Response to Reviewer #2

We thank the reviewer for their detailed and insightful comments. We appreciate the recognition of the practical relevance and computational efficiency of the proposed framework. Below, we address each comment and describe how the manuscript has been revised accordingly.

We indicate the reviewer's comments in **bold** and the text modifications in blue.

General Comments

1) There are some distribution shifts between training and testing datasets that may undermine the model generalizability.

According to Table 1 and Figure 6, the training flood volumes are 220.1±131.4×10^6 m^3, while test location ND111 produces volumes of 395.71-974.55×10^6 m^3, far exceeding the training range. The MAE at ND111 reaches 247.81×10^{-2} m, nearly ten times higher than the training error.

(Also, Table 1 has two rows labeled "test"; please clarify how these correspond to test cases/locations.)

Please discuss the model's applicable range and extrapolation limits.

If feasible, include higher-volume training samples or explore domain adaptation for out-of-distribution conditions.

We thank the reviewer for highlighting this important issue.

Indeed, the testing volume ranges are far greater than the training ones. The test dataset was dependent on the availability of official pre-computed simulations from the VNK project. We leveraged this high variability to assess the model's response to extreme events, showing that, as expected, the model tends to underestimate them. However, the model still understands the spatio-temporal evolution for these floods, as indicated by the high CSI values.

Regarding the duplicate "test" rows, one provides the aggregate summary throughout all the test samples, while the second "row" provides a more detailed analysis of each scenario. We updated the table caption to better reflect this distinction as:

"Training, validation, and testing metrics for the mSWE-GNN on dike ring 41, reporting mean and standard deviation for the mean absolute error (MAE) for water depth and unit discharge and critical success index for a water depth threshold tau (CSI_τ) These metrics are reported

also for the three test locations HD073, ND234, and ND111, for the different return periods (RP). ..."

In terms of application range and extrapolation, we evaluated the model on a purposefully wide range of conditions and locations in Section 4.2, showing that the model performs best in a range that is similar to the training one, while showing progressively less reliable predictions outside of it. The reasons why a single model cannot capture well the system's response to all volume ranges are 1) that the training process converges to the mean, leading to worse performance towards the extremes and 2) the flow propagation speeds across different boundary conditions and time steps can be too different for a single model to capture them effectively. In lines 436-437, we propose a possible solution to this problem, involving the use of mixture-of-experts models:

"Moreover, future works could explore the use of mixture-of-experts models to improve the model performance across a wider range of boundary conditions, for example by combining or selecting the output of different models, each trained with a smaller range of conditions."

Regarding the reviewer's request to include an event with a higher volume, this would not necessarily improve the overall performance, for the reason explained above that the model will converge to the mean. One solution might be to train on a much larger dataset and use a bigger model to cover a broader range of interactions. With the given dataset, when focusing on a smaller range of conditions, preliminary experiments showed that the model performance increases for the given range, whether these events are very small or very large. To prove this, we carried out a small fine-tuning experiment, where we re-trained the best model with the single largest training flood event, using a smaller learning rate (0.0005) and few epochs (5) to avoid overfitting.

The results, reported only here in Table R2, show that the overall testing performance improves for the largest flood events, reducing the water depth MAE from 247.81 to 170.22, from 152.26 to 87.56, and from 55.51 to 39.34 for test location ND111. However, it also equivalently drastically reduces the performance of smaller events (locations ND234 and HD073). For this reason, we decided not to include further larger events and instead focus on a broader training range.

Table R2. Testing metrics for the mSWE-GNN fine-tuned on a single large flood event, reporting mean and standard deviation for the mean absolute error (MAE) for water depth and unit discharge and critical success index for a water depth threshold tau (CSI_τ). These metrics are reported also for the three test locations HD073, ND234, and ND111, for the

different return periods (RP).

ID	RP	Total volume	MAE ↓		CSI_τ [%] ↑	
	[yrs]	[10^6 m^3]	h [10^{-2} m]	$\|q\|$ [10^{-2} m^2 s^{-1}]	τ=0.05 m	τ=0.3 m
-	-	323.4 ± 349.2	98.36 ± 37.95	4.30 ± 0.78	63.49 ± 22.73	61.16 ± 22.73
HD073	10000	89.70	96.29	3.62	62.23	57.4
	1000	48.86	99.45	3.61	48.24	45.3
	100	15.99	90.46	3.22	36.15	36.29
ND234	10000	299.45	62.37	4.42	70.73	73.19
	1000	170.48	104.32	4.08	57.93	58.1
	100	62.21	135.21	4.32	42.52	39.55
ND111	10000	974.55	170.22	5.67	85.66	78.27
	1000	734.98	87.56	5.2	89.5	83.27
	100	395.71	39.34	4.51	78.42	79.07

---

2) The theoretical foundation of the ARME metric requires further clarification.

