Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving Deep Learning-Based Seismic Phase Picking by Addressing Label Imbalance

Shinya Katoh
The University of Tokyo
Masaaki Imaizumi
The University of Tokyo
Yoshihisa Iio
Kyoto University
Hiromichi Nagao
nagaoh@eri.u-tokyo.ac.jp

The University of Tokyo

Article

Keywords:

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10439246/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Improving Deep Learning-Based Seismic Phase Picking by Addressing Label Imbalance

Shinya Katoh[1], Masaaki Imaizumi[2,3,4], Yoshihisa Iio[5], and Hiromichi Nagao[1,3,6,*]

[1]Earthquake Research Institute, The University of Tokyo, Tokyo 113-0032, Japan
[2]Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo 153-8902, Japan
[3]RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan
[4]Division of Physics and Astronomy, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan
[5]Disaster Prevention Research Institute, Kyoto University, Kyoto 611-0011, Japan
[6]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan
[*]Corresponding author: Hiromichi Nagao (nagaoh@eri.u-tokyo.ac.jp)

ABSTRACT

Deep learning-based phase-picking models have significantly improved seismic waveform analyses efficiency. However, detection performance degrades when multiple seismic phases occur within a single waveform trace, particularly during intense seismic sequences such as aftershock activity. This may be attributed to pronounced label imbalance in semantic-segmentation-based phase picking, where narrow P- and S-phase label regions are overwhelmed by background label. To address this, we introduce an area-weighted soft cross-entropy loss that emphasizes phase-arrival regions during training without modifying model architecture. The proposed loss function was evaluated using two phase-picking architectures, SegPhase and PhaseNet. Experiments on the test dataset demonstrated that the weighted loss function reduced the scatter of arrival time residuals and improved the temporal consistency of predicted picks. Weighted-loss models also exhibited more recall-oriented behavior, reducing missed arrivals, though at the cost of increased false positives. Application to continuous waveform data from the 2019 Ridgecrest earthquake sequence showed that the weighted-loss models captured the main spatial and temporal characteristics of the seismicity while detecting more events than the original-loss models. These results demonstrate that loss function design constitutes an effective approach for improving the sensitivity of deep learning-based phase picking and enhancing event recovery in continuous waveform analysis.

Introduction

Phase picking is one of the most fundamental and important analytical procedures in seismology. Accurate travel time information is essential for event detection, hypocenter determination, and a broader understanding of seismic activity. However, conventional analysis requires experts to manually determine arrival times by visually inspecting waveform data, which requires substantial time and effort. In recent years, the rapid increase in the amount of waveform data acquired through advances in observational technology has highlighted the limitations of manual analysis. To overcome this problem, many automatic phase-picking models based on deep learning have been developed, and their effectiveness has been widely reported[1–8]. These models have achieved an accuracy comparable to that of human experts, while providing a substantially higher computational efficiency. Because of this high efficiency, the application of deep learning models to continuous waveform data has enabled the detection of earthquakes that are difficult to identify by manual analysis, thereby contributing to a more detailed understanding of seismic activity[9–14].

Deep learning-based phase picking has the potential to significantly advance the automation of fundamental seismic analyses. However, an important problem remains: when multiple seismic phases are present within a single trace, it is difficult to detect all arrival times accurately[7,15]. In addition,[15] reported that the number of events detected by deep learning models was substantially lower than that identified by template matching. These limitations may not be critical for routine monitoring of ordinary seismicity, but they become a significant problem in situations where the number of earthquakes per unit time increases, such as aftershock sequences following large earthquakes or intensely induced seismicity associated with fluid

injection experiments. Therefore, improvement of deep learning models is necessary to accurately identify arrival times, even when multiple seismic phases are included in a single trace.

To address this problem,[15] proposed a method in which the input window was finely shifted and inference was repeatedly performed on the same waveform. This approach is effective because changing the input position facilitates the detection of multiple seismic phases. However, when applied to long, continuous waveforms, the computational cost is extremely high. Although[15] attempted to mitigate this problem by modifying the inference procedure, in this study, we focus on the training stage rather than the inference stage, and propose a model improvement approach to address this issue.

In this study, we considered the possibility that one factor responsible for the deterioration of detection performance in traces containing multiple seismic phases is label imbalance[16]. In recent years, phase picking has often been formulated as a semantic segmentation problem. Here, we describe the label representation assumed in this study using the example shown in Fig. 1. The input consists of three-component waveform data containing seismic signals, and the target labels consist of P- and S-phase labels represented by Gaussian distributions centered at their arrival times with a standard deviation of 0.1 s, together with a background (N) label representing the remaining region (Fig. 1(a), (b), (c), and (d)). The model outputs a probability distribution over the classes for each time sample, and arrival time is determined from the sample in which the output probability reaches its maximum. For example, in this setting, when a 30-second waveform sampled at 100 Hz is used, the label-area ratios during training are approximately 1% for the P-phase label, 1% for the S-phase label, and 98% for the N label. In other words, most of the label area is occupied by the N label, whereas the P- and S-phase labels, which are essential for phase picking, occupy only limited regions. Consequently, under a standard loss function, the training loss is dominated mainly by the N label, and the contributions of the P- and S-phase labels are relatively small. Consequently, under such N-dominant conditions, the model may become more likely to miss signals corresponding to P- and S-phase arrivals, leading to reduced sensitivity for arrival-time detection.

