Responses by authors in blue:

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

Review by Referee #1

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

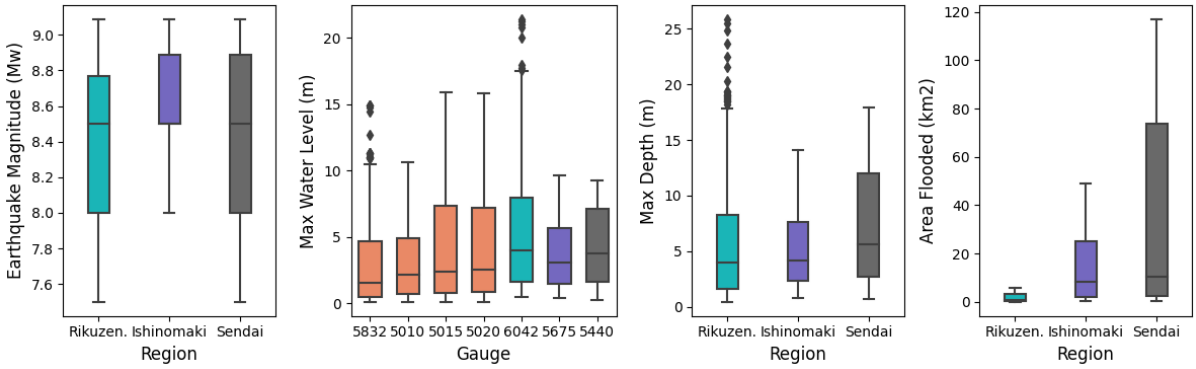
The paper does not provide sufficient details on the features of the test sets used for model validation. Specific characteristics and parameters of the test sets should be clearly stated to assess the model's generalizability. Additionally, the rationale behind selecting specific test sets is only briefly 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 much larger simulation database than in this study. Line 473-476 in revised manuscript discusses the benefit of such approaches in the discussion section.

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:

Section 2.2, especially 2.2.2 with line 308-318 explains well how our training approach tackles overfitting tendencies and prevents leakage.

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 have been updated in the revised manuscript.

Review by Referee #2

The study uses a surrogate approach based on a variational encoder-decoder (VED) to predict the tsunami time series at different depths and maximum inundation depths at three coastal sites in Japan. The surrogate accuracy is validated against historical rupture scenarios. I add some comments below that I believe could strengthen the work presented.

Comments:

The design of experiments is not very clear in terms of number of scenarios and input variables. The authors mention 559 events split to 383 and 176 depending on the nature of the rupture. More information is needed on how these numbers were selected, and also the parameter ranges that led to the variation in magnitudes (length, width, displacement). Furthermore, more clarification is needed on whether there are any other input variables beside the moment magnitude (e.g. location of the event) that are varied in the surrogate development.

Response:

Thank you for highlighting the need for a more detailed description of the design of the experiment (DOE). We updated the section 2.1.2 providing more details as reads below.

The design of the experiment consists of a total of 559 hypothetical events of two categories of rupture, distinguished by their geometry and slip distribution: (a) Type A - ruptures represented by a single rectangular planar surface with homogeneous slip and (b) Type B - ruptures that combine numerous smaller rectangular planar surfaces (i.e., sub-ruptures), each of which has homogeneous slip, such that the rupture surface can bend and the slip distribution can be heterogeneous. The number of events in the DOE is constrained by available computational resources and our goal to use a feasible number of training events and maximise the efficiency of the surrogate model. The source scenarios are modelled by adapting procedures previously applied by Gusman et al. (2014) and Mulia et al. (2018).

A total of 119 locations were selected as the top centre of the faults for modelling hypothetical tsunamigenic earthquakes of Type A. These events span M_w 7.5 - 9.0 at an interval of 0.5 and are uniformly distributed over the Tohoku subduction interface, see Figs. 4a and b. This results in a potential 476 events (119 locations x 4

magnitudes). The M_w 9.0 events are restricted to locations where the centre of the rupture's top edge is shallower than 16 km. Deeper events cause unrealistic uplift on large inland portions of the study region and are unlikely to cause tsunami. To ensure realistic modelling and prohibit these events from adversely affecting the quality of the surrogate training, these events were excluded, leaving 383 Type A events.