In Equation (5), when the predicted volume may be negative in early simulation stages due to subtracting V_0 in Equation (4), or when V_t approaches zero, ARME may produce numerically unstable or meaningless results.

Please discuss ARME's behavior during initial simulation phases and whether such issues were handled in the results, as they may bias plausibility assessment.

We appreciate the reviewer's examination of the ARME formulation.

In principle, ARME could become numerically unstable when the total volume V_t is close to zero. However, the ARME is computed starting from time step $t = 1$(Eq. 5) and the simulations start only once boundary conditions are non-zero, as they would be trivial otherwise. As a result, there is always positive inflow entering the domain. Furthermore, V_t is a strictly positive variable that does not approach zero unless the domain completely empties out. Moreover, while predicted volumes could in theory be lower than the initial volume V_0, this does not cause numerical issues because (i) the absolute value in Eq. (5) prevents sign-related instabilities and (ii) water depths are physically constrained to be non-negative, preventing unphysical negative volumes.

We noticed that this relevant physical component (already present in the original mSWE-GNN model) was not reported in the manuscript. As such, we updated Eq. 1 to be $\widehat{U}^{t+1} = \sigma(U^t + \Phi(X_s, U^{t-p:t}, \varepsilon))$ and changed lines 112-114 to:

"where $\widehat{U}^{t+1}$ is the predicted hydraulic variables, Ut are the hydraulic variables (water depth [m] and unit discharge [$m^2 s^{-1}$]) at time t, Φ (•) is the model for a fixed time step, X_s are static node features, $U^{t-p:t}$ are dynamic node features for time steps t–p to t, σ is a rectified linear unit (ReLU) used for guaranteeing positive hydraulic variables, and E are edge features."

As regards the ARME behavior during initial simulation phases, we did not encounter any issues. While some instabilities might have originated from a low total volume V_t, i) the model's predictions were generally the best right after the beginning of the flood event and ii) a high instability would entail a bad prediction and hence should result in a high ARME.

---

3) The finding that approximately 50% of simulations are classified as implausible (ARME>0.4) raises concerns about operational applicability.

How can users determine prediction reliability when encountering new boundary conditions in practical applications?

Does this high discard rate introduce bias in probability distribution estimates?

The authors should analyze whether there are systematic differences in spatial distribution or boundary conditions between discarded and retained simulations, and discuss strategies for improving the plausibility rate.

We thank the reviewer for this important point of discussion.

When encountering new boundary conditions, we can always determine prediction reliability thanks to the ARME metric. If this value is too high, we can just discard the corresponding simulations and run more, with small differences in boundary conditions. This compromises only speed, as we have to execute the model more times, but not reliability, as we can always calculate the ARME for new boundary conditions.

As for the high discard rate on the probability distribution estimates, selected simulations tend to be more central in terms of volume ranges, so the probability distributions will also be less spread out than considering the full set of simulations.

We modified lines 365-366 to clarify this: "Selecting only plausible results skews the probability distributions to be less spread out with respect to the complete set, giving a clearer distribution estimate."

To analyse the differences in spatial distributions and boundary conditions, we carried out the experiments in Section 4.2 which on purpose have much wider distributions than the training one to provide a more complete overview of the model's response. This is also the

main reason why 50% of the simulations are not reliable. As shown in Figures 9 and 10, most unplausible simulations have a total flood volume much smaller or higher than that of the training range. Similarly, some locations (Figure 12b) have a much lower percentage of plausible simulations which negatively influences the overall performance, as explained in lines 353-358. We clarified these points in the manuscript in lines 315-316: "We designed this range to be purposefully much wider than for a practical scenario, to assess a more complete response of the model to different boundary conditions and locations."

Minor Comments

1) What criteria guided the selection of dike-breach locations, and are they representative and physically justified?

For the test dataset, all three locations are representative and physically justified as they are taken from selected homogeneous dike segments that are potentially prone to breach, based on a geotechnical evaluation. All other locations in training, validation, and large-scale testing are randomly selected by considering approximately equidistant points along the part of the dike that is in contact with a river. Because of the presence of the river, we assume all locations to be potential breach points, and thus also representative and physically justified, even if in practice they might have a low probability of failure. The purpose of selecting equidistant locations is to evaluate the general functioning of the model. We clarified this in lines 234-236 as:

"For training and validation data, we used, respectively, 30 and 10 numerical simulations performed with Delft3D (Deltares, 2025), each with a different breach location, selected to be approximately equidistant along the dike ring boundary, and a different dike outflow hydrograph over time as boundary conditions."

