

# Seismic-phase detection using multiple deep learning models for global and local representations of waveforms

Tomoki Tokuda<sup>1</sup> and Hiromichi Nagao<sup>1,2</sup>

<sup>1</sup>*Earthquake Research Institute, The University of Tokyo, Tokyo 113-0032, Japan*

<sup>2</sup>*Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-0033, Japan. E-mail: [nagaoh@eri.u-tokyo.ac.jp](mailto:nagaoh@eri.u-tokyo.ac.jp)*

Accepted 2023 July 3. Received 2023 July 2; in original form 2022 November 17

## SUMMARY

The detection of earthquakes is a fundamental prerequisite for seismology and contributes to various research areas, such as forecasting earthquakes and understanding the crust/mantle structure. Recent advances in machine learning technologies have enabled the automatic detection of earthquakes from waveform data. In particular, various state-of-the-art deep-learning methods have been applied to this endeavour. In this study, we proposed and tested a novel phase detection method using deep learning, which is based on a standard convolutional neural network in a new framework. The novelty of the proposed method is its separate explicit learning strategy for global and local representations of waveforms, which enhances its robustness and flexibility. Prior to modelling the proposed method, we identified local representations of the waveform by the multiple clustering of waveforms, in which the data points were optimally partitioned. Based on this result, we considered a global representation and two local representations of the waveform. Subsequently, different phase detection models were trained for each global and local representation. For a new waveform, the overall phase probability was evaluated as a product of the phase probabilities of each model. This additional information on local representations makes the proposed method robust to noise, which is demonstrated by its application to the test data. Furthermore, an application to seismic swarm data demonstrated the robust performance of the proposed method compared with those of other deep learning methods. Finally, in an application to low-frequency earthquakes, we demonstrated the flexibility of the proposed method, which is readily adaptable for the detection of low-frequency earthquakes by retraining only a local model.

**Key words:** Machine learning; Neural networks, fuzzy logic; Time-series analysis; Body waves; Computational seismology.

## 1 INTRODUCTION

The detection of earthquakes is a fundamental prerequisite for seismology and contributes to various research areas, such as forecasting earthquakes and understanding the crust–mantle structure. Using highly sensitive seismometers installed throughout the world in various locations, modern seismology allows us to detect large earthquakes and infinitesimal ones that may be imperceptible in normal circumstances.

Several methods that focus on human-selected characteristic features have been proposed for earthquake detection using seismic waveform data such as amplitude and frequency (Zhu *et al.* 2022). A classic method is based on short-term amplitude (STA) over long-term amplitude (LTA) averages of waveforms (STA/LTA method; Stevenson 1976; Allen 1978; Baer & Kradolfer 1987), which detects the onset of an earthquake as a sudden change in the STA/LTA ratio. Combining the STA/LTA method with an autoregressive model

enables the effective inference of the *P*-wave's (Primary wave) arrival time based on the Akaike information criterion (AR-AIC) method (Baer & Kradolfer 1987). Furthermore, frequency-based methods (Lomax *et al.* 2012; Mousavi & Langston 2016) detect earthquakes by focusing on specific frequency domains that characterize the dominant frequency of earthquakes.

Recently, deep learning-based methods have gained much attention for earthquake detection (Mousavi & Beroza 2022). Instead of human-selected characteristic features, a deep learning model learns specific waveform features in a data-driven manner without prior knowledge. First, using training data consisting of waveforms and seismic phase labels, the model parameters for a neural network are optimized. Subsequently, based on the trained model, earthquakes are detected for a new instance of the waveform. Several methods have been proposed based on different forward neural network architectures, such as the convolutional neural network (CNN; Ross *et al.* 2018; Perol *et al.* 2018; Yang *et al.* 2021) and U-net (Zhu &

Beroza 2019; Woollam *et al.* 2019), which have been widely used in fields other than earthquake detection, such as image classification and image segmentation. Furthermore, methods based on recurrent neural networks have been proposed (Zhou *et al.* 2019; Mousavi *et al.* 2020; Soto & Schurr 2021). These methods aim to capture contextual information by preserving the time-sequential memory of a waveform. In particular, the method proposed in Mousavi *et al.* (2020) incorporates the attention mechanism (Bahdanau *et al.* 2014) into the recurrent neural network, which allows for effective feature extraction. Moreover, a hybrid method that involves U-net and an attention mechanism has been proposed (Liao *et al.* 2022). A recent review paper (Münchmeyer *et al.* 2022) on various deep learning-based methods suggests superior performance in seismic-phase detection using the generalized phase detection (GPD) method (Ross *et al.* 2018), PhaseNet (Zhu & Beroza 2019) and earthquake transformer (EQT; Mousavi *et al.* 2020).

Note that the deep learning approach typically does not reveal the relevant features of the waveform for earthquake detection. This is a general problem of the deep learning approach, which is widely known as a ‘black box’ problem (Rudin 2019). The attention mechanism (Yang *et al.* 2016; Vaswani *et al.* 2017) is a promising bail-out strategy for this problem. Implicitly, the attention mechanism focuses on a specific part of the data, effectively learning relevant features and making it explainable for those extracted features. For earthquake detection, EQT adopts this approach, which, for seismic detection, focuses on specific parts of the waveform in a data-driven manner (Mousavi *et al.* 2020).

In this paper, we report on a novel method for phase detection of Primary wave (P-phase), Secondary wave (S-phase) and noise, explicitly focusing on specific parts of the waveform. In contrast to EQT, we predetermine the focal parts of the waveform. We consider both global and local representations of the waveform for a 4-s time window. For global representation, we focus on the whole waveform, whereas for local representations, we focus on the first and second half of the time window (Fig. 1). For each representation, we independently train a detection model using a CNN deep-learning architecture similar to the GPD model. For a new instance of a waveform, we combine the outputs of these different models in the form of a product of phase probabilities. This separate learning strategy adds more information on the features of the waveform, which makes our method more robust for earthquake detection and flexible in feature extraction. To our knowledge, for seismic-phase detection, there has been no neural network method that similarly uses a separate learning strategy.

The strategy of using both global and local representations of data is closely related to context modelling for object detection in the deep learning literature (Liu *et al.* 2020). The fundamental idea of context modelling is that a physical object coexists with the surrounding background and other objects; hence, context plays an important role in object detection. Conventional CNN-based deep learning is supposed to capture contextual information implicitly with multiple levels of abstraction (Liu *et al.* 2020), but it can potentially overlook local details (Farabet *et al.* 2012; Zeng *et al.* 2016). Therefore, it is beneficial to explicitly model a network structure that considers both global and local contexts. The multi-region CNN model (Gidaris & Komodakis 2015) is a pioneering work that uses this strategy to extract features from several different regions, such as half regions, border regions and central regions. These extracted features were combined in the final layer of the neural network for object detection. Moreover, several variants related to this method have been proposed. The gated bidirectional CNN

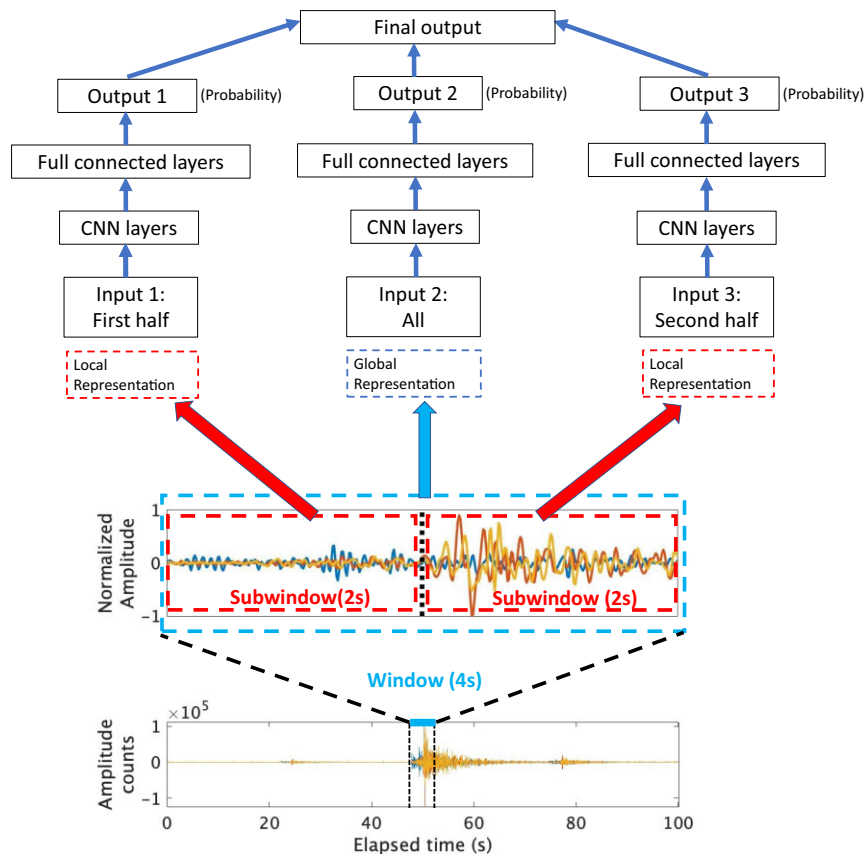
method (Zeng *et al.* 2016) allows interactions between multiscale context regions in feature extraction. In doing so, local contexts are expected to complement each other in validating the CNN’s feature extraction. The attention to context CNN (Li *et al.* 2017; Zhu *et al.* 2017) method considers different architectures for global and local contexts, which allows for the effective capture of contextual locations.

In the following sections, we first introduce the concepts of global and local waveform representations. Local representations were identified in a data-driven manner using multiple clustering analyses. Subsequently, we developed our method, in which the phase probability is defined by the product of the phase probabilities of these representation models. Next, we show that the proposed method outperforms the GPD model when the data are contaminated by noise. For application to continuous seismic waveform data, we demonstrate the high performance of our method compared with the GPD model, PhaseNet and EQT on the 2016 Bombay Beach swarm. Furthermore, we show that our method can be adapted for the detection of low-frequency earthquakes (LFEs) by retraining a local representation model without a real LFE waveform. These results imply the robustness and flexibility of our proposed method, which makes the best use of global and local information on waveform by a separate learning strategy for different focal parts.

## 2 GLOBAL AND LOCAL REPRESENTATIONS

The main idea of our proposed method is to incorporate both global and local waveform information explicitly into a phase-detection model. In this study, ‘global representation’ refers to features related to the entire waveform, whereas the ‘local representation’ refers to features in a local part of the waveform. To identify local representations, we reviewed the general procedure of seismic-phase detection. In conventional practice, a particular part of the waveform (4–60 s, depending on the method) is extracted from continuous waveform data, which are further normalized (i.e. divided) by the maximum absolute amplitude (e.g. the GPD model) or the standard deviation (e.g. EQT). This normalization step plays a crucial role in phase detection by modulating various scales of the waveform. Importantly, its effect is not limited only to the contrast of the waveform before and after the onset, but differences among the three phases also appear in the data points before and after the onset, respectively (Fig. 2). This suggests the possibility that we may better classify the three phases, combining information on the entire part, the before-onset part and the after-onset part. Thus, the question arises whether these focal parts intrinsically contribute to phase detection. If this is the case, it would be beneficial to make the best use of the information of different focal parts for phase detection.

We examined this question in a more general framework, in which local representations may be derived in a data-driven manner. To this end, we applied a multiple clustering method (Tokuda *et al.* 2017, 2018) to waveform data. The multiple clustering method optimally partitions features (data points in our context) into several subsets and clusters instances of the waveform using each subset of features. Hence, if  $K$  subsets of features are yielded,  $K$  cluster solutions are identified for instances of the waveform, in which  $K$  is estimated in a data-driven manner using the Dirichlet process (Gelman *et al.* 2013; Li *et al.* 2019). Importantly, the cluster solutions are assumed to be independent in terms of probabilistic distributions



**Figure 1.** Architecture of the proposed method. We focus on a time window (4 s) in the 3-component seismic waveform. The waveform in the time window is normalized by the maximum absolute amplitude of the waveform, which constitutes a global representation. Furthermore, we consider subwindows (2 s) for the first and second halves of the data points, which constitute local representations. The waveform data in each window is passed to a convolutional neural network as an input (inputs 1–3 in the illustration), and the phase probability for the P-phase, S-phase and noise is yielded as an output (outputs 1–3). Importantly, each neural network independently undergoes supervised training. For a test instance of the waveform, outputs 1–3 are combined as a probabilistic product, which yields a final phase probability. Note that outputs 1–3 are the same for a training instance, whereas for a test instance, these are generally different.

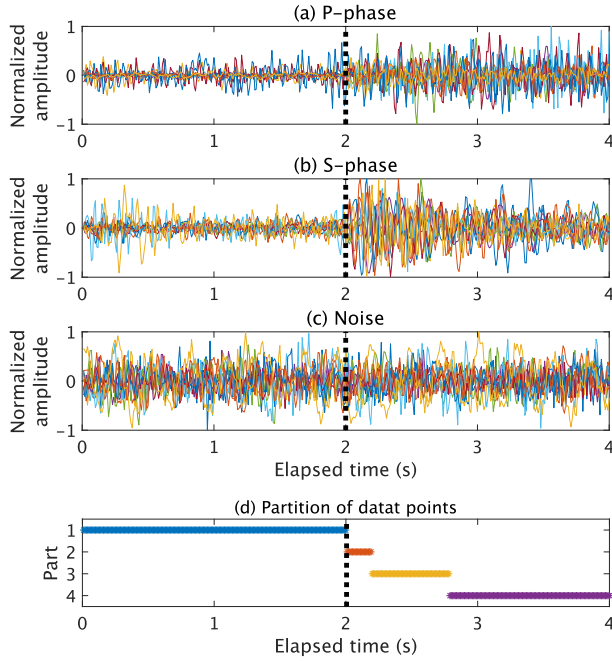
of features. In other words, each cluster solution captures different clustering patterns of instances, which in turn suggest different information on instances (for more details, please refer to Fig. S1 and Section ‘Multiple clustering’ in Supporting Information). We applied the multiple clustering method to a data set of 1000 instances of a 4-s waveform (100 Hz, N–S component) including the P-phase, S-phase and noise. Note that as regards the time length and sampling rate, we followed the default setting for the GPD model (Ross *et al.* 2018). These instances were randomly selected from Southern California Seismic Network (SCSN) data (Ross *et al.* 2018), which have been previously used to train the GPD model. Importantly, in the SCSN data, the data point for the analyst’s selection of P- and S-phases is centred for each instance, as shown in Fig. 2. Hence, it is expected that a common structure of local representations over instances can be derived using the multiple clustering method.