Multi-fault ruptures of Type B were created using a combination of 6 to 12 planar sub-faults similar to the unit sources used in NOAA SIFT database (Gica et al. 2008) of length 100 km and width 50 km are created as in Fig.4c. The event magnitudes range M_w 8.68 - 9.08, and the ruptures are distributed along the shallow section of the Tohoku subduction interface. The bottom centre of the rupture edges are at depths between 17-28 km. The slip distributions are modelled as a skewed normal distribution where the average combined slip value is between 10 and 20 m. The scenarios varied by the number of faults involved: scenarios with 6 faults were assigned a slip of 10 meters, scenarios with 8 faults had slips of 10 and 15 meters, 10 faults with 15 meters, and 12 faults with slips of 15 and 20 meters. This systematic variation led to a total of 176 different Type B earthquake scenarios.

Information on depth, slip, strike, and dip (see Table 2) is derived from the Slab2 model of the Japan trench (Hayes et al., 2018), and the rake is always set at 90 degrees (Aki and Richards, 2002). The seafloor deformation is analytically modelled assuming homogeneous slip for the rupture or sub-ruptures using Okada solution (Okada, 1985) with the value of rupture length, width, and slip scaled (see Table 2) based on the magnitude of the event (Strasser et al., 2010). We consider that the co-seismic displacement is instantaneous and equivalent to the sea surface displacement generating the tsunami. This initial sea surface displacement is modelled to match the base resolution of the tsunami model at 0.01215 degrees grids.

In summary, the DOE for training the surrogate model considered two main factors: (A) moment magnitude, which determines the profile of displacement (length, width, and slip amount) based on the moment magnitude-area scaling relationship (Strasser et al., 2010), and (B) the location of the events where fault parameters such as depth, dip, rake, and strike are derived from the Slab2 dataset.

Table 2. Earthquake rupture parameters for Type A and Type B, the rake value is always 90 degrees.

Type	Mw	Length (km)	Width (km)	Displ. (m)	Depth (km)	Dip (degrees)	Strike (degrees)
Type A Min	7.5	81.37	56.29	0.24	10.2	5.54	187.20
Type A Max	9	613.76	189.23	3.30	45.7	17	225.78
Type B Min	8.68	300	100	4.72	17.01	8.37	188.72
Type B Max	9.08	600	100	17.36	28.98	16.53	222.27

I would suggest adding an outline of the times 1) for building the two ML surrogates, 2) for prediction and 3) to run the deterministic model. Possibly in the form of a matrix, this should showcase the benefits of using a surrogate approach.

Response: This was previously provided as the supplement Tables S5 and S6; we briefly touched on this at line 60 and 460 in the original manuscript. In the revised manuscript we have moved this to the section 5. (Discussion and conclusions) and present it as part of the discussion, see lines 500-509.

In 310 and elsewhere in the manuscript please replace observations/observed with model/modelled or simulations/simulated as it can be confused with physical observations of the event.

Response: Thanks for suggesting this, we have updated this in the revised manuscript.

In table 5 there seems to be a lot of variance regardless the number of the fold. In some cases, increasing the fold reduces the SME but in other cases it increases the SME. Is this variance random or based on certain conditions?