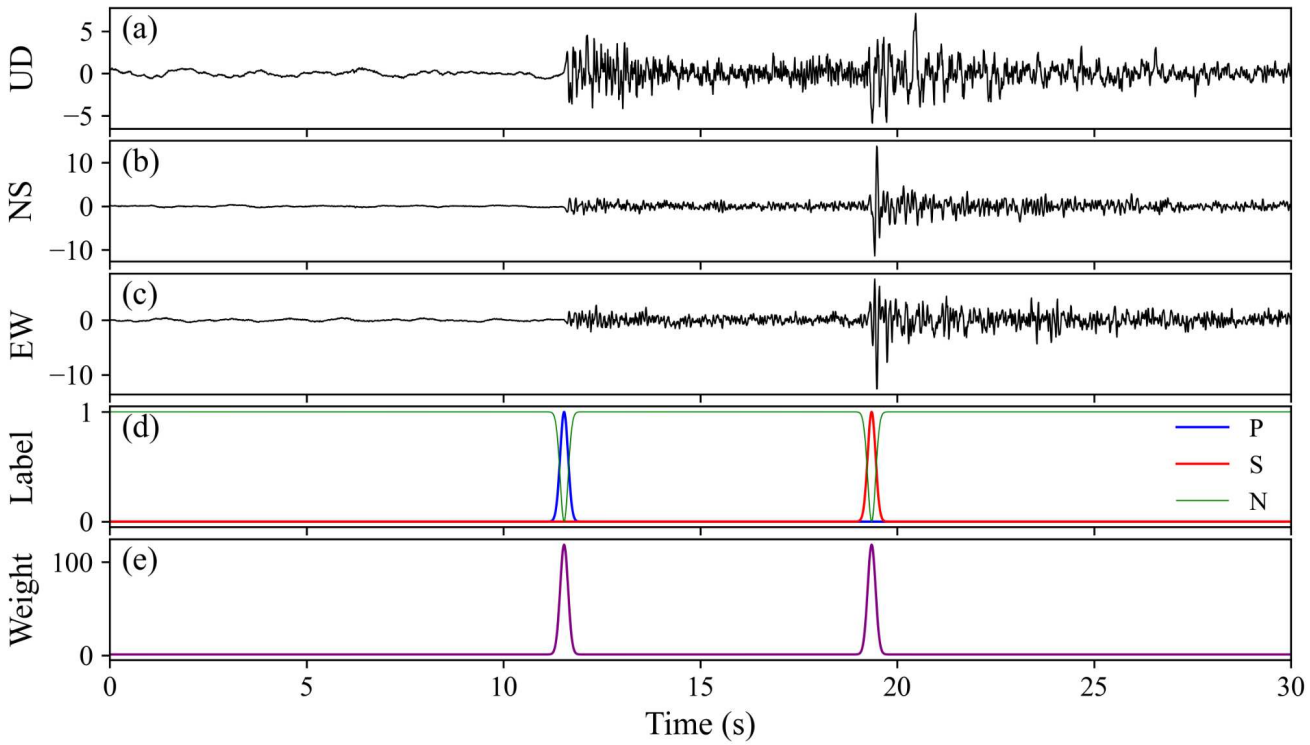


Figure 1. Target labels and weighting scheme used for the weighted loss. Example of waveform data, target labels, and time-dependent weights used in the proposed area-weighted soft cross-entropy loss. Panels (a), (b), and (c) show the three-component input waveform normalized by Z-score normalization. Panel (d) shows the Gaussian-shaped target labels for the P phase, S phase, and N(background) class. Panel (e) shows the corresponding time-dependent weight used in the proposed loss function.

To address this label imbalance and improve model sensitivity to the P and S phases, we introduced a weighted cross-entropy loss function. In standard cross-entropy loss, all classes and time samples are treated equally; thus, the loss tends to be dominated by the N label. In contrast, the weighted cross-entropy loss introduced in this study considers that the temporal areas of the P- and S-phase labels are substantially smaller than those of the N label, and assigns larger weights to the relatively small label regions. Thus, the contribution of samples around the phase arrival times is emphasized without changing the model architecture. First, we evaluated phase-picking performance using models trained with the standard and weighted cross-entropy losses. The trained models were then applied to continuous waveform data during the aftershock period of the 2019 Ridgecrest earthquake to perform event detection. Finally, we examined the effectiveness of the proposed loss function by comparing the number of detected events with that obtained through template matching.

Methods

In this study, we investigated whether the performance degradation of deep learning-based phase-picking models in traces containing multiple seismic phases can be mitigated by modifying the loss function. SegPhase[7] and PhaseNet[2] were used as the base model, and the standard cross-entropy loss was compared with a weighted loss function designed to compensate for the imbalance between the narrow P- and S-phase label regions and the dominant background label. The waveform dataset, preprocessing procedure, and data augmentation followed those described in[7]. Therefore, only components directly relevant to the present comparison are described in detail below.

Target-label construction

Target-label construction was based on[7]. For each waveform, Gaussian-shaped labels were assigned to manually picked arrival times. Let μ denote the manually picked arrival times. The target label was defined using a Gaussian function normalized to a maximum value of 1:

$$G(t;\mu,\sigma) = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right), \tag{1}$$

where the maximum value is obtained at $t = \mu$, which corresponds to the picked arrival time. Using this function, the P- and S-phase labels were defined as:

$$P_{\text{label}}(t) = G(t;\mu_P,\sigma_P), \tag{2}$$

$$S_{\text{label}}(t) = G(t;\mu_S,\sigma_S), \tag{3}$$

where μ_P and μ_S are the manually picked arrival times of the P and S phases, respectively. In this study, the standard deviations were set to $\sigma_P = \sigma_S = 0.1$ s. This setting differs slightly from that of[7] in that $\sigma = 0.1$ s was also used for the P-phase label in the present study. The P- and S-phase labels were first defined as soft labels using Gaussian functions, and the background label was then defined so that the sum of the three class labels became one at each time sample. Specifically, the background label was defined as follows:

$$N_{\text{label}}(t) = 1 - \{P_{\text{label}}(t) + S_{\text{label}}(t)\}. \tag{4}$$

Thus, the labels satisfy $P_{\text{label}}(t) + S_{\text{label}}(t) + N_{\text{label}}(t) = 1$ at each time sample.

Loss function

In this study, phase picking was formulated as a semantic segmentation problem. Let $y_{b,c,t}$ and $\hat{y}_{b,c,t}$ denote the target label and model output probability, respectively, for sample $b \in \{1,\ldots,B\}$, class $c \in \{P,S,N\}$, and time index $t \in \{1,\ldots,T\}$. Here, P, S,

and N represent the P-phase, S-phase, and background classes, respectively, and $\hat{y}_{b,c,t}$ is the class probability obtained using the softmax function.

The standard soft cross-entropy loss is given by:

$$\mathscr{L}_{\mathrm{CE}} = -\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{c\in\{P,S,N\}} y_{b,c,t}\log\hat{y}_{b,c,t}. \tag{5}$$

This formulation assigns equal importance to all time samples. However, when Gaussian labels are used for phase picking, the temporal areas of the P- and S-phase labels are significantly smaller than those of the background label. Consequently, the loss is dominated by time samples distant from the phase arrivals, and the contribution of the phase-arrival regions becomes relatively small. This imbalance may reduce the sensitivity of the model to phase-arrival signals, particularly when multiple seismic phases are included within a single trace. Since most time samples are labeled as background, the standard loss effectively encourages the model to maintain high background probabilities over the majority of the waveform. This background-dominant training can make the model conservative in producing P- or S-phase probabilities outside the most prominent arrivals. As a result, the model may have difficulty forming multiple clear probability peaks within a single trace, especially for weak or overlapping arrivals.

To mitigate this problem, we introduced an area-weighted soft cross-entropy loss with a weighted mean over time. First, the temporal area of each class label was defined as:

$$A_c^{(b)} = \sum_{t=1}^{T} y_{b,c,t}, \qquad c \in \{P,S,N\}. \tag{6}$$

Using these areas, the weighting coefficients for the P- and S-phase labels were defined as:

$$\alpha_P^{(b)} = \frac{A_N^{(b)}}{A_P^{(b)}+\varepsilon}, \qquad \alpha_S^{(b)} = \frac{A_N^{(b)}}{A_S^{(b)}+\varepsilon}. \tag{7}$$

Here, ε denotes a small positive constant used for numerical stability and to prevent division by zero. These coefficients become larger when the temporal area of the phase label is smaller relative to that of the background label, thereby increasing the contribution of underrepresented phase-arrival regions.

Using these coefficients, the time-dependent weight was defined as follows:

$$w^{(b,t)} = 1 + \alpha_P^{(b)} y_{b,P,t} + \alpha_S^{(b)} y_{b,S,t}. \tag{8}$$

As the Gaussian label reaches its maximum at the picked arrival time and gradually decreases away from it, this weighting scheme places greater emphasis on the central part of the P- and S-phase labels (Fig. 1 (e)). In contrast, the time samples distant from the phase arrivals are assigned relatively small weights.

The soft cross-entropy at each time sample was defined as

$$\ell^{(b,t)} = -\sum_{c\in\{P,S,N\}} y_{b,c,t}\log\hat{y}_{b,c,t}. \tag{9}$$

The final weighted loss was then defined as

$$\mathscr{L}_{\mathrm{WCE}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{t=1}^{T} w^{(b,t)}\,\ell^{(b,t)}}{\sum_{t=1}^{T} w^{(b,t)}}. \tag{10}$$

Compared to the standard cross-entropy loss, the proposed loss increases the contribution of the time samples around the P- and S-phase arrivals according to their relative temporal areas and label amplitudes. In addition, because the weighted sum of the time-sample losses is divided by the sum of the weights for each waveform trace, the loss scale is normalized over the entire time window of each training example, which stabilizes the training process.