The multiple clustering results suggest that the data points were optimally partitioned into four parts (Fig. 2d): The data points in the first half of the segment constitute a single partition, whereas those in the second half constitute three partitions. Furthermore, it was found that the cluster solutions of these four parts are completely different and that they are also rather different from the cluster solution using all data points in terms of the Adjusted Rand Index (ARI; Hubert & Arabie 1985; see Table S1). These results imply that the four partitions of the data points provide different

information characterizing the instances. This observation motivated us to develop a phase-detection method based on both global (all data points) and local representations (partitions of data points). We expect that a detection method of this kind would be more robust against noise because of the additional information on the waveform. In this study, instead of the four partitions, we consider local representations characterized by the first and second halves of the data points, which facilitates the implementation of our proposed method in terms of training and testing. It should be noted that these specific local representations coincide with our initial observation of the waveform shown in Fig. 2, as discussed in the first paragraph of this section.

### 3 MODEL

Based on the results in Section 2, we developed a phase detection model. In addition to the global representation, we considered local representations that focused on the first and second halves of the waveform. As a basic model, we used the convolutional neural network (CNN) of the GPD model (Ross *et al.* 2018). For each representation, a different phase detection model was trained. For a new waveform, the phase probability was evaluated as a product of the phase probabilities of each model, assuming that these models should yield consistent phase.



**Figure 2.** Examples of waveforms in the training data (Ross *et al.* 2018) and clustering results of data points. Ten waveforms with up–down components were randomly selected and overlaid for P-phase (panel a), S-phase (panel b) and noise (panel c), respectively. For the P- and S-phases, the onset time is centred, which is denoted by a black dotted line. The horizontal axis denotes elapsed time (s), whereas the vertical axis denotes normalized amplitude. The normalization was performed by dividing all records by the maximum of the absolute amplitude in any of the three components (i.e. up–down, north–south and east–west; Ross *et al.* 2018). The number of data points is 400 (i.e. a sampled frequency of 100 Hz). Panel (d): the optimal partition of data points using multiple clustering. We applied multiple clustering for 1000 instances of waveform randomly selected from the training data set (Ross *et al.* 2018). We ran the multiple clustering algorithm for 30 random initial configurations, and selected the optimal solution based on MAP (maximum *a posteriori*) estimation. The horizontal axis denotes the same data points as in panels (a)–(c), whereas the vertical axis the partition of data points (denoted as ‘Part’). A single point in the panel corresponds to a data point, which identifies the part to which the data point belongs. Since the majority of adjacent data points belong to the same part, the plotted points appear as a line.

### 3.1 Basic model

We based our model on the GPD model, which consists of six layers of neural networks: four convolution layers and two fully connected layers (Table 1). The convolution layers extract relevant features for phase detection, whereas the fully connected layers classify the waveform based on these extracted features. Notably, the method is insensitive to the specific architecture of the deep learning model whether it is shallower or deeper and the GPD model’s performance was superior to many other possible choices (Ross *et al.* 2018). The input is the 4-s waveform data with a sampling rate of 100 Hz, whereas the output, which is denoted by  $P_G(\cdot)$  (the suffix ‘G’ denotes ‘global’), is a 3-D vector of probabilities for the seismic phases: P-phase (‘p’), S-phase (‘s’) and noise (‘n’). For the P- and S-phases, the phase initiation point was assumed to be at the centre of the window. Hence, the training waveform data should thus be sampled. Because of this training design, for test waveform data, a large phase probability is expected when the (true) initiation point is at the centre, whereas the phase probability decreases as the (true) phase initiation point moves away from the centre.

### 3.2 Proposed model

Inspired by the multiple clustering results in Section 2, we considered separately developing phase detection models for both global and local representations. For global representation, we adopted the GPD model, in which the waveform of all data points was taken as the input. Conversely, for local representations, we considered a model in which the waveform of the first or second half of the data points was taken as the input (Fig. 1). For these local models, the architecture of the neural network was basically identical to the global one (Table 1), but the filter size was reduced by half because the window size was narrower. For these local models, we denoted the phase probabilities for the first and second halves of the data points as  $P_{L1}(\cdot)$  and  $P_{L2}(\cdot)$  (the suffix ‘L’ denotes ‘local’), respectively. Finally, we defined the final output  $P(x)$  as the product of the three probabilities yielded by the global model and the two local models:

$$P(x) \stackrel{\text{def}}{=} P_G(x)^{w_G} \times P_{L1}(x)^{w_{L1}} \times P_{L2}(x)^{w_{L2}}, \quad (1)$$

where  $x \in \{‘p’, ‘s’, ‘n’\}$ ;  $w_G, w_{L1}, w_{L2} \in \{0, 1\}$ . The binary parameters  $w_G, w_{L1}$  and  $w_{L2}$  allowed us to widen the scope of the model representations. The proposed model, hereafter referred to as the ‘GL model’, can be expressed by setting binary parameters as follows:

(i) GL model:  $(w_G, w_{L1}, w_{L2}) = (1, 1, 1)$ .

Similarly, the three components of the GL model can be expressed as the following submodels:

(i) G model:  $(w_G, w_{L1}, w_{L2}) = (1, 0, 0)$

(ii) L1 model:  $(w_G, w_{L1}, w_{L2}) = (0, 1, 0)$

(iii) L2 model:  $(w_G, w_{L1}, w_{L2}) = (0, 0, 1)$ ,

where model G is the global model, L1 is the local model for the first half of the data points and L2 is the local model for the second half of the data points. The G model is identical to the GPD model. These submodels are separately trained using the training data (inputs 1–3 and outputs 1–3 in Fig. 1), whereas the final output in eq. (1) is evaluated only for the test data.

Because of the formulation of the GL model, the summation of probabilities is not always one, that is,  $P(‘p’) + P(‘s’) + P(‘n’) \neq 1$ . One may consider normalizing the probabilities, but in this study, we retained this formulation as it is because an interpretation of probability  $P(x)$  is straightforward. It should be noted that the normalization does not change the large and small relationships among the probabilities of  $x \in \{‘p’, ‘s’, ‘n’\}$  (because  $a \leq b \Leftrightarrow a/(a+b) \leq b/(a+b)$  for  $a > 0$  and  $b > 0$ ). Hence, the classification results (Sections 4.1, 4.2 and 4.4) based on the maximum likelihood principle remained unchanged regardless of whether normalization was performed.

### 3.3 Model training

We trained the G, L1 and L2 models using the waveform data set of the Southern California Seismic Network (SCSN), which was used to train the GPD model in Ross *et al.* (2018). Based on these sub-models, we constructed the GL model, as shown in eq. (1). The training data set consists of 4.5 million instances, which are equally distributed for the P-phase, S-phase and noise. All the waveforms were sampled at 100 Hz for 4 s. For the P-phase and S-phases, the waveform was centred on the analyst’s selection. We randomly split the data into three subsets: the training data (75 per cent), validation data (15 per cent) and test data (10 per cent). The validation data were used for early stopping, the



**Table 1.** The deep learning model architecture for both global and local representations. They differ only in the convolutional layer's filter size, which is shown in the fourth row of the table. We follow the abbreviations from Ross *et al.* (2018): 'CBP' denotes convolution, batch normalization and pooling (in this order); 'FB' denotes fully connected and batch normalization (in this order); 'F' denotes fully connected.

Layer	1	2	3	4	5	6	7
Stage	CBP	CBP	CBP	CBP	FB	FB	F
Number of channels	32	64	128	256	200	200	3
Filter size (Global/local)	21/10	15/7	11/5	9/4	-	-	-

conventional mechanism used in deep learning to prevent overfitting (Goodfellow *et al.* 2016; Géron 2019). To train the model, we followed the same settings as in Ross *et al.* (2018): 480 batch size, batch normalization, early stopping with six patience epochs, the cross-entropy loss function and the Adam optimization algorithm. It was found that only a few epochs were sufficient for the training to converge (four, six and seven epochs for G, L1 and L2 models, respectively), which prevented overfitting of the model by the early stopping based on the validation data (Fig. S2). Using four NVIDIA GPUs in our computational environment, the training took less than 30 min.

## 4 APPLICATION

Based on the trained models described in the previous section, we applied the proposed method to various types of waveform data. First, we report the classification results of the test data and the Stanford Earthquake Dataset (STEAD; Mousavi *et al.* 2019) using the GL, G, L1 and L2 models. Furthermore, we report on the phase-picking performance for the STEAD data. Subsequently, we examined the classification performance when the test data were noise-contaminated. Thirdly, we applied the proposed method to continuous seismic waveform data, taking the example of the 2016 Bombay Beach swarm. Finally, we applied the method to continuous seismic waveform data in Japan, which includes LFEs. LFE is a specific type of earthquake that dominates lower frequency component (2–8 Hz). In the latter two cases, the performance was compared with that of the GPD model, PhaseNet and EQT.

### 4.1 Classification of test data

We report on the classification results of the three phases for the test data set in Section 3.3. The classification is based on the maximum likelihood principle by which the phase that provides the maximum probability is assigned. We applied the GL, G, L1 and L2 models to the test data, which were evaluated in terms of recall and precision (Figs 3a and b). Here, we adapted the recall and precision metrics (Buckland & Gey 1994) for multiclass classification, that is, ' $p$ ', ' $s$ ' and ' $n$ ': We extended the definition of completeness of retrieval (recall) and purity of retrieval (precision) by focusing on a particular class (for more details, please refer to Appendix A). It was found that the GL and G models performed equally well, with the recall and precision of both being  $>0.98$ . The performance of the L2 model was also good, with a recall and precision of  $>0.95$ , whereas the recall and precision of the L1 model deteriorated slightly, scoring between 0.90 and 0.95. These results suggest that the global (i.e. the G model), and the local model (i.e. the L1 and L2 models) offer considerable information for phase detection, as shown in Fig. 2.

### 4.2 Classification of STEAD data

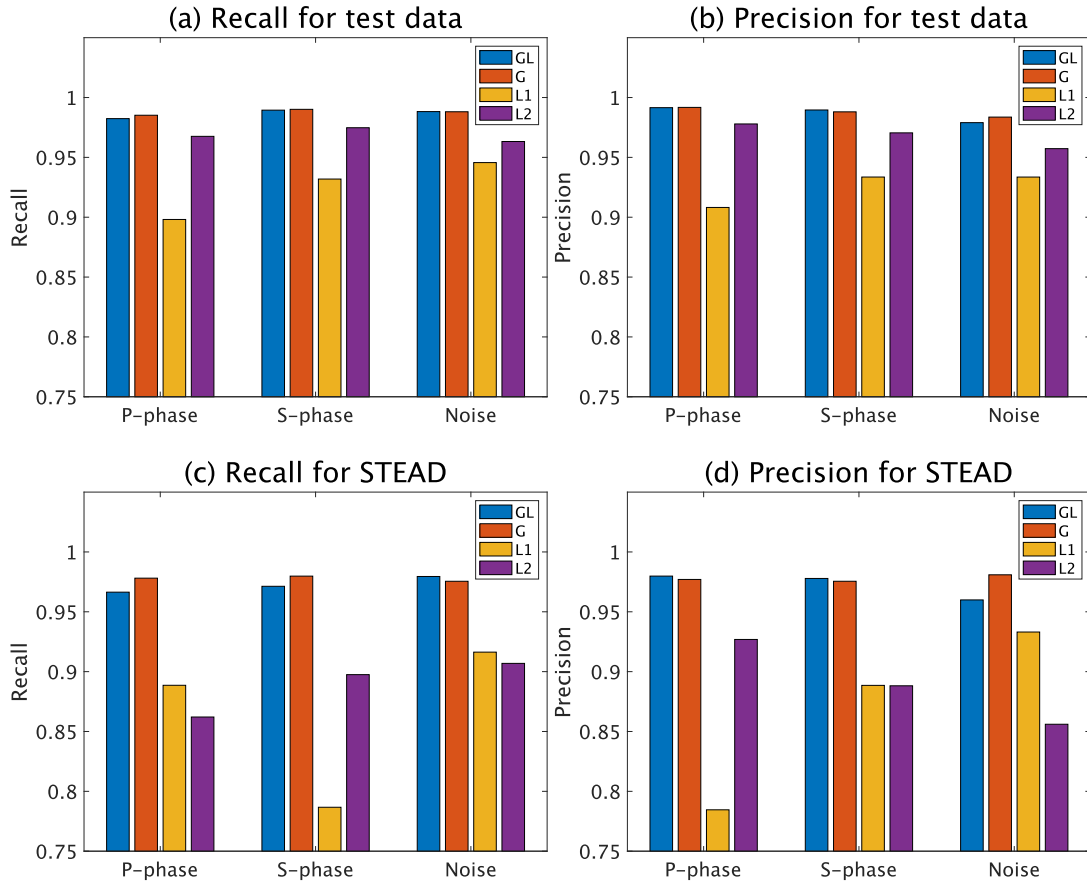
Similarly, we report on the classification results for the STEAD data, which is often used as a benchmark data set for evaluating phase detection methods (Mousavi *et al.* 2020; Münchmeyer *et al.* 2022). The objective is to conduct a cross-domain analysis for the classification performance of the trained models. The STEAD data is a large-scale global data set with over 1 million instances of a 60-s P-phase, S-phase and noise waveform, which were obtained from various seismometers in the world (Mousavi *et al.* 2019). From the STEAD data, we randomly chose 10 000 instances for each phase, followed by the extraction of a 4-s waveform with the onset time being centred (for the noise phase, we extracted the 4-s waveform centred at the 30-s datapoint). The performance results are summarized in Figs 3(c) and (d). The GL and G models performed equally well, with the recall and precision of both being between 0.96 and 0.98 (a little better for the G model). These results were compatible with that of the test data. Conversely, the performances of the L1 and L2 models were moderate, with a recall and precision of between 0.78 and 0.92 for the L1 model and between 0.85 and 0.93 for the L2 model. Compared with the test data, their performance somewhat deteriorated.

