Dear Reviewer,

We sincerely thank you for your careful reading of our manuscript and for your thoughtful and constructive comments. We are grateful for your positive assessment of the core idea and for recognizing the potential value of the proposed framework.

Your comments helped us identify several important areas where the manuscript needed clarification and improvement, including the data split and generalization strategy, the distinction between rapid scenario simulation and real-time forecasting, the interpretation of physical consistency, the evaluation metrics, and the reproducibility of the model implementation. We have carefully considered each comment and revised the manuscript accordingly.

Below, we provide a point-by-point response.

Comment 1: Data Split Clarity and Generalization Beyond Random Perturbations

Response:
We thank the reviewer for pointing out this ambiguity. We agree that the original manuscript did not clearly describe the data split and used inconsistent wording across Table 1, Section 3.3, and the Results section. We clarify that, for each case study, **150 realizations were generated: 100 were used for training, and 50 were kept completely untouched as held-out test scenarios, unseen during training, for final evaluation**. The results reported in Tables 2–4 were computed only over these **50 held-out test scenarios**, not over training events.

In the revised manuscript, we will explicitly state this in Table 1, Section 3.3, and the Results section. We will also revise phrases such as “training and validation,” “validation and testing,” and “across all events” to consistently indicate **“100 training events and 50 held-out test events”** and **“across all held-out test events.”**

Regarding the reviewer’s concern about random splitting and out-of-distribution generalization, we clarify that the test cases were **not used during training** and were generated independently from the training cases using the controlled hydrograph-perturbation procedure described in Appendix D. Although these scenarios are derived from a reference hydrograph, they are **not expected to produce only small or linearly interpolated hydraulic responses**. Because the shallow-water equations are **highly nonlinear**, variations in hydrograph magnitude, timing, and limb shape can lead to substantially different flood responses through changes in **wave propagation, floodplain storage activation, wetting and drying, and terrain-controlled connectivity**.

Therefore, the held-out test set evaluates generalization to **unseen and hydraulically distinct forcing realizations within the designed scenario space**. We agree that independent historical events or deliberately withheld forcing ranges would provide a **stronger extrapolation test**. In the revised manuscript, we will clarify this distinction, avoid

overstating generalization beyond the sampled forcing envelope, and identify independent-event and withheld-range testing as important future work.

Comment 2: Temporal causality and forecast-conditioned simulation

Response:
We clarify that the current framework is **conditioned on the prescribed spatiotemporal forcing over the simulation horizon**. This is consistent with standard hydrodynamic modeling practice, where SWE-based solvers such as HEC-RAS, ANUGA, LISFLOOD-FP, or Delft3D also require boundary inflow hydrographs or forecasted forcings over the simulation/forecast period.

We agree that the FNO temporal convolutions are **not causal in the autoregressive sense**. The model does not independently forecast future inflows from past observations. Instead, it rapidly predicts the hydraulic response of the domain given an externally provided forcing trajectory, which may come from observed hydrographs, design scenarios, hydrologic forecasts, meteorological forecasts, or ensemble boundary conditions.

In the revised manuscript, we will explicitly state this distinction. We will describe the framework as a **rapid hydrodynamic emulator / forecast-conditioned scenario simulator**, rather than as a standalone real-time forecasting model. We will also revise wording related to “forecasting” and “flood warning” to clarify that the method can support such workflows only when future forcing is supplied by an external forecasting system.

Comment 3: Optional global rescaling and mesh-to-grid averaging in Appendix C:

Response:
Appendix C was included by mistake and does not describe a procedure used in the reported experiments. The global rescaling step was **not applied** to the final Magnifier outputs. In the revised manuscript, we will remove Appendix C and revise the methods to describe only the actual preprocessing and interpolation steps used in this study.

Comment 4: Peak-period definition and signed mean errors in Table 4

Response:
Table 4 should indeed more clearly distinguish between **signed bias** and **the magnitude of absolute error**.

In the current formulation, negative values are not errors in the calculation; they indicate direction. A negative peak-value error means that the model underestimates the reference peak-period volume, while a positive value indicates overestimation. Similarly, a negative peak-timing error means the predicted peak occurs earlier than the reference peak, while a positive value indicates a later peak.

