Response to Reviewers

Deep learning of flood forecasting by considering interpretability and physical constraints

Ting Zhang *, Ran Zhang, Jianzhu Li, Ping Feng

State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin 300072, China

Corresponding author: Ting Zhang (zhangting_hydro@tju.edu.cn)

Dear editor,

Thank you for coordinating the review process and conveying the constructive feedback. We sincerely appreciate the reviewers' recognition of our manuscript improvements. We have provided detailed, point-by-point responses to the reviewers' comments in the following pages. Note that the reviewers' comments are presented in *italics*, and our responses are in Times New Roman and blue font. In addition, all the line numbers in the responses refer to the revised manuscript. All changes made to the manuscript are marked in red font. Please do not hesitate to contact us if you have any questions or require any additional information. Thank you for your consideration.

Sincerely,

Ting Zhang

To the **Reviewer #1's** comments, we make the following responses and changes in the manuscript:

1. The authors have addressed all the points raised during the first round of review in a very satisfactory manner, significantly improving the quality and clarity of the manuscript. However, I believe that one important aspect has not yet been fully resolved: the issue concerning the division of the dataset into training and test sets.

Specifically, the selection of events in the two subsets appears to have been carried out according to a systematic but not entirely objective criterion. No analysis has been conducted to evaluate how sensitive the results are to the specific assignment of events to the two groups. The absence of techniques such as cross-validation or bootstrapping makes it difficult to assess the robustness and stability of the model with respect to different dataset partitions.

While I understand the constraints due to limited data availability, I suggest that this limitation be at least acknowledged in the discussion or conclusions. It should be clarified that the presented results refer to a specific configuration of the training and test sets, and that no investigation was carried out on the influence of different event assignments.

Such a clarification would strengthen the scientific approach of the study and provide greater methodological transparency.

Response: Thank you for your insightful comment. We have supplemented the division method between the training set and the validation set in the available dataset. The division of the two datasets took into account the main characteristics of floods, including temporal occurrence, peak discharge, and flood duration, to ensure that both datasets comprehensively cover diverse flood

characteristics as much as possible.

In the revised manuscript, Page 14, Line 327-332:

The partitioning of training and validation sets was designed to ensure balanced representation of flood characteristics across both datasets, specifically considering temporal occurrence, peak discharge, and flood duration. This stratification achieves comprehensive inclusion of major, moderate, and minor flood magnitudes while encompassing diverse hydrograph types—including both single-peak and multi-peak events—to maintain hydrological process representativeness.

As for the limitations that may arise from insufficient data, we have supplemented this part in the **Discussion** section and clarified that the results provided are based on the specific flood event classification of the training and validation sets, and we have not studied the impact of different dataset classifications on the results.

In the revised manuscript, Page 30-31, Line 649-660:

Furthermore, the dataset was partitioned solely into training and validation sets primarily due to limitations in available historical flood events—only 30 events were utilized, most with relatively short durations. This resulted in a limited sample size and insufficient additional floods for model testing; future data acquisitions will be incorporated to enhance robustness. To maximize coverage of flood diversity and capture spatiotemporal heterogeneity, we partitioned data based on temporal occurrence, peak discharge, and flood duration. This methodology follows established precedents (e.g., Lv et al., 2020; Read et al., 2019; Xie et al., 2021; Jiang et al., 2020) where dual-set partitioning is widely adopted beyond flood forecasting applications. Crucially, our results are contingent upon the specific flood event partitioning of training and validation sets detailed in

Table 1, with no investigation of alternative partitioning impacts. Future research could employ cross-validation or bootstrapping to evaluate model robustness and stability across different dataset divisions.