### 4.3 Onset time picking

Furthermore, we report on the performance of the onset time picking for both P-phase and S-phase. As the SCSN data are available only in 4-s waveforms, the picking time cannot be easily evaluated. As an alternative, we used the STEAD data, randomly choosing 10 000 instances for each phase as in Section 4.2. For these instances, we extracted 10-s waveforms as in Münchmeyer *et al.* (2022), including the true onset time between 2 and 8 s of the 10-s extracted instance for the P- and S- phases. To evaluate the onset time picking, we defined the onset time as a peak in the probability distribution for each phase (Ross *et al.* 2018). We evaluated the time residuals between the true and predicted onset times (Münchmeyer *et al.* 2022): a histogram for visual inspection; the fraction of instances with high absolute residuals  $>0.45$  s (OUT); the root mean square (RMSE); mean absolute error (MAE). It was found that the performance of the GL model slightly deteriorated for both the P- and S-picks compared with the G model (Fig. 4): The residual distribution became more dispersed and biased. Section 5 discusses an alternative GL model approach for better performance in picking the onset time.

### 4.4 Classification of noise-contaminated data

Now, the question arises concerning whether there is any difference between the GL and G models in terms of phase detection capability. Because the GL model contains additional information in the first and second halves of the waveform, it was hypothesized that the GL model may provide more robust phase detection than the G



**Figure 3.** Performance of classification for the test data and the STEAD data. Panels (a) and (c): recall. Panels (b) and (d): precision. These panels summarize the P-phase, S-phase and noise detection performance. The colour bars denote the performance for the GL model ('GL'), G model ('G'), L1 model ('L1') and L2 model ('L2'), respectively. For more classification results details, please refer to Tables S2 and S3.

model. To test this hypothesis, we performed a simulation study in which noise was added to the test data. Here, we denote 'noise' by a noise-phase waveform. Denoting  $\gamma$  as the noise proportion ( $0 \leq \gamma \leq 1$ ), we considered transforming the waveform by:

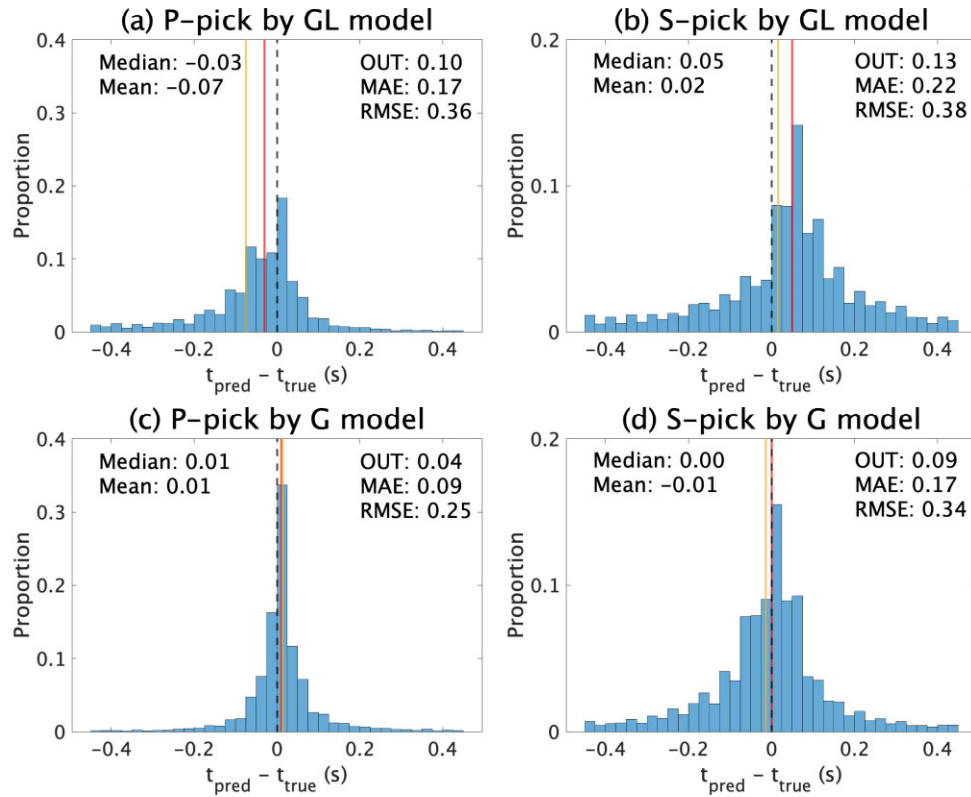
$$\mathbf{x} \leftarrow (1 - \gamma) \times \mathbf{x} + \gamma \times \mathbf{n}, \quad (2)$$

where  $\mathbf{x}$  is the waveform data with arbitrary phase and  $\mathbf{n}$  noise. This additive noise model is a natural and simplified approximation, but it is tractable and good enough for the test. For this simulation study, we generated the following data sets: First, 1000 waveforms were randomly selected from the test data. Secondly, each instance was transformed into eq. (2) using instance  $\mathbf{n}$ , which was randomly and independently sampled from the test data that were not selected for the aforementioned 1000 instances. We manipulated the value of noise proportion  $\gamma$  from zero to one in increments of 0.25. Moreover, to evaluate the effect of the locus of noise contamination, we considered the following three cases: all data points, the first half of the data points (DP1) and the second half of the data points (DP2). In the case of contamination in DP1, the data transformation in eq. (2) is performed only for DP1, whereas DP2 remain unchanged. Similarly, in the case of DP2, contamination is limited to DP2, whereas DP1 remain unchanged. Hereby, we generated 100 data sets for each value of noise proportion  $\gamma$  and the locus of contamination. For these data sets, we applied the GL, G, L1 and L2 models to evaluate their performance in terms of accuracy, which summarizes overall classification performance (please refer to Appendix A for the definition of accuracy).

#### 4.4.1 Results

When noise was added to all data points, the accuracy largely decreased for all models as the noise proportion  $\gamma$  increased (Fig. 5a). In this case, the performance of the GL model (blue in the figure) was slightly worse than that of the G model (red in the figure) for a noise proportion of 0.5 with accuracy 0.72 and 0.65, respectively. Similarly, when noise was added only for the first half of the data points, the accuracy tended to decrease for the GL, G and L1 models as the noise proportion  $\gamma$  increased. However, the accuracy of the L2 model remained high irrespective of the noise proportion  $\gamma$  because noise no longer influences the L2 model (Fig. 5b). Importantly, the performance of the GL model was significantly better than that of the G model: For a noise proportion of 0.75 the accuracy of the GL model and G model was 0.88 and 0.65, respectively. A similar observation was made when noise was added to the second half of the data (Fig. 5c).

Next, we examined waveform instances that the G model misclassified and the GL model correctly classified. As a typical case, we focused on the setting in which noise was added to the first half of the data points with a noise proportion of 0.5. In this setting, such instances most frequently occurred in the noise phase (the average number of such instances was 31.8, 39.4 and 69.1 out of 1000 instances for P-phase, S-phase and noise, respectively). We visually investigated the waveforms of these instances (Fig. 6). It was found that there were differences in amplitude between the first and second halves of the data points in which the second half had



**Figure 4.** Performance of onset time picking. Panel (a): distribution of the time residuals between the true and predicted onset time for P-arrival by the GL model. The median and the mean are shown with red and orange lines, respectively. The black dotted line denotes the zero residual. The results of the evaluation indices are displayed in the panel, including the fraction of instances with high absolute residuals  $>0.45$  s (OUT), the root mean square (RMSE) and the mean absolute error (MAE). Note that the range of the histogram is between  $-0.45$  and  $0.45$  s, whereas those evaluation indices are based on all residuals. Panel (b): for S-arrival by the GL model. Panel (c): P-arrival by the G model. Panel (d): S-arrival by the G model.

a larger amplitude. Possibly because of this apparent contrast in amplitude between these two parts, the G model misclassified the noise as either the P- or S-phase. Conversely, the GL model worked well for these instances because it had information on the L1 and L2 models, which correctly classified the noise phase. It is noted that apparently similar instances to these contaminated waveforms were included in the training data (Figs S3 and S4). However these instances were correctly identified as noise by the G model, which suggests that these are of different type of noise from those in Fig. 6.

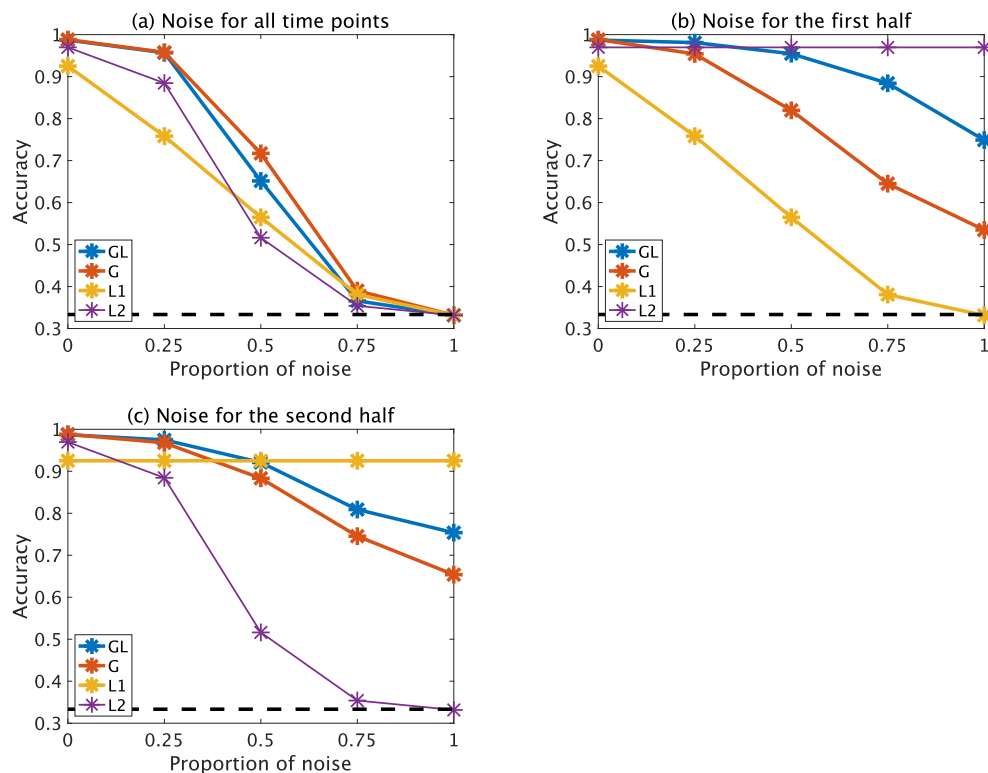
#### 4.5 2016 Bombay Beach swarm

Subsequently, we applied the proposed method (GL model) to the 2016 Bombay Beach swarm that occurred in California, USA (Ross *et al.* 2018; McBride *et al.* 2020). The objective was to evaluate the performance of the proposed method for continuous seismic waveform data, which was further compared with the performances of the GPD model and two other state-of-the-art methods based on deep neural networks: PhaseNet (Zhu & Beroza 2019) and EQT (Mousavi *et al.* 2020). For these methods, we used publicly available open-source models from Woollam *et al.* (2022), by which the GPD model was trained using the original training data from Southern California, whereas PhaseNet and EQT were trained using the SCEDC (Southern California Earthquake Data Center) data set. Thus, the geometric region for the training data coincided with the location of the Bombay Beach swarm for all these methods, including the proposed method. Hence, it was expected that these methods

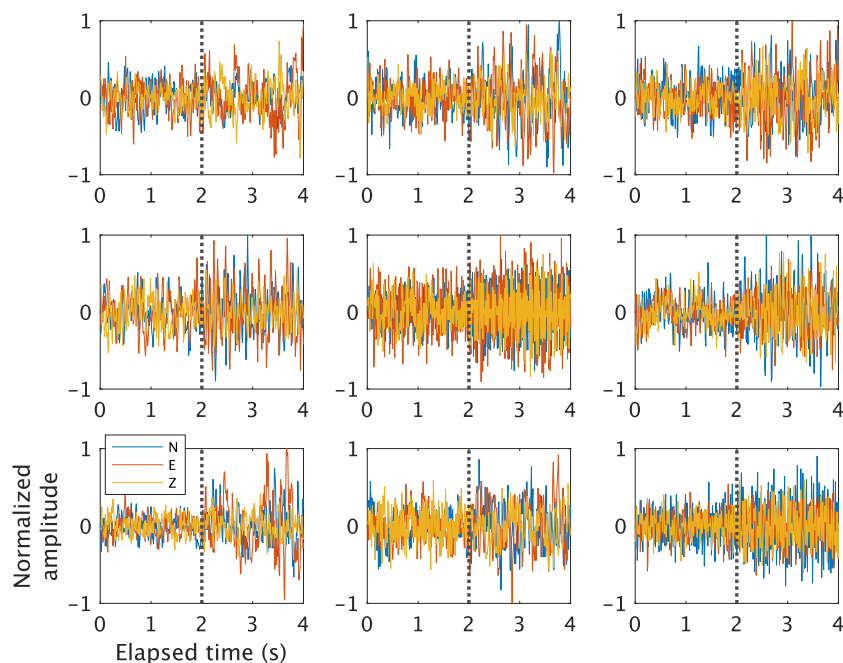
would be appropriate for detecting the 2016 Bombay Beach swarm phases.

We set the detection thresholds for both the P- and S-phases as follows: For the GL model, we set the phase probability to 0.5. This criterion implies that all three sub-probabilities on the right-hand side of eq. (1) should be greater than 0.5. If these probabilities are identical, each probability becomes 0.79. For the remainder of the methods, we adopted thresholds from their original papers: 0.98 for the GPD model (Ross *et al.* 2018), 0.5 for PhaseNet (Zhu & Beroza 2019) and 0.3 for EQT (Mousavi *et al.* 2020). It is noted that in their original papers, no clear reasons were given for these thresholds other than empirical ones. The window size for phase detection followed the original setting: 4 s for the GL and GPD models, 30 s for PhaseNet and 60 s for EQT. Furthermore, we set the shift width (stride) to 0.1 s for all methods.

