

Authors' Responses to Comments from Reviewer #2:

General Comments

This manuscript presents a differentiable hybrid hydrological modeling framework for short-term flood forecasting by integrating process-based hydrological representations with data-driven learning components. The authors introduce a physics-informed neural architecture search (PINAS) routing scheme and combine it with differentiable runoff generation, baseflow recession, channel routing, and future meteorological information derived from radar echo extrapolation. The topic is highly relevant to improving the accuracy, physical consistency, and interpretability of short-term flood forecasting, particularly for small and medium-sized basins. However, several important issues need to be addressed before the manuscript can be considered for publication.

Response : We sincerely thank the reviewer for the careful and constructive evaluation of our manuscript. We greatly appreciate the reviewer's recognition of the relevance of combining differentiable hydrological processes, PINAS routing, and radar-derived future meteorological information for short-term flood forecasting. We agree that the manuscript would benefit from a clearer mathematical description of the differentiable hydrological components, stronger evidence regarding the physical role of the coupled components, and more cautious interpretation of the data-dependency and SHAP analyses. In the revised manuscript, we will address these issues through focused internal-process diagnostics, limited component ablation, methodological clarification, re-analysis of existing results, and improvements to the presentation of figures, tables, and mathematical notation.

Major Comments

1. Although the authors claim that the proposed framework combines mechanistic hydrological structures with learnable components, the physical consistency of the fully coupled model and the contributions of individual components are not sufficiently demonstrated. Since the final forecasts are jointly influenced by the PINAS routing scheme, radar-derived future forcing, and data-driven residual correction, it remains

unclear whether the learned corrections maintain the physical consistency of the original hydrological processes. Further analysis of the model behavior and internal processes is needed to better support the physical interpretation of the proposed framework.

Response 1: We sincerely thank the reviewer for this important comment. We fully agree that the physical consistency of the complete coupled model should not be inferred solely from the presence of physically motivated individual modules. We will address this issue from two complementary aspects. First, we will add process-level diagnostics of the PINAS routing component using the existing trained models. These diagnostics will include learned effective travel time, routing-kernel shape and dispersion, peak attenuation, reach-scale mass-balance residuals, and the stability of these quantities for flood events excluded from calibration and architecture selection. These analyses will allow the internal routing behavior to be evaluated directly rather than inferring physical plausibility from outlet NSE or RMSE alone. Second, we will quantify the relationship between the physical-prior discharge Q_{phys} and the final corrected forecast. The magnitude of the correction will be examined at different forecast lead times and hydrograph stages, including the rising limb, flood peak, and recession period. A focused ablation comparison will also be added using three configurations: the physical hydrological core without the final forecasting correction, the complete hybrid model, and the forecasting head without Q_{phys} . This minimum set of comparisons will allow us to determine whether the final forecasting skill reflects meaningful interaction between the physical and data-driven components rather than the residual component effectively replacing the hydrological prior. The corresponding interpretation in the manuscript will also be revised so that structural constraints and empirically evaluated hydrological process realism are clearly distinguished.

2. Section 2.2.2: The manuscript repeatedly emphasizes that the hydrological modules are reformulated into differentiable forms and trained in an end-to-end manner. However, the specific mathematical formulations and model architecture of these

differentiable processes remain unclear. The authors should provide the detailed equations and clarify how the conventional hydrological processes are modified to enable differentiable optimization, which is essential for understanding the physical basis and reproducibility of the proposed framework.

Response 2: We agree with the reviewer that the current description of the differentiable hydrological modules is insufficient for fully understanding and reproducing the proposed framework. In the revised Section 2.2.2, we will provide the mathematical formulations of the principal hydrological components, including runoff generation, baseflow recession, sub-basin response, channel routing, and their coupling within the end-to-end forecasting framework. For each component, we will clarify the learnable parameters, parameter initialization, physically admissible parameter ranges, and the differentiable transformations used to maintain these constraints during gradient-based optimization. The corresponding state-update equations and information flow among the hydrological modules will also be described more explicitly. For the runoff-generation component, the revised description will include the complete SCS-CN formulation, parameter ranges, initialization, antecedent-moisture treatment, and its recursive implementation at the 30-min model time step. For the PINAS routing component, the manuscript will be revised so that the description of the architecture/search procedure is fully consistent with the actual implementation, including the searchable parameters, candidate settings, validation criterion, physical regularization terms, and final selected configuration.

3. The authors investigate the influence of training data size on model performance by gradually reducing the training samples from 100% to 25%. However, this analysis only evaluates the sensitivity of the proposed PINAS-based framework and lacks comparisons with benchmark models. It remains unclear whether the observed performance variation with training data size is a specific advantage of the proposed hybrid framework, or a general characteristic shared by other hydrological forecasting approaches. In addition, the manuscript does not provide sufficient details regarding

the division of training, validation, and testing datasets, including the number of flood events and the corresponding temporal periods used for each subset.

Response 3: We appreciate this comment and agree that the current training-data-size experiment should not be interpreted as evidence of a model-specific sample-efficiency advantage. In the revised manuscript, the 25%, 50%, 75%, and 100% experiment will be described as a training-data sensitivity analysis rather than as a demonstration that PINAS intrinsically requires less training data than other models. We will clearly document how the reduced training subsets were constructed and will explicitly discuss the potential influence of differences in flood-event composition. In addition, we will substantially improve the description of the dataset partitioning. A table will be added to summarize, for each basin, the number and temporal periods of flood events used for training, validation, and independent testing, together with radar-data availability where relevant. We will also clarify that dataset division is performed at the flood-event level so that overlapping 30-min sequences from the same flood event are not distributed across different subsets.

