Responses by authors in blue:

We thank the Referee#2 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 revised manuscript are highlighted as track change and black text shows the referee comments.

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 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 is provided as the supplement Tables S5 and S6; we briefly touch on this at line 60 and 460. In the revised manuscript we will move this to the section 5. (Discussion and conclusions) and present it as part of the discussion.

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 will update 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:

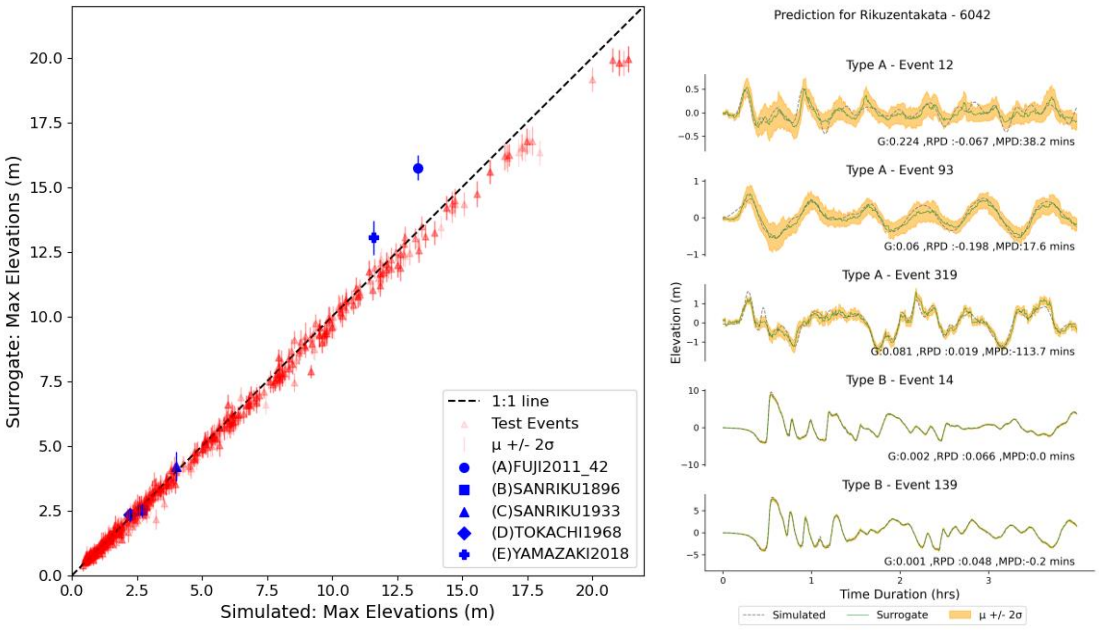
For the nearshore surrogate, the relatively consistent MSE across folds suggests that the model is robust, with no significant overfitting and an overall good fit to the data. However, for the onshore surrogate, there is noticeable sensitivity to the training data, as evidenced by slightly higher MSE 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.