Response: Thank you for your observation. We would like to clarify that we conducted k-fold cross-validation with 5 folds (k=5) exclusively. Each column in Table 5 represents the results from the evaluation using the withheld test set for each fold iteration. Consequently, each column corresponds to a unique combination of random events used for training and testing, with no repetition across fold iterations. The primary purpose of calculating the Mean Squared Error (MSE) across different folds is to assess the model's sensitivity, identify poor fits, and detect signs of overfitting for that specific portion of the training dataset. The variance in MSE observed across folds can indeed reflect differences in the performance of the surrogate model depending on the subset of data used for training. We will update the text to include the below explanations in section 4.1 (General fit):

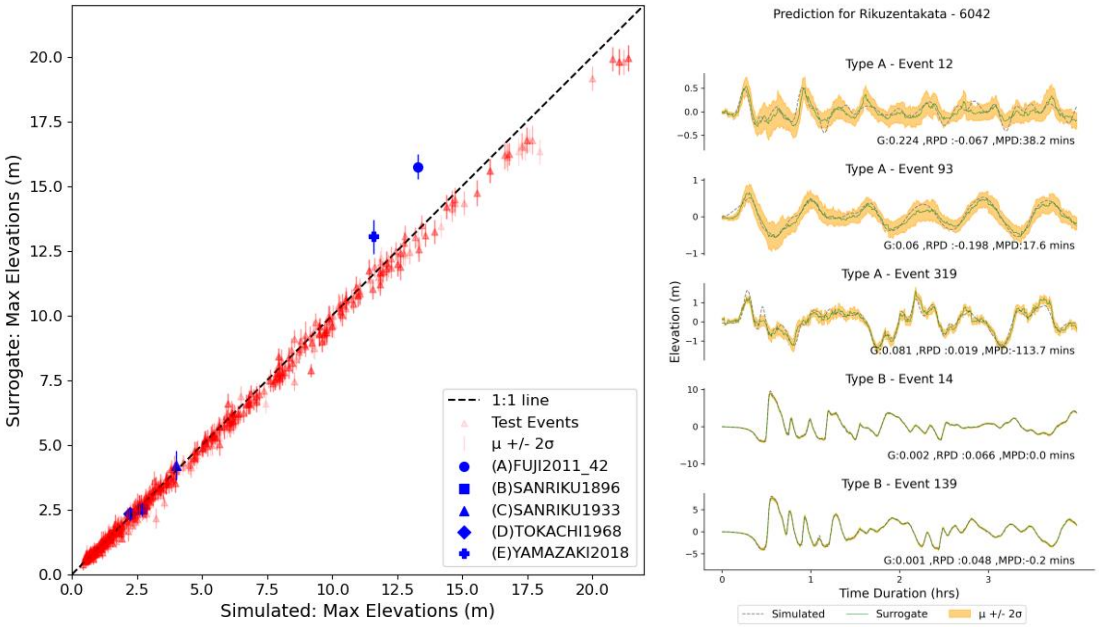
For the nearshore surrogate, the relatively consistent MSE across folds suggests that the model is robust, with no significant overfitting and an overall good fit to the data. However, for the onshore surrogate, there is noticeable sensitivity to the training data, as evidenced by slightly higher MSE values for certain locations, such as Rikuzentakata (fold 2) and Ishinomaki (fold 0). This increased sensitivity could be attributed to two factors, namely the training size and the complexity of the output. The smaller size of the training set may lead to higher variance in model performance, as the model has less data to learn from, making it more sensitive to the specific events included in each training fold. The onshore surrogate is tasked with predicting inundation, which is inherently more complex and variable compared to nearshore waveforms. This complexity can further lead to greater variation in model performance across different folds, especially with limited training data.

In summary, the observed variance in MSE across folds is not random but is influenced by the size and complexity of the training data. For the nearshore surrogate, the model demonstrates stable performance, while the onshore surrogate shows some sensitivity, which is expected given the challenging nature of the inundation predictions.

The legends in the figures should be more descriptive, especially in figures 10, 11 the red, blue symbols, lines and black dotted line. In these figures I would assume that these are simulated outputs instead of observed? In a similar manner for figures 12 and 13 for dotted lines, uncertainty bounds etc.

Response: Yes, you are correct. We will update the figure with better legend to prescribe that, this is a comparison between the values simulated by GEOCLAW numerical simulation and the prediction from surrogate along with its uncertainty bounds.

