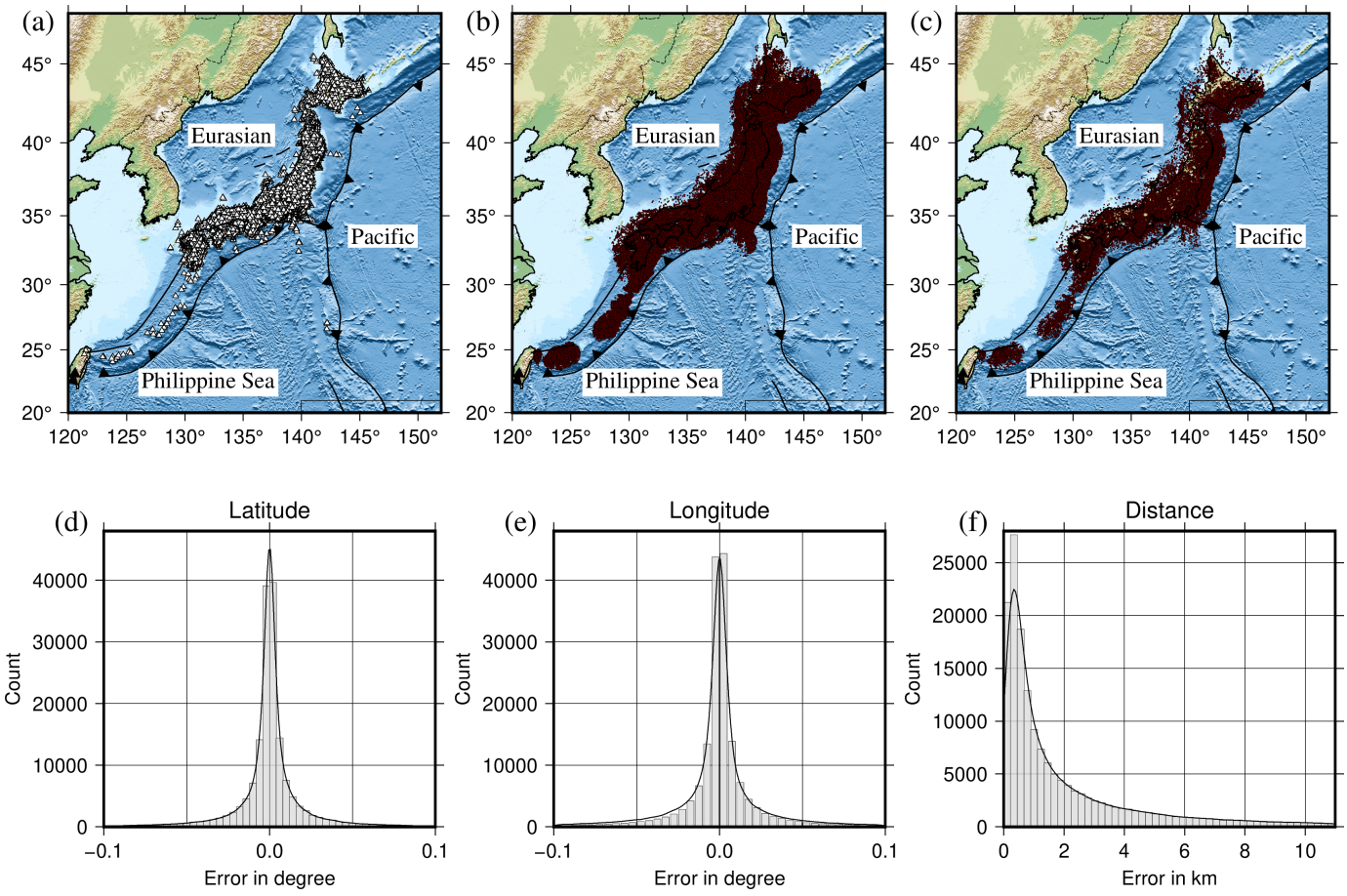

Fig. 2: (a) The distribution of the Japanese seismic stations. (b) The distribution of the training events. (c) The distribution of the testing events. (d) The error distribution between the catalog latitude and the estimated latitude corresponding to the proposed algorithm. (e) The error distribution between the catalog longitude and the estimated longitude corresponding to the proposed algorithm. (f) The error distribution between the catalog location and the estimated location corresponding to the proposed algorithm.

