Author comments on RC4- egusphere-2025-1161

Comment 1: *The question of what model structures (including those that incorporate LSTM functionality) can best represent the relevant hydrologic processes across a continental scale domain with varying landscape and hydroclimatic characteristics is an important one. Providing useful answers to this sort of question relies on some interrogation and analysis of which model structures are associated with better performance. The authors compile a dataset of basin characteristics but make no connection between that and model performance variation. This manuscript would be improved with more explanation and interpretation of how known differences in model structures relate to better or worse model performance - and better address the stated goal of providing "scientific guidance for selection and application of hydrologic models" [line 166-167]*

Response:

Thank you for this constructive and insightful comment. We fully agree that understanding how model structures interact with basin-specific characteristics is crucial for advancing hydrological modeling and providing meaningful guidance for model selection. In this study, although we did not classify basins based on individual catchment attributes, we performed our analysis by dividing the 544 basins according to China's nine major river systems and six climatic zones. These regional divisions already reflect important differences in landscape and hydroclimatic conditions and offer spatially coherent insights for model comparison across diverse hydrological regimes.

The main reasons we did not further stratify basins using specific catchment attributes (slope, soil type, land cover) are as follows: (1) the current attribute data are relatively limited in terms of quantity and resolution, and (2) there is no widely accepted or standardized method for catchment classification based on multivariate attributes (Jehn et al., 2020; He et al., 2023; Yang et al., 2023). Moreover, attribute-based classification may lead to spatial discontinuities, which would reduce its usefulness for practitioners aiming to apply modeling strategies in specific regions.

In further research, we plan to expand the Chinese large-sample watershed dataset by incorporating additional static watershed attributes and developing classification methods based on these attributes. This expansion aims to refine our model selection framework and evaluate the effectiveness of various watershed classification schemes in improving runoff simulation accuracy. A discussion of these future directions has been included in the revised manuscript to highlight the potential for further enhancing the robustness and generalizability of our approach.

We have added discussion on this point in the revised manuscript, which reads as follows:

"…

Although this study does not explicitly group catchments based on physiographic or land surface attributes, the analysis framework incorporates regional differentiation by organizing the basins according to the nine major river systems and six major climatic zones in China. These divisions reflect inherent hydroclimatic and geographic diversity, and enable a meaningful comparison of model performance under varied environmental conditions.

Nonetheless, exploring how model performance varies with catchment characteristics remains an important direction for future research. With the improvement of catchment attribute datasets and the development of robust clustering methods, future studies can expand on the current work by examining how model structure suitability depends on physiographic, climatic, or hydrological conditions. This line of inquiry may provide more refined guidance for regional model selection and enhance the interpretability and generalizability of hybrid modeling approaches.

…"

Comment 2: *The performance of models appears to be evaluated by comparing to the results of a different (VIC) model. The practical need for this approach (lack of consistent and comprehensive streamflow data) is understandable, but a rationale that justifies how this provides meaningful insight is not provided. For example - does the analysis reflect which model structures are most similar to VIC, or do they provide some broader insight about a "true" or "best" hydrologic model for different watersheds and regions? Additional explanation and justification is needed to clarify this component of the study.*

Response:

Thank you for this insightful comment. We fully agree that ideally, model evaluation should be based on observed streamflow data. However, due to the scarcity and inconsistency of long-term, high-resolution observed runoff data across all 544 basins in China, which are particularly pronounced in remote or ungauged basins, it is not currently feasible to construct a comprehensive nationwide benchmark completely based on observations. In light of this limitation, we adopted the

VIC-CN05.1 runoff dataset as a reference product, considering both its spatial-temporal continuity and its demonstrated reliability in previous hydrological research in China (Miao et al., 2020; Ma et al., 2024; Wang et al., 2024; Yu et al., 2025).