---

2) Figure 6(b) caption reads "Training and testing discharge hydrographs…". Given validation in Figure 6(a), consider "Training and validation discharge hydrographs…" (or include testing if applicable).

Thank you for noting this inconsistency. We change the caption as:

"… (b) Training, validation, and testing discharge hydrographs used as boundary conditions for the simulations."

---

3) Pearson's r (Line 16, Figure 11, Table 2) is outlier-sensitive and assumes linearity. Given Figure 11's scatter, please justify Pearson's r or report Spearman's rank as a complementary robust measure.

We agree with the reviewer that Pearson's correlation can be sensitive to outliers and assumes linear dependence. In the experiments, all data pairs for which a clear correlation exists (i.e., between CSI and ARME) tend to follow a linear relationship, as also suggested by similar values of both coefficients. In the revised manuscript, we now report Spearman's ρ both in line 16 and Figure 11. In Table 2, we kept both metrics to present a more complete picture of the correlations.

---

4) Lines 30–32 contain many references for one statement on dike-breach uncertainties; retain the most representative and recent.

We have reduced the number of citations in this section by retaining at most up to two references per uncertainty variable.

These lines now read as:

"Building probabilistic hazard maps remains challenging as the number of uncertain variables can be large, particularly for dike breaching, where additional geotechnical properties must be considered. Uncertainties include breach location (D'Oria and Maranzoni, 2019; Westerhof et al., 2023), breach width (Mazzoleni et al., 2014; de Moel et al., 2014), breach development time (Apel et al., 2006; Ferrari et al., 2020), failure time (D'Oria and Maranzoni, 2019), and failure mechanism (D'Oria and Maranzoni, 2019; Mazzoleni et al., 2014)."

---

5) Cite more recent references (ideally last 10 years) when discussing computation costs.

We updated the references in lines 34-36, where we talk about computational costs of standard approaches, as follows:

"Estimating output uncertainty may require up to hundreds of thousands of simulations, making standard numerical flood models computationally prohibitive, unless using large high-performance clusters (Gibbons et al, 2020)."

Reference:

Gibbons, S. J., Lorito, S., Macías, J., Løvholt, F., Selva, J., Volpe, M., Sánchez-Linares, C., Babeyko, A., Brizuela, B., Cirella, A., Castro, M. J., de la Asunción, M., Lanucara, P., Glimsdal, S., Lorenzino, M. C., Nazaria, M., Pizzimenti, L., Romano, F., Scala, A., Tonini, R., Manuel González Vida, J., and Vöge, M.: Probabilistic Tsunami Hazard Analysis: High Performance Computing for Massive Scale Inundation Simulations, Frontiers in Earth Science, Volume 8 - 2020, https://doi.org/10.3389/feart.2020.591549, 2020.510

---

6) The 8-hour output resolution in Section 3.1 may be coarse for rapid flood-front dynamics. Briefly discuss how this choice affects ARME and CSI.

We agree with the reviewer on this observation. We selected this output temporal resolution to match that of the official VNK simulations.

In terms of effect on ARME, this choice does not affect its performance as this metric only considers the total volumes at each time step, which are independent of their spatial distribution.

Regarding the CSI, increasing the temporal resolution might improve the performance. This is because most errors occur at the front of the flood wave. A higher temporal resolution would correspond to fewer spatio-temporal variations, making it easier for the model to learn the correct flood spreading.

---

7) Line 344: fix “outlier Contrarily” to “Contrarily” (or rephrase).

We thank the reviewer for spotting this typo. The phrase “outlier Contrarily” has been corrected to “Contrarily”.

---

8) The computational efficiency claim of "10,000 times faster" requires more detail, including the specific Delft3D configuration (number of CPU cores, parallelization), whether data I/O time is included.

We have expanded the description of the computational configuration for the numerical simulations to include more configuration details.

Lines 246-247 now read as:

“All numerical simulations are run on an AMD Ryzen 7 5700X 8-Core Processor (3.40 GHz) CPU, using four OpenMP threads.”

In terms of speed-up calculations, as mentioned in line 288, “Both times exclude mesh creation, data pre-processing and post-processing.”