we observe relatively small errors in predicted locations if the first recording station is closer to the event. The distance errors of the estimated locations for all testing events show an overall small uncertainty of less than 0.1° ($\approx$11.2 km), with slightly larger errors observed near the coastal regions (Figure 3c). This pattern is primarily caused by the varying station density and azimuthal station coverage, which is relatively sparse in the offshore region and limits the accuracy of location prediction. In the future, we will investigate how azimuthal distribution of stations will affect the accuracy of localization. Figure 3d shows the spatial distribution of ray density. The ray path density of each cell in a 250×250 grid is calculated by counting the number of rays intersecting that cell. Comparing Figures 3c and 3d, it is clear that the lowest ray-density area (e.g., around the coast) has the largest location-prediction error. For those high ray-density areas, e.g., the interior of Japan, the prediction errors are generally small. The effects of station density (e.g., number and distance to earthquake) on event localization will be examined in detail in the discussion section. We perform a K-fold cross-validation test, where the dataset is randomly split into K partitions, with (K-1) partitions used for training and the remaining (one) partition for testing. This process is repeated K times, with each time picking a different partition for testing. The average MAE from all K (herein K=10) testing sets shows a distance error of 2.865 km and an accuracy rate of 94.44%, which are comparable with the results of the original training. Based on these testing results, we suggest that, despite varying training datasets, the proposed algorithm can accurately locate the earthquake with limited P-wave arrival time information, which is crucial for designing an effective EEW system. Additionally, we assess the proposed algorithm by training the RF model with a varying degree of training/testing splitting ratio, e.g, from 10% to 90%. The resulting average MAEs of the distance errors indicate a robust performance even for the pessimistic case of 10% of the training data (169,279 events), achieving an MAE value of 4.468 km (Figure 4e).

IV. Discussion

A. Depth estimation

In this section, we propose to estimate the depth using two approaches. The first approach is adding a third output for the current framework (Figure 1), which denotes the depth. The second approach is designing a separate RF model for the depth estimation using the same input of the proposed framework in Figure 1. In the latter case, both models (depth and latitude/longitude models) can work in parallel to produce the hypocenter location of the detected earthquake. We train both models separately using the same training and testing sets in Section III(B), and both models obtain the same test results. For both models, the depth error of the testing set has an MAE of 3.8 km compared to the catalog depth.

B. Effect of network density and application to other areas

A dense network, such as the Hi-net investigated in this study, is a fundamental requirement for an effective EEW system. To assess the robustness of the proposed method, we test less ideal application scenarios with relatively sparse seismic networks by randomly eliminating 20%, 40%, and 60% from 1,541 Hi-net seismic stations (Figures 4a-c). The overall prediction error (measured as MAE) decreases semi-linearly with an increasing number of stations in the network. The spatial distribution of MAE also reflects a similar trend where smaller values generally exist in regions with a denser ray path coverage. The test with all stations preserved leads to the smallest average MAE of 3.481 km (Figure 4d), with more than 80% percent of land areas achieving an MAE value less than 0.1°, which suggests a robust prediction ability given the current setup of Hi-net. The average station spacing for the four cases in Figures 4a-d) are 38 km, 34 km, 29 km , and 24 km, respectively. A list of prediction errors measured by MAE in latitude, longitude, and distance is summarized in Table I. All these station density tests do not explicitly consider regional velocity structures, as often required in earthquake location studies. This highlights an advantage of the proposed algorithm that we may bypass the velocity model building provided a sufficiently dense network and well-distributed events. In fact, the velocity model is hidden inside the arrival time difference and can be inverted using a similar framework, though at the risk of introducing systematic errors. The network density test can be considered as a stress test, where it shows when the proposed algorithm fails. The station density is a crucial factor for the proposed algorithm, i.e., a larger station density leads to robust learning capability of the proposed algorithm. According to Table I, when the number of stations decreases to 616 (40% of the whole stations), the MAE dramatically increases to 15.983 km. Thus, we can conclude that the proposed algorithm fails when the network density is small. For areas with less dense station coverage, it could be beneficial to include more knowledge into the machine learning framework about the subsurface structure and ray paths for compensating the deficiency caused by the shortage of stations.