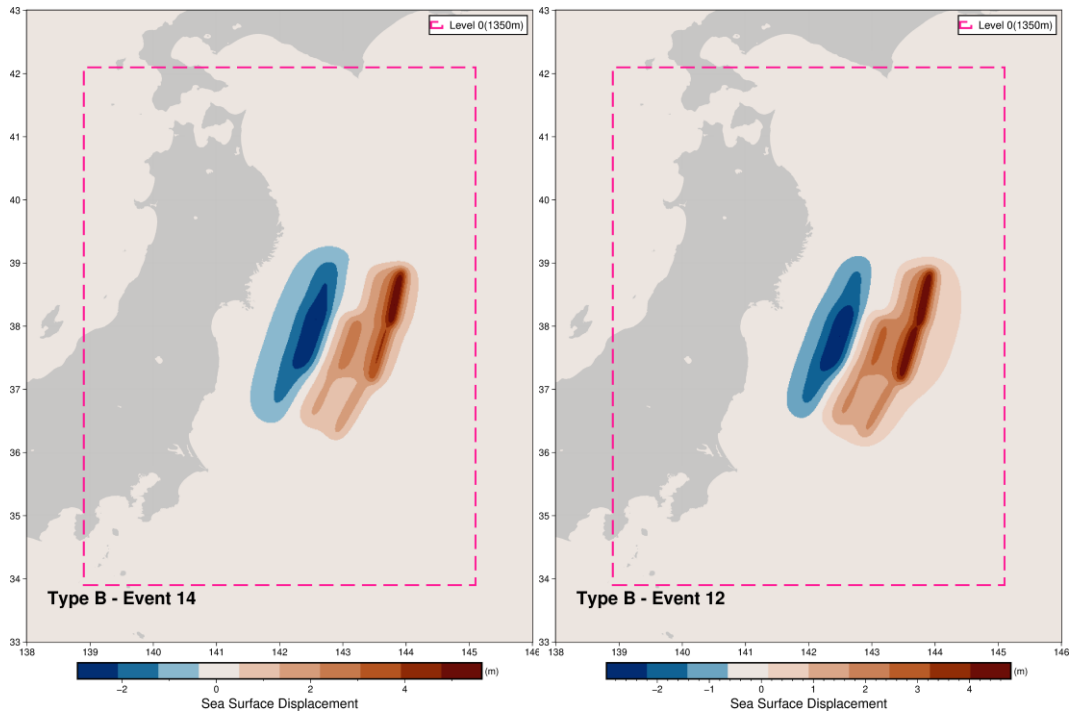
As below:



In figure 12, the predictions of events 14, 139, 102, 87 and 96 match nearly identically to the simulations with very small uncertainty bounds. Are those events in close proximity to other events in the training dataset?

Response: Yes, there are events in close proximity in the training events occurring at same location occur but with a different slip and magnitude. Our surrogate model also has a tendency to fit better for events with larger values as function the loss function built with the MSE component which penalises larger misfit more (also noticeable in the prediction results in Fig.14 available in the next comment) and discussed in the section Generalisation Ability(Line 430).

When considering, for example event 14 of type B for Rikuzentakata. The events in proximity from the training set are events where the same 6 faults rupture but with a different slip distribution. Event 12 is the closest match we found shown in the figures below, additional comments added with line 372-374 to explain such examples.