We applied these detection methods to three-component waveform data recorded at a frequency of 100 Hz at the BOM station in the Southern California Seismic Network (California Institute of Technology 1926; SCEDC 2013). From the waveform record, we extracted 6-hr waveforms before and after the onset of the swarm. However, to avoid uncertainty in the swarm's onset time, we removed the waveform data from 1 hr before and after the onset, which produced a 5-hr waveform for each time segment. To alleviate the possible effect of earthquakes that occurred before the onset of the swarm, we discarded the detection results 1 min after the earthquakes in the SCEDC catalogue. In this catalogue, 12 earthquakes were recorded during this period, 11 of which had magnitudes of less than 1.5, with the hypocentre more than 65 km away



**Figure 5.** Classification performance for noise-contaminated test data. Panel (a): the noise waveform is added to all data points. Panel (b): the noise waveform is added to the first half of the data points. Panel (c): it is added to the second half of the data points. The horizontal axis denotes noise proportion, whereas the vertical axis denotes mean accuracy for 100 data sets of 1000 randomly selected instances. The black dashed line denotes the expected accuracy at the chance level (i.e.  $1/3$ ). The colour lines denote performance for the GL model ('GL'), G model ('G'), L1 model ('L1') and L2 model ('L2'), respectively. For the mean values and standard deviations, please refer to Table S4.



**Figure 6.** Examples of waveforms that the GL model correctly identified as noise that the G model misidentified. These waveforms are contaminated by noise with a noise proportion of 0.5 in eq. (2). The horizontal axis denotes elapsed time (4 s), whereas the vertical axis denotes normalized amplitude [we did not renormalize the waveform after the transformation in eq. (2)]. The dotted line denotes the middle point of the time window. For the legend, 'N' denotes N-S component; 'E' E-W component; 'Z' up-down component. These examples were randomly selected from instances used in Section 4.4.



from the BOM station. One earthquake had a magnitude of 6.0, but it took place 10 440 km away from the station. Because of these data selection procedures, in the following analysis, we assumed that there was no effect attributable to the earthquakes/swarm before the onset.

#### 4.5.1 Results

First, a visual inspection of the phase detection suggests substantial differences in the P- and S-phase detections before and after the onset of the swarm (Figs 7, 8 and 9). For all methods, more P- and S-phases were detected after the onset than before it. However, among these methods, there were considerable differences in the phase-detection efficacy. Such differences after the onset are clearly shown in Fig. 8. Furthermore, we quantified the number of phase detections before and after the onset (Table 2). With regard to the P-phase, the GL model yielded two detections before onset, whereas 923 detections occurred after onset. In contrast, the GPD model yielded 18 detections before onset, whereas 1336 detections occurred after onset. PhaseNet yielded no detections before onset, whereas 67 detections occurred after onset. Finally, EQT yielded seven detections before the onset, whereas 353 detections occurred after the onset. As the number of detections does not necessarily denote the number of events, it is not straightforward to evaluate these results. Furthermore, among these methods, there were differences in how a detection probability was assigned to non-peak times. Nonetheless, the ratio of the number of detections before and after the onset (column 'Ratio' in Table 2) suggests that the GL model and PhaseNet performed very well, with a ratio of 461 and  $\infty$ , respectively. Note that the ratio before and after the onset implies a relative scale of phase detections, which is comparable among the different methods. However, notably, PhaseNet tended to yield somewhat conservative detection, which did not detect the P-phase in the period of 2 hr after the onset (the 8th column in Table 2). Regarding the S-phase, the GL model worked as intended without any detection before the onset of the swarm, whereas it yielded a large number of detections after onset. Since no detection was observed before onset, those detections after onset can be considered to be related to the swarm, rather than false positive cases.

Secondly, we analysed the phase-detection performance in terms of phase probability. Because the number of phase detections changes depending on the probability threshold, we evaluated the performance without specifying the threshold. We focused on probabilities at least 0.1, discarding almost zero probabilities of the vast majority of the phase detections by these methods. The distributions of phase probabilities before and after the onset of the swarm imply that there are differences in performance depending on methods (Fig. 10). To quantify these differences, we evaluated the area under the curve (AUC) of a receiver operating characteristic curve (Fawcett 2006; Hand 2012) as follows: First, the label 'before' was assigned to the probabilities that were yielded before the onset, whereas the label 'after' was assigned to those after the onset. Next, a receiver operating characteristic curve was constructed for a binary classifier of 'before' and 'after' with various thresholds of probabilities, which were subsequently used for the AUC evaluation. For the P-phase, the AUC values for the GL model, the GPD model, PhaseNet and EQT were 0.78, 0.66, 0.76 and 0.76, respectively. For the S-phase, the AUC values were 0.91, 0.79, 0.67 and 0.63, respectively, for the same models. These results imply that, for both phases, the GL model yielded the most

separable probability distribution before and after the onset of the swarm.

#### 4.6 Low-frequency earthquake in Japan

Finally, we examined the detection performance for LFEs, which differ in type from conventional earthquakes, in which the low-frequency component (2–8 Hz) dominates in the seismic wave (Ide *et al.* 2007). From both the theoretical and observational perspectives, it is inferred that an LFE may be caused by Brownian motion throughout the area hosting the shear slip (Ide 2008; Ide & Maury 2018). In particular, LFEs have recently gained considerable attention to gain a better understanding of earthquake mechanisms and for their possible connection to large earthquakes (Obara 2002; Shelly *et al.* 2007; Peng & Gomberg 2010; Kocharyan 2021; Kato & Ben-Zion 2021). Hence, it is particularly important to correctly detect LFEs, which provide valuable information for such research. However, detecting LFEs is challenging because they manifest very weakly in waveforms with orders of magnitude below those of typical earthquakes (Thomas *et al.* 2021). Moreover, owing to the low frequency of the waveform, a detection model trained using typical EQ data is not readily applicable for LFE detection (Thomas *et al.* 2021). Hence, the current practice of LFE detection is based on human analysts making manual selections (Kurihara & Obara 2021) or on a matched-filter technique in which a new LFE is identified by cross-correlation analysis of waveforms with known LFEs (Shelly 2017; Kato & Nakagawa 2020). In this study, as a novel attempt, we aimed to detect LFEs effectively using the proposed method's flexible framework.

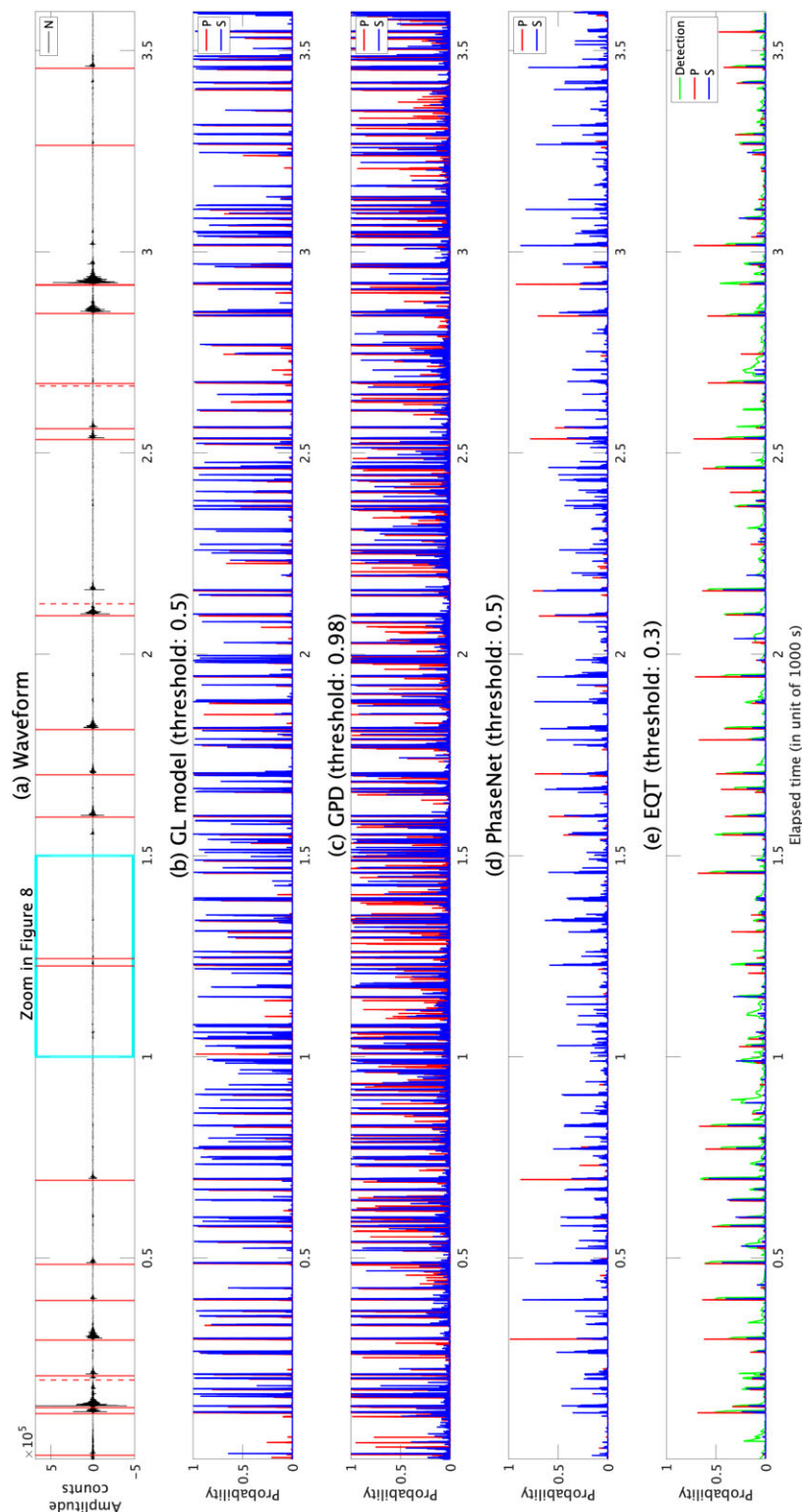
In this experiment, we focused on the LFEs in the Tohoku region of Japan (latitude 37°N to 41°N and longitude 139°E to 142°E), as recorded in the JMA catalogue from 2015 to 2019 (Japan Meteorological Agency 2022). We evaluated whether our method could detect the catalogued LFEs. For better demonstration, we selected the LFEs from the JMA catalogue that met the following conditions:

- (i) Condition 1: There was no occurrence of other LFEs 10 min before and after a target LFE.
- (ii) Condition 2: In the same period as in Condition 1, there were no occurrences of conventional earthquakes with magnitudes of less than 3 within 100 km from the epicentre of the target LFE.
- (iii) Condition 3: Also, in the Condition 1 period, there were no occurrences of earthquakes with magnitudes larger than 3 in the Tohoku region.

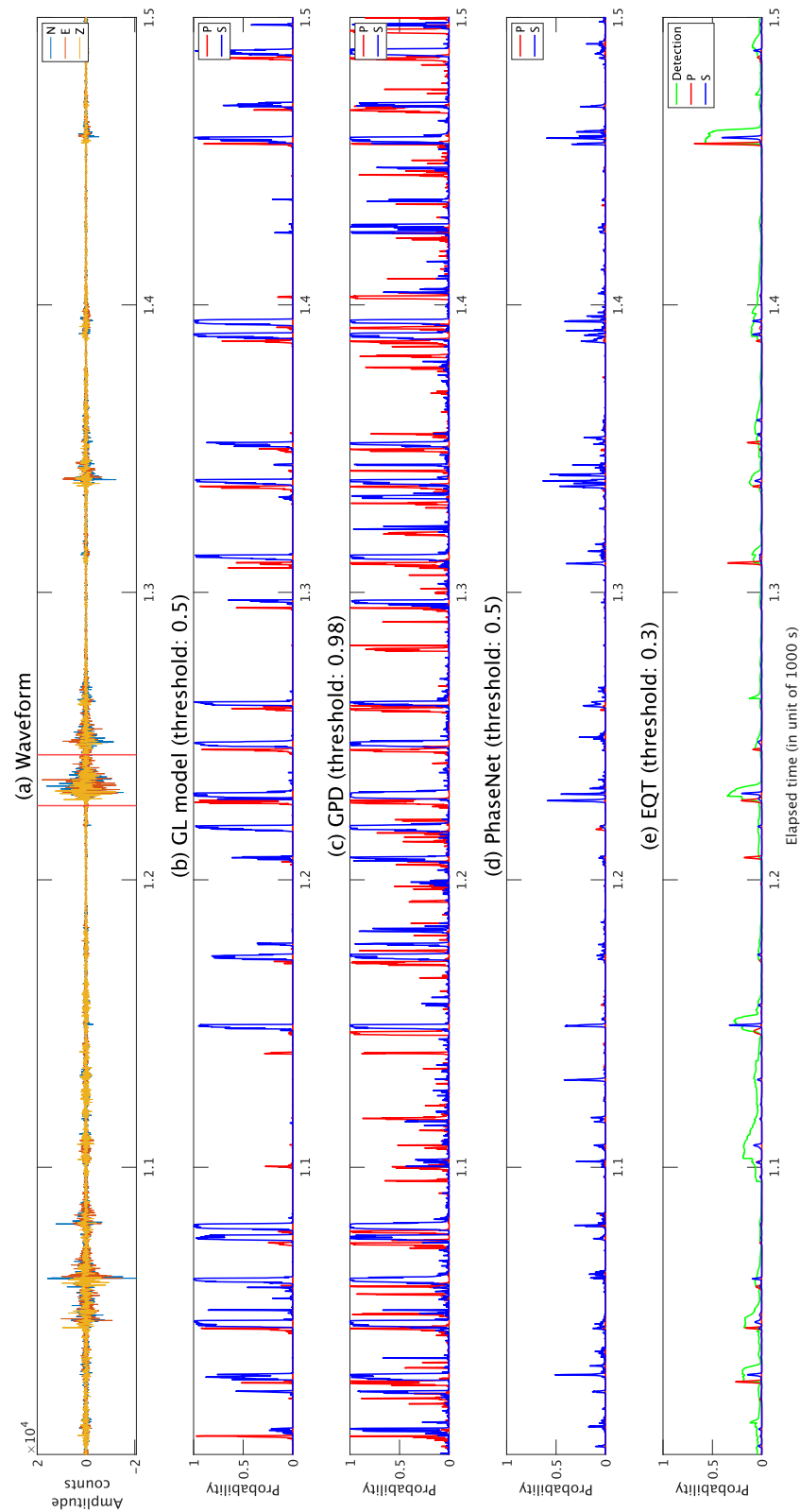
These conditions allowed us to assume the absence of influences by other earthquakes 10 min before and after the target LFE. Furthermore, to be more rigorous, we retrieved the waveform 5 min before and after the target LFE.