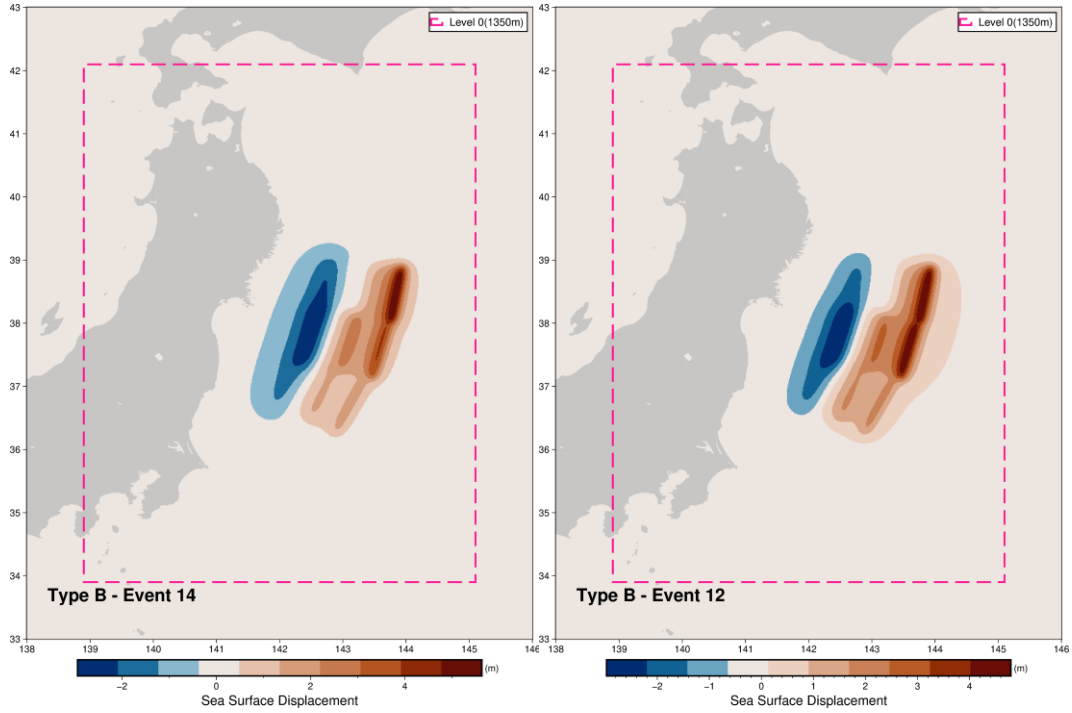
As below:



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).

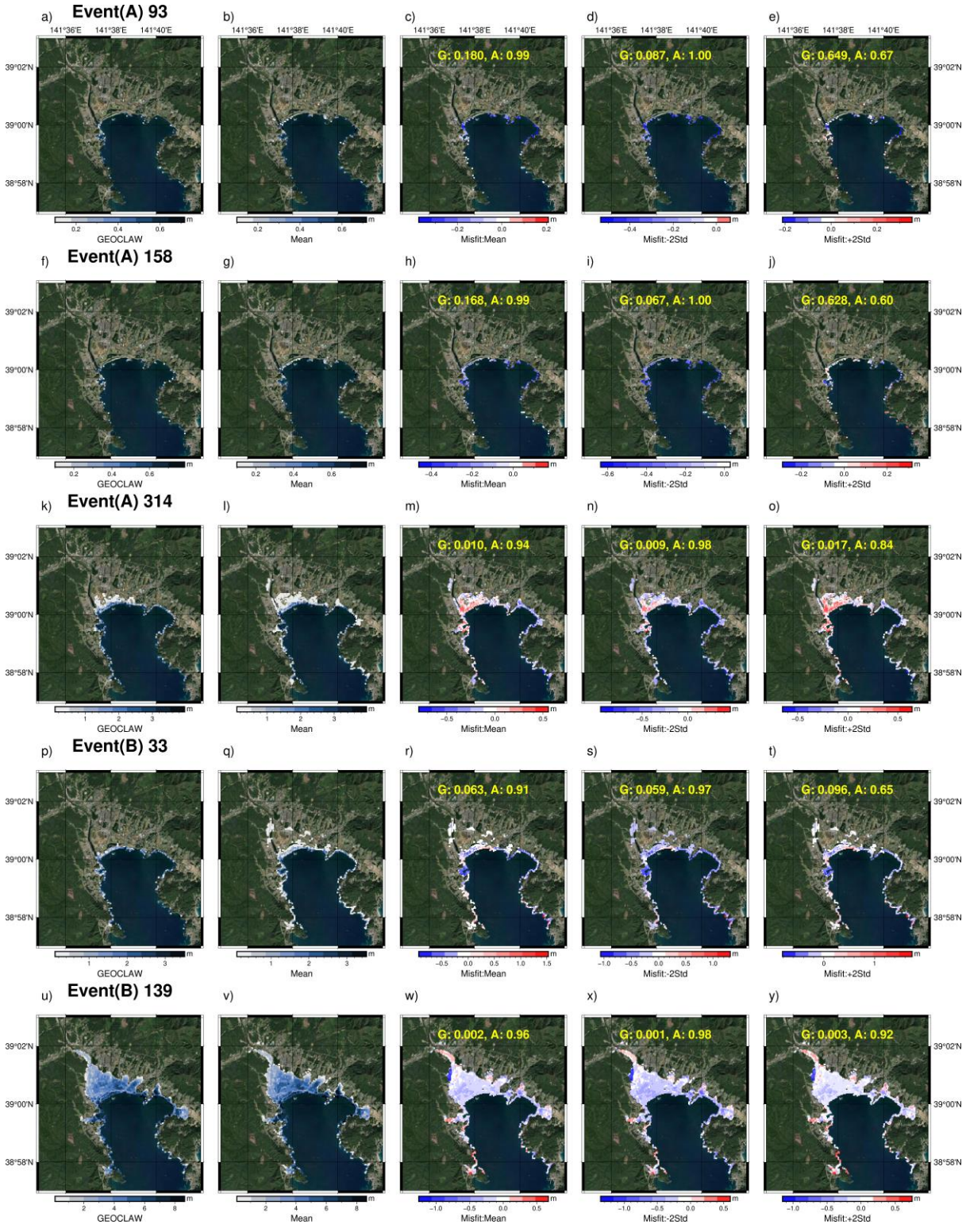
When considering, for example event 14 of type B for Rikuzentakata. The events in proximity from the training set are events where the same 6 faults rupture but with a different slip distribution. Event 12 is the closest match we found shown in the figures below.