C. The Effect of the Number of Available Stations

The number of stations is a critical factor that determines the data availability and prediction accuracy. The proposed RF model takes the arrival times of P waves recorded at multiple stations as the input, hence a more stringent requirement of simultaneous recording at an increased number of stations lowers the availability of qualified events. To properly evaluate the effect of the station quantity on the prediction, we select a subset of the events (425,802 earthquakes) that are recorded simultaneously by 20 stations, the maximum value considered in our tests. This ensures the same baseline (i.e., training and testing dataset) for testing cases with varying station numbers, which minimizes the potential prediction bias caused by a different spatial distribution of events. To mimic the real situation in an EEW system, we use the first few stations, which vary from 3 to 20, that record an earthquake for location prediction. The MAEs of predicted locations show that the distance errors decrease from 2.586 km to 1.964 km along with an increasing number of stations (Figure 4f). The most significant drop-off of MAE is observed at station number

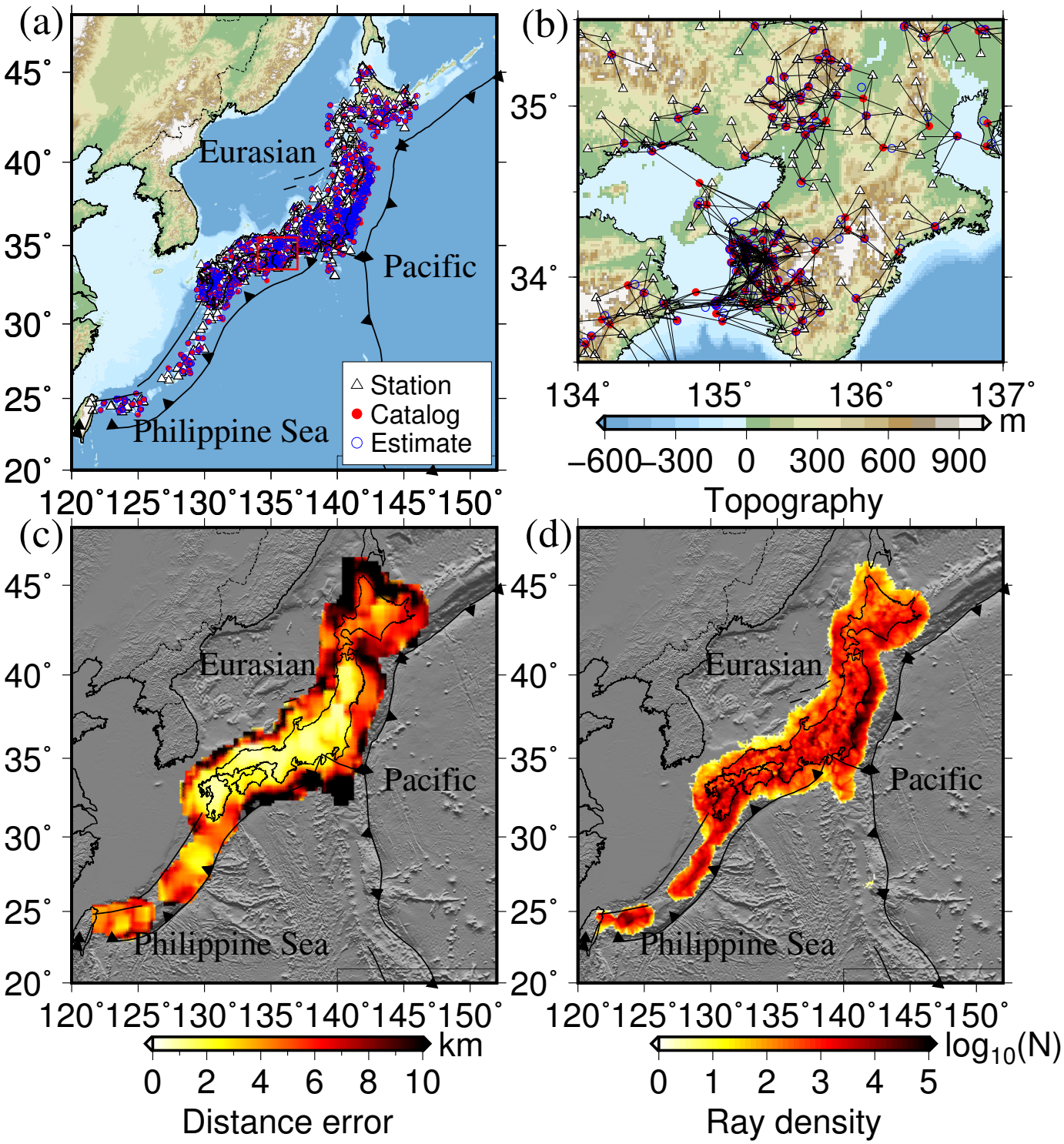

Fig. 3: (a) A comparison between predicted (blue circles) and catalog locations (red dots) for 1000 randomly selected testing events. The white triangles show the Hi-net stations. (b) A zoomed map from (a), with its location highlighted by the red rectangle. The black lines connect a catalog location to the five nearby stations. (c) The spatial distribution of location prediction errors across the study region. (d) The spatial distribution of ray path density.

5, beyond which the MAE converges quickly with only a minimal decrease of 0.011 km at the station number of 20. The accuracy of the test cases is inversely related to the MAE, showing an increasing trend from 95.65% to 97.69% with a similar turning point defined at the station number of 5 (Figure 4f). The distance error generally shows a similar distribution for all test cases (Figure 4(g)). Based on these test results, we suggest that our RF algorithm is capable of making useful predictions once at least five stations have been triggered.