Model architecture and training settings

The models used in this study were PhaseNet and SegPhase. For PhaseNet, we used the model architecture implemented in SeisBench[17]. SegPhase followed the architecture proposed by[7], with several minor refinements. Specifically, learnable positional embeddings were replaced with one-dimensional sinusoidal positional encoding, reflection padding was incorporated into several convolutional operations, the normalization layer in the output module was revised from batch normalization to layer normalization, the activation function was updated from ReLU to GELU. These changes were implementation-level refinements and did not alter the overall hierarchical architecture of SegPhase. The same modified implementation of SegPhase was used for both loss functions.

Training was performed using the RAdam optimizer with a learning rate of 1×10^{-3} and a batch size of 32 for a maximum of 100 epochs. No weight decay or learning-rate scheduler was used. Early stopping was applied based on the validation loss, and training was terminated when the minimum validation loss was not updated within 10 consecutive epochs. The model with the lowest validation loss in each epoch was selected as the optimal model.

To evaluate the effect of the proposed loss function as directly as possible, the same training and validation split, random seed, and hyperparameter settings other than the loss function were used in both experiments. Therefore, the performance differences described below mainly reflected the effects of the loss function under the same basic training conditions.

Results and Discussion

This section evaluates the effect of the proposed area-weighted soft cross-entropy loss function from two complementary viewpoints: phase-picking performance on the test dataset and applicability to continuous seismic waveform analysis. First, we examined how the weighted loss function changed the residual distributions and threshold-dependent detection characteristics of SegPhase and PhaseNet. We then applied the trained models to continuous waveform data from the 2019 Ridgecrest earthquake sequence to investigate whether the increased sensitivity to phase arrivals led to improved event recovery under realistic monitoring conditions.

Evaluation of travel-time picking performance using the test dataset

The test dataset evaluation was conducted to clarify how the proposed weighted loss function changed the phase-picking behavior compared to that of the original loss. To ensure a fair comparison, the same model architecture, training data split, random seed, and hyperparameter settings were used; only the loss function was changed. Following Katoh et al.[7], a predicted P- or S-wave pick was counted as a positive detection when the predicted probability exceeded a given threshold. A detection was counted as a true positive (TP) when the predicted arrival time was within 0.1 s of the manually picked arrival time, whereas detections outside this tolerance were counted as false positives (FP). Missed manual picks were counted as false negatives (FN). Based on these counts, the precision, recall, and F1 scores were defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

and

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

These metrics were evaluated across probability thresholds ranging from zero to one in increments of 0.01.

Residual analysis showed that the weighted loss function reduced the scatter of arrival-time residuals for both SegPhase and PhaseNet (Fig. 2; Table 1). For both the P- and S-wave picks, the residual distributions of $\Delta T = T_{\text{True}} - T_{\text{Pred}}$ were centered near 0 s, indicating that the weighted loss did not introduce a systematic bias in the predicted arrival times. The mean residuals were 0.00 s across all cases. In contrast, the spread of the residuals was notably reduced when the weighted loss function was used. For SegPhase, the standard deviation decreased from 0.20 to 0.07 s for P-wave picks and from 0.27 to 0.05 s for S-wave picks. For PhaseNet, the standard deviation decreased from 0.38 to 0.11 s for P-wave picks and from 0.34 to 0.11 s for S-wave picks. These reductions were also consistent with the sharper peaks at approximately 0 s observed in the residual histograms (Fig. 2).

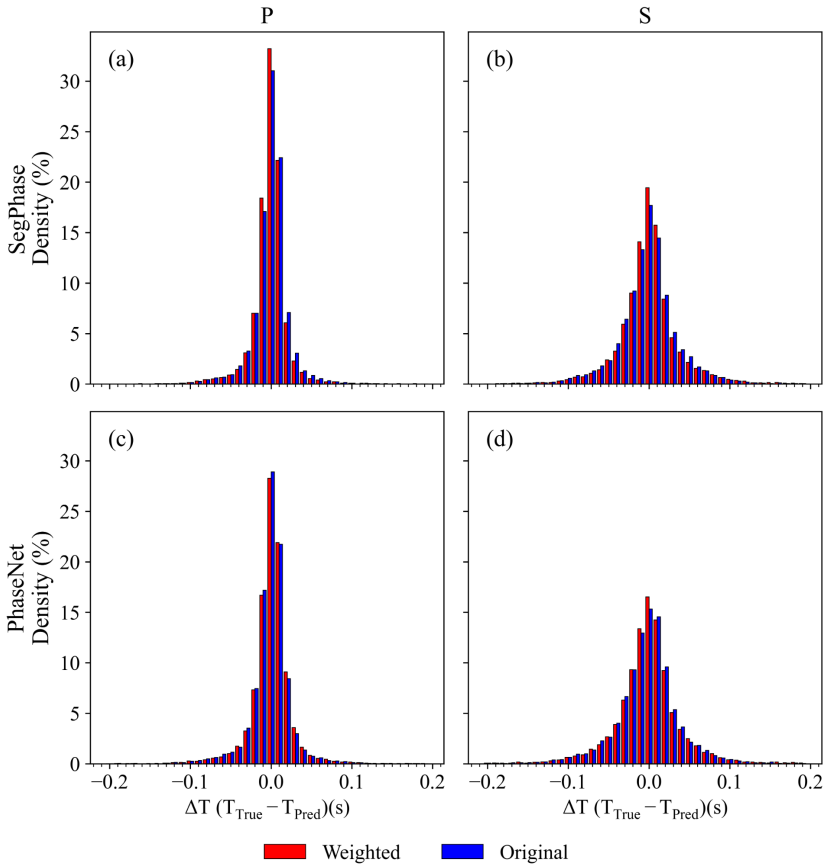


Figure 2. Distributions of arrival-time residuals for P- and S-wave picks. Distributions of arrival-time residuals for P- and S-wave picks obtained using the original- and weighted-loss models. The residual is defined as $\Delta T = T_{\text{True}} - T_{\text{Pred}}$, where T_{True} is the manually picked arrival time and T_{Pred} is the predicted arrival time. Panels (a) and (b) show the residual distributions for SegPhase P- and S-wave picks, respectively. Panels (c) and (d) show the corresponding results for PhaseNet. Red and blue histograms indicate the weighted- and original-loss models, respectively.

Table 1. Mean and standard deviation of arrival-time residuals for P- and S-wave picks.

Model	Loss	P-wave		S-wave	
		Mean (s)	Std. (s)	Mean (s)	Std. (s)
SegPhase	Original	0.00	0.20	0.00	0.27
SegPhase	Weighted	0.00	0.07	0.00	0.05
PhaseNet	Original	0.00	0.38	0.00	0.34
PhaseNet	Weighted	0.00	0.11	0.00	0.11

This reduction in residual scatter can be attributed to the manner in which the proposed loss function redistributes the training objective over time. Under the standard soft cross-entropy loss, all time samples are weighted equally, and the optimization is dominated by the background class, as most samples are located far from the P- and S-phase arrivals. Therefore, errors near the true onset times contribute to only a small fraction of the total loss. In contrast, the proposed weighted loss function reallocates a larger fraction of the optimization pressure to the narrow P- and S-phase label regions. Because the Gaussian labels assume their largest values at manually picked arrival times, the central parts of the arrival regions are emphasized most strongly during training. This redistribution of the training objective encourages the model to more precisely capture the temporal structure of the phase onset, yielding sharper probability peaks and narrower residual distributions.

