Authors' response to Referees' comments

Dear Editor,

We would like to thank the referees for their comments, observations, and suggestions regarding our manuscript. We found their feedback both informative and constructive. Below, you will find our detailed responses to all the comments.
Referees' comments are reported in black, and our replies are provided in blue.

In response to the reviewers' comments, we have revised the manuscript to clarify the scope and novelty of our approach. The introduction has been reframed to better contextualize our contribution to ground motion prediction and to provide clearer explanations of GCNs and the masking mechanism. Quantitative performance improvements were added to the abstract, and limitations of using ShakeMap-derived data as targets were addressed. We tested a pre-trained PhaseNet encoder, which improved both model performance and computational efficiency. While experiments with per-event dynamic graphs were conducted, we retained the fixed graph topology for its greater training stability. The Results section was reorganized for better readability, with more informative paragraph titles. Additional details were added on data thresholds, GCN architecture, and training hyperparameters. Minor edits were also made to improve clarity, address figure suggestions, and respond to technical remarks.

Reviewer A

This manuscript presents an enhanced version of the masked graph neural network model for ground motion intensities prediction and applied to the INSTANCE benchmark dataset. The authors develop a Masked GCN architecture that integrates convolutional and graph convolutional layers and includes a masking mechanism to overcome the obstacles of rapid response built upon prior work. The model is clearly described and comprehensively benchmarked, it is appreciated that the code is made available online. Overall, the manuscript is well-written, methodologically sound. However, some aspects—particularly related to the framing of the application and interpretation of the model's capabilities—would benefit from clarification. While the authors suggest that the model has potential applications in Earthquake Early Warning (EEW), its current design and evaluation are more aligned with ground motion intensity prediction rather than EEW system. I would therefore encourage the authors to frame this work as a flexible and fast IM prediction model, potentially useful for EEW systems but not yet validated as such (more test results like time window less than 5 seconds, restricted number of stations). I recommend minor revisions.
Abstract and Introduction:

1. Line 24 – 27: The authors state the performance is improved by masking stations and integrating additional information, but quantitative metrics of improvements would be more suitable in here.

   We agree that including quantitative metrics strengthens the clarity of the abstract. As suggested, we have revised it to include specific improvements in model performance, both compared to the Bindi et al. (2011) GMM and to the original TISER-GCN (see lines 23-26).

2. Line 31 – 44: The discussion of the "blind zone" and proximity-based EEW challenges is not directly relevant to the capabilities of the Masked GCN, which predicts shaking at stations further away from the source. Consider revising or refocusing this paragraph to better align with the model's capability.

   In the revised manuscript, we have replaced the discussion of the "blind zone" and proximity-based EEW challenges by listing other approaches for the rapid prediction of ground shaking, aiming to better present the landscape of machine learning methods in the literature and provide a clearer overview of the current state of the art.

3. Line 44 – 57: The prior work review is helpful, but the authors should also briefly highlight limitations in existing models to better contextualize the novelty of this work.

   Thank you for the suggestion. In the revised introduction, we have included a brief presentation of GCNs, highlighting their key strengths, such as the ability to model spatial relationships between stations and robustness to missing data, which directly address several limitations of existing methods. Additionally, we emphasize that the masking method adopted in our work helps overcome the typical constraint of fixed station geometry, further enhancing the flexibility and applicability of the model. These additions aim to better contextualize the novelty and advantages of our approach (see lines 63-84).

4. Line 58 – 59: The paragraph briefly explains CNNs. However, since the core innovation is the use of Masked GNN, a clearer introduction to GNN and masker mechanisms and their relevance in seismology would be more appropriate. Overall, the Introduction part should be reframed.

   Thank you for your comment. As noted in our response to the previous comment, we have added a clearer introduction to Graph Convolutional Networks (GCNs) in the revised manuscript. In addition, a new part has been included to explain the masking mechanism and its specific relevance to our application. (See lines 63-84).