To ensure that the VIC-CN05.1 runoff can reasonably represent actual streamflow dynamics, we conducted an independent verification in the Supplementary Material (Figure S2), where 15 representative basins with available observed daily runoff data (from January 1, 2015 to December 31, 2015) were compared to the corresponding VIC-CN05.1 simulated runoff series. The Pearson correlation coefficients between the VIC runoff and observed data exceeded 0.8 in nearly all basins, indicating strong agreement in temporal variability. This supports the notion that the VIC-CN05.1 product can be used as a proxy benchmark for comparative model evaluation, especially in large-sample settings where observational coverage is limited.

It is important to emphasize that the goal of our study is not to assess the fidelity of models in emulating VIC per se, but rather to understand the relative predictive skill, generalization capacity, and physical consistency of different modeling paradigms (process-based, data-driven, and hybrid) under a unified reference runoff product. While we acknowledge the limitations inherent in using simulated runoff, this framework enables a fair, consistent, and spatially extensive performance comparison across models.

To address your concern and avoid possible misunderstanding, we have clarified this rationale in the revised manuscript and explicitly stated that using simulated runoff introduces a trade-off between spatial completeness and observational accuracy. We have also acknowledged that future studies should aim to validate the models further using observed streamflow datasets as more consistent data become available. The details of our revised content are as follows:

"…

About Data: used to clarify the sources and characteristics of runoff data

The streamflow data used for model training and evaluation in this study are derived from the VIC-CN05.1 runoff product, which was generated by the VIC hydrological model driven by CN05.1 precipitation. It should be noted that this is not observational streamflow data, but a simulated product that serves as a consistent and spatially complete surrogate in the absence of publicly available observed daily streamflow records for a large number of catchments in China. While VIC-CN05.1 has been evaluated in previous studies, its use as a reference introduces potential structural bias, particularly when comparing models forced by the same precipitation product.

About Discussion: add discussion corresponding to runoff data

One limitation of this study lies in the use of a simulated runoff product (VIC-CN05.1), generated using the CN05.1 precipitation data, as the reference for model evaluation. This introduces a structural bias that may favor models driven by CN05.1 due to internal consistency between inputs and targets. As a result, the observed performance advantage of CN05.1-driven models may not necessarily indicate the intrinsic quality of the precipitation product. The meteorological forcing comparison in this study should therefore be interpreted as an exploration of input-output consistency effects rather than a definitive evaluation of forcing product accuracy. Future studies incorporating additional meteorological datasets and observed streamflow records will be important for validating and extending these findings.

…"

Comment 3: *It is not entirely clear what the value of using the ERA5 and CN05.1 forcing datasets is when the evidence suggests the CN05.1 dataset provides better consistency with local conditions and better performance (albeit in comparison to a model also forced with CNO5.1). It seems that if the performance were being evaluated against true runoff or streamflow observations, model performance (such as with NSE as in Figure 7) would provide a meaningful basis for interpretation. As presented, the inclusion of models driven by ERA5 forcing complicates (and potentially obscures) a clear interpretation of the appropriateness of model structures. For example - how does training/calibrating a model that uses the ERA5 forcing against a target based on the CN05.1 forcing generate reliable information?*

Response:

Thank you for this insightful comment. We acknowledge that using ERA5-Land forcing data as input while using runoff outputs from the VIC-CN05.1 product as target may lead to a mismatch in the input-target consistency.

The primary motivation for including both CN05.1 and ERA5-Land precipitation as model inputs was not to determine which meteorological product is superior in an absolute sense, but rather to investigate the sensitivity and robustness of different modeling approaches (process-based, deep learning, and hybrid) to variations in forcing data. This approach allows us to assess how input data quality and consistency influence model performance under otherwise identical settings. It also highlights the degree to which models can generalize across datasets of different spatial and temporal characteristics. Indeed, CN05.1-based models are expected to perform better when compared against VIC-CN05.1 runoff, since both originate from the same meteorological forcing. However, this does not invalidate the value of ERA5-Land -driven experiments. In fact, these models help us understand how well each modeling paradigm can accommodate inconsistencies or mismatches between input data and target outputs, which is highly relevant in practical applications, especially in data-sparse regions or when switching between data sources is necessary.

