

Authors' Responses to Comments from Reviewer #1:

General Comments

I have carefully evaluated the manuscript entitled “Differentiable Hybrid Hydrological Model for Short-Term Flood Forecasting with Future Meteorological Information.” The study addresses a scientifically relevant problem at the intersection of differentiable hydrological modelling, radar-based precipitation nowcasting, and operational flood forecasting. Its main strengths include the development of a modular hybrid architecture, the comparison of three differentiable routing formulations, the introduction of a physics-informed neural architecture search scheme, and testing across basins with markedly different drainage areas and response times. The combination of long-term rainfall-runoff records, recent radar observations, multi-lead forecasts, data-dependency experiments, and feature-attribution analyses gives the work potential significance for the emerging literature on physics-guided hydrological learning.

Despite these strengths, important weaknesses in the experimental design, description of the model, interpretation of physical consistency, and statistical evaluation currently limit the strength and reproducibility of the conclusions. I therefore recommend Major Revision.

Response: We sincerely thank the reviewer for the careful evaluation of our manuscript and for recognizing the scientific relevance and potential significance of the proposed differentiable hybrid hydrological forecasting framework. We greatly appreciate the constructive comments regarding the physical interpretation of PINAS, data processing and partitioning, model reproducibility, the role of future radar information, statistical evaluation, and model interpretability.

We agree that several aspects of the current manuscript require clearer methodological descriptions and more rigorous interpretation. In the revised manuscript, we will address these concerns through targeted methodological clarification, process-level diagnostics, ablation experiments, re-analysis of the existing results, and a more cautious discussion of the scope and limitations of the proposed framework.

Major Comments

1. Lines 75-107: The motivation for PINAS assumes that searchable routing architectures provide greater transferability and physical plausibility than manually specified operators, but the manuscript does not define what constitutes physical plausibility beyond non-negativity, approximate volume consistency, and smoothness. These constraints do not ensure realistic travel times, attenuation, causality at the river-network scale, or preservation of wave dynamics. The scientific question should therefore distinguish numerical constraints from hydrological process realism, and the evaluation should include learned travel times, reach-scale mass-balance residuals, kernel shapes, and stability outside the calibration events rather than inferring physical plausibility primarily from outlet hydrograph accuracy.

Response 1: We sincerely thank the reviewer for this important comment. We agree that non-negativity, kernel smoothness, and approximate volume consistency should be regarded as structural or numerical constraints rather than sufficient evidence of hydrological process realism. In the revised manuscript, we will explicitly distinguish structural or numerical constraints from hydrological process realism. Specifically, using the existing trained models, we will supplement the current outlet-based evaluation with reach-scale diagnostics of learned effective travel time, routing-kernel shape and dispersion, peak attenuation, and reach-scale mass-balance residuals. We will also examine the stability of these quantities across available flood events excluded from calibration and architecture selection. These analyses will be used to assess the internal routing behavior directly rather than inferring physical plausibility from outlet NSE or RMSE alone.

2. Lines 150-159: The event-selection and data-partitioning procedures are insufficiently documented and may introduce substantial selection and information-leakage biases. Selecting events only when they exhibit “distinct and complete” rising and recession phases can exclude poorly observed, compound, truncated, or operationally difficult floods, thereby yielding an optimistic assessment. The manuscript should report objective event thresholds, event separation rules, missing-

data treatment, the number of radar-supported events actually used in training, validation, and testing, and whether splitting was chronological or event-wise. The stated data period ends in 2023, yet a 2024 Jumahe event is evaluated later, which must also be reconciled. All samples from a single event must remain within one partition because random division of overlapping 30 min sequences would leak nearly identical hydrograph and radar states across subsets.

Response 2: We thank the reviewer for pointing out the need for a more transparent description of flood-event selection and dataset partitioning. In the revised manuscript, we will provide the objective criteria used to identify and separate flood events, the treatment of missing observations, and the number of events with available radar information. We will also add a table summarizing the training, validation, and testing events for each basin, including their temporal periods and radar availability. We will clearly state that the dataset partitioning is performed at the flood-event level, so that overlapping 30-min sequences generated from the same event are not divided among training, validation, and testing subsets. We will also recheck the complete data-processing workflow to ensure that this principle is consistently maintained. Finally, the inconsistency between the stated data period and the Jumahe event mentioned later in the manuscript will be checked and corrected.