The threshold-dependent metric curves further show that the weighted loss function changes the balance between precision and recall rather than uniformly improving all aspects of phase-picking performance (Fig. 3). For both the P- and S-wave picks, the weighted-loss models maintained a higher recall over a broad range of probability thresholds, whereas the original-loss models generally exhibited higher precision. Consequently, the probability threshold that maximized the F1 score shifted to a higher value for the weighted-loss models than for the original-loss models. For both SegPhase and PhaseNet, the original-loss models achieved higher or comparable peak F1 scores, whereas the weighted-loss models occupied a higher recall and lower precision region in the precision–recall plane. This tendency was particularly evident for the S-wave picks.

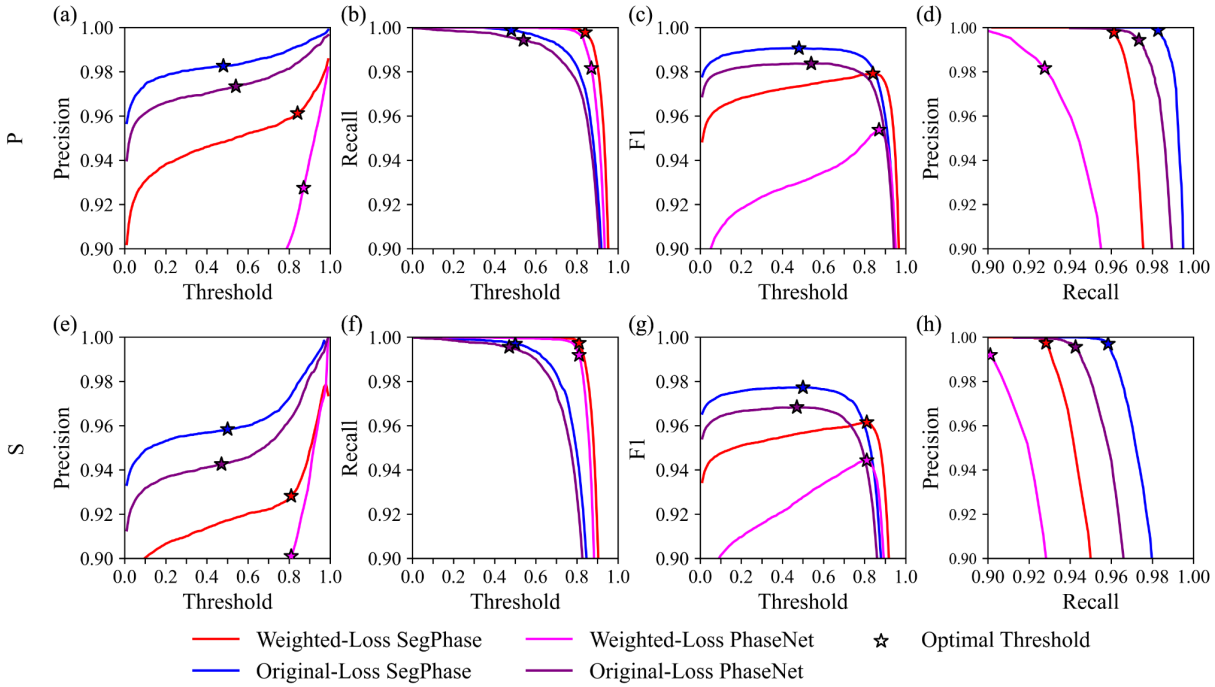


Figure 3. Threshold-dependent phase-picking performance. Threshold-dependent precision, recall, F1 score, and precision–recall relationships for P- and S-wave picks. Panels (a)–(d) show the results for P-wave picks, and panels (e)–(h) show the results for S-wave picks. Red, blue, magenta, and purple curves indicate weighted-loss SegPhase, original-loss SegPhase, weighted-loss PhaseNet, and original-loss PhaseNet, respectively. Star symbols denote the probability thresholds that maximize the F1 score.

The TP, FP, and FN curves provide a direct explanation for this behavior (Supplementary Fig. 1). For both phases, the weighted-loss models suppressed the decrease in TP counts as the threshold increased, indicating that detections corresponding to manually labeled arrivals tended to retain higher predicted probabilities. At the same time, the FP counts were larger for the weighted-loss models across a substantial portion of the threshold range. Because precision is defined by Eq. (11), this increase in FP directly reduces precision. The FN counts were also generally smaller for the weighted-loss models over a broad range of thresholds, which explains their higher recall. These results indicate that the weighted loss function shifts the models toward a more recall-oriented behavior by increasing the number of candidate detections, including both true-positive and false-positive detections.

This precision–recall trade-off is consistent with the mechanism of the proposed weighting. By assigning larger penalties to errors around the P- and S-phase label regions, the weighted loss changes the effective decision behavior of the model. In other words, the model becomes less conservative in responding to waveform features that resemble true phase arrivals. In this sense, the effective decision boundary is shifted toward a more recall-oriented regime, in which weak or partially overlapping arrivals are more likely to generate detectable probability peaks. This behavior is advantageous for suppressing missed picks, because even relatively weak arrivals can still be detected. At the same time, however, this increased sensitivity is not limited only to correctly labeled arrivals. Waveform features such as coda waves, noise fluctuations, or other non-target signals can partially share characteristics with true phase onsets, including short-duration amplitude changes or changes in waveform shape. Therefore, once the model is encouraged to respond more strongly to phase-like features, these ambiguous signals may also more easily produce elevated probabilities, which leads to an increase in false positives and a reduction in precision.

Overall, the test dataset results demonstrated that the proposed weighted loss function systematically changed the behavior of both SegPhase and PhaseNet. Its main effect was not a uniform improvement across all evaluation metrics, but rather a shift in detection behavior caused by redistributing the training objective away from the background-dominated part of the waveform toward the narrow phase-arrival regions. Consequently, the weighted-loss models produced more temporally concentrated picks after detection, fewer missed arrivals, and more false detections. Thus, the proposed loss makes the models more sensitive and recall-oriented. This change in behavior is important for the next subsection, where we examine whether such increased sensitivity is beneficial in continuous waveform analysis, in which phase picks are used as intermediate inputs for subsequent association and hypocenter determination.

Application to continuous seismic waveforms for the 2019 Ridgecrest earthquake

The practical usefulness of the proposed approach was further examined using continuous waveform data from the 2019 Ridgecrest earthquake sequence. In this experiment, phase picking was applied to continuous waveforms from 36 seismic stations from 15:00 on July 4, 2019, to 23:59 on July 9, 2019. For each model, the probability threshold was set to a value that maximized the F1 score, as shown in Fig. 3. Phase associations were performed using REAL [18]. Candidate events were required to satisfy more than four P-wave arrivals, four S-wave arrivals, eight total arrivals, and four stations with both P- and S-wave picks. Hypocenter determination was subsequently performed using Hypomh PS [19].

The detected epicenters showed that all models captured the main spatial distribution of seismicity associated with the Ridgecrest earthquake sequence (Fig. 4). Although the models were trained using Japanese seismic waveform data, they reproduced the broad aftershock zone from continuous waveform records in California. The obtained epicenters delineated the northwest–southeast to northeast–southwest trending aftershock zones, which is broadly consistent with the seismicity patterns reported in previous studies [20,21].