To avoid potential misinterpretation, we have revised the relevant statements in the Results section to avoid implying that CN05.1 is "better" than ERA5-Land. Instead, we now emphasize that CN05.1 yields better consistency with the target runoff used in this study (VIC-CN05.1), and that ERA5-Land experiments primarily serve to test model robustness and adaptability. We have also added the following clarification in the revised manuscript:

"…

It should be noted that while CN05.1-based models naturally align more closely with the VIC-CN05.1 runoff product used as reference, the inclusion of ERA5-Land forcing data in this study is intended to evaluate the sensitivity of model performance to meteorological input differences. This helps highlight the robustness and generalization capabilities of different modeling paradigms when faced with inconsistent or lower-fidelity inputs, conditions that are often encountered in practice, especially in ungauged or data-sparse basins. Therefore, the comparison between CN05.1 and ERA5-Land inputs should be interpreted as an assessment of model adaptability, rather than an absolute evaluation of forcing quality.

…"

Comment 4: *The water budget analysis is a valuable complement to the runoff performance comparisons. However, it is unclear exactly how the closure error should be interpreted. The water balance presented implies the closure error may include changes in watershed storage (groundwater, snow, deep soil, etc) that may or may not be well represented in the models. Some further* *clarification and explanation that justifies the interpretation that the "smaller value of epsilont, the better the water budget balance of the basin" [lines 580-581]*

Response:

Thank you for the reminder. We agree that the presented water balance formulation, expressed as $\varepsilon = / P - ET - Q /$, simplifies the full hydrological water balance by not explicitly including the change in storage (ΔS), which may encompass components such as soil moisture, groundwater, snowpack, or deep aquifer storage. In our study, this simplified formulation was intended as a diagnostic metric to evaluate the apparent water budget imbalance using the available data products, under the assumption that storage changes over longer periods are comparatively small or tend to average out. However, we acknowledge that this assumption may not hold for shorter time windows or in basins with strong seasonal storage dynamics (snow-affected basins, deep soils, or those influenced by significant groundwater).

To avoid confusion and over-interpretation of the ε, we have revised the manuscript to clarify its limitations and now explicitly state that ε captures both the imbalance in modeled fluxes and the effects of neglected or unmodeled storage changes. Therefore, while a smaller ε suggests a more internally consistent model in terms of the fluxes represented, it does not imply perfect mass closure in a strict hydrological sense. The relevant sentence in the manuscript has been updated as follows:

"…

It should be noted that the ε used in this study does not account for changes in water storage (ΔS) such as snow accumulation, soil moisture, or groundwater levels. Thus, ε represents a combination of actual water balance error and the effect of unmodeled or unobserved storage variations. A smaller ε value generally indicates a more internally consistent representation of fluxes (P, ET, and Q), but it should not be interpreted as strict mass balance closure.

…

***Comment 5:** In general the figures are well done and informative. Their effectiveness could be improved in many cases by 1) enlarging the axis and label text and 2) providing more informative and more complete captions and annotations. In many figures (e.g. Figure 11) it is difficult directly discern the many types of information being presented.*

Response:

Thank you for your constructive suggestion regarding the clarity and readability of the figures. We appreciate your positive comments about their overall quality and agree that improvements in text size and caption detail would enhance their interpretability.

In the revised manuscript, we have carefully reviewed all figures and made the following modifications as per your recommendation:

Axis and label text sizes have been enlarged uniformly across all figures to ensure clarity and readability, particularly when viewed in print.

Figure captions have been rewritten or expanded to provide more detailed descriptions of what each panel represents, the data sources used, and how to interpret key patterns or comparisons

For figures with multiple subplots (Figure 11), we have:

Corrected any sub-figure labeling issues (duplicate (b) labels). Added explicit explanations in the captions about what each subplot shows. Included improved legends or annotations within the figures where appropriate to aid interpretation.