3. Lines 153-155: Temporally interpolating rainfall-runoff records to 30 min resolution may create artificial timing information precisely where the study evaluates peak timing and short-lead predictability. The original resolutions of rainfall and discharge, interpolation methods, handling of accumulated rainfall, discharge-rating uncertainty, and temporal alignment between gauges, radar scans, and outlet flow are not provided. Linear interpolation of coarse discharge observations can smooth peaks, while interpolation of rainfall totals can alter intensity and antecedent wetness. The authors should document the native sampling frequencies and demonstrate that conclusions at 0.5-2 h lead times are not artifacts of temporal resampling.

Response 3: We sincerely thank the reviewer for raising this important issue. We agree that the temporal processing of rainfall and discharge should be described more clearly,

particularly because the study focuses on short forecast lead times and flood-peak timing. We would like to clarify that the original rainfall and discharge observations are available at temporal intervals finer than the 30-min model time step. Therefore, the 30-min dataset was not generated by disaggregating coarse hourly observations. In the preprocessing procedure, rainfall observations within each 30-min interval were accumulated to preserve precipitation totals, whereas the higher-frequency discharge observations were resampled onto regular 30-min timestamps using temporal interpolation between neighboring observations. We will revise the Methods section to report the original observation frequencies, the exact resampling procedures for rainfall and discharge, the treatment of missing records, and the temporal alignment among rainfall, radar, and discharge observations.

4. Lines 191-222: The claimed spatial and physical representation of runoff generation is difficult to reconcile with assigning generated runoff to subbasins according only to area fractions. This formulation appears unable to represent spatial differences in rainfall, soils, land cover, antecedent moisture, or runoff coefficients, even though spatial radar information is central to the study. The SCS-CN implementation also lacks equations, parameter ranges, initialization, antecedent-moisture treatment, and an explanation of how an event-based cumulative formulation is applied recursively at 30 min resolution. Without these details, neither runoff volumes nor the attributed routing improvements can be independently reproduced, and errors assigned to routing may actually originate from an overly homogeneous runoff-generation module. The consequences of dynamic land-surface conditions for runoff attribution should also be discussed in relation to 10.5802/crgeos.179.

Response 4: We agree that the current description of the runoff-generation component is insufficient and may give the impression that a higher degree of spatial heterogeneity is represented than is actually included in the present model. In the revised manuscript, we will provide the complete SCS-CN formulation, parameter ranges, initialization, antecedent-moisture treatment, and the recursive implementation used at the 30-min time step. We will also clarify explicitly how generated runoff is distributed among

sub-basins. Importantly, we will revise the interpretation of the runoff-generation component to acknowledge that the current implementation represents spatial heterogeneity in a simplified manner and does not explicitly resolve all spatial differences in soil properties, land cover, or dynamically varying runoff coefficients. The implications of this simplification will be discussed in relation to the literature suggested by the reviewer. Because all three routing formulations are evaluated using the same runoff-generation module, their relative routing comparison is made under an identical runoff-generation assumption. Nevertheless, we agree that uncertainty in runoff generation can affect the absolute interpretation of routing performance. We will therefore avoid attributing all model differences exclusively to routing and will state this limitation explicitly.

5. Lines 250-281: The PINAS procedure is not described at the level required to reproduce or assess the proposed methodological contribution. The manuscript should specify the number of layers, candidate kernel sizes and dilation factors, architecture parameters, relaxation or sampling method, search and retraining stages, validation criterion, complexity penalty, physical regularization formulas and weights, stopping rules, random seeds, and the final architecture selected for each basin or reach. It is also unclear whether architecture selection used any events subsequently included in testing. Because architecture search increases model-selection flexibility, comparing its best searched configuration with single fixed Muskingum and CNN-Unit configurations may confer an unfair advantage. Parameter counts, search costs, training times, repeated runs, and equivalently tuned baselines are needed to separate architectural benefit from greater optimization effort.