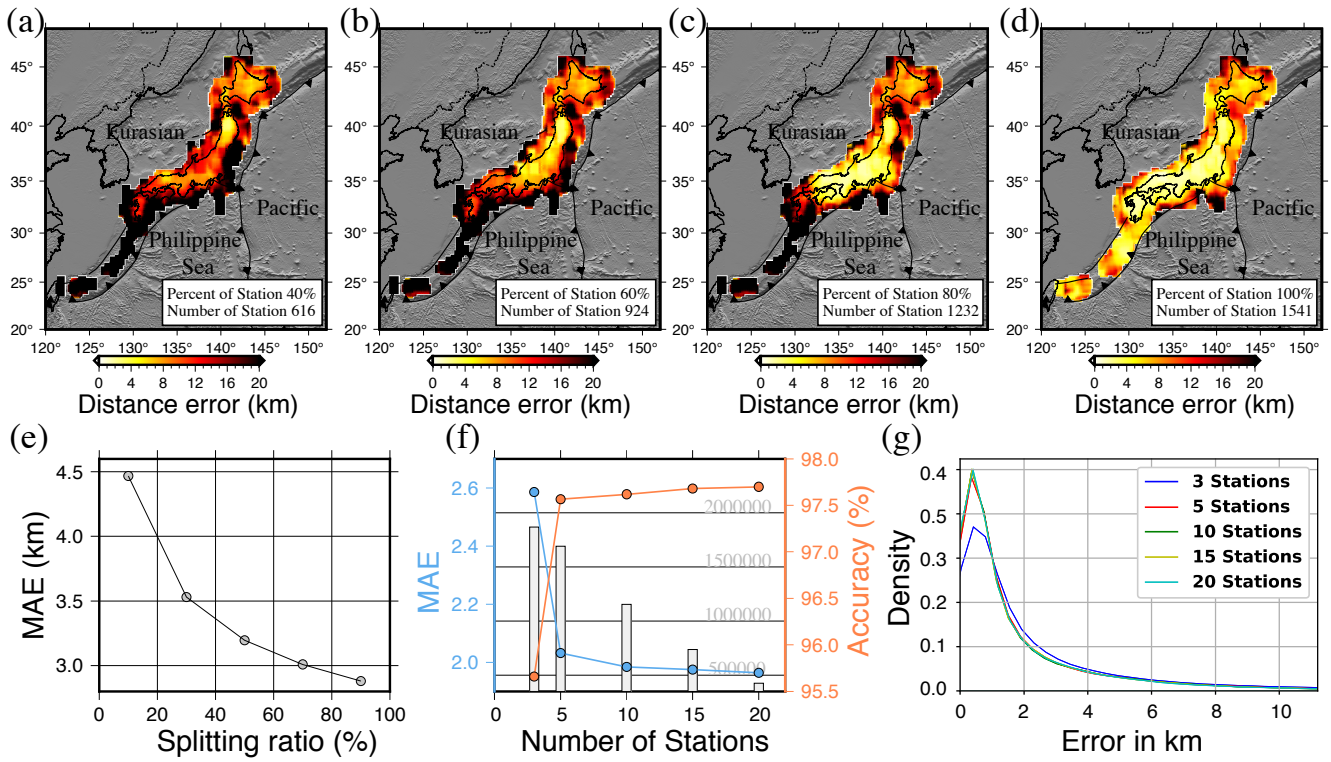

Fig. 4: The network station density test using 40% (a), 60% (b), 80% (c), 100% (d) of the total stations for training. The number of testing events is 100,000 and remains fixed for all test cases. (e) Variation of MAE as a function of splitting ratio of the training data. (f) Variation of MAE and accuracy of the predicted location with different numbers of stations. The histogram corresponds to the number of available events for each test case. (g) The distribution of the distance error for varying numbers of stations.

TABLE I: The results of the RF model using fewer seismic stations for training. The number of testing events is 100,000.

Stations Number	Training Events Number	Latitude MAE (°)	Longitude MAE (°)	Distance MAE (km)
1,541(100%)	1,592,787	0.019	0.025	3.481
1,232(80%)	865,068	0.049	0.058	8.552
924(60%)	257,205	0.065	0.084	11.855
616(40%)	133,798	0.088	0.116	15.983

D. The Effect of the P-wave Arrival Time

The proposed algorithm exploits the P-wave arrival times at the first few recording stations, hence its performance is dependent on the accuracy of the picked P-wave arrivals. We perform a hypothetical test by simulating erroneous P-wave picks, whereby random travel-time perturbations from -0.2s to 0.2s are added to P-wave arrival times. This set of new observations are used to predict the event location following the same procedure. The average MAE from ten simulations is 3.545 km, which accounts for an approximately 23% increase of the result (2.879 km) from a set of otherwise accurate P-wave travel times. Similar tests are conducted with travel-time perturbations with magnitudes up to 0.5, 1.0, and 2.0 seconds. The respective MAEs for these three cases are 4.778 km, 6.251 km, and 7.266 km. Based on these test results, we suggest that an average P-arrival travel time error of less than 0.2 sec is required to ensure a robust location estimate. Such a level of accuracy is generally achievable with recently proposed picking methods [7].

E. Robustness to false picking

In practice, false triggers may be an issue to cause the defect of the training and testing datasets. If the false trigger is caused in a single station (e.g., electrical spike noise), due to the multiple stations constraining the picks, this single-station false trigger will not be included for location prediction. If the false trigger is an inaccurate picking of the P-arrival of a real earthquake, the false trigger will either be too late from a nearby station compared with the reasonable pick, which will make the station be neglected since we only choose the earliest stations that record the event, or too early from a faraway station, which will not be included since we limit the stations used for prediction to be close to each other. As a result, the picks used for location prediction will be close in time due to the constraint of multiple stations. Even if there are errors in the picks, they should be small, and will not affect the result greatly. Besides, using a robust phase association method can reduce the false alarms produced by the picker and enhance the results of the proposed method.