The revised Figure 11 and its caption are as follows:

"….

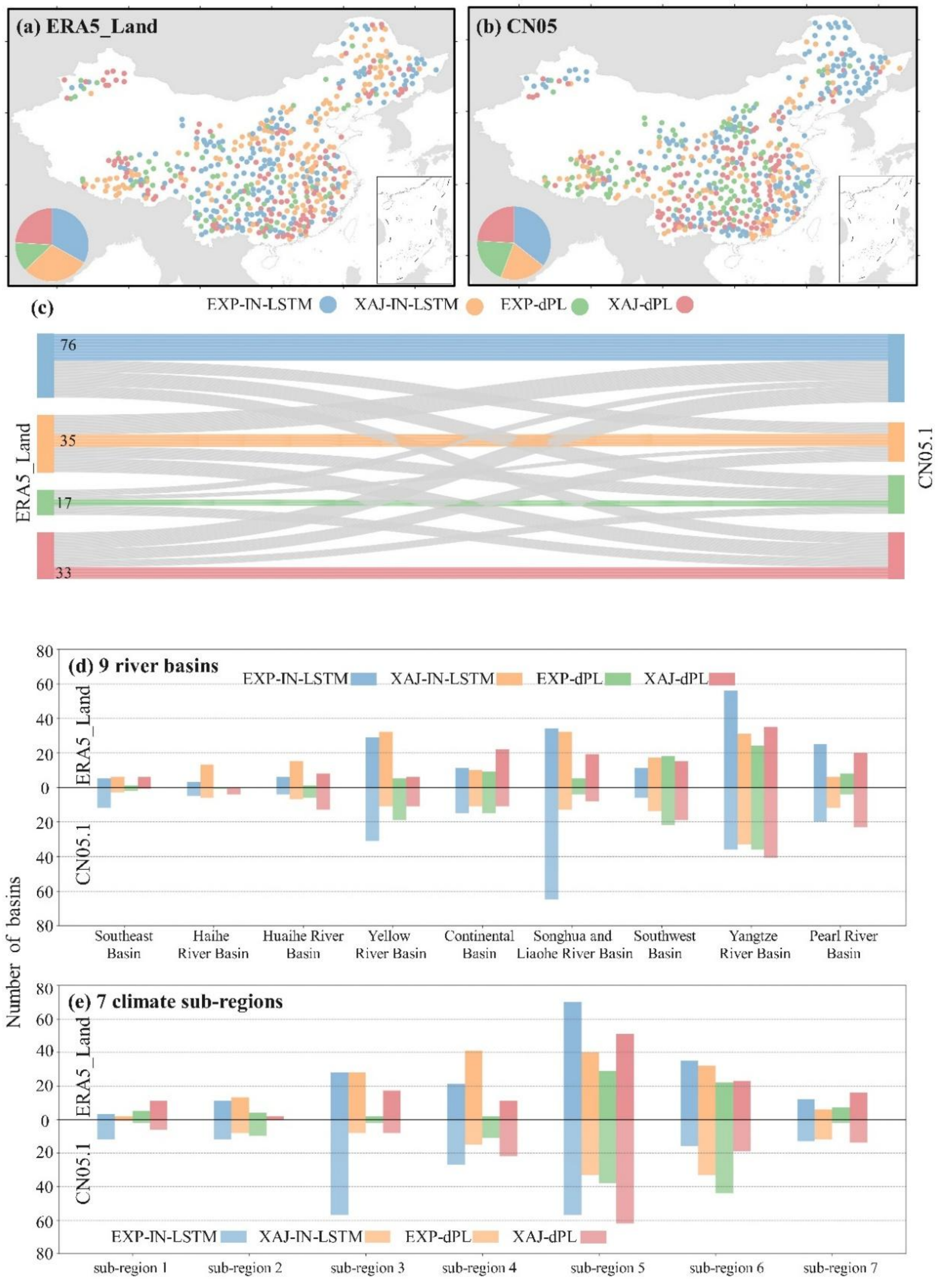

Figure 11. Comparison of hybrid model performances across basins using ERA5-Land and CN05.1 precipitation data. (a) Spatial distribution of the best-performing hybrid models under ERA5-Land forcing. (b) Spatial distribution under CN05.1 forcing. (c) Sankey diagram showing the consistency of best-performing models across datasets. (d) Number of best-performing basins for each model in the 9 major river basins. (e) Number of best-performing basins in the 7 climate sub- regions.

…."

Comment 6: *The methods and interpretation used for the "prediction in ungaged basins (PUB)" portion of the analysis is a bit confusing. Some more specific explanation that covers 1) the intent of this analysis and 2) how it is different from the other performance comparisons would make the paper much more effective.*

Response:

Thank you for the reminder. To improve the clarity and effectiveness of the manuscript, we have revised the relevant sections to more clearly explain both the motivation behind the PUB test and how it differs from the other modeling strategies. The PUB test in this study is designed to evaluate the extrapolation capability of various models in situations where no observed runoff data are available for the target basin, which is an issue frequently encountered in practical hydrological modeling (Hrachowitz et al., 2013; Sivapalan et al., 2003). This setup reflects a real-world scenario in which models must rely entirely on information learned from other basins, without any local calibration, to make predictions in ungauged regions.

To simulate this condition, we employed a leave-one-cluster-out strategy. In this strategy, each cluster is randomly grouped. During the training phase, one cluster is withheld and used solely for testing. This ensures that the model's performance in these basins reflects its generalization ability rather than its fit to seen data. In contrast, the regional modeling experiment includes all basins during training and measures performance within a cross-validation framework, while the basin-specific experiment calibrates and tests models individually at each site.