Response 5: We sincerely thank the reviewer for this comment. We agree that the current description of the PINAS search procedure is not sufficiently detailed for reproducibility. In the revised manuscript, we will ensure that the methodological description and terminology are fully consistent with the actual implementation. We will report the searchable architecture/hyperparameter space, candidate kernel settings, search or sampling strategy, number of search trials, search-stage training settings,

validation criterion, complexity penalty, physical regularization terms and weights, stopping criteria, random seed, and the final selected configuration for each basin. We will also explicitly state which data are used for architecture selection and confirm that independent testing events are not involved in this process. To improve the fairness and transparency of the model comparison, we will report the parameter counts and approximate search/training costs of the different routing formulations and clarify the tuning procedure used for the fixed baselines. The comparison will therefore distinguish the benefit of the selected routing structure from differences arising solely from optimization effort.

6. Lines 282-325: The residual forecasting head can potentially overwrite the hydrological prior, so the final prediction is not necessarily constrained by runoff generation, routing, or water balance despite the physical interpretation assigned to it. Non-negative output alone does not preserve mass, timing, recession dynamics, or the contribution of the physical core. The authors should quantify the magnitude of the residual relative to Q_{phys} evaluate physical-prior forecasts before correction, and report whether corrections systematically dominate during peaks or specific lead times. Ablations using the physical core alone, the forecasting head without Q_{phys} , and alternative correction constraints are essential to demonstrate that the final skill arises from meaningful hybridization rather than from an LSTM learning discharge while treating the physical simulation as an additional predictor.

Response 6: We agree that this issue is important for demonstrating that the final forecasting skill results from meaningful hybridization rather than from a data-driven forecasting head that effectively overrides the physical simulation. In the revised manuscript, we will first quantify the magnitude of the total correction relative to the physical-prior discharge (Q_{phys}). The correction will be examined for different forecast lead times and hydrograph stages.

We will also add a focused ablation comparison consisting of three configurations:

- (1) the physical hydrological core without the final forecasting correction;
- (2) the complete hybrid model;

(3) the forecasting head without Q_{phys} .

This set of comparisons will allow us to evaluate whether the physical prior provides meaningful information to the final forecast and whether the data-driven correction systematically dominates the prediction. Depending on the resulting correction-magnitude diagnostics, we will further assess whether an explicit constraint on the correction magnitude is necessary. We will also revise the manuscript to avoid implying that non-negative final discharge alone guarantees physical consistency.

7. Lines 333-368: Radar reflectivity is advected directly in dBZ space and supplied as future meteorological information, but the manuscript does not establish the hydrometeorological quality or independent value of these extrapolations. Details are needed on radar calibration, clutter and attenuation correction, beam blockage, scan interval, spatial resolution, gauge adjustment, missing echoes, PyrLK settings, and extrapolation verification. More importantly, no benchmark isolates the contribution of future information. Comparisons should include persistence of the latest echo, models without extrapolated radar, models with observed radar only, and, where possible, an upper-bound experiment using subsequently observed rainfall or radar. Otherwise, performance differences cannot be attributed specifically to meteorological foresight rather than to additional correlated predictors.

Response 7: We thank the reviewer for this important comment. We agree that the reliability of radar extrapolation and its incremental contribution to hydrological forecasting should be evaluated separately. In the revised Methods section, we will provide a more complete description of the radar data, including the available spatial and temporal resolution, missing-data handling, available quality-control procedures, and the actual echo-extrapolation algorithm and settings used in the implementation.

we will compare three configurations:

- (1) observed radar information without future extrapolation;
- (2) persistence of the latest radar field;
- (3) the radar extrapolation used in the proposed framework.

This comparison will distinguish the contribution of future meteorological information from the effect of simply adding correlated radar predictors and will also quantify the additional value of the extrapolation method relative to a simple persistence benchmark. Radar extrapolation skill at different lead times will also be reported in the Supplementary Material.

8. Lines 359-373 and 412-427: The training objective and evaluation design are internally unclear. Table 4 describes a quantile-based data loss, but the model is subsequently evaluated as a single deterministic forecast, and the quantile levels, weighting matrix, error function, regularization coefficients, ramp-up epoch, and hyperparameter-selection procedure are omitted. In addition, NSE and RMSE calculated over strongly autocorrelated time steps within only a few events do not provide independent evidence of superior generalization. Results should be aggregated across all withheld events, with event-level distributions and confidence intervals obtained from repeated training seeds or paired resampling. Peak-timing error should also be reported because much of the interpretation concerns phase displacement, yet the defined (T_p) variables are not used in the metrics.