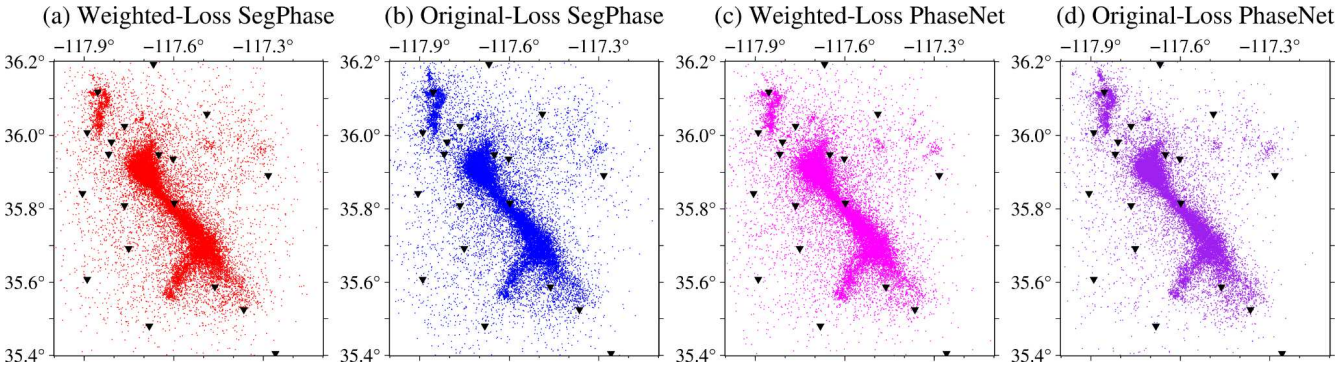


Figure 4. Epicentral distributions of detected earthquakes in the Ridgecrest application. Epicentral distributions of earthquakes detected from continuous waveform data during the 2019 Ridgecrest earthquake sequence. Panels (a) and (b) show the results for weighted- and original-loss SegPhase, respectively. Panels (c) and (d) show the corresponding results for weighted- and original-loss PhaseNet, respectively. Colored dots indicate detected earthquakes, and black inverted triangles indicate the seismic stations used in the analysis.

The temporal variation in the number of detected events further demonstrates the effect of the weighted loss function (Fig. 5). For SegPhase, the weighted-loss model detected 34,008 earthquakes during the analysis period, whereas the original-loss model detected 26,914. For PhaseNet, the weighted-loss model detected 32,997 earthquakes, whereas the original-loss model detected 23,622. For comparison, the catalog of [20] contains 29,385 events during the same period, and the catalog of [21] contains 16,778. Thus, for both model architectures, the weighted loss function increased the number of detected earthquakes, and the resulting event counts exceeded those of the two reference catalogs developed using the template matching method considered here. In addition, the two-hour event-count time series showed that the weighted-loss models detected more earthquakes than the original-loss models during most of the analysis period, while reproducing the overall temporal evolution of the Ridgecrest sequence (Fig. 5).

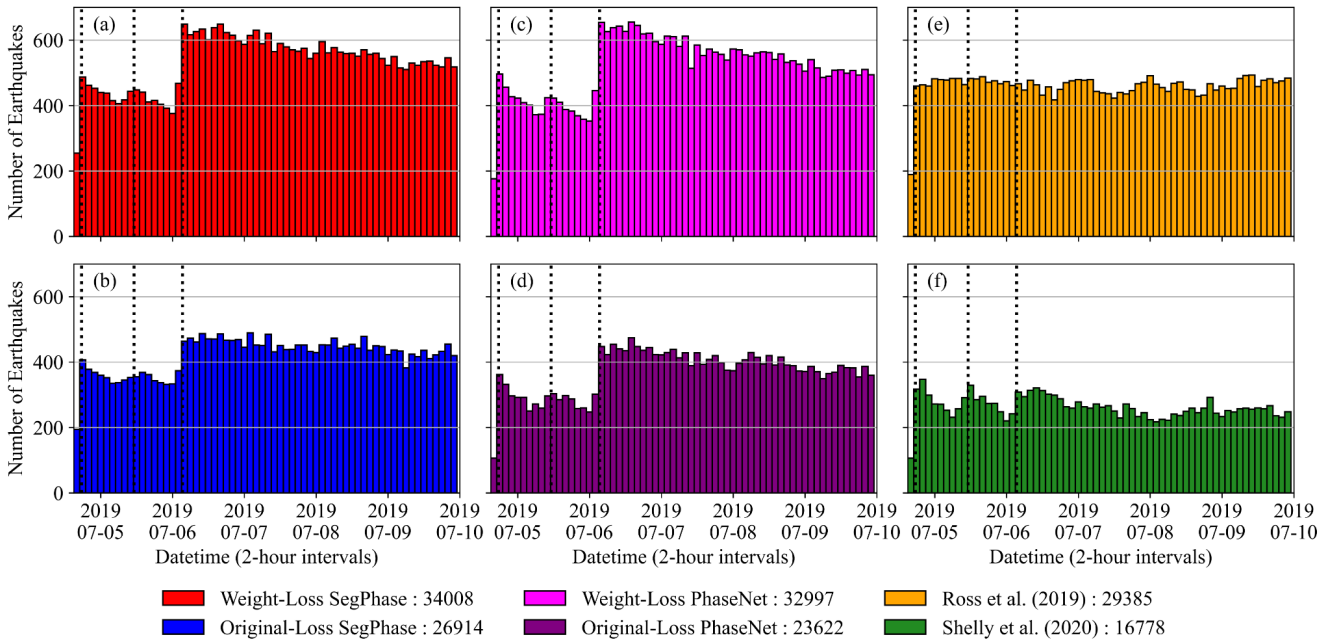


Figure 5. Temporal variation in the number of detected earthquakes during the 2019 Ridgecrest earthquake sequence. Two-hour event-count time series for earthquakes detected during the 2019 Ridgecrest earthquake sequence. Panels (a) and (b) show the results for weighted- and original-loss SegPhase, respectively. Panels (c) and (d) show the corresponding results for weighted-loss and original-loss PhaseNet, respectively. Panels (e) and (f) show the event counts from the reference catalogs[20,21], respectively. Vertical dotted lines indicate the occurrence times of the major Ridgecrest earthquakes: from left to right, the Mw 6.4 event of July 4, 2019 (17:33:48), the Mw 5.4 event of July 5, 2019 (11:07:53), and the Mw 7.1 event of July 6, 2019 (03:19:52).

This improvement in event recovery can be understood in terms of the structure of the downstream detection workflow. In the test-dataset evaluation, phase picking is assessed at the level of individual arrivals, and false positives directly reduce precision. In contrast, in the continuous waveform analysis, individual phase picks are not the final output. These intermediate candidates are later filtered by phase association and hypocenter determination. Event detection by REAL requires a sufficient number of mutually consistent P- and S-wave picks across multiple stations within physically plausible travel-time relationships[18]. Therefore, this workflow is often more tolerant to moderate increases in pick-level sensitivity than missed arrivals. For weak earthquakes near the detection limit, a small increase in the number of usable phase picks can significantly affect event recovery, as such events may fail to be associated when only one or two critical arrivals are missing. From this perspective, the weighted loss function improves event recovery not simply because it produces more picks, but because it increases the likelihood that sufficient compatible picks survive across the network to satisfy the association criteria. This explains why the recall-oriented behavior of the weighted-loss models is particularly useful in continuous seismic monitoring.

Figure 6 provides an example of the phase-picking results for continuous Ridgecrest waveforms and illustrates how the weighted loss changes the model response to real seismic signals. In the original-loss model, strong probability responses were mainly limited to relatively clear arrivals. In contrast, the weighted-loss model produced stronger probability responses for a larger number of time samples and stations, resulting in more candidate phase picks (Fig. 6). This behavior is consistent with the test-dataset results, where the weighted–loss models showed higher recall and fewer false negatives. In continuous waveform analysis, such a reduction in missed picks is particularly important because event detection by phase association requires a sufficient number of arrivals across multiple stations. Even a moderate increase in reliable candidate picks can therefore increase the number of event hypotheses that satisfy the association conditions.