We obtained waveform data from a high-sensitivity seismograph network (Hi-net) operated by the National Research Institute for Earth Science and Disaster Resilience (NIED), Japan (Okada *et al.* 2004). Hi-net observation stations are densely situated throughout Japan with a 20-km mesh and routinely collect 3-component waveform data. First, the Hi-net station nearest to the epicentre of the target LFE was identified. Subsequently, we extracted 10-min waveforms with three components 5 min before and after the occurrence of the target LFE. If a complete waveform was not available, the LFE was discarded. As a result, we obtained full waveforms for 446 LFEs. No pre-processing was performed for these waveforms.

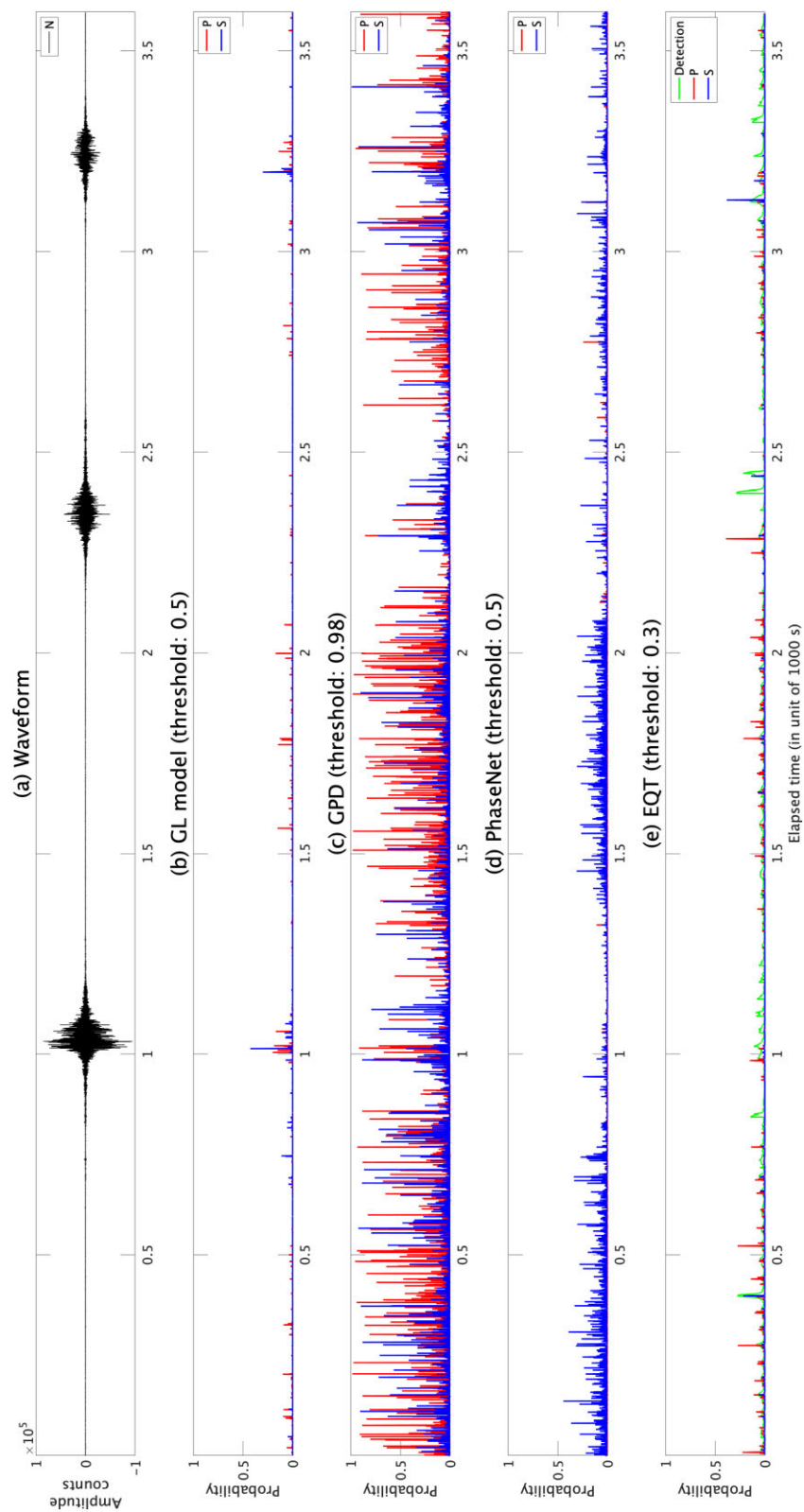
We applied the GL model, the GPD model, PhaseNet and EQT to the waveform data to detect the P-phases of the LFEs. For each



**Figure 7.** Waveform and phase detections *after* the onset of 2016 Bombay Beach swarm, California, USA. Panel (a): waveform at the BOM station with the HHN channel. The horizontal axis denotes the 1-hr time course starting at 8:00. Pacific daylight time (PDT), 5 hr after the onset of the swarm, whereas the vertical axis denotes amplitude counts. The vertical red line denotes the origin time for the earthquakes that occurred less than 10 km from the BOM station in the SCEDC catalogue (SCEDC 2013), whereas the red dashed lines denote those further than 10 km. The region enclosed by the cyan lines is zoomed in Fig. 8. Panels (b)–(e): phase detection results for GL model, the GPD model, PhaseNet and EQT, respectively. Phase detection is based on three components of waveform with the HHZ, HHN and HHE channels. We pre-processed the waveform data by detrending and high-pass filtering above 2 Hz as in Ross *et al.* (2018). For these panels, the vertical axis denotes the *P*-wave (in red) and *S*-wave (i.e. Secondary wave, in blue) probabilities. For EQT, the detection probability is also plotted in green. In the title of each panel, the threshold for phase detection is displayed. Except for the GL model, we used the open-source models in the Seisbench toolbox Woollam *et al.* (2022).



**Figure 8.** Waveform and phase detections *after* the onset of the 2016 Bombay Beach swarm, which zoomed the time duration from 1000 to 1500 s in Fig. 7, which is enclosed by the cyan lines in the figure. For the legend in panel (a), ‘N’ denotes the N–S component (HHN channel); ‘E’ E–W component (HHE channel); ‘Z’ up–down component (HHZ channel).

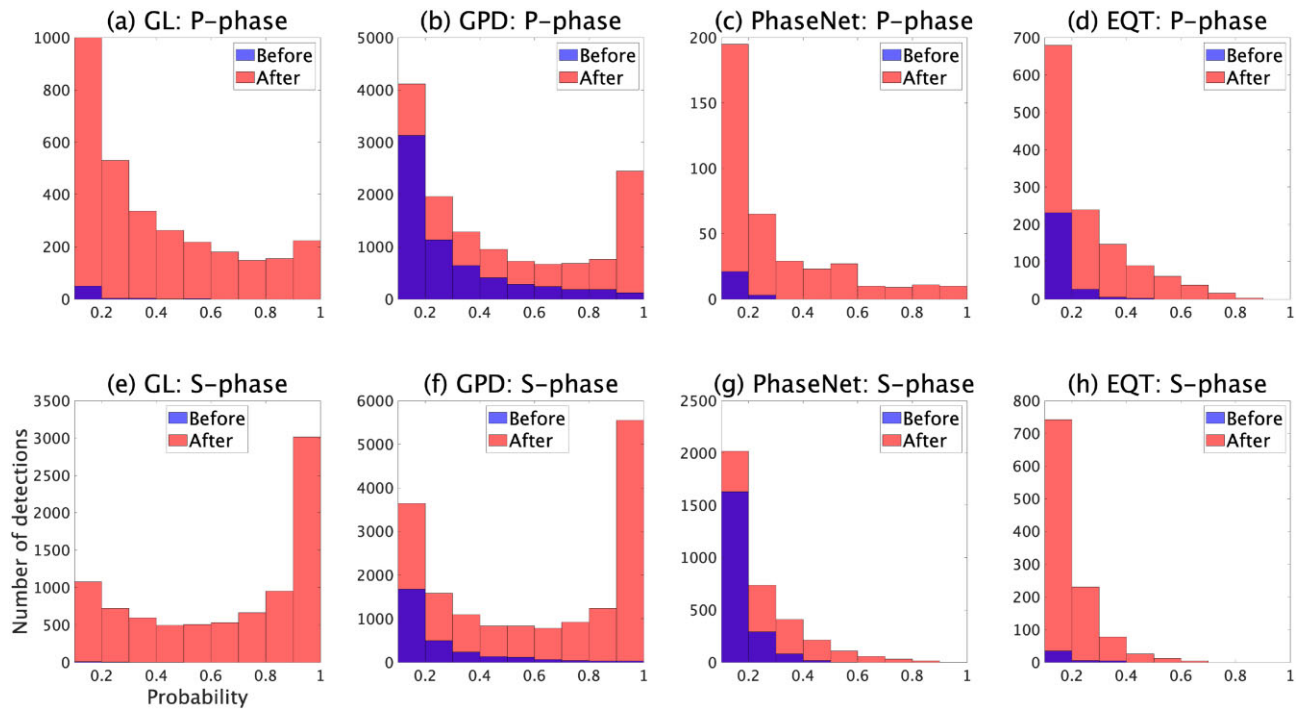


**Figure 9.** Waveform and phase detection *before* the onset of 2016 Bombay Beach swarm. This figure is similar to Fig. 7, but the time coverage starts 5 hr before the onset of the swarm, that is at 23:00 on September 25 (PDT).



**Table 2.** Summary of phase detections for the 2016 Bombay Beach swarm. The numbers of phase detections per hour are summarized for the time coverages of 2 and 6 hr before and after the onset of the swarm, respectively. In the table, the period before the onset is shown as negative, whereas the period after the onset is positive. The numbers of detections for the P- and S-phases are separately presented. The summation (column ‘Sum’) is taken over each time segment, that is before and after the onset (‘before-segment’ and ‘after-segment’). The ratio (column ‘Ratio’) is evaluated as the number of detections in the after-segment over the number of detections in the before-segment.

Method	Before-segment						After-segment						After/before
Period (hr)	–6	–5	–4	–3	–2	Sum	2	3	4	5	6	Sum	Ratio
P-phase detection													
GL	0	0	1	0	1	2	45	88	210	356	224	923	461
GPD	11	1	1	3	2	18	51	66	447	430	342	1336	74
PhaseNet	0	0	0	0	0	0	0	9	5	32	21	67	∞
EQT	4	3	0	0	0	7	22	36	42	147	106	353	50
S-phase detection													
GL	0	0	0	0	0	0	168	594	1583	1772	1550	5667	∞
GPD	0	1	0	3	0	4	110	400	914	1114	943	3481	870
PhaseNet	0	0	0	1	0	1	8	22	59	59	66	214	214
EQT	1	4	0	0	0	5	9	14	20	4	37	120	24



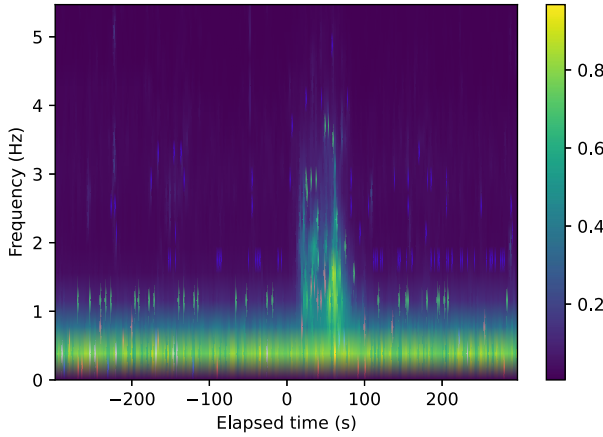
**Figure 10.** The probabilities for P- and S-phase detections before and after the onset of the Bombay Beach swarm. Each panel denotes a histogram for the number of phase probabilities in the 5-hr period before the onset (blue bar) and in the 5-hr period after the onset (red bar). The horizontal axis denotes the probability for phase detection, and the vertical axis denotes the number of detections. The first and second rows are for the P-phase and S-phase, respectively. Panels (a) and (e) are for the GL model, panels (b) and (f) are for the GPD model, panels (c) and (g) are for PhaseNet and panels (d) and (h) are for EQT. These panels display probabilities larger than 0.1.

method, we set the same detection threshold (0.5, 0.98, 0.5 and 0.3 for GL model, GPD model, PhaseNet and EQT, respectively) and window width (4, 4, 30 and 60 s for GL model, GPD model, PhaseNet and EQT, respectively), as in the Bombay Beach swarm analysis described in Section 4.5. Because information on the exact timing of the P-phase reaching an observation station was not available, we set the maximum time delay time to 100 s. In other words, we considered detection to be successful when the method in question detected the P-phase within 100 s after the LFE’s time of occurrence. This presumed delay time is based on the observation that high-intensity waveform power of between 1 and 4 Hz exclusively occurred in this period (Fig. 11). For more details, please

refer to the fourth paragraph of Section 5). Similarly, we considered detection to be false when the method detected the P-phase within 100 s before the LFE’s time of occurrence.