4. The interpretability analysis based on SHAP requires further consideration. The proposed framework contains multiple coupled components. Due to the heterogeneous inputs and the complexity of the model structure, SHAP values calculated from the final discharge prediction may mainly reflect the contribution of different variables to the residual correction component rather than their actual roles in the underlying hydrological processes. Some SHAP-based interpretations thus may not fully correspond to hydrological understanding. For example, the statement that “in the Jumahe Basin, the contribution of areal rainfall gradually increased while radar echo contribution decreased with increasing lead time” requires further justification.

Response 4: We sincerely thank the reviewer for this important clarification. We agree that SHAP values derived from the final discharge prediction should not be interpreted directly as contributions to the underlying hydrological processes. We will revise the role of SHAP in the manuscript and define it explicitly as an analysis of the predictive

attribution of the final forecasting model, rather than as a direct attribution of hydrological-process causality. The revised Methods section will provide additional information on the SHAP implementation, including the algorithm used, the background/reference data, and the feature-processing procedure. We will conduct grouped permutation analysis for major hydrometeorological input categories, particularly rainfall, currently observed radar information, and extrapolated future-radar information. The resulting performance changes will be used to assess whether the SHAP-based importance patterns are also reflected in predictive sensitivity. The statement regarding the increase in areal-rainfall attribution and decrease in radar attribution with increasing lead time in the Jumahe Basin will therefore be re-examined using these additional diagnostics. The terminology used throughout the Results and Discussion will also be revised to avoid equating SHAP attribution with hydrological causation.

Specific Comments

1. Line 116-117: The statement that “the practical application value of radar in hydrological forecasting is still unclear” appears to be overstated.

Response 1: We agree with the reviewer that the original wording is too broad and does not adequately acknowledge the established value of weather radar in hydrological forecasting. We will revise this statement so that the scientific question is framed more specifically. Rather than questioning the general practical value of radar, the revised manuscript will focus on the incremental value of short-term radar extrapolation beyond currently observed hydrometeorological information, particularly across different basin scales and forecast lead times.

2. Line 121-122: “A hybrid multiscale flood forecasting framework is proposed”, what is the multiscale?

Response 2: Thank you for pointing out this ambiguity. We agree that the term “multiscale” is not sufficiently defined in the current manuscript. We will re-examine its use throughout the manuscript. If the revised framework does not contain a clearly defined and explicitly demonstrated spatial or temporal multiscale formulation, we will remove the term “multiscale” and replace it with a more precise description, such as “differentiable hybrid flood forecasting framework.”

3. it would be better to utilize the table to introduce three basins.

Response 3: We agree with the reviewer. In the revised Section 2.1, we will add a summary table presenting the principal characteristics of the Shahe, Jumahe, and Liulin basins. The table will summarize the available basin characteristics, hydrometeorological observation periods, relevant monitoring information, selected flood events, and radar-data availability. The existing textual description will be shortened accordingly to avoid unnecessary repetition.

4. Fig. 1 could be further improved to provide a clearer overview of the spatial relationship among the three study basins and the radar location.

Response 4: We agree with the reviewer. Figure 1 will be revised to more clearly present the spatial relationship among the three study basins, their outlet locations, and the radar location and/or coverage. The map layout, labels, scale information, and visual hierarchy will also be improved so that the geographical relationship among the study basins and the radar observations can be understood more directly.

5. Fig. 2 It was unclear what is the precipitation forcing, radar or rain gauge?

Response 5: We thank the reviewer for identifying this ambiguity. Figure 2 will be revised to clearly distinguish the different precipitation-related information used in the framework, including measured rainfall information, currently observed radar information, and extrapolated future radar information. The corresponding Methods text will also be revised to ensure that the figure and model description are fully consistent.

6. The manuscript contains a large number of mathematical formulations in the main text, and some detailed equations that are not essential for understanding the overall framework could be moved to the Appendix or Supplementary Material to improve readability

Response 6: We agree with the reviewer. In the revised manuscript, we will reorganize the mathematical presentation to improve readability while maintaining reproducibility. Only equations necessary for understanding the principal hydrological processes, coupling relationships, and methodological contributions will remain in the main text. Detailed derivations, auxiliary regularization terms, and implementation-specific formulations will be moved to the Appendix or Supplementary Material.

7. Since future meteorological forcing derived from radar echo extrapolation is an important component of the proposed framework, the authors are encouraged to provide additional evaluation results of the radar forecasting performance in the Appendix/Supplementary Material to better demonstrate the reliability of the meteorological inputs used for flood forecasting.

Response 7: We agree that the reliability of radar-derived future meteorological information should be evaluated independently from the hydrological forecasting model. We will add a concise radar-nowcasting evaluation in the Appendix or Supplementary Material. Extrapolated radar information at different lead times will be compared with the corresponding subsequently observed radar fields using appropriate verification metrics supported by the available radar data.

8. Please standardize the typography of all mathematical notations in Table 5 and throughout the text to ensure consistency.

Response 8: We thank the reviewer for pointing this out. We will systematically check and standardize the mathematical notation throughout the manuscript, including Table 5. The formatting of scalar variables, vectors or matrices where applicable, subscripts,

superscripts, Greek symbols, mathematical operators, and units will be made consistent, and all symbols will be defined at their first occurrence.