F. Considerations on EEW application

A network EEW system requires both location and magnitude, based on which the ground motion prediction equation (GMPE) and ground motion to intensity conversion equation (GMICE) can be used to predict the intensity. In this work, we focus on the location problem while leaving the magnitude prediction problem a future topic based on a similar efficient machine learning framework. The magnitude prediction is mainly based on P-wave amplitude and thus requires more waveform data for a prediction as contrary to the P-arrival times in this work. The proposed framework can supplement some existing efficient magnitude estimation methods based on either deterministic methods or deep learning methods proposed recently [8]. Since larger earthquakes (e.g., above M4) are of primary concern in EEW systems, we also investigate the model performance for larger earthquakes. We conduct another test by only selecting those catalog events that are above M4 or M5. For this subset of data, the MAEs for the M4 and M5 events are 4.950 km and 4.271 km, respectively. The error is relatively larger for M4 and M5 events because there are fewer training samples for larger earthquakes. Specifically, most the M4 events are located close to the coastal regions (offshore), where the station denisty is relatively sparse. To deal with the issue of insufficient training data sets, there are some potential ways, e.g., weighting the objective function, augmenting the training dataset, or conducting synthetic tests. To estimate the overall time required to obtain the earthquake location, we obtain the difference time between the origin time and the P-wave arrival time of the fifth station. Since the average station spacing is 24 km, the P travel-time to 5 nearby stations is roughly within 5 s. Additionally, the picking algorithms need around 1 s after the first P arrival to verify a pick, and there is likely data transmission latency of around 1 s. Besides, the RF model consumes 0.107 s to predict the earthquake location. Thus, the total estimated time required by the proposed model is 7.107 s. On the other hand, the deep learning approach [4] for earthquake localization consumes 0.179 s to locate the earthquake, besides, it needs two seconds of data after the arrival time. Thus, the deep learning approach [4] has a total time of 8.179 s. In future seismic monitoring, the distribution of stations could be much denser, which makes the proposed method more suitable.

V. Conclusions

We use the P-wave arrival time differences and the location of the seismic stations to locate the earthquake in a real-time way . Random forest (RF) has been proposed to perform this regression problem, where the difference latitude and longitude between the earthquake and the seismic stations are considered as the RF output. The Japanese seismic area is used as a case of study, which demonstrates very successful performance and indicates its immediate applicability. We extract all the events having at least five P-wave arrival times from nearby seismic stations. Then, we split the extracted events into training and testing datasets to construct a machine learning model. In addition, the proposed method has the ability to use only three seismic stations and 10% of the available dataset for training, still with encouraging performance, indicating the flexibility of the proposed algorithm in real-time earthquake monitoring in more challenging areas. Despite the sparse distribution of many networks around the world, which makes the random forest method difficult to train an effective model, one can use numerous synthetic datasets to compensate for the shortage of ray paths in a target area due to insufficient catalog and station distribution.

References

[1] Q. Kong, R. M. Allen, L. Schreier, and Y.-W. Kwon, "Myshake: A smartphone seismic network for earthquake early warning and beyond," *Science advances*, vol. 2, no. 2, p. e1501055, 2016.

[2] T.-L. Chin, K.-Y. Chen, D.-Y. Chen, and D.-E. Lin, "Intelligent real-time earthquake detection by recurrent neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5440–5449, 2020.

[3] T.-L. Chin, C.-Y. Huang, S.-H. Shen, Y.-C. Tsai, Y. H. Hu, and Y.-M. Wu, "Learn to detect: Improving the accuracy of earthquake detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8867–8878, 2019.

[4] O. M. Saad, A. G. Hafez, and M. S. Soliman, "Deep learning approach for earthquake parameters classification in earthquake early warning system," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[5] X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, and W. Li, "Locating induced earthquakes with a network of seismic stations in oklahoma via a deep learning method," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza, "Earthquake transformeran attentive deep-learning model for simultaneous earthquake detection and phase picking," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[8] S. M. Mousavi and G. C. Beroza, "A Machine-Learning Approach for Earthquake Magnitude Estimation," *Geophysical Research Letters*, vol. 47, no. 1, p. e2019GL085976, 2020.