Data:

1. Line 106-109: the authors should discuss why choose a conservative way and how conservative this is?

   The choice to apply a conservative selection threshold—specifically, limiting velocity values to within ±7,000,000 counts—was driven by the need to avoid saturation or clipping effects in high-amplitude recordings, which are known to distort ground motion features. Although 24-bit data loggers typically support integer ranges up to ±8,388,608 (2^{23}), a safety margin of approximately 17% below full scale was deliberately introduced. This accounts for potential issues such as signal amplification, preprocessing inconsistencies, gain misalignment, DC offset, or transient spikes.

   In addition, an upper limit of 100 cm/s² was imposed on the derived acceleration values as a further safeguard. This constraint aims to exclude physically implausible or mechanically distorted records, thereby ensuring the dataset emphasizes high-quality input.

   Taken together, these conservative thresholds prioritize data reliability—potentially at the cost of discarding a small fraction of high-amplitude signals—and contribute to improved model stability by minimizing the risk of overfitting to noisy or clipped inputs.
   This rationale is explicitly clarified in lines 118-122.

Method:

1. Line 153-158: I understand the structure of graph neural network might be similar with (Bloemheuvel et al. 2023), but some basic details of the structure of GNN should be discussed here, like: how did you determine the edges in GNN? Are all stations connected? Does the graph change dynamically given different events?

   We appreciate the reviewer's observation. While our GCN architecture shares some conceptual similarities with Bloemheuvel et al. (2023), we now provide additional clarification regarding the graph construction in the revised version of the manuscript.

   Specifically, basic details about the GCN structure—including how nodes and edges are defined—have been added to the updated introduction (lines 63–70) and further elaborated in Section 3.1 (lines 181-186), where we describe how the adjacency matrix is constructed. In our approach, each node represents a seismic station, and edges are constructed based on inter-station distances using a fully connected graph with distance-based weighting. This ensures that all stations are connected, but the contribution of each connection is modulated by spatial proximity, allowing the model to emphasize local relationships.

   Regarding whether the graph changes dynamically across different events: this is a key aspect of our method and is discussed in detail in our response to Reviewer B, Question 2. In brief, while the node topology remains fixed (i.e., the set of stations), the input features associated with each node vary from event to event based on the recorded waveforms. Therefore, although the graph structure in terms of connectivity is static, the signal-dependent nature of the input allows the GCN to adapt its processing dynamically. This is further clarified in Section 3.2 of the manuscript, where we describe how event-specific input features are handled.

Discussion:

1. As 89% training data label comes from the ShakeMap, the limitation and potential biases should be extensively discussed here. Also, how this might affect generalization and whether performance against observed data differs. In addition, the calculation time should be compared with the real-time ShakeMap too.

   We thank the reviewer for the insightful comment. Indeed, the target values are composed of 11% observed data and 89% ShakeMap-calculated data, which can introduce potential limitations and biases. This aspect is now explicitly discussed in the revised manuscript (see lines 155-161).

   We have already verified that this does not have a significant impact on the model during training, nor does it compromise its ability to generalize to observational data. In fact, by performing the comparative evaluation in the new Section 4.4 using the subset of test data containing only observed IMs (11%), we indirectly assessed the impact of potential biases introduced by using ShakeMap-derived targets during the training phase. In this evaluation, we compared our model's predictions with those obtained from the GMM by Bindi et al. (2011), which is specifically calibrated for Italy. For this comparison, we excluded the ShakeMap predictions used in the absence of observed data. Our analysis shows that the median values of log(IM_obs / IM_pred) are close to zero, indicating that the model performs well and is capable of accurately predicting observed ground motion values.

   With respect to computational performance, it is important to first clarify that ShakeMap-calculated data were used exclusively during the training phase and will not be adopted in any operational setting. Furthermore, our method

achieves significant improvements in inference time compared to the standard real-time ShakeMap pipeline: once the waveform features are extracted, the trained GCN produces spatial ground motion estimates in under 1 second, whereas real-time ShakeMap generation can take several minutes, depending on data volume and processing complexity. This highlights the suitability of our approach for rapid applications such as early warning or immediate post-event assessment, although it is not intended to replace ShakeMap in detailed post-processing scenarios.