#### 4.6.1 Results

The detection results obtained using these settings are summarized in Table 3. The GL model correctly identified 94 out of 446 LFEs (21 per cent), whereas the GPD model, PhaseNet and EQT correctly identified 213 (48 per cent), 38 (8.5 per cent) and 94 (21 per cent) LFEs, respectively (third column in Table 3). For visual inspection, Fig. 12 shows randomly sampled waveforms that were correctly



**Figure 11.** Mean dynamic spectrogram of the waveform's up-down component for 446 LFEs in study. The horizontal axis denotes the time points in which the LFE's origin time is set to 0 and the vertical axis, frequency (Hz). The colour denotes the mean intensity of the power. The spectrogram was generated as follows. First, for each LFE, the 10-min (600 s) up-down component of the waveform was standardized. Secondly, we generated a spectrogram for each LFE using the Python function 'scipy.signal.spectrogram' with the default setting (Tukey window with a shape parameter of 0.25; a segment length of 256 data points; no overlapping of segments) (Virtanen *et al.* 2020). Lastly, the mean spectrogram was evaluated for all 446 LFEs.

detected by the GL model. Conversely, the false detection cases were 27 (6.1 per cent), 97 (21 per cent), 42 (9.4 per cent) and 57 (12 per cent) for the GL model, GPD model, PhaseNet and EQT (second column in Table 3), respectively. In terms of the ratio of correct to false detections, the GL model performed the best (ratio = 3.5), followed by the GPD model (2.2), EQT (1.6) and PhaseNet (0.90) (fourth column in Table 3).

#### 4.6.2 Adapted models

In the next experiment, we adapted the GL model specifically to detect the LFE P-phase. As in eq. (1), the GL model consisted of three submodels: G, L1 and L2. Among these submodels, we retrained the L2 model using synthetic data. To mimic the LFE waveform, we contaminated the training waveform data (i.e. SCSN data) with synthetic LFE waveforms, as follows:

$$\mathbf{x}^{(p)} \leftarrow 0.5 \times \mathbf{x}^{(p)} + 0.5 \times \mathbf{s}, \quad (3)$$

where  $\mathbf{x}^{(p)}$  denotes an up-down component of the P-phase waveform, and  $\mathbf{s}$  is the synthetic waveform of the LFE. We generated the synthetic waveform  $\mathbf{s}$  based on the Brownian motion model in Ide (2008), which was subsequently bandpass filtered (for more details, please refer to Appendix B and Fig. S5). We considered the following bands for the bandpass filters: 1–2 Hz, 2–4 Hz, 4–6 Hz and 6–8 Hz. These low-frequency bands presumably reflect the dominant frequency of the LFE. Using these contaminated data (with a sample size of 150 000), we retrained the L2 model, which was incorporated into a new GL model in eq. (1) (hereafter referred to as the 'adapted GL model'); the remainder of the submodels, including the G and L1 models, remained unchanged. The performance results of the adapted GL models are summarized in Table 3 (the sixth to the ninth rows). The number of correct detection cases was 22 (4.9 per cent), 44 (9.9 per cent), 32 (7.2 per cent) and 17 (3.8 per cent), respectively for the 1–2 Hz, 2–4 Hz, 4–6 Hz and 6–8 Hz bandpass filters. In contrast, the false detection cases were 1 (0.2 per cent), 3 (0.7 per cent), 6 (1.3 per cent) and 8 (1.8 per cent)

for the 1–2 Hz, 2–4 Hz, 4–6 Hz and 6–8 Hz bandpass filters, respectively. Concerning the ratio of correct and false detections, the adapted model with 1–2 Hz bandpass filters performed the best (ratio = 22), followed by the adapted models with the 2–4 Hz (15), 4–6 Hz (5.3) and 6–8 Hz (2.1) bandpass filters. For comparison with the GL model, the adapted models with the 1–2 Hz, 2–4 Hz and 4–6 Hz bandpass filters significantly outperformed the GL model.

Finally, we analysed the performance of the P-phase detection regarding detection probability. We evaluated the maximum probability of the P-phase before and after the LFE's origin time. In this manner, two probabilities were assigned for a single waveform. The distributions of these probabilities are shown in Fig. 13. It is observed that the GL model and the adapted GL models with 1–Hz and 2–4 Hz bandpass filters separate the two segments to some extent. However, PhaseNet and EQT did not discriminate well between the two segments. Notably, for the GPD model, an effective separation between two segments is mainly observed at a probability larger than 0.9 but is not clearly visible in Fig. 13. Furthermore, to summarize separability based on detection probabilities, we evaluated the AUC (Fig. 14). The adapted GL model with a bandpass filter of 1–2 Hz performed the best (AUC = 0.72), followed by the adapted GL model with a bandpass filter of 2–4 Hz (0.71), GL model (0.68), GPD model (0.65), adapted GL model with a bandpass filter of 4–6 Hz (0.65), adapted GL model with a bandpass filter of 6–8 Hz (0.61), EQT (0.59) and PhaseNet (0.51).

## 5 DISCUSSION

We propose a novel method for seismic-phase detection. The proposed method is an extension of the existing GPD method that considers local waveform information as well as global information. The method is based on deep learning using CNN architectures, which separately model the global and local representations of the waveform. The novelty of our method is that it explicitly focuses on both local and global representations, and subsequently combines them as a probability product. To our knowledge, there has been no neural network method that uses such a separate learning strategy for seismic-phase detection. For other application fields, the separate learning strategy has been successfully used for pedestrian detection (Wang *et al.* 2018) and action detection (Peng & Schmid 2016; Tu *et al.* 2018). In the field of image detection, such a strategy is a promising approach for better performance in object detection. The framework of our method is similar to the attention mechanism adapted by the EQT method (Mousavi *et al.* 2020), which implicitly determines the focal data points in a data-driven manner. The attention mechanism determines the attention weights for each instance (i.e. a waveform in our context); hence, these weights are not necessarily common to different instances. Moreover, the interpretability of the results yielded by the attention mechanism remains controversial owing to the uncertainty of the relationships between the attention weights and model outputs (Jain & Wallace 2019; Niu *et al.* 2021). In contrast, we explored prior knowledge by applying a multiple clustering method to waveform data in which the analyst's pick for the P- and S-phases is centred. It was implied that the first and second halves of the data points convey different information on the seismic phases. This prior knowledge provides an explicit and consistent framework for modelling the local structure of a waveform, which in turn enables robust phase detection and flexible modelling of local information.

The classification performance using the SCSN data set demonstrates the robustness of our method. When the waveform data were

**Table 3.** Summary of P-phase detections for the LFE 100 s before and after the origin time. The column ‘Before-segment’ denotes the results for the time segment 100 s before LFE occurrence, whereas ‘After-segment’ denotes the 100 s after an LFE occurred. The numbers in parenthesis denote percentages of detection over the total number of LFEs (446). The column ‘After/Before’ denotes the ratio of detections before and after the origin time (i.e. the values in column ‘After-segment’ over the values in column ‘Before-segment’). The methods ‘adapted GL’ denote adapted GL models in which submodel L2 was trained using the contaminated data with synthetic LFE waveforms in eq. (3). The frequency band for the bandpass filter to generate synthetic LFE is indicated in parentheses after the method name ‘adapted GL’ in the first column of the table.

Method	Before-segment	After-segment	After/before (ratio)
GL	27 (6.1 per cent)	94 (21 per cent)	<b>3.5</b>
GPD	97 (21 per cent)	213 (48 per cent)	<b>2.2</b>
PhaseNet	42 (9.4 per cent)	38 (8.5 per cent)	<b>0.90</b>
EQT	57 (12 per cent)	94 (21 per cent)	<b>1.6</b>
Adapted GL (1–2 Hz)	1 (0.2 per cent)	22 (4.9 per cent)	<b>22</b>
Adapted GL (2–4 Hz)	3 (0.7 per cent)	44 (9.9 per cent)	<b>15</b>
Adapted GL (4–6 Hz)	6 (1.3 per cent)	32 (7.2 per cent)	<b>5.3</b>
Adapted GL (6–8 Hz)	8 (1.8 per cent)	17 (3.8 per cent)	<b>2.1</b>

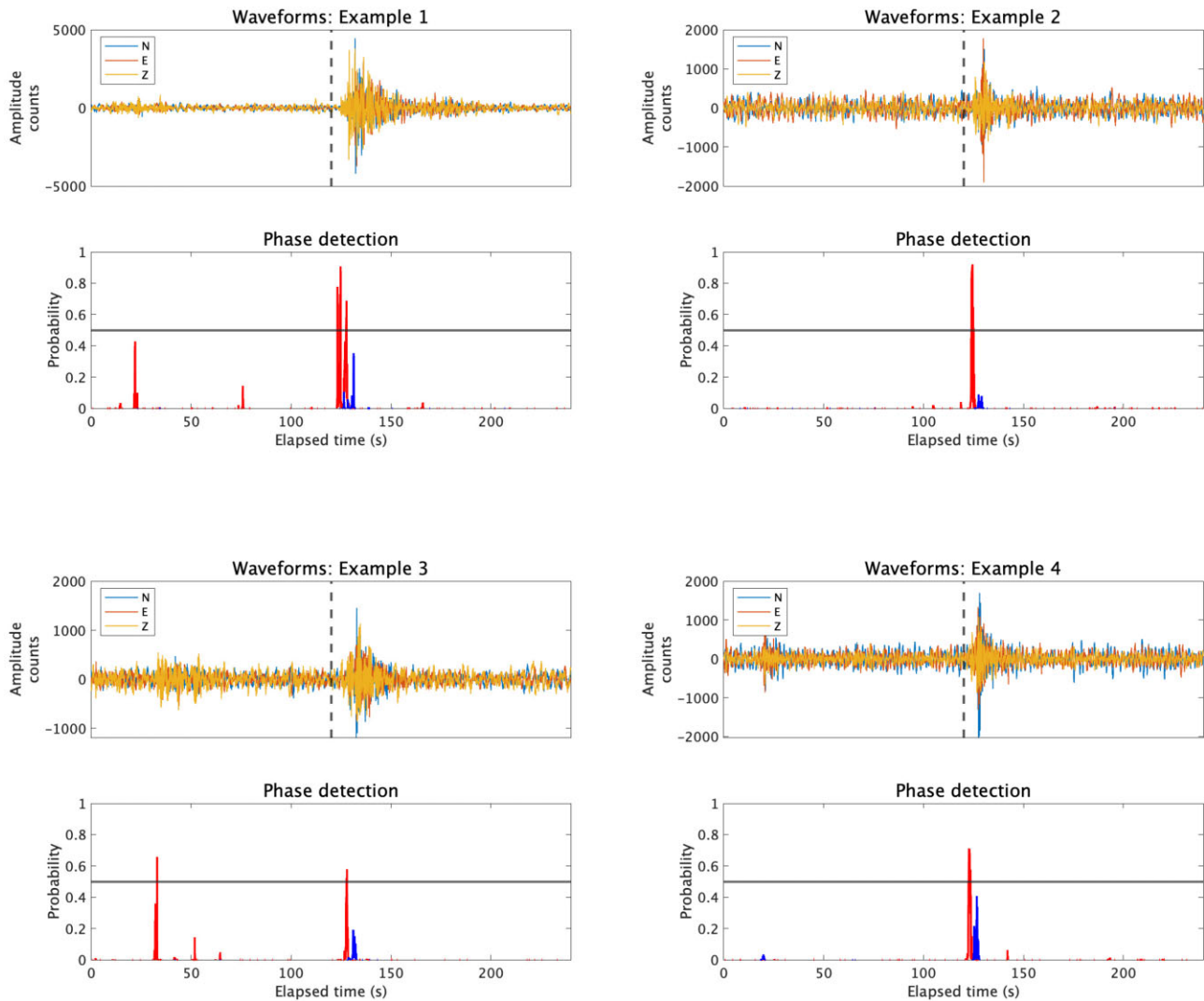
not contaminated with noise, the separate learning strategy of the proposed method performed almost equally to the G model for both test data and STEAD data. Apparently, there would be no gain for the proposed method, which further required the additional training cost for L1 and L2 models. However, when the waveform data were partially contaminated with noise, our proposed method performed excellently, outperforming the global representation method (e.g. for a noise proportion of 0.75 in the first half of data points, the accuracy of the GL model and G model was 0.88 and 0.65, respectively). In particular, it was found that when there was a difference in noise type between the first and second halves of the data points, the proposed method significantly improved the performance of noise phase detection. This was because the performance of neither the L2 model nor the L1 model was influenced by the noise proportion in the first or second halves of the data points, which in turn contributed to the high performance of the GL model. These results suggest that the proposed method can potentially reduce the false-positive cases caused by noise. It is noted that non-seismic waveform with such an abrupt change of amplitude in Fig. 6 was reported in several papers (Yang *et al.* 2022; Zhu *et al.* 2019). The sources for such noise are generally seldom obvious, but in the Green *et al.* (2017) study, the effect of subway trains was clearly identified as a source. In general, waveform data obtained from observation stations located in urban areas such as the Metropolitan Seismic Observation network (MeSO-net) in Japan (Aoi *et al.* 2021) may potentially suffer from urban noise. It is expected that the proposed method may effectively reduce false positive cases arising from urban noise.

The robustness of the proposed method was verified by applying it to the 2016 Bombay Beach swarm. Our method identified a considerable number of seismic phases after the onset of the swarm, whereas it identified only a few cases before the onset. Overall, the proposed method outperformed the other methods, including the GPD model, PhaseNet and EQT. This observation is further supported by the result of the AUC analysis of the probability distributions, which suggests that the proposed method outperforms the other methods irrespective of the detection threshold. Moreover, in this application, a marked difference between the GPD model and the proposed method is evident. In Fig. 9 (before the swarm onset), the GPD model yielded a large number of high probabilities for the P- and S-phases, whereas the proposed method yielded few.

By setting a high probability threshold (0.98), the GPD model reduces false-positive cases. Nonetheless, the interpretation of such a large threshold is not straightforward. In contrast, the proposed method sets the threshold to 0.5, which facilitates the theoretical interpretation of phase probability.