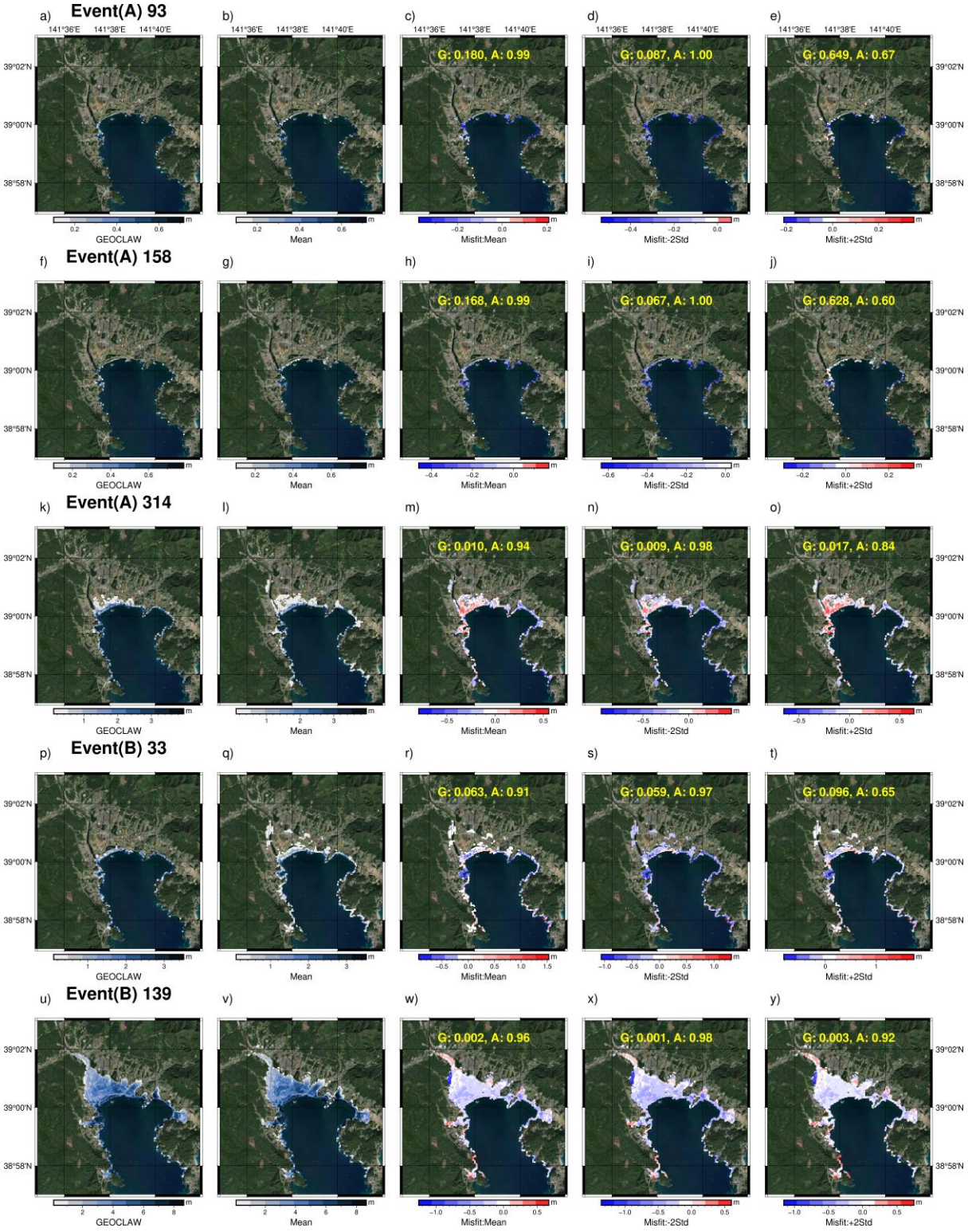
In figure 14, the misfit between predictions and simulations and +-2 standard deviations and simulations (columns 3,4 and 5) seem to be very close in terms of values, possibly because the standard deviations are small? Can the authors provide an example with values and how including the standard deviation reduces the misfit between prediction and simulation?

Response:

Thank you for this question regarding the Fig.14. We have updated the events selection to present more diverse examples; see updated Figures below. In the prediction examples from the DOE test set here, the relatively small standard deviation reflects the high accuracy and low variance of the surrogate's prediction ensemble. This variance reduces especially as the events get bigger in magnitude, driven by the loss function as discussed above.

