Responses by authors in blue:

We thank the Referee#1 for their valuable comments and suggestions which improves the quality of the manuscript. Detailed responses are provided to your questions. Blue text shows our response, updates which will be incorporated in the manuscript are highlighted in red as track change and black text shows the referee comments.

The paper presents an advance in tsunami hazard approximation, offering a machine learning (ML)-based solution to the computational challenges of traditional methods. However, for publication, major revisions are necessary to address the following concerns.

Details on Test Sets

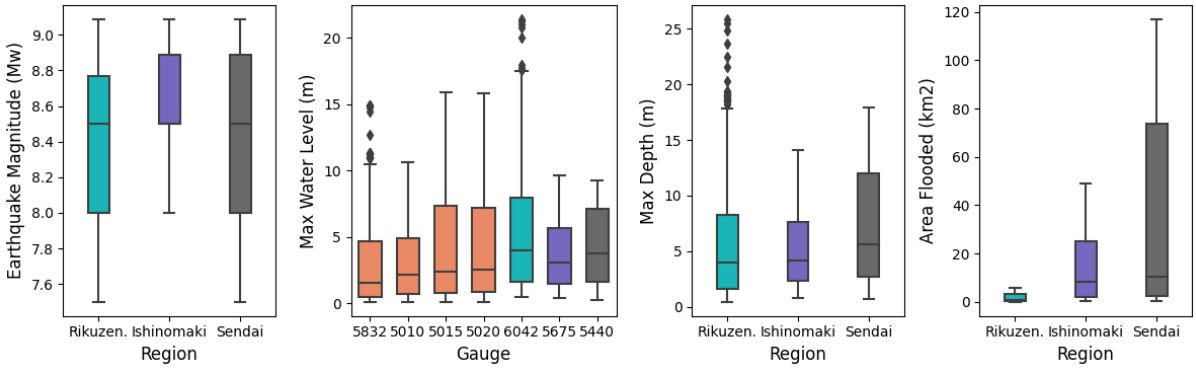
The paper does not provide sufficient details on the features of the test sets used for model validation. Specific characteristics and parameters of the test sets should be clearly stated to assess the model's generalizability. Additionally, the rationale behind selecting specific test sets is only briefly mentioned. A more detailed justification for their selection is necessary. Include a detailed description of the test sets, including their characteristics. Explain how these sets represent the diversity of potential tsunami scenarios and justify their choice to enhance the study's credibility.

Response:

We assume that the reference here is to the test sets used from the design of experiments (DOE) during the k-fold cross-validation (CV) study. K-fold CV is widely used for ML model evaluation, especially when dealing with limited data, as it ensures that each data point is used for both training and validation, enhancing the robustness of the evaluation. In our study, we utilize a 5-fold CV approach, where the dataset is randomly partitioned into five equal-sized folds. Each fold is used once as the test set while the remaining four folds are used for model training. This process is repeated five times, with each fold serving as the test set exactly once. The characteristics and parameter ranges for each test fold as when all the test folds are combined, mirror those of the overall training dataset. This is summarized in Table 2 and Figure 6.

To address the reviewer's request for specific details, we updated the following information into the revised manuscript:

Updated Figure 6. shared below to also describe the earthquake magnitude in the training set for the three regions, along with outliers.



In section 2.2.2, Line 296 is modified as:

"Given the relatively small size of our training dataset, we employ K-fold cross-validation with five folds to fully use the available data set for training and testing. The data set is randomly partitioned into five equal-sized folds. Each fold is used once as the test set while the remaining four folds are used for model training. This process is repeated five times, with each fold serving as the test set exactly once. The characteristics and parameter ranges for each test fold as when all the test folds are combined together mirror that of the overall training dataset as found in Table 2 and Figure 6. This helps in assessing the model's generalisation ability and sensitivity to overfit with such limited training information (Mulia et al., 2020)."

Reference: Mulia, I. E., Gusman, A. R., and Satake, K.: Applying a Deep Learning Algorithm to Tsunami Inundation Database of Megathrust Earthquakes, Journal of Geophysical Research: Solid Earth, 125, e2020JB019 690, https://doi.org/10.1029/2020JB019690, eprint:https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020JB019690, 2020.

Model Training and Weights

The paper does not discuss whether different weights were assigned to various types of events during training. Assigning different weights could help ensure the model does not overfit to more frequent, less severe events, thus improving its predictive performance for rare, high-impact events. Elaborate on the training process, including whether different weights were assigned to events and how this impacted model performance. If not used, consider implementing and discussing the potential benefits of such weighting schemes.

Response: In traditional classification problems, approaches such as over sampling, under sampling, and reweighting are commonly used to mitigate class imbalance problems. However, for our tsunami surrogate models, the definition of imbalance is more complex and multifaceted. It can be based on various factors, including earthquake characteristics (e.g., magnitude and hypocentre location), waveform input characteristics (e.g., maximum amplitude), or output characteristics (e.g., maximum inundation depths and coverage).

Understanding and addressing the influence of such imbalances in the training set is crucial. While assigning weights to different types of events is a common approach, it requires a detailed understanding of how these weights would impact the sensitivity and performance of the model. This is a subject of our forthcoming work, where we will investigate these aspects in detail using evaluation on a large simulation database.