Therefore, we will retain the signed metrics because they provide useful information about systematic bias. However, we agree that signed mean errors alone may mask event-level variability due to cancellation between positive and negative errors. In the revised manuscript, we will add absolute peak magnitude and timing errors, together with event-level summary statistics, to better represent the actual error magnitude.

Comment 5: Need for additional physical diagnostics beyond depth and volume

Response:
We agree that velocity fields, cross-section discharge, and local continuity residuals would provide stronger independent diagnostics of hydrodynamic consistency.

In the present study, the surrogate model is formulated as a **water-depth/inundation regressor**, and the current training targets do not include velocity fields or discharge. Therefore, we cannot directly evaluate velocity-based or momentum-related quantities from the present model outputs. We agree that the close agreement in total water volume should be interpreted only as a supporting diagnostic, not as complete evidence of preservation of shallow-water physics.

In the revised manuscript, we will moderate claims related to hydrodynamic consistency and clarify that the current validation focuses on water depth, inundation extent, peak volume, and timing. We will also note that, if future datasets include velocity components, discharge, or flux information from the hydrodynamic solver, these variables could be added as additional output channels or diagnostic targets, enabling direct evaluation of cross-section discharge, local continuity residuals, and broader SWE consistency.

Comment 6: Validation against ANUGA versus observed flood data

Response:
Good point. The primary evaluation presented in this study is a **surrogate-to-ANUGA comparison**, and therefore demonstrates the ability of the proposed deep-learning model to reproduce ANUGA-generated flood-depth and inundation fields under the tested configurations.

We clarify that the ANUGA base simulation was checked against available observational inundation information. However, the main objective of this manuscript is not to present a full hydrodynamic calibration study, but to evaluate the proposed deep-learning surrogate framework. Therefore, the reported metrics focus on how accurately the surrogate reproduces ANUGA-generated depth and inundation fields.

In the revised manuscript, we will make this distinction clearer and moderate any wording that may imply direct validation of the surrogate against real flood observations. We will state that observation-based hydrodynamic validation supports the reference simulation setup, while the core contribution and evaluation of this paper focus on the deep-learning surrogate.

Comment 7: Additional depth-error metrics and threshold sensitivity

Response:
Agree that pooled **NSE** and **relative RMSE** alone may not fully capture model performance, especially in shallow-water and wetting/drying regions.

We note that the current **Table 2 already separates performance between shallow** ($h<0.5$m) **and deeper** ($h\geq0.5$m) **flow regimes**. However, we agree that additional distribution-based diagnostics would make the evaluation clearer.

In the revised manuscript, we will retain **Table 2** for exact summary values and add **two complementary violin plots** to show the distribution of key depth-performance metrics across the held-out test cases. These plots will better show the spread and variability of model performance beyond single pooled values.

We will also add a **CSI-versus-threshold curve** for binary inundation performance using wet/dry thresholds of **0.01, 0.05, 0.10, and 0.30 m**. This will show whether the inundation skill remains robust across different flood-depth thresholds.

Comment 8: Missing implementation and reproducibility details

Response:
Indeed the manuscript should provide more complete implementation details for reproducibility.

In the revised manuscript, we will add a dedicated **implementation and reproducibility appendix** that reports the missing training, architecture, preprocessing, inference, and hyperparameter settings for the FNO, Magnifier, and joint fine-tuning stages. This will make the workflow more transparent and easier to reproduce.

Minor Comments

Response:
We thank the reviewer for these helpful minor comments.

1. We will verify the mesh-node and triangle numbers reported in **Table 1** and correct any duplicated or erroneous values.
2. We will revise **Figure 12** to include the evaluated timestamp, the inundation threshold used for the binary comparison, and the interpolation baseline.
3. We agree that the inverted **CSI** axis in **Figure 11** makes interpretation less intuitive. We will redesign the figure using a standard axis orientation.
4. We agree that H(t) is more accurately described as a **domain-integrated water-volume time series**, rather than a hydrograph. We will revise the wording accordingly where appropriate.

We sincerely thank the reviewer again for the careful and constructive evaluation of our manuscript. Your comments were very helpful in improving the clarity, transparency, technical completeness, and reproducibility of the paper. We believe that the revisions made in response to your suggestions have strengthened the manuscript and improved the overall rigor and quality of the work.