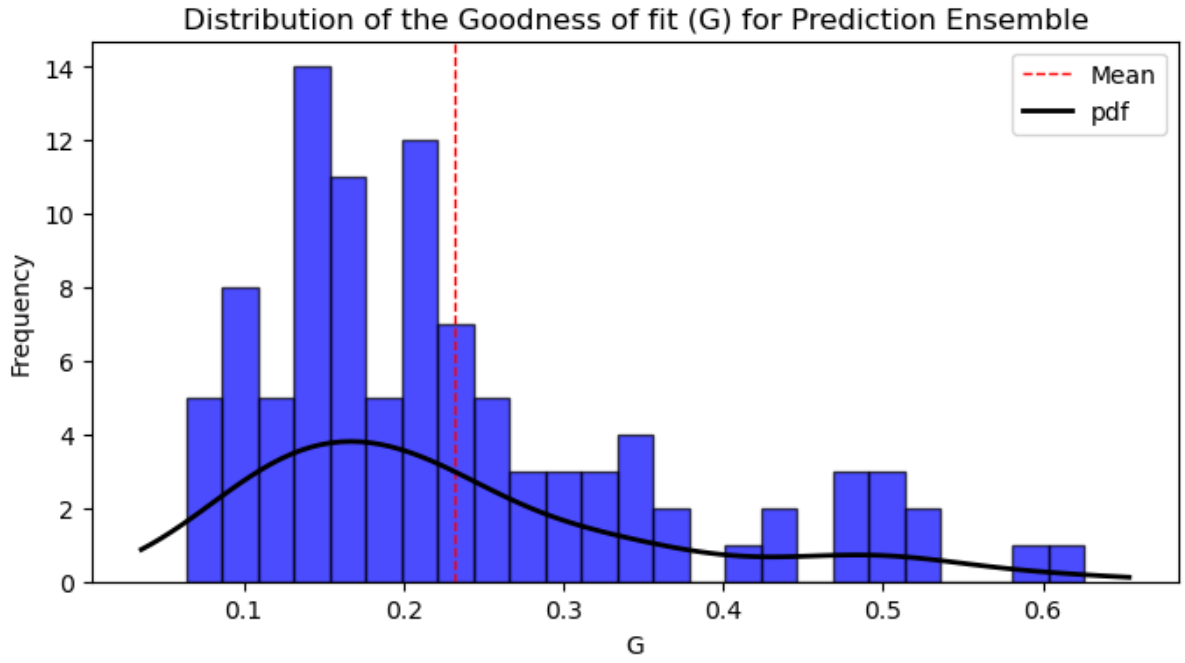
To illustrate how including the standard deviation affects the misfit and accuracy we provide the measure of G and A as an indicator along with the examples for the mean, +- 2 standard deviation Fig.14. Here higher accuracy (A values close to 1) represents good prediction of the inundation extent while good fit (G close to 0) represents good prediction in the inundation depth. The general tendency is that