In the current study, we did not explicitly assign different weights to various events. Instead, we focused on the robust design of the surrogate model and implemented several regularization techniques to handle the potential overfitting and inherent imbalance. We employ shallow neural network layers, with regularization schemes such as max-pooling operations, dropout, and variational latent spaces to ensure the model can scale appropriately for different magnitudes of events and their associated input and output parameters. Our model's performance, particularly for higher water levels in time series and water depths in inundation, shows stability and significant skill above an L2Norm value of 10, as illustrated in Figure 17(a). Additionally, Figures 16(a) and 16(b) demonstrate that the accuracy (A) and goodness of fit (G) metrics tend to improve for higher depth classes in both DOE test sets and historical events. This also indicates that the model does not overfit to more frequent, less severe events.

Furthermore, in the initial stages of our work, we explored other approaches to mitigate imbalance or scaling issues, which were not all beneficial:

1. **Scaling and normalizing** the dataset based on training data.
2. **Curriculum learning scheme**, gradually train the model from lower magnitude (Mw) events to higher ones.
3. **Alternative output parameters**, considered using max inundation height as the output parameter for the onshore surrogate.
4. **Supplement the training dataset** with more events(type B), initially our dataset only consisted of type A events.

In conclusion, while we did not use explicit weighting schemes in the current study, we acknowledge the potential benefits of such approaches. Our future work will delve deeper into understanding the impact of event weighting and other balancing schemes to enhance model performance and generalizability as suggested in Section.5 L441.

Information Leakage

The paper lacks details on measures taken to prevent information leakage, which can lead to overly optimistic performance estimates if the test data inadvertently influences the training process. Clearly outline the steps taken to ensure strict separation between training and test data. Discuss any data augmentation techniques and how they are managed to avoid information leakage. This

transparency will strengthen the reliability of the reported results. What do you think about the possibility of information leakage for the 2011 Tohoku test case?

Response: To prevent information leakage, but also overfitting our study adhered to the following protocol:

1. **No data augmentation**: We did not use data augmentation techniques in our study.
2. **Data splitting and cross-validation**: As discussed in our response to the comment above on "Details on Test Sets," we employed a 5-fold cross-validation (CV) approach. In this method, the dataset is randomly partitioned into five equal-sized folds. Each fold is used once as the test set, while the remaining four folds are used for training. This process ensures that every data point is used for both training and validation, and it helps prevent any inadvertent information leakage that could occur with a single train-test split.
3. **Ensemble of variational encoder-decoder (VED) Models**: To mitigate the risk of overfitting, sensitivity to parts of the training data, information leakage, we used an ensemble of models for the generalisation tests on historical events. Each model in the ensemble was trained on different subsets of the training data derived from the 5-fold CV process. The final predictions were generated by aggregating the results from these models. This ensemble approach helps capture the overall uncertainty due to variability in the training dataset and the stochastic nature of model parameterization in the latent space, enhancing the robustness and reliability of the results.

This ensures that there is no contamination of the training data by the test data, providing an unbiased evaluation of the model's generalization capabilities for the historic events like 2011 Tohoku test case.

Conclusions

The results shown in Figure 15 suggest that the quality of training sets is more important than the quantity. For the test cases outside the design of experiments, there are clear discrepancies between the prediction and observation, which is an intrinsic nature of ML algorithms. Include a discussion on the design of experiments, including the limitations. This will provide a more comprehensive understanding of the model's strengths and areas needing improvement.

Response: That is an important remark, that designing a training dataset limited in size but with sufficient variability is vital for training such tsunami surrogate. We will add an additional paragraph to the discussion section as follows:

"The training information from our DOE is constrained by both the quality and quantity of events, recognizing its limitations is crucial for guiding future improvements in the experimental design and training dataset along with advances in the model architecture and training. First, the geographic focus is limited to the Tohoku subduction source region and modelled with a simple scheme, restricting the diversity of the training data and impacting the model's ability to generalize to other regions or varied tsunami scenarios, such as the historic test events (b, c, and d). Second, there is an event imbalance for inundation, particularly for the onshore surrogate, where more events of large inundation are needed to provide sufficient training scenarios at locations far from the coast, which are rarely inundated in the training dataset. Finally, the generalization test on the onshore surrogate highlights varying prediction accuracies across different test sites, at Rikuzentakata, Sendai, and Ishinomaki. These variations reflect the complexities and limitations of the DOE, where certain test sites with more complex inundation patterns are not well-represented in the training data, leading to less accurate predictions."

Minor comments:

L15 Tsunami -> **Tsunamis**

L16 USD 280 billion damage -> USD 280 billion **in** damage

L29 a past historical event -> historical events

L91 machine learning-based -> ML-based; You may replace machine learning with ML in other places of the paper.

L193 **Inconsistency in labeling the test events.** Make sure the labels are consistent including Test A&E in the 2011 Tohoku case, Figure 5 & 15.

L458 remove "it contains"

Response:

Thank you for your observations, listed correction will be updated in the revised manuscript.