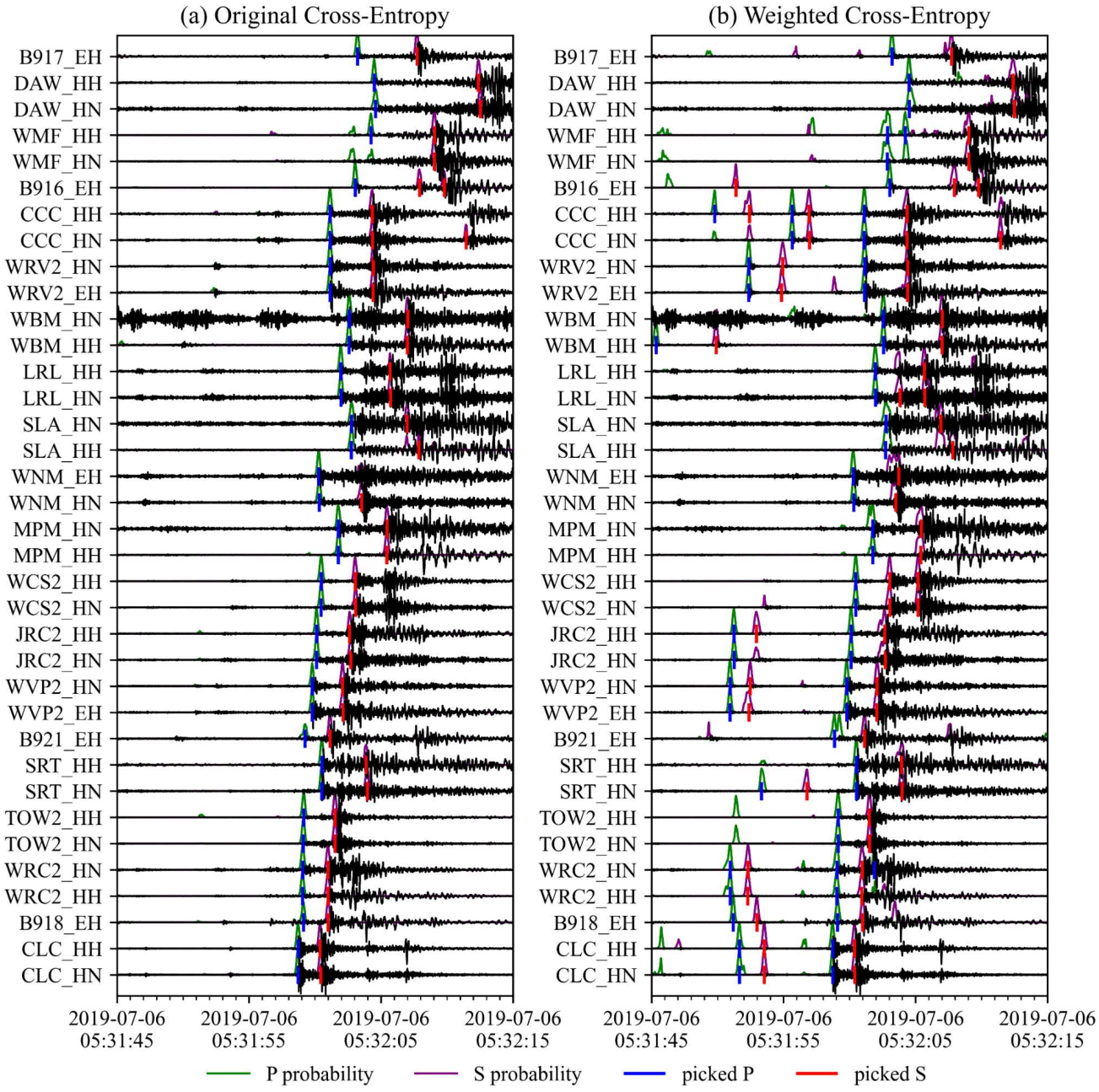


Figure 6. Example of phase picking applied to continuous waveforms from the 2019 Ridgecrest earthquake sequence. Example of phase-picking results for continuous Ridgecrest waveform data. Panel (a) shows the results obtained using the original cross-entropy loss model, and panel (b) shows the results obtained using the weighted cross-entropy loss model. Black traces indicate waveforms obtained by stacking the three-component records at each station. Green and purple curves indicate the predicted P- and S-phase probabilities, respectively. Blue and red vertical lines indicate the picked P- and S-wave arrival times, respectively.

Simultaneously, the weighted-loss model produced additional probability peaks that were not present in the original-loss result (Fig. 6). This behavior was consistent with the increase in false positives observed in the threshold-dependent test-dataset evaluation. In the Ridgecrest application, this shift likely contributed to an increase in the number of associated events because more phase candidates were available for REAL to construct event hypotheses. However, this also indicates that appropriate probability thresholds and association constraints remain necessary to suppress spurious picks. In this respect, the weighted

loss should be interpreted as a method that improves sensitivity at the pick-generation stage, whereas the later association step remains essential for controlling final catalog quality.

Overall, the Ridgecrest application demonstrates that the proposed weighting strategy is effective not only in the test dataset, but also in a realistic continuous-data setting. For both SegPhase and PhaseNet, the weighted-loss models produced catalogs with spatial distributions consistent with the known aftershock zones and detected more earthquakes than the corresponding original-loss models. These results indicate that improving phase-arrival sensitivity during training can directly increase event recovery in continuous waveform analysis, even when the target sequence is outside the region used for model training.

Remaining limitations in the coda of large earthquakes

Figure 7 shows that the increase in events detected by the weighted-loss model was not uniform over all magnitudes and times. The additional detections were mainly concentrated in the low-magnitude range, particularly around $M \sim 1$–2, after the major Ridgecrest earthquakes. This result suggests that the weighted loss enhances the sensitivity of the models to weak phase arrivals associated with small earthquakes. However, immediately after the largest event, the number of detected small-magnitude earthquakes remained limited even for the weighted-loss model (Fig. 7).

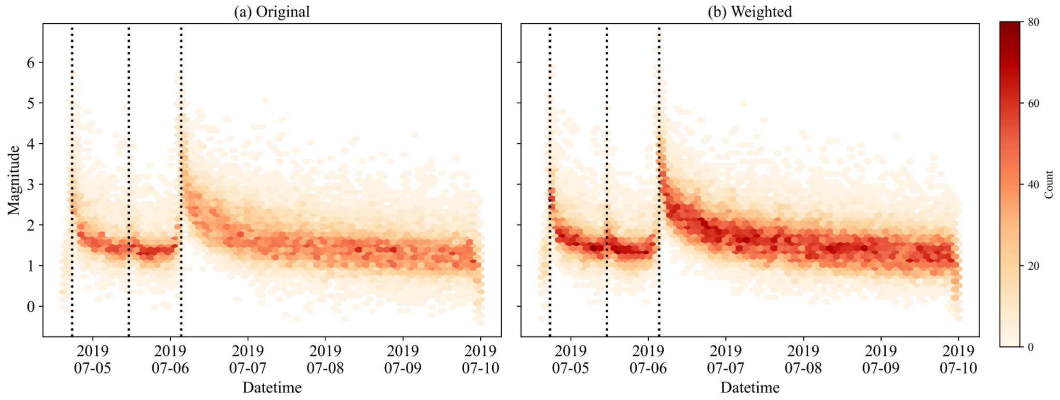


Figure 7. Magnitude–time distributions of detected Ridgecrest earthquakes. Magnitude–time density plots of earthquakes detected during the 2019 Ridgecrest earthquake sequence. Panels (a) and (b) show the results obtained using the original- and weighted-loss models, respectively. The horizontal axis represents event origin time, and the vertical axis shows magnitude. Color indicates the number of detected earthquakes in each magnitude–time bin. Vertical dotted lines indicate the occurrence times of the major Ridgecrest earthquakes: from left to right, the Mw 6.4 event of July 4, 2019 (17:33:48), the Mw 5.4 event of July 5, 2019 (11:07:53), and the Mw 7.1 event of July 6, 2019 (03:19:52).