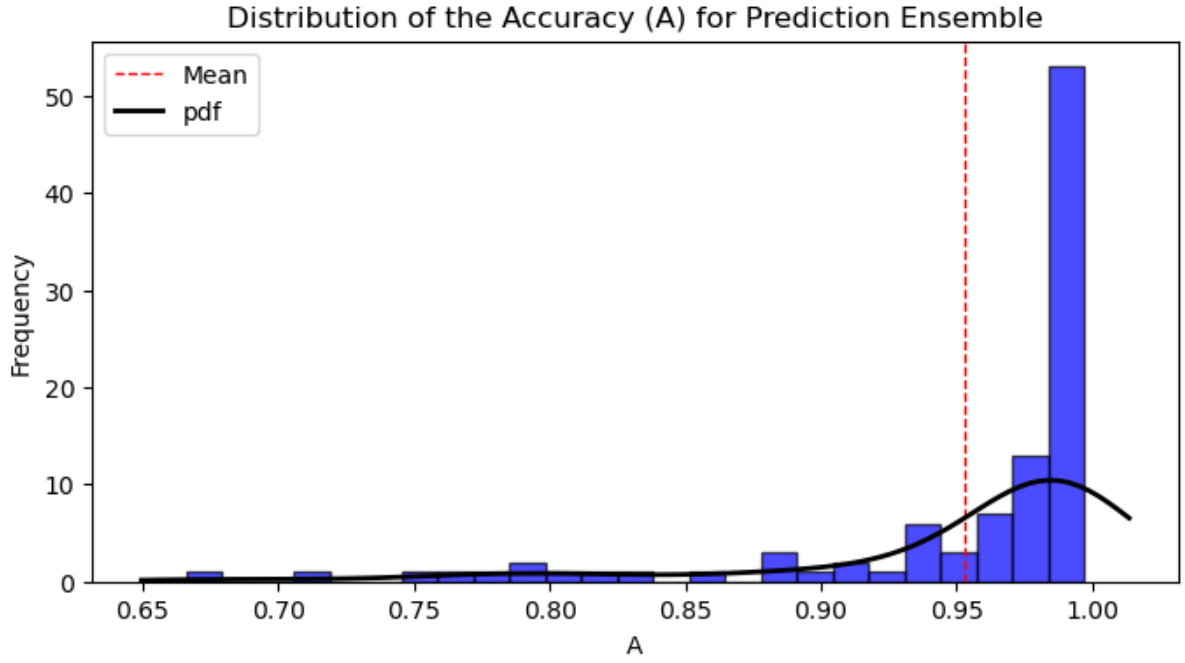
misfit reduces when using mean - 2 standard deviation value, highlighting some overestimation in the mean prediction.



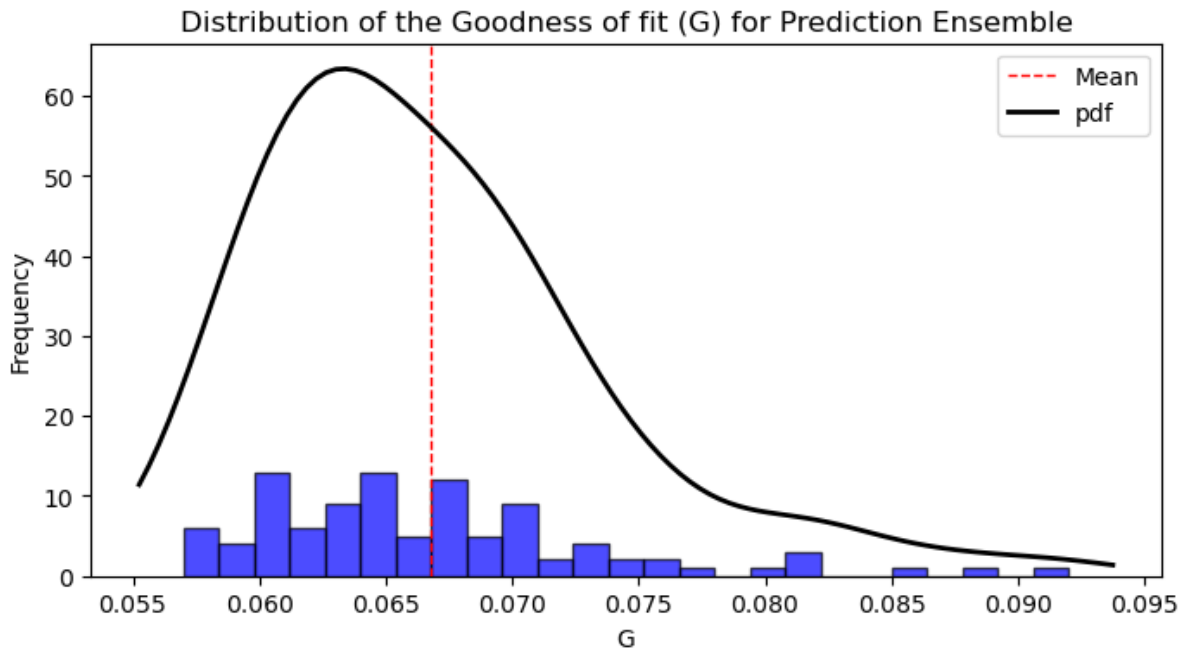
We additionally plot the distribution of the performance metric G and A for two events using the ensemble of the predictions, to show the how the ensemble captures predictions close to the desired simulation results. This is discussed in Line 392 -394 with figure added to the supplement section of the manuscript.