We have now incorporated these clarifications into the Methods section of the revised manuscript to emphasize the unique purpose and implications of the PUB analysis. The revised specific content is as follows:

"….

Furthermore, to further assess the generalization performance of the models, a five-fold cross-validation method is implemented. Specifically, the 544 basins are divided into five relatively even clusters (with each cluster containing either 109 or 108 basins, as shown in Figure 2). The validation process is as follows: the model is trained using the training period data from the basins in four of the clusters, and its performance is validated on the test period data from the remaining cluster. This operation is repeated in a loop, with each iteration designating a different cluster as the ungauged basin, thereby allowing for the evaluation of the predictive performance of each basin treated as an ungauged basin. This experiment design effectively simulates a Prediction in Ungauged Basins (PUB) scenario, which is a key research topic in large-sample hydrology. In this setting, no runoff observations from the target basins are used during model training, ensuring a strict spatial independence between training and testing data. The purpose of this analysis is to assess the models' ability to generalize hydrological knowledge from gauged to ungauged basins, which is critical for practical applications in data-scarce regions. Compared with regional modeling experiments, the PUB test can evaluate model robustness and transferability. By averaging the performance across five folds, this strategy offers a reliable estimate of model's generalization capability.

…."

Comment 7: *Lines 96-98: "type of model excels in data collaboration....This type of model performs well in data-driven collaboration..."*

Response:

Thank you for your reminder. The two sentences convey overlapping ideas and lead to redundancy. In the revised manuscript, we have removed the repetitive expression for clarity and conciseness. The revised sentence is as follows:

"….

This type of model excels in capturing complex patterns from large-scale data and is particularly suitable for hydrological modeling in data-rich environments (Fang et al., 2022; Kratzert et al., 2019a; Tsai et al., 2021).

…."