This behavior indicates a limitation that cannot be resolved using loss reweighting alone. Immediately after a large earthquake, continuous waveforms are dominated by high-amplitude and long-duration coda waves that persist over a wide time range and are coherently recorded across many stations. Under such conditions, the phase onsets of subsequent small earthquakes may be partially buried in the complex and nonstationary signal energy generated by the preceding large event. Even if the proposed loss increases the sensitivity of the model to phase-arrival regions, the model can respond only to information that remains distinguishable in the waveform. When the onset of a small event is strongly masked by the coda of a large earthquake, the difficulty is not only insufficient emphasis during training, but also limited separability of the signal itself. This difficulty is consistent with previous studies showing that overlapping seismic signals remain a challenge in automatic phase picking and event recovery[22].

In addition, coda waves often contain sharp fluctuations and waveform complexities that may resemble parts of the true phase arrivals. Therefore, increasing the sensitivity can also increase the ambiguity between small-event onsets and coda-related features. This interpretation is consistent with the present results; the weighted-loss model improved the recovery of many successive small earthquakes, but it did not fully recover small events immediately after the largest Ridgecrest earthquakes (Fig. 7). These observations suggest that the remaining difficulty is controlled not only by label imbalance, but also by waveform overlap, dynamic range, and the intrinsic separability of small-event signals in strong coda environments.

Therefore, the proposed method should be regarded as a partial but meaningful improvement rather than a complete solution to the problem of overlapping seismic signals. The proposed weighted loss is effective for enhancing sensitivity to weaker and successive arrivals; however, additional developments are needed to detect small earthquakes embedded in the coda of large events. Possible directions include training with more examples containing strong coda waves, introducing coda-aware data augmentation, using a longer temporal context, or combining the phase picker with postprocessing methods that explicitly consider overlapping events.

Conclusions

In this study, we investigated whether the performance of deep learning-based seismic phase picking can be improved by addressing the label imbalance inherent in semantic-segmentation-based arrival-time estimation. Because the P- and S-phase labels occupy only narrow temporal regions compared to those of the dominant background label, the standard soft cross-entropy loss may place insufficient emphasis on phase-arrival samples during training. To reduce this imbalance, we introduced an area-weighted soft cross-entropy loss that increases the contribution of the P- and S-phase label regions while preserving the model architecture.

The test-dataset results show that the proposed weighted loss improves the temporal concentration of predicted arrival times after detection. For both SegPhase and PhaseNet, the weighted-loss models produced residual distributions that were more sharply concentrated around the manually picked arrival times than those of the original-loss models, while maintaining mean residuals of 0.00 s across all cases (Fig. 2; Table 1). This result suggests that increasing the optimization pressure within the narrow arrival regions enables the model to fit the temporal structure of the onset more precisely.

The threshold-dependent evaluation further showed that the weighted loss systematically changed the detection characteristics of the models (Figs. 3 and 4). The weighted-loss models maintained a higher recall over a broad threshold range and suppressed the decrease in true-positive counts, indicating improved sensitivity to phase arrivals. Simultaneously, this increased sensitivity was accompanied by an increase in false positives, which reduced precision. Therefore, the proposed loss function does not uniformly improve all performance measures. Instead, it shifts the models toward a more recall-oriented detection behavior by redistributing the training objective away from the background-dominated portions of the trace toward the narrow P- and S-phase label regions. The higher F1-optimal thresholds required by the weighted-loss models are a direct consequence of this shift in sensitivity.

Application to continuous waveform data from the 2019 Ridgecrest earthquake sequence demonstrated the practical value of this change in model behavior (Figs. 5, 6 and 7). Although the models were trained using Japanese seismic waveform data, they reproduced the main spatial distribution and temporal evolution of Ridgecrest seismicity. For both SegPhase and PhaseNet, the weighted-loss models detected more earthquakes than the corresponding original-loss models did. This improvement can be understood from the structure of continuous-data processing: event detection depends on whether sufficient mutually consistent picks are available across the network; therefore, a reduction in missed picks can directly increase the number of events that satisfy the association criteria[18]. Recall-oriented picking is particularly advantageous in this context.

Simultaneously, the magnitude–time analysis revealed an important limitation (Fig. 7). Although the weighted loss improved the recovery of many small and successive earthquakes, it did not fully solve the difficulty of detecting small events immediately after large earthquakes. In such periods, phase arrivals may be strongly masked by high-amplitude and long-duration coda waves, and the remaining difficulty is controlled not only by label imbalance, but also by waveform overlap and limited signal separability. This result indicates that improving the loss function alone is insufficient for the complete recovery of small events in the most complex parts of dense aftershock sequences.

Overall, this study demonstrates that loss-function design is an effective means of improving recall-oriented phase picking in dense seismic sequences. The proposed weighted loss increases sensitivity to P- and S-phase arrivals and improves the temporal consistency of predicted picks, which, in turn, increases event recovery in continuous waveform analysis. However, the proposed method does not fully resolve the difficulty of detecting small events whose phase arrivals are strongly masked by high-amplitude coda waves after large earthquakes, and the increased sensitivity can also lead to more false-positive picks. Therefore, appropriate probability thresholding, association constraints, and further developments for strong-coda environments are important for constructing more reliable automatic earthquake catalogs.

List of abbreviations

CE Cross-entropy

FN False negative

FP False positive

JMA Japan Meteorological Agency

N Background or noise class

NIED National Research Institute for Earth Science and Disaster Resilience

REAL Rapid Earthquake Association and Location

SCEDC Southern California Earthquake Data Center

TP True positive

WCE Weighted cross-entropy

Declarations

Availability of data and materials

Data from the stationary observation network are available at the Data Management Centre of the National Research Institute for Earth Science and Disaster Resilience (NIED) https://hinetwww11.bosai.go.jp/auth/?LANG=en, the Japan Meteorological Agency (JMA) http://www.data.jma.go.jp/svd/eqev/data/bulletin/index_e.html and http://www.data.jma.go.jp/svd/eqev/data/daily_map/index.html, and the Earthquake Research Institute, the University of Tokyo http://tkypub.eri.u-tokyo.ac.jp/harvest. Please contact the author for the use of other data. The continuous waveforms of the 2019 Ridgecrest earthquake can be accessed from the Southern California Earthquake Data Center[23]. The deep learning model and Loss function are available at the author's GitHub repository https://github.com/ShinyaKatoh/.

Competing interests

The authors declare no competing interests.

Funding

This study was supported by the MEXT Project for Seismology Toward Research Innovation with Data of Earthquake (STAR-E Project) JPJ010217; the MEXT Project for Seismology Toward Research Innovation for Earthquake Evaluation with Exact Data (STAR-E NEXT Project) JPJ013735; the MEXT The Third Earthquake and Volcano Hazards Observation and Research Program (Earthquake and Volcano Hazard Reduction Research); the JST NEXUS, Japan No. JPMJNX25C5; JSPS KAKENHI Grant-in-Aid for Research Activity Start-up No. 23K19061; Grant-in-Aid for Scientific Research (A) No. 23H00466; Grant-in-Aid for Fund for the Promotion of Joint International Research (International Collaborative Research) No. 23KK0181; Grant-in-Aid for Challenging Research (Pioneering) No. 25K21806; Grant-in-Aid for Early-Carrer Scientists No. 25K17482; and Earthquake Research Institute, University of Tokyo Joint Research ERI JURP 2026-A-03, 2024-B-01, and 2025-B-01.