Reviewer B

This paper proposes a deep-learning-based method to perform rapid prediction of PGA values in Italy through a graph neural network, along with a masking mechanism to dynamically select seismic waveforms where the P-wave arrival falls within 10 seconds of the earthquake origin time. The authors conducted three main tests: (1) they evaluated the impact of incorporating additional event information, (2) they assess the effectiveness of the method against baseline approaches, and (3) they test whether a smaller window would degrade the prediction. Overall, I enjoyed reading this manuscript and I believe that it is suitable for publication in Seismica after a few moderate-to-minor reviews.

Major comments

1. Model. You adopt a "time then space" strategy. I was surprised to see that you use convolution kernels of size 125 time steps for your CNN encoder. Why did you choose such large filters? What about using the pretrained encoder from Bloemheuvel et al.? Or, maybe even better, why didn't you choose a pre-trained temporal encoder such as PhaseNet, which has been trained on 2M seismic waveforms? Would this be relevant for your case?

   Regarding the use of convolutional kernels of size 125 time steps: we shared your initial concern about the unusually large filter size. However, we decided to follow the original configuration from Bloemheuvel et al. (2023), as our baseline model closely replicates their architecture. Our goal was to maintain comparability and assess the performance of their design before introducing modifications.

   Regarding the use of a pre-trained temporal encoder: this was indeed part of our original research plan for future work, but your suggestion prompted us to anticipate this step and run additional experiments. Based on your recommendation, we selected PhaseNet, as it is a larger and more robust model trained on a much broader dataset. We used the pre-trained version of PhaseNet provided by SeisBench, trained on the STEAD dataset.

   Incorporating the pre-trained PhaseNet encoder yielded noticeably better performance (see Table 1 attached below), confirming the benefits of improved generalization. Additionally, it brought significant computational advantages: GPU memory usage decreased from 9.2 GB to 2.584 GB.

   We did not include the pre-trained encoder from Bloemheuvel et al.(2023) in our comparisons because the dataset they used for pre-training is a subset of our own, which could result in an unfair advantage and compromise the fairness of the evaluation.

   Table 1. MSE results for the experiments on our data set incorporating the pre-trained PhaseNet encoder. The letters A, B, and C indicate the addition of the coordinates of the first station that records the P-wave (A), interstation distances (B), and maximum amplitude information (C), respectively, as metadata inserted into the flattened layer.

Experiment	A	B	C	PGA	PGV	SA(0.3)	SA(1.0)	SA(3.0)
1	no	no	yes	0.183	0.188	0.188	0.229	0.256
2	yes	yes	no	0.196	0.204	0.200	0.247	0.278
3	yes	no	yes	0.183	0.189	0.189	0.230	0.257
4	yes	no	no	0.192	0.203	0.196	0.247	0.279
5	no	yes	yes	0.183	0.189	0.189	0.230	0.257
6	no	yes	no	0.187	0.188	0.196	0.229	0.267
7	no	no	no	0.192	0.193	0.196	0.236	0.263

2. Masking strategy. In section 3.2, you say that you prefer keeping a "consistent graph structure" in lieu of having distinct graphs for each event. Why this choice? Later, on line 201, you say: "graph convolutional networks [...] typically struggle to handle changes in node structures". I am not fully convinced by this statement. Graph neural networks were originally applied to protein classification, where indeed several proteins are modeled as distinct graphs. Why couldn't you build a graph for each event instead of forcing to zero nodes with no P-wave arrival within 10 seconds?

   Thank you for your insightful comment. In response to your suggestion, we experimented with training the model using per-event dynamic graphs, where the graph structure is reconstructed for each event based on P-wave arrivals. However, we encountered several limitations (see Table 2 below) that ultimately led us to retain the fixed graph approach in the main selected experiments.