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?

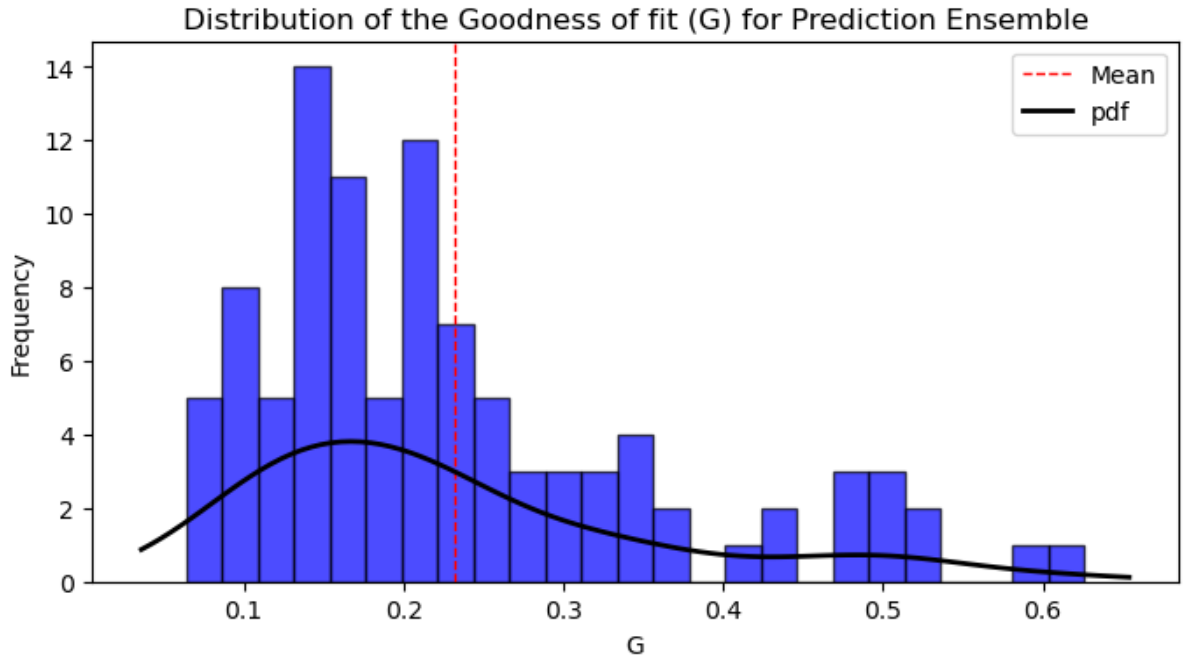
Response:

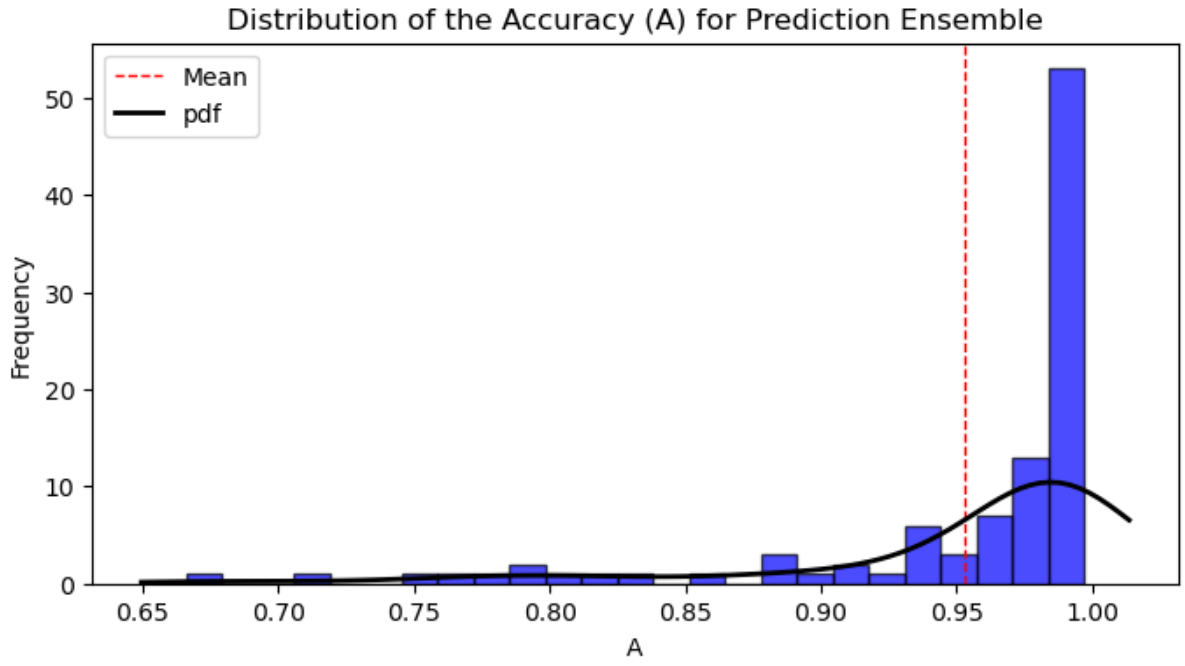
Thank you for this question regarding the Fig.14. We have updated the events selection to present more diverse examples; see updated Figures below. In the prediction examples from the DOE test set here, the relatively small standard deviation reflects the high accuracy and low variance of the surrogate's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