Authors' contributions

S.K. performed the analyses and drafted the manuscript. M.I., H.N., and Y.I. contributed to manuscript preparation. M.I. proposed the concept of the proposed method. Y.I. maintained the seismic stations and supervised data collection. H.N. contributed to funding acquisition, resources, and project administration. All authors have read and approved the final version of the manuscript.

Authors' information

S.K. is affiliated with the Earthquake Research Institute, The University of Tokyo, Japan. M.I. is affiliated with the Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, the RIKEN Center for Advanced Intelligence Project, and the Division of Physics and Astronomy, Graduate School of Science, Kyoto University, Japan. Y.I. is affiliated with the Disaster Prevention Research Institute, Kyoto University, Japan. H.N. is affiliated with the Earthquake Research Institute, The University of Tokyo, the RIKEN Center for Advanced Intelligence Project, and the Graduate School of Information Science and Technology, The University of Tokyo, Japan.

Acknowledgements

We appreciate the support provided by the "Manten" project team, Mie Tanaka, and Rie Mizushima in seismic observations and data processing. We also thank the staff of the Disaster Prevention Research Institute for their assistance with the seismic observations. We gratefully acknowledge the seismic data provided by the National Research Institute for Earth Science and Disaster Prevention (NIED), National Institute of Advanced Industrial Science and Technology (AIST), Japan Meteorological Agency (JMA), Kyoto University, Kyushu University, and the Earthquake Research Institute of the University of Tokyo. Generic Mapping Tools[24] and Matplotlib[25] were used to create the figures. Obspy[26] was used for waveform processing. PyTorch[27] was used to create the deep learning models.

References

1. Ross, Z. E., Meier, M. A., Hauksson, E., and Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901. doi:10.1785/0120180080

2. Zhu, W. and Beroza, G. C. (2018). PhaseNet: a deep-neural-network-based seismic arrival time picking method. *Geophysical Journal International*. doi:10.1093/gji/ggy423

3. Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. (2020). Earthquake Transformer: an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1), 3952. doi:10.1038/s41467-020-17591-w

4. Tokuda, T. and Nagao, H. (2023). Seismic-phase detection using multiple deep learning models for global and local representations of waveforms. *Geophysical Journal International*, *235*(2), 1163–1182. doi:10.1093/gji/ggad270

5. Sun, H., Ross, Z. E., Zhu, W. and Azizzadenesheli, K. (2023). Phase neural operator for multi-station picking of seismic arrivals. *Geophysical Research Letters*, **50**, e2023GL106434. doi:10.1029/2023GL106434

6. Naoi, M., Tamaribuchi, K., Shimojo, K., Katoh, S., and Ohyanagi, S. (2024). Neural phase picker trained on the Japan Meteorological Agency unified earthquake catalog. *Earth, Planets and Space*, *76*(1). doi:10.1186/s40623-024-02091-8

7. Katoh, S., Iio, Y., Nagao, H., Katao, H., Sawada, M., and Tomisaka, K. (2025). SegPhase: development of arrival time picking models for Japan's seismic network using the hierarchical vision transformer. *Earth, Planets and Space*, *77*(1). doi:10.1186/s40623-025-02249-y

8. Katoh, S., Nagao, H., Iio, Y. (2025). PoViT-UQ: P-wave polarity and arrival time determination using vision transformer with uncertainty quantification. *Geophysical Journal International*, *243*(1). doi:10.1093/gji/ggaf324

9. Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., Chiaraluce, L., Beroza, G. C., and Segou, M. (2021). Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 Central Italy sequence. *The Seismic Record*, *1*(1), 11–19. doi:10.1785/0320210001

10. Kim, A., Nakamura, Y., Yukutake, Y., Uematsu, H., and Abe, Y. (2023). Development of a high-performance seismic phase picker using deep learning in the Hakone volcanic area. *Earth, Planets and Space*, *75*(1). doi:10.1186/s40623-023-01840-5

11. Wilding, J. D., Zhu, W., Ross, Z. E., and Jackson, J. M. (2023). The magmatic web beneath Hawai'i. *Science*, *379*(6631), 462–468. doi:10.1126/science.ade5755

12. Tamaribuchi, K., Kudo, S., Shimojo, K., and Hirose, F. (2023). Detection of hidden earthquakes after the 2011 Tohoku earthquake by automatic hypocenter determination combined with machine learning. *Earth, Planets and Space*, *75*(1). doi:10.1186/s40623-023-01915-3

13. Kato, A. (2024). Implications of fault-valve behavior from immediate aftershocks following the 2023 Mj 6.5 earthquake beneath the Noto Peninsula, Central Japan. *Geophysical Research Letters*, *51*(1). doi:10.1029/2023GL106444

14. Suzuki, R., Uchida, N., Zhu, W., Beroza, G. C., Nakayama, T., Yoshida, K., Toyokuni, G., Takagi, R., Azuma, R., and Hasegawa, A. (2025). The forearc seismic belt: a fluid pathway constraining down-dip megathrust earthquake rupture. *Science*, *389*(6756). doi:10.1126/science.adt6389

15. Park, Y., Beroza, G. C., and Ellsworth, W. L. (2023). A mitigation strategy for the prediction inconsistency of neural phase pickers. *Seismological Research Letters*. doi:10.1785/0220230003

16. Saini, M. and Susan, S. (2023). Tackling class imbalance in computer vision: a contemporary review. *Artificial Intelligence Review*, *56*(S1), 1279–1335. doi:10.1007/s10462-023-10557-6

17. Woollam, J., Rietbrock, A., Bueno, A., and De Angelis, S. (2019). Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network. *Seismological Research Letters*, *90*(2A). doi:10.1785/0220180312

18. Zhang, M., Ellsworth, W. L., and Beroza, G. C. (2019). Rapid earthquake association and location. *Seismological Research Letters*, *90*(6), 2276–2284. doi:10.1785/0220190052

19. Kawanishi, R., Iio, Y., Yukutake, Y., Shibutani, T., and Katao, H. (2009). Local stress concentration in the seismic belt along the Japan Sea coast inferred from precise focal mechanisms: Implications for the stress accumulation process on intraplate earthquake faults. *Journal of Geophysical Research: Solid Earth*, *114*(B1). doi:10.1029/2008JB005765

20. Ross, Z. E., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M., Wang, X., Zhan, Z., Simons, M., Fielding, E. J., Yun, S.-H., Hauksson, E., Moore, A. W., Liu, Z., and Jung, J. (2019). Hierarchical interlocked orthogonal faulting in the 2019 Ridgecrest earthquake sequence. *Science*, *366*(6463). doi:10.1126/science.aaz0109

21. Shelly, D. R. (2020). A high-resolution seismic catalog for the initial 2019 Ridgecrest earthquake sequence: foreshocks, aftershocks, and faulting complexity. *Seismological Research Letters*, *91*(4). doi:10.1785/0220190309

22. Enescu, B., Mori, J. and Miyazawa, M. (2007). Quantifying early aftershock activity of the 2004 mid-Niigata Prefecture earthquake (Mw6.6). *Journal of Geophysical Research: Solid Earth*, **112**(B4). doi:10.1029/2006JB004629

23. SCEDC (2013). Southern California Earthquake Data Center. Caltech [Dataset]. doi:10.7909/C3WD3xH1

24. Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., and Tian, D. (2019). The Generic Mapping Tools version 6. *Geochemistry, Geophysics, Geosystems*, *20*(11), 5556–5564. doi:10.1029/2019GC008515

25. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi:10.1109/MCSE.2007.55

26. Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010). ObsPy: A Python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530–533. doi:10.1785/gssrl.81.3.530

27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv*. doi:10.48550/arXiv.1912.01703

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WeightSciRepsupplementary.pdf