   You are correct that GNNs can, in principle, operate on batches of distinct graphs, as is common in domains like molecular classification. However, in our specific context, we opted for a fixed graph structure across all events for both methodological and practical reasons:

   1) Training stability: We observed that using per-event dynamic graphs (i.e., reconstructing the graph for each event based on P-wave arrivals) led to highly unstable training. This was mainly due to batch inconsistency (rapidly changing topologies between samples), small computational graphs (few active nodes), and loss of station identity across events. These factors resulted in high gradient variance, poor convergence, and lower overall performance, likely due to limited message passing and disrupted spatial context.

   2) Efficiency: Dynamic graph construction is also computationally expensive. For example, training with graphs built on-the-fly increased the epoch time from 0.08 seconds (fixed graph) to 166 seconds per epoch. To address this, we precompute the dynamic graphs for each event, which makes training time comparable to the fixed graph setup. Notably, the dynamic graph version reduces GPU memory usage from 2.584 GB (fixed) to 1.8 GB (dynamic), which is a practical advantage.

   In summary, while dynamic graphs are theoretically possible and offer some computational benefits, in practice they resulted in less stable and less effective training in our scenario. We believe this is due to both instability and reduced message passing. For future work, we plan to explore more sophisticated dynamic graph approaches, such as edge reweighting or attention mechanisms, to better balance stability and local adaptivity.

   It should also be noted that the information on which stations are masked or not (i.e. triggered or not) brings additional constraints on the earthquake location. While the model does not provide the epicenter location explicitly, it is expected that it needs to implicitly understand where the earthquake is to adequately calculate the ground motion propagation to more distant stations. This could be another (physically explainable) factor which could affect the results of the dynamic graph experiments.

   Table 2. MSE results for the experiments on our data set using per-event dynamic graphs. The letters A, B, and C indicate the addition of the coordinates of the first station that records the P-wave (A), interstation distances (B), and maximum amplitude information (C), respectively, as metadata inserted into the flattened layer.

Experiment	A	B	C	PGA	PGV	SA(0.3)	SA(1.0)	SA(3.0)
1	yes	yes	yes	1.097	0.539	1.198	0.797	0.491
2	yes	yes	no	1.098	0.541	1.199	0.799	0.494
3	yes	no	yes	1.091	0.539	1.192	0.796	0.492
4	yes	no	no	1.107	0.544	1.210	0.806	0.497

3. Results. This section could benefit from some reworking. In particular, the order in which you present the tests is not entirely logical in my opinion. I suggest that the authors rearrange the paragraphs as follows: (1) Model performance, (2) Example of event data prediction, (3) Baseline models, (4) Additional knowledge, (5) window length. This way, the overall results are presented first, then the model is compared against baseline methods (reversing the order of these two works as well). After the model has been validated, it makes more sense to test how additional knowledge and window length affect the results. Also, it would be beneficial to provide more informative titles for paragraphs 3, 4, and 5 (e.g., something like "comparison with baseline models" instead of "baseline models" alone).

   We have revised the order of the paragraphs as recommended and we have updated their titles to be more informative. The only exception is the first paragraph (Section 4.1), which has been kept in its original position. This is because it describes the preliminary evaluation used to define the final configuration of the model adopted throughout the study, and therefore logically precedes the presentation of the results and validation and comparative analyses.

4. Follow-up comment on the masking strategy. I may be missing something, so I would appreciate it if the authors could clarify. In section 4.5, you test a smaller window and you relax the constraint on the masking mechanism in order to include more stations regardless of their distance from the first recording station. As a result, you find comparable

performance with respect to the “10-second” case. Does that imply that your masking mechanism is somehow redundant and that your GNN extracts spatial features that are sufficient for the task? Would it perhaps be worth testing whether your results remain unchanged if you force the mask to 1 in the 10-second scenario.