Response 8: We sincerely thank the reviewer for pointing out this ambiguity. We agree that the current manuscript does not clearly distinguish between the role of the quantile-based loss during model training and the deterministic forecasting objective used for model evaluation. We would like to clarify that the present study is evaluated as a deterministic multi-lead discharge forecasting problem. During training, the 0.1, 0.5, and 0.9 quantile outputs are jointly optimized using a pinball-loss term together with the error term applied to the median prediction. In this study, however, the quantile formulation is used only as an auxiliary robust training constraint, whereas the median prediction ($q_{0.5}$) is treated as the single deterministic discharge forecast for all hydrological performance evaluations. The present study does not attempt to evaluate probabilistic forecasting skill or prediction-interval reliability. In the revised manuscript, we will clarify this distinction explicitly and revise the corresponding Methods text so that the role of the quantile loss is described as a training

regularization/robustness component rather than as evidence of probabilistic forecasting capability. We also agree that NSE and RMSE calculated over strongly autocorrelated time steps should not alone be interpreted as independent evidence of model generalization. Therefore, in addition to NSE, RMSE, REP, and REV, we will report peak-timing error.

9. Lines 396-402 and 480-527: The data-dependency experiment cannot presently support claims of sample efficiency because the method used to select the 25%, 50%, and 75% subsets is unspecified. Different fractions may contain different flood severities and types, making sample composition inseparable from sample quantity, as the authors themselves acknowledge. Nested subsets should be generated repeatedly using stratification by basin, peak magnitude, season, and hydrograph type, with identical independent test events and uncertainty across repetitions. A learning curve based on a single subset per fraction can produce non-monotonic behaviour simply through sampling variability or stochastic optimization and should not be interpreted as evidence that physical constraints alleviate data scarcity without direct comparison against the other models under the same reductions.

Response 9: We agree with the reviewer that the current 25%, 50%, 75%, and 100% training-data experiment cannot, by itself, distinguish the influence of sample quantity from differences in flood-event composition. Our original intention with this experiment was to illustrate the sensitivity of the proposed model to the amount of available training information rather than to establish a rigorous comparative measure of sample efficiency. We acknowledge that the current wording overinterprets the results in this respect. Therefore, in the revised manuscript, we will redefine this section as a training-data sensitivity analysis and will no longer use the non-monotonic learning curve as evidence that PINAS intrinsically requires less training data or that physical constraints necessarily alleviate data scarcity. We will clearly document how the reduced subsets were constructed and discuss the potential influence of event composition.

10. Lines 532-589: The SHAP results are interpreted too strongly as evidence that radar echoes or measured rainfall “control” basin forecasts and demonstrate the necessity of hydrometeorological fusion. These inputs are highly correlated across time and between gauge, areal-rainfall, observed-radar, and extrapolated-radar groups, under which SHAP attribution can be redistributed among redundant variables and depends on the background distribution and explainer. Summing absolute contributions also removes directionality and does not demonstrate hydrological causation or incremental predictive information. The authors should identify the SHAP algorithm and reference dataset, assess sensitivity to correlated-feature grouping, report signed and event-wise contributions, and corroborate the interpretation through input-ablation or permutation experiments. The basin-scale explanations should remain hypotheses unless supported by such predictive tests.

Response 10: We fully agree that SHAP attribution should not be interpreted as direct hydrological causation or proof of the necessity of a particular predictor. In the revised manuscript, we will explicitly report the SHAP algorithm, background and feature-processing procedure. We will also supplement the global absolute SHAP results with signed and event-level information. To address the effect of correlated predictors, we will conduct a grouped permutation analysis for major hydrometeorological input categories, such as rainfall, observed radar, and future-radar information. Changes in forecasting performance after perturbing each input group will provide an independent predictive check on the SHAP interpretation. Throughout the manuscript, causal expressions such as “control,” “determine,” or “demonstrate necessity” will be replaced by more appropriate terms.