Moreover, the application of the proposed method to LFE detection in Japan demonstrated the usefulness of the separation learning strategy. In this experiment, based on the Brownian motion model, we generated synthetic LFE waveforms, which were further bandpass filtered. Subsequently, we retrained the L2 model using contaminated data with synthetic LFEs. This adapted procedure obtains better performance from the adapted models with 1–2 Hz and 2–4 Hz bandpass filters than the other methods, including the GL model, the GPD model, PhaseNet and EQT. Overall, this result is consistent with the nature of the LFE waveform, in which the lower frequencies dominate. To clarify this, we performed a supplementary analysis to evaluate the mean dynamic spectrogram of 10-min waveforms of 446 LFEs (Fig. 11). It was observed that before the origin time, a frequency band of less than 1 Hz (background noise) dominated. In contrast, a high-power intensity for 1–4 Hz was observed between the 0- and 100-s period, after the origin time. This suggests that an LFE occurred during this time with a characteristic frequency of 1–4 Hz. This frequency that dominated the waveform data may be attributed to the better performance of the adapted models with 1–2 Hz and 2–4 Hz bandpass filters. Notably, the P-phase detection of the test data (conventional EQ) by these adapted models deteriorates considerably (recall is 0.035 without noise contamination, Fig. S6), which suggests that the observed good performance may be limited to the LFEs. Furthermore, theoretically, this application demonstrates that the proposed method has a specific form of transfer learning in which the learned model (conventional EQ) is adapted to a different context (LFE) by retraining a submodel (i.e. the L2 model). This type of transfer learning is a unique feature of the proposed method owing to the separate learning strategy of the submodels. Moreover, this application suggests the possibility of developing an LFE phase-detection model without using the LFE waveform, which is often not readily available with phase labels (Kurihara & Obara 2021).

As regards the onset time picking, the performance of the GL model slightly degraded compared with the G model (i.e. GPD; Fig. 4). A possible reason for the deterioration is due to a lower

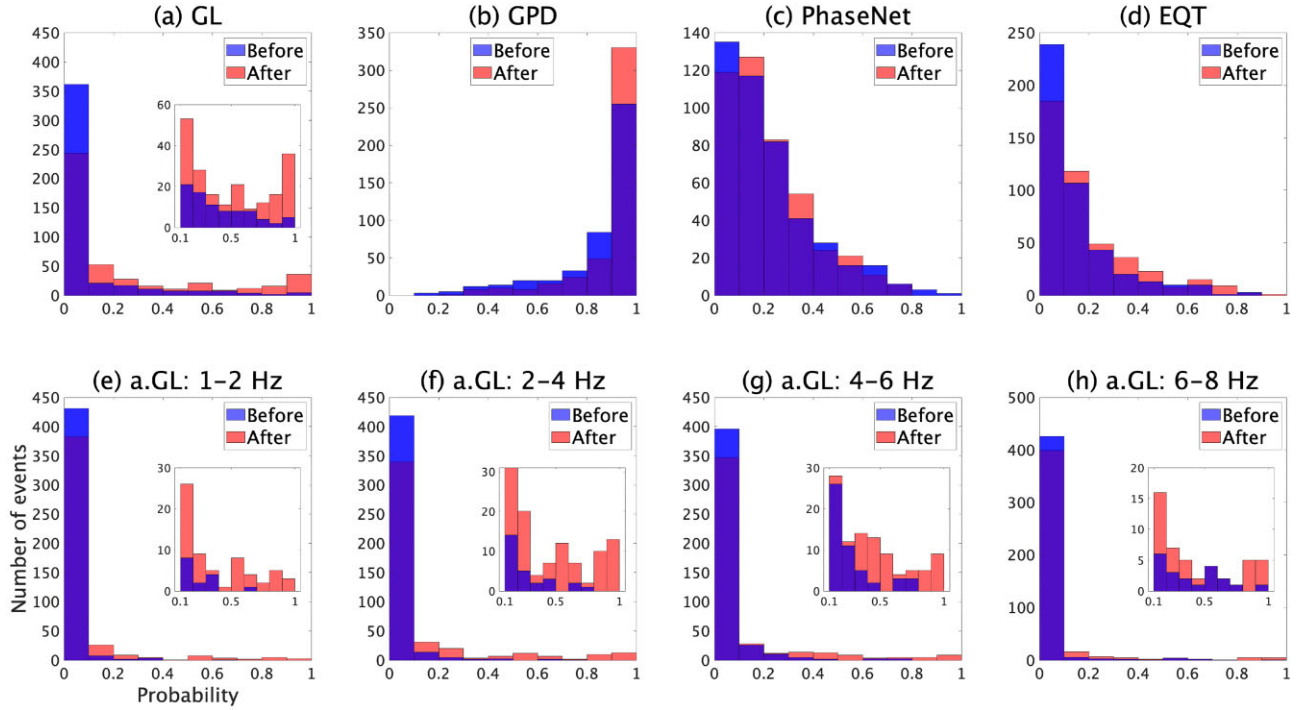


**Figure 12.** Examples of LFE detection by the GL model. The 4-min (240 s) waveform of three components and phase detection probabilities are displayed for the upper subpanel and lower subpanel, respectively. For phase detection, the red line denotes a probability for P-phase, whereas the blue line is for S-phase. The horizontal black line denotes the threshold (0.5) for phase detection. The displayed examples were randomly sampled from those cases (78 cases) in which the GL model detected a P-phase within 30 s after the origin time. The origin time is centred, which is denoted by a dashed line in the upper panel. Note that there are some time lags between the origin time and P-phase detection, possibly due to the geometrical distance between the hypocentre and an observation station. Furthermore, note that there were no earthquakes before the onset time due to the selection criterion of LFE as described in the second paragraph of Section 4.6. Hence, before the onset time, the events with a probability of  $\geq 0.5$  are false positive cases, whereas those with a probability of  $< 0.5$  are correctly identified as noise.

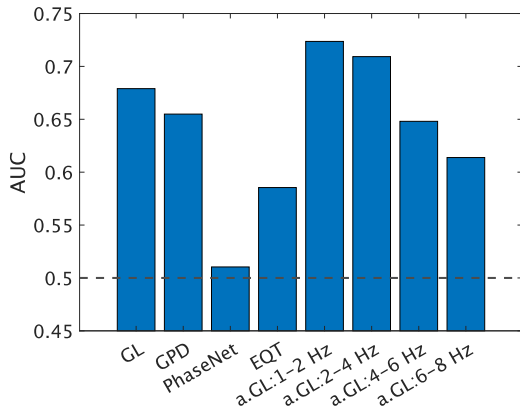
capability of identifying an onset time by L1 and L2 models. Conversely, the general performance of picking the onset time by the GPD model is very good in the study of Münchmeyer *et al.* (2022). Hence, an effective approach to picking the onset time would be to optimize using the high capability of the G model in the framework of the GL model. A possible approach would be to identify the onset time (referred to as ‘tentative onset time’) by the GL model. Setting a small time-window centred on the tentative onset time, we would identify the onset time using the G model within this time-window, which is taken as the final onset time. We re-evaluated the performance of the phase picking following this procedure with the time-window 0.9 s (i.e.  $2 \times 0.45$  s; Fig. S7). As had been expected, the performance obviously improved, which became compatible with that of the G model. It is noted that the performance may depend on the choice of the time-window, which requires optimization. This is left for a future study.

Next, we discuss further theoretical aspects of the proposed method, which are useful for feature research on the development of phase-detection methods. The key idea of our method is based on the observations that the first and second halves of the waveform *per se* have sufficient information to discriminate among the P-phase, S-phase and noise (Fig. 3). Existing detection methods can potentially improve their performance by incorporating local information as an additional component in the model. Secondly, the proposed method can be improved by designing completely different neural network architectures for local information. For local models, we simply adopted the same CNN architecture as the global model; however, it would be worth considering choosing a CNN (possibly with different filter size) or other architecture that is best suited for the local model. For instance, one may consider a Long Short-Term Memory Network (LSTM; Hochreiter & Schmidhuber 1997; Sherstinsky 2020) for a L2 model, which can potentially capture





**Figure 13.** The maximum probabilities for P-phase detection before and after the LFE's origin time. Each panel denotes a histogram for the number of maximum P-phase probabilities in the 100-s period before the LFE (blue bar) and in the 100-s period after the LFE (red bar). The horizontal axis denotes the maximum probability for P-phase detection, and the vertical axis denotes the number of LFE events (out of 446 events). Panels (a)–(h) are for the GL model, the GPD model, PhaseNet, EQT and adapted GL models, respectively, with the 1–2 Hz, 2–4 Hz, 4–6 Hz and 6–8 Hz bandpass filters. For Panels (a), (e), (f), (g) and (h), the inset figure shows the same histogram as the main figure, which focuses on a maximum probability larger than 0.1.



**Figure 14.** The AUC evaluation of the detection differences before and after the LFE. For each LFE (out of 446), we evaluated the maximum probability of P-phase detection before and after the origin time. As a result, there are 446 maximum probabilities for the time segment ‘before’ and the time segment ‘after’, respectively. The distributions of these maximum probabilities are displayed in Fig. 13. For these maximum probabilities of P-phase detection, the AUC is evaluated for binary time segment classification ‘before’ and ‘after’, which, between two time segments, measures the separability of the maximum probabilities. The horizontal axis denotes detection methods, whereas the vertical axis denotes AUC value. The dashed line denotes the chance level of the AUC value, that is 0.5. Abbreviations for adapted GL model: ‘a.GL: 1–2 Hz’, adapted GL model with the 1–2 Hz bandpass filter; ‘a. GL: 2–4 Hz’, adapted GL model with the 2–4 Hz bandpass filter; ‘a. GL: 4–6 Hz’, adapted GL model with the 4–6 Hz bandpass filter; ‘a. GL: 6–8 Hz’, adapted GL model with the 6–8 Hz bandpass filter.

different time-dependency among three phases. Thirdly, one may generalize the binary power weight in eq. (1), which allows for real-valued power weight. The quality of information for classification may differ between the global and local models, which may in turn depend on the waveform data in question. Hence, it is a promising approach to determine these weights in a data-driven manner by setting a user-defined criterion.

Finally, we discuss the limitations of this study. We evaluated the performance of phase detection for the 2016 Bombay Beach swarm before the onset of the swarm by removing the detection results from 1 min after the earthquakes in the SCEDC catalogue. Nonetheless, we cannot rule out the possibility of non-catalogued earthquakes. The same limitation applies to the performance evaluation of low-frequency earthquakes in Japan. Moreover, in the latter case, there is another limitation to the performance analysis: We assumed that all LFEs were detectable by waveform data from a single station. This assumption may result in an underestimation of the detection capabilities of these methods. Further detailed analysis of the waveform could potentially overcome some of the limitations, which are worth further investigation. Second, we did not take into account the possible impact of filtering waveform data. The phase-detection models in this study were trained with different default filtering settings: high-pass filtered with 2 Hz for the G- and GL models; bandpass filtered with 1–45 Hz for the EQT; no bandpass filtering for PhaseNet. Because of the large overlap of frequency bands for pre-processing, for simplicity, we applied a high-pass filter with 2 Hz for 2016 Bombay Beach swarm, whereas we did not apply any bandpass filtering for LFEs. Nonetheless, we cannot rule out a possible impact on the results of this simplification.

## SUPPORTING INFORMATION

Supplementary data are available at [GJI](#) online.

## suppl.data

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

## ACKNOWLEDGMENTS

This study was supported by the MEXT Project for Seismology Toward Research Innovation with Data of Earthquake (STAR-E) grant JPJ010217. The key ideas in this study were derived from the activities of JST CREST under grants JPMJCR1761 and JPMJCR1763, JSPS KAKENHI Grant-in-Aid for Scientific Research (A) No. 23H00466, Grant-in-Aid for Challenging Research (Exploratory) No. 20K21785, Grant-in-Aid for Scientific Research (S) No. 19H05662 and Earthquake Research Institute, University of Tokyo Joint Research ERI JURP 2023-A-03, 2021-B-01 and 2022-B-06.

## 6 DATA AVAILABILITY

Southern California Seismic Network data (SCSN) used for the training, validation and test is available at: <https://scedc.caltech.edu/data/deeplearning.html>. Stanford Earthquake Dataset (STEAD) is publicly available through: <https://github.com/smousavi05/STEAD>. The continuous data for Bombay Beach swarm is available (<https://scedc.caltech.edu/index.html>), whereas that for low-frequency earthquakes at: <https://www.hinet.bosai.go.jp/?LANG=en>. Catalog of LFEs are available at: <https://www.data.jma.go.jp/eqev/data/bulletin/index.html> (in Japanese).