We thank the reviewer for this insightful observation. As shown in Figure 9, relaxing the masking constraint and reducing the input window to 5 seconds results in a slight increase in MSE across all IMs, ranging from +0.75% to +9.39%. This demonstrates that while the GCN layers are indeed effective at extracting spatial features and mitigating the impact of incomplete data, the masking mechanism still plays a relevant role, particularly for shorter-period IMs like PGA and SA(0.3), where up to a 9% increase in error was observed.

Moreover, it is worth noting that relaxing the masking constraints results in the inclusion of signals from stations that may have not yet received meaningful earthquake waveforms. These early-time signals—especially from distant stations can introduce additional noise or irrelevant patterns, which may lead the model to learn spurious correlations. While this may not strongly affect performance on the validation set, it could harm generalization, particularly for unseen seismic events or different geographic configurations.

Thus, the dynamic masking mechanism serves a dual purpose: (1) enforcing a realistic temporal constraint aligned with EEW requirements, and (2) acting as a regularization mechanism by limiting input to physically meaningful, time-consistent data. We agree that further sensitivity testing—such as enforcing a full mask of 1 in the 10-second scenario—would be valuable and plan to explore this in future work.

5. Training set. Could you please provide the size of the training set? Is that 975 waveforms? If that is the case, I am having a hard time understanding how the training of your 53-million-parameter model converged with only 975 training samples. Also, 53M parameters seem quite an overshoot to me. Did you actually need all those parameters?

   Thank you for your question and for pointing out a potential misunderstanding.

   While it is correct that we have 975 seismic events, each event is recorded across up to 565 stations, resulting in a total of 35078 waveform instances. Therefore, the effective training dataset is substantially larger than 975 samples. These data are split into a training set (80%) and a test set (20%). For validation, we use 1/5 of the training set in rotation, as we adopt the k-fold cross-validation technique with 5 folds. We included this information in the main text chapter 3.1 (lines 196-199).

   Regarding the 53 million parameters model: while the parameter count is nontrivial, the data volume, when considered at the station-waveform level, is sufficient to support model convergence. Additionally, our architecture includes domain-specific constraints such as structured masking, graph-based inductive biases, and physical priors (e.g., arrival time, distance), which all contribute to better generalization and reduce overfitting risk.

   With the updated version of our model, in which we replace the initial CNN layers with the pretrained PhaseNet feature extractor, the total number of trainable parameters is increased to 55 million parameters.

Minor comments

6. In the introduction, at lines 58-59, I would also add that the strength of CNNs is the capability of extracting patterns from raw data.
   This point has been added to the introduction (see lines 55-59).
7. Line 63: you mention Random Forests while talking about deep learning. This is not the place for that, and it should rather be moved to line ~52 when discussing machine learning methods.
   We thank the reviewer for the comment. The mention of Random Forests has been moved to line 45, where machine learning methods are introduced.
8. Line 70-71: I am puzzled by these two lines: what is the point of talking about a method that talks about the graph adjacency matrix?
   The sentence referring to the method involving the graph adjacency matrix (Bloemheuvel et al. (2024)) has been removed to avoid confusion and maintain focus on the relevant approaches discussed.
9. Figure 1: you could plot the events as points with a size dependent on the magnitude.
   The figure has been updated accordingly: the events are now plotted as points with sizes scaled proportionally to their magnitudes.
10. Line 148: “with several modifications”. Which ones?
    The main architectural modifications are now explicitly specified in the revised manuscript in lines 170-172.
11. Line 157. You should briefly recall how you built the adjacency matrix here, as you never state it explicitly.
    A brief explanation of how the adjacency matrix is constructed following the methodology proposed by Bloemheuvel et al. (2023) has now been included in Section 3.1 (lines 181–186).