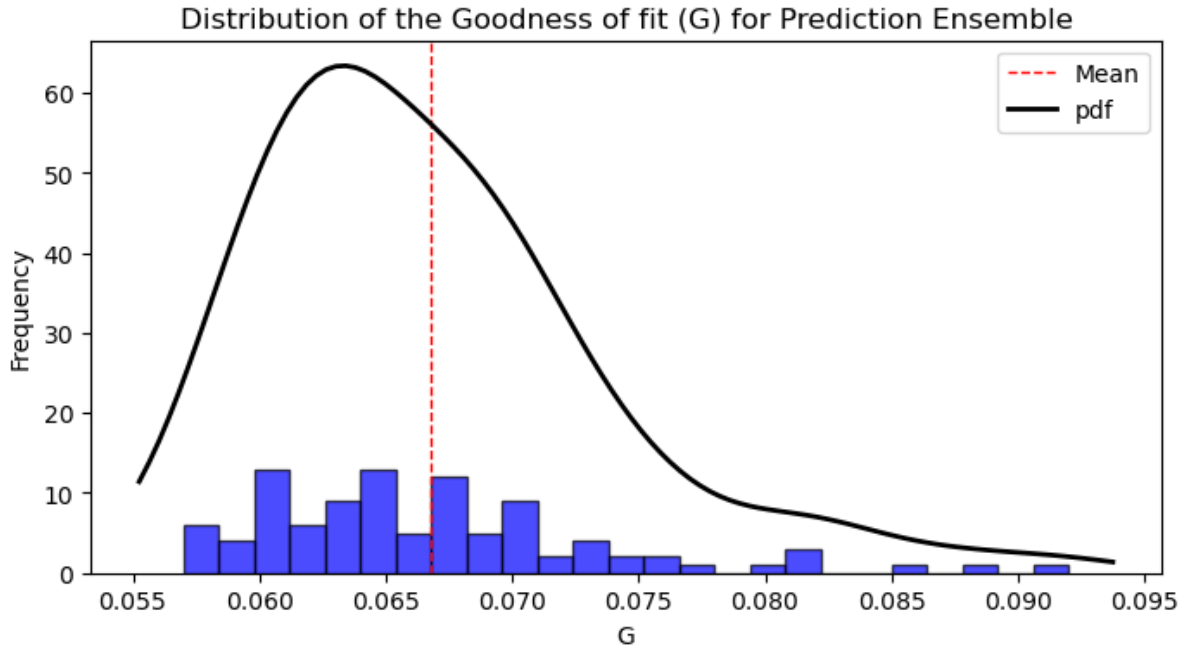
We additionally plot the distribution of the performance metric G and A for two events using the ensemble of the predictions, to show the how the ensemble captures predictions close to the desired simulation results.

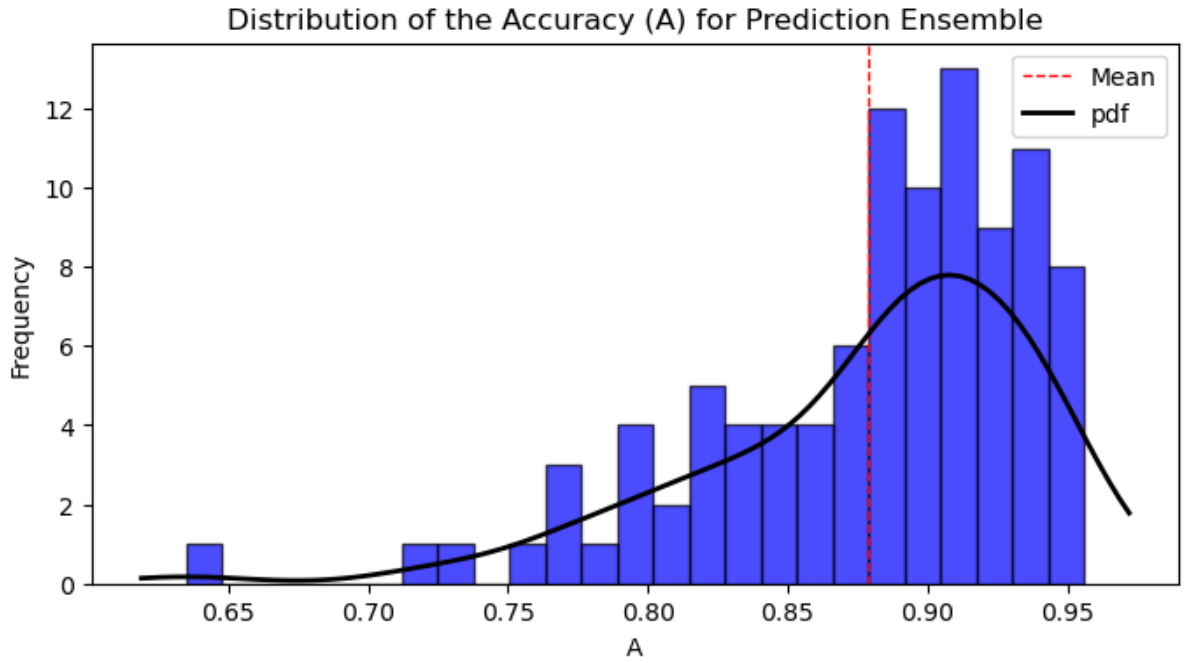
Event 158(Type A)





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