Comment 8: *Lines 127-128: "..neuralizes the process-based model and adjusts model parameters by back propagating gradients based on daily prediction results.."*

Response:

Thank you for your reminder. The original manuscript has lacked clarity and could be confusing to readers unfamiliar with the hybrid modeling framework. In the revised manuscript, we have rephrased the sentence to provide a more precise and comprehensible explanation of how the differentiable hybrid model integrates neural networks with process-based model components using gradient-based optimization. The revised sentence is as follows:

"….

Specifically, this hybrid modeling approach transforms the process-based model into a differentiable form by embedding it within a neural network framework. This allows the model parameters to be optimized through gradient backpropagation using daily runoff prediction errors.

…."

Comment 9: *Lines 321-322: "With the continuous advancement of deep learning technology, its applications in the field of hydrology are also expanding."*

Response:

Thanks for your comment. The original manuscript was not clear enough and failed to provide specific insights. We have revised the sentence in the manuscript to better reflect the relevance and current significance of deep learning applications in hydrology, especially in the context of large-sample hydrological modeling and hybrid model development. The revised sentence is as follows:

"….

Deep learning techniques have recently gained significant attention in hydrological modeling due to their ability to learn complex, nonlinear relationships directly from data without relying on explicit physical assumptions. In this study, the long short-term memory (LSTM) network is adopted as a representative purely data-driven model.

…."

We would like to thank the editors and reviewers once again for their valuable suggestions on our manuscript. We have incorporated these suggestions into the revised manuscript. Looking forward to hearing from you.

Chunxiao Zhang

Corresponding author

E-mail address: zcx@cugb.edu.cn

References:

Fang, K., Kifer, D., Lawson, K., Feng, D., Shen, C., 2022. The Data Synergy Effects of Time Series Deep Learning Models in Hydrology. Water Resour. Res. 58, e2021WR029583. https://doi.org/10.1029/2021WR029583

He, Z., Shook, K., Spence, C., Pomeroy, J. W., and Whitfield, C., 2023. Modelling the regional sensitivity of snowmelt, soil moisture, and streamflow generation to climate over the Canadian Prairies using a basin classification approach, Hydrol. Earth Syst. Sci., 27, 3525–3546, https://doi.org/10.5194/hess-27-3525-2023.

Hrachowitz, M., Savenije, H., Blöschl, G, McDonnell, J., Sivapalan, M., Pomeroy, J., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. Hydrological sciences journal, 58(6), 1198–1255.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., and Houska, T., 2020. Using hydrological and climatic catchment clusters to explore drivers of catchment behavior, Hydrol. Earth Syst. Sci., 24, 1081–1100, https://doi.org/10.5194/hess-24-1081-2020.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research 55, 11344–11354. https://doi.org/10.1029/2019WR026065

Ma X , Wang A ., 2024. Evaluation and Uncertainty Analysis of the Land Surface Hydrology in LS3MIP Models Over China. Earth & Space Science, 11(7). https://10.1029/2023EA003391.

Miao Y , Wang A ., 2020. Evaluation of Routed-Runoff from Land Surface Models and Reanalyses Using Observed Streamflow in Chinese River Basins.Journal of Meteorological Research, 34(1):73-87. https://10.1007/s13351-020-9120-z.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological sciences journal, 48(6), 857–880.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., Shen, C., 2021. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nat. Commun. 12, 5988. https://doi.org/10.1038/s41467-021-26107-z

Wang Z , Li M , Zhang X., 2024. Prediction of long-term future runoff under multi-source data assessment in a typical basin of the Yangtze River.Journal of Hydrology: Regional Studies, 56(000). https://10.1016/j.ejrh.2024.102053.

Yang M , Olivera F. , 2023. Classification of watersheds in the conterminous United States using shape-based time-series clustering and Random Forests. Journal of Hydrology, 620(Pt.A). https://10.1016/j.jhydrol.2023.129409.

Yu X , Zhang Q , Zeng X., 2025. The distribution and driving climatic factors of agricultural drought in China: Past and future perspectives.Journal of Environmental Management, 377. https://10.1016/j.jenvman.2025.124599.