12. Line 159: “tanh activation function to constrain the representation”. What do you mean? Hyperbolic tangent activations are usually used to add non-linearities as well as to bound the activations between -1 and 1. Is that what you mean?
    We agree with the reviewer that the original sentence was unclear. We have revised it as follows:
    “The output of the GCN layers is passed through a tanh activation function to introduce non-linearity and bound the values within [-1, 1] (lines 186-187).
13. Line 162: “we consider the maximum amplitude value from the vertical component of the waveform data”. What do you use that for? Unclear.
    The sentence has been revised to improve clarity (lines 190-191)
14. Line 163-164: “we verified that the same results were obtained when incorporating the maximum amplitude value across all three components, which is the solution to be adopted in the operational phase”. What do you mean?
    The sentence has been revised to improve clarity (lines 191-193)
15. Line 170: “A learning rate scheduler, starting at 0.0001, was employed to dynamically adjust the learning rate during training”. Which type of scheduling did you use? Linear, cyclic, warm-up strategy, ...? What is the minimum/maximum value of the learning rate that you are exploring across the iterations?
    We used a ReduceLROnPlateau scheduler that starts with a learning rate of 0.0001, and reduces it by a factor of 0.5 if the validation loss does not improve for 3 consecutive epochs. The minimum learning rate explored is 1e-6. We have updated the text to clarify the type of scheduler and its parameters. We have updated the text (Chapter 3.1), to clarify the type of scheduler used and its associated parameters (lines 199-201).
    Additionally, we confirm that we used early stopping with a patience of 8 epochs, based on the validation loss, to prevent overfitting and ensure training efficiency. The training loop terminates early if no improvement in validation loss is observed over 8 consecutive epochs.
16. Line 186: I was wondering why you use 10 seconds? Does this come from some tradeoffs between data quality, regional characteristics, etc.?
    We thank the reviewer for the question. In the study by Jozinović et al. (2020) conducted for earthquakes in the same region, the authors investigated the optimal window length after the origin time required to make reliable predictions of the maximum values of IMs. This work describes the same deep convolutional neural network (CNN)-based technique to predict intensity measures (IMs) of earthquake ground shaking that we adopt in our model.
    They tested window lengths of 7, 10, and 15 seconds. For the 7-second window, the mean squared error (MSE) of the residuals between the base-10 logarithms of observed and predicted IMs [i.e., $\log_{10}$(IM_true / IM_predicted)] was 0.228. This value decreased to 0.176 for a 10-second window and further to 0.165 for a 15-second window, indicating a significant drop in performance with shorter windows.
    Based on these findings, Jozinović et al. (2020) adopted a 10-second window as a good trade-off between prediction accuracy and timeliness. We followed the same approach in our work.
17. Line 288: “inconsistencies in the input waveform patterns”. What do you mean?
    The sentence has been revised to improve clarity (lines 300-302)
18. Figure 5: I see that the model has a slight tendency to overpredict (underpredict) smaller (larger) values for the quantities that you compare here (basically, the linear regression computed on the densest points seems to have a slightly smaller slope than the 1:1 line). How do you comment on that? Is that relevant?
    We appreciate the reviewer’s observation. As noted, the model shows a slight tendency to overpredict lower IM values and underpredict higher ones. However, we do not consider this trend to be particularly relevant, for several reasons.
    First, the behavior is common in models trained to minimize the overall mean squared error, especially when the training data are unbalanced across the range of target values—as is often the case with IMs, where low-to-moderate amplitudes are far more frequent than high ones. Second, as discussed in the manuscript, this tendency does not significantly affect the practical performance of the model. The residual distributions remain centered, the scatter remains acceptably low, and importantly, the model retains robust accuracy across the entire range of predicted values.
    For these reasons, and in line with previous studies using similar learning-based approaches, we interpret this as a minor artifact of the optimization process, not one that substantially compromises the reliability or usefulness of the predictions.
19. Line 332: “indicating its ability to learn seismic wave propagation patterns”. This is yet to be demonstrated. I would maybe just say that the model learns spatial relationships between the stations, conditioned on the data.
    The sentence has been updated to incorporate the suggestion and to provide a more accurate description of the model’s behavior (lines 368-371).

Giuseppe Costantino