## REFERENCES

- Aki, K. & Richards, P.G., 2002. *Quantitative Seismology*, 2nd edn, University Science Books.
- Allen, R.V., 1978. Automatic earthquake recognition and timing from single traces, *Bull. seism. Soc. Am.*, **68**(5), 1521–1532.
- Aoi, S., Kimura, T., Ueno, T., Senna, S. & Azuma, H., 2021. Multi-data integration system to capture detailed strong ground motion in the Tokyo Metropolitan Area, *J. Disaster Res.*, **16**(4), 684–699.
- Baer, M. & Kradolfer, U., 1987. An automatic phase picker for local and teleseismic events, *Bull. seism. Soc. Am.*, **77**(4), 1437–1445.
- Bahdanau, D., Cho, K. & Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate, in *3rd International Conference on Learning Representations Conference Track Proceedings*, San Diego, CA, USA, 7–9 May 2015, preprint (arXiv:1409.0473).
- Buckland, M. & Gey, F., 1994. The relationship between recall and precision, *J. Am. Soc. Inform. Sci.*, **45**(1), 12–19.
- California Institute of Technology, 1926. Southern California Seismic Network. International Federation of Digital Seismograph Networks. Other/Seismic Network. doi:10.7914/SN/CI.
- Farabet, C., Couprie, C., Najman, L. & LeCun, Y., 2012. Learning hierarchical features for scene labeling, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**(8), 1915–1929.
- Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recog. Lett.*, **27**(8), 861–874.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B., 2013. *Bayesian Data Analysis*, 3rd edn, Chapman and Hall/CRC.
- Géron, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media.
- Gidaris, S. & Komodakis, N., 2015. Object detection via a multi-region and semantic segmentation-aware CNN model, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1134–1142.
- Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning*, MIT Press.
- Green, D.N., Bastow, I.D., Dashwood, B. & Nippress, S.E., 2017. Characterizing broadband seismic noise in Central London, *Seismol. Res. Lett.*, **88**(1), 113–124.
- Hand, D.J., 2012. Assessing the performance of classification methods, *Int. Stat. Rev.*, **80**(3), 400–414.
- Hochreiter, S. & Schmidhuber, J., 1997. Long short-term memory, *Neural Comput.*, **9**(8), 1735–1780.
- Hubert, L. & Arabie, P., 1985. Comparing partitions, *J. Classif.*, **2**(1), 193–218.
- Ide, S., 2008. A Brownian walk model for slow earthquakes, *Geophys. Res. Lett.*, **35**(17), doi:10.1029/2008GL034821.
- Ide, S., Beroza, G.C., Shelly, D.R. & Uchide, T., 2007. A scaling law for slow earthquakes, *Nature*, **447**(7140), 76–79.
- Ide, S. & Maury, J., 2018. Seismic moment, seismic energy, and source duration of slow earthquakes: application of Brownian slow earthquake model to three major subduction zones, *Geophys. Res. Lett.*, **45**(7), 3059–3067.
- Jain, S. & Wallace, B.C., 2019. Attention is not explanation, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, NAACL, Minneapolis, MN, Association for Computational Linguistics.
- Japan Meteorological Agency, 2022. Japanese seismic catalog (in Japanese), available at: <https://www.data.jma.go.jp/eqev/data/bulletin/index.html>, accessed on 01.08.2022.
- Kato, A. & Ben-Zion, Y., 2021. The generation of large earthquakes, *Nat. Rev. Earth Environ.*, **2**(1), 26–39.
- Kato, A. & Nakagawa, S., 2020. Detection of deep low-frequency earthquakes in the Nankai subduction zone over 11 years using a matched filter technique, *Earth, Planets Space*, **72**(1), 1–9.
- Kocharyan, G., 2021. Nucleation and evolution of sliding in continental fault zones under the action of natural and man-made factors: a state-of-the-art review, *Izvest., Phys. Solid Earth*, **57**(4), 439–473.
- Kurihara, R. & Obara, K., 2021. Spatiotemporal characteristics of relocated deep low-frequency earthquakes beneath 52 volcanic regions in Japan over an analysis period of 14 years and 9 months, *J. geophys. Res.*, **126**(10), e2021JB022173, doi:10.1029/2021JB022173.
- Li, Y., Ouyang, W., Zhou, B., Wang, K. & Wang, X., 2017. Scene graph generation from objects, phrases and region captions, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1261–1270.
- Li, Y., Schofield, E. & Gönen, M., 2019. A tutorial on Dirichlet process mixture modeling, *J. Math. Psychol.*, **91**, 128–144.
- Liao, W.-Y., Lee, E.-J., Mu, D. & Chen, P., 2022. Toward fully autonomous seismic networks: backprojecting deep learning-based phase time functions for earthquake monitoring on continuous recordings, *Seismol. Soc. Am.*, **93**(3), 1880–1894.
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X. & Pietikäinen, M., 2020. Deep learning for generic object detection: a survey, *Int. J. Comput. Vis.*, **128**(2), 261–318.
- Lomax, A., Satriano, C. & Vassallo, M., 2012. Automatic picker developments and optimization: Filterpicker—a robust, broadband picker for real-time seismic monitoring and earthquake early warning, *Seismol. Res. Lett.*, **83**(3), 531–540.
- McBride, S.K., Llenos, A.L., Page, M.T. & Van Der Elst, N., 2020. #EarthquakeAdvisory: exploring discourse between government officials, news media, and social media during the 2016 Bombay beach swarm, *Seismol. Res. Lett.*, **91**(1), 438–451.
- Mousavi, S.M. & Beroza, G.C., 2022. Deep-learning seismology, *Science*, **377**(6607), doi:10.1126/science.abm4470.

- Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y. & Beroza, G.C., 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Commun.*, **11**(1), 1–12.
- Mousavi, S.M. & Langston, C., 2016. Fast and novel microseismic detection using time-frequency analysis, in *SEG Technical Program Expanded Abstracts 2016*, pp. 2632–2636, Society of Exploration Geophysicists.
- Mousavi, S.M., Sheng, Y., Zhu, W. & Beroza, G.C., 2019. STanford EArthquake Dataset (STEAD): a global data set of seismic signals for AI, *IEEE Access*, **7**, 179 464–179 476.
- Münchmeyer, J. et al., 2022. Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers, *J. geophys. Res.*, **127**(1), e2021JB023499, doi:10.1029/2021JB023499.
- Niu, Z., Zhong, G. & Yu, H., 2021. A review on the attention mechanism of deep learning, *Neurocomputing*, **452**, 48–62.
- Obara, K., 2002. Nonvolcanic deep tremor associated with subduction in southwest Japan, *Science*, **296**(5573), 1679–1681.
- 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**(8), xv–xxviii.
- Peng, X. & Schmid, C., 2016. Multi-region two-stream R-CNN for action detection, in *European Conference on Computer Vision*, pp. 744–759, Springer.
- Peng, Z. & Gombert, J., 2010. An integrated perspective of the continuum between earthquakes and slow-slip phenomena, *Nat. Geosci.*, **3**(9), 599–607.
- Perol, T., Gharbi, M. & Denolle, M., 2018. Convolutional neural network for earthquake detection and location, *Sci. Adv.*, **4**(2), e1700578, doi:10.1126/sciadv.1700578.
- Ross, Z.E., Meier, M.-A., Hauksson, E. & Heaton, T.H., 2018. Generalized seismic phase detection with deep learning, *Bull. seism. Soc. Am.*, **108**(5A), 2894–2901.
- Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.*, **1**(5), 206–215.
- SCEDC, 2013. Southern California Earthquake Center. Caltech. Dataset. doi:10.7909/C3WD3xH1.
- Shelly, D.R., 2017. A 15 year catalog of more than 1 million low-frequency earthquakes: tracking tremor and slip along the deep San Andreas Fault, *J. geophys. Res.*, **122**(5), 3739–3753.
- Shelly, D.R., Beroza, G.C. & Ide, S., 2007. Non-volcanic tremor and low-frequency earthquake swarms, *Nature*, **446**(7133), 305–307.
- Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Phys. D*, **404**, doi:10.1016/j.physd.2019.132306.
- Soto, H. & Schurr, B., 2021. Deepphasepick: a method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks, *J. geophys. Int.*, **227**(2), 1268–1294.
- Stevenson, P.R., 1976. Microearthquakes at Flathead Lake, Montana: a study using automatic earthquake processing, *Bull. seism. Soc. Am.*, **66**(1), 61–80.
- Thomas, A.M., Inbal, A., Searcy, J., Shelly, D.R. & Bürgmann, R., 2021. Identification of low-frequency earthquakes on the San Andreas Fault with deep learning, *Geophys. Res. Lett.*, **48**(13), e2021GL093157, doi:10.1029/2021GL093157.
- Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., Yamawaki, S. & Doya, K., 2017. Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions, *PLoS one*, **12**(10), e0186566, doi:10.1371/journal.pone.0186566.
- Tokuda, T., Yoshimoto, J., Shimizu, Y., Okada, G., Takamura, M., Okamoto, Y., Yamawaki, S. & Doya, K., 2018. Identification of depression subtypes and relevant brain regions using a data-driven approach, *Sci. Rep.*, **8**(1), 1–13.
- Tu, Z., Xie, W., Dauwels, J., Li, B. & Yuan, J., 2018. Semantic cues enhanced multimodality multistream CNN for action recognition, *IEEE Trans. Circuits Syst. Video Technol.*, **29**(5), 1423–1437.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I., 2017. Attention is all you need, in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.
- Virtanen, P. et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Methods*, **17**, 261–272.
- Wang, S., Cheng, J., Liu, H. & Tang, M., 2018. PCN: Part and context information for pedestrian detection with CNNs, preprint (arXiv:1804.04483).
- Woollam, J. et al., 2022. Seisbench—a toolbox for machine learning in seismology, *Seismol. Soc. Am.*, **93**(3), 1695–1709.
- Woollam, J., Rietbrock, A., Bueno, A. & De Angelis, S., 2019. Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network, *Seismol. Res. Lett.*, **90**(2A), 491–502.
- Yang, L., Liu, X., Zhu, W., Zhao, L. & Beroza, G.C., 2022. Toward improved urban earthquake monitoring through deep-learning-based noise suppression, *Sci. Adv.*, **8**(15), eabl3564.
- Yang, S., Hu, J., Zhang, H. & Liu, G., 2021. Simultaneous earthquake detection on multiple stations via a convolutional neural network, *Seismol. Res. Lett.*, **92**(1), 246–260.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. & Hovy, E., 2016. Hierarchical attention networks for document classification, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489.
- Zeng, X., Ouyang, W., Yang, B., Yan, J. & Wang, X., 2016. Gated bi-directional CNN for object detection, in *European Conference on Computer Vision*, pp. 354–369, Springer.
- Zhou, Y., Yue, H., Kong, Q. & Zhou, S., 2019. Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks, *Seismol. Res. Lett.*, **90**(3), 1079–1087.
- Zhu, W. & Beroza, G.C., 2019. PhaseNet: a deep-neural-network-based seismic arrival-time picking method, *J. geophys. Int.*, **216**(1), 261–273.
- Zhu, W., Mousavi, S.M. & Beroza, G.C., 2019. Seismic signal denoising and decomposition using deep neural networks, *IEEE Trans. Geosci. Remote Sens.*, **57**(11), 9476–9488.
- Zhu, W., Tai, K.S., Mousavi, S.M., Bailis, P. & Beroza, G.C., 2022. An end-to-end earthquake detection method for joint phase picking and association using deep learning, *J. geophys. Res.*, **127**(3), e2021JB023283, doi:10.1029/2021JB023283.
- Zhu, Y., Zhao, C., Wang, J., Zhao, X., Wu, Y. & Lu, H., 2017. CoupletNet: coupling global structure with local parts for object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4126–4134.

## APPENDIX A:

### RECALL/PRECISION/ACCURACY

Conventionally, recall and precision terminologies are used to evaluate the binary classification performance. In this study, we extended this definition to multiclass classification performance.  $n_{i,j}$  denotes the number of instances of the true phase  $i$  that are detected as phase  $j$  by a phase detection method. Here, we assume three phases in the context:  $i, j \in \{‘p’, ‘s’, ‘n’\}$ . We define the recall and precision of phase  $i$  as follows:

$$\text{Recall} = \frac{n_{i,i}}{\sum_{j=1}^3 n_{i,j}} \quad (\text{A1})$$

$$\text{Precision} = \frac{n_{i,i}}{\sum_{j=1}^3 n_{j,i}}. \quad (\text{A2})$$

Similarly, we define accuracy as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^3 n_{i,i}}{\sum_{i=1}^3 \sum_{j=1}^3 n_{i,j}}. \quad (\text{A3})$$

## APPENDIX B: SYNTHETIC LFE WAVEFORM

To generate a synthetic LFE waveform, we followed the Brownian walk model proposed by Ide (2008). Here, we summarize the model, including the specifications of the relevant parameters.

First, we consider a circular fault with radius  $r_t$  that changes dynamically with time  $t$ . We assume that the radius  $r_t$  follows a differential equation:

$$dr_t = -\alpha dt + \sigma dB_t, \quad (\text{B1})$$

where  $dB_t$  is a random variable with a Gaussian distribution  $N(0, dt)$ ,  $\alpha$  is the damping coefficient and  $\sigma$  is the diffusion coefficient. Furthermore, assuming that shear slip occurs for this circular fault with a constant velocity  $v_0$ , the seismic moment rate  $\dot{M}_0$  is given by

$$\dot{M}_0 = \mu \times \pi r_t^2 \times v_0, \quad (\text{B2})$$

where  $\mu$  is the rigidity.

Hence, for a small value of  $dt$ ,

$$(\ddot{M}_0)^2 \sim 4\pi^2 \mu^2 v_0^2 \sigma^2 r_t^2 / dt \propto r_t^2. \quad (\text{B3})$$

In Ide (2008), both intermediate-field and far-field were considered for  $S$ -wave generation. Here, for simplicity, we considered only the far field for  $P$ -wave generation. The first derivative of the  $P$ -wave

displacement  $u$  is given by eq. (4.32) in Aki & Richards (2002)

$$\dot{u}(t) = \frac{1}{4\pi\rho v_p^2 r} A^{FP} \ddot{M}_0(t - r/v_p), \quad (\text{B4})$$

where  $r$  denotes the distance from the source to the station.  $\rho$  density,  $v_p$  velocity of the  $P$  wave,  $A^{FP}$  radiation patterns for the far-field  $P$  waves. Our aim is to generate a waveform in a single component with an arbitrary amplitude (the amplitude is normalized later). Therefore, we further simplify eq. (B4) while ignoring the time delay.

$$\dot{u}(t) \propto \ddot{M}_0(t). \quad (\text{B5})$$

We consider that the waveform generated using eq. (B5) mimics the LFE.  $\ddot{M}_0(t)$  is generated based on eq. (B3), where the sign of  $\ddot{M}_0(t)$  is determined by the sign of  $\dot{M}_0(t + dt) - \dot{M}_0(t)$ . We set the relevant parameters as in Ide (2008).  $\alpha = 0.02 \text{ s}^{-1}$ ;  $\sigma = 400 \text{ m s}^{-1/2}$ ;  $dt = 0.01 \text{ s}$ . Using these formulations, we stochastically updated  $r_t$  based on eq. (B1) for 2 s starting with  $r_t = 0$ . For each time point, we evaluate  $\dot{u}(t)$  in eq. (B5), which constitutes a single LFE waveform. Because each time step  $dt = 0.01$ , this procedure generates 200 data points of the LFE with a frequency of 100 Hz. Subsequently, we applied bandpass filters to the obtained waveform (1–2 Hz, 2–4 Hz, 4–6 Hz and 6–8 Hz). We repeat this procedure to generate a synthetic waveform  $s$  for each instance of  $\mathbf{x}^{(p)}$  in eq. (3). Lastly, we normalized  $s$  such that the maximum absolute amplitude became identical to that of  $\mathbf{x}^{(p)}$ .