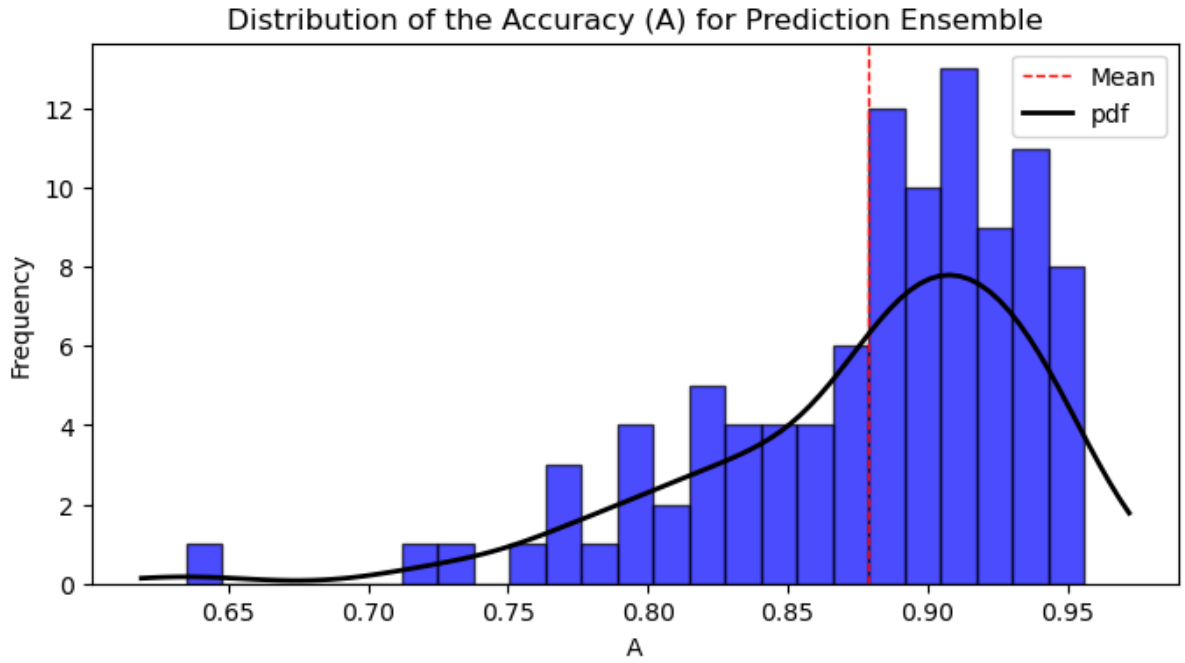
Event 158(Type A)





Event 33(Type B)





References

Gusman, A.R. & Tanioka, Y., 2014. W phase inversion and tsunami inundation modelling for tsunami early warning: case study for the 2011 Tohoku event, Pure appl. Geophys., 171(7), 1409–1422.

Mulia, I. E., Gusman, A. R., and Satake, K.: Alternative to non-linear model for simulating tsunami inundation in real-time, Geophysical Journal International, 214, 2002–2013, https://doi.org/10.1093/gji/ggy238, 2018.

Gica, E., Spillane, M. C., Titov, V. V., Chamberlin, C. D., & Newman, J. C. (2008). Development of the forecast propagation database for NOAA's Short-term Inundation Forecast for Tsunamis (SIFT). *NOAA technical memorandum OAR PMEL*, **139**. https://repository.library.noaa.gov/view/noaa/11079