Author comments on RC1- egusphere-2025-1161

Comment 1: *For assessing the performance of the different experiments, the authors compare the simulated streamflow to streamflow from VIC-CCN5.1… which is also simulated streamflow. This choice is justified, although I guess that VIC-CCN5.1 had to be evaluated against observed streamflow, so why not using it. To add to the confusion, several of the experiments come from models forced by CCN5.1. That induces a bias in the conclusions that can be drawn.*

Response:

Thank you for your careful review and comments. Your suggestion is very professional. For our research, using the observed runoff data of each basin as the target variable is the most rigorous. However, the reality is that we cannot obtain the daily runoff data of hundreds of basins in full. On the one hand, most of the runoff simulation studies in China are only conducted in specific areas, which is one of the reasons that prompted us to conduct large-sample hydrology studies in China. The standards for demarcating basin boundaries in these studies are not uniform, so even if we can obtain daily runoff data in full, we cannot guarantee that these data have good consistency. On the other hand, the data start and end times of the above studies are not the same, and even most of the time periods of the studies do not overlap, which is very unfavorable for the establishment of a large-sample basin data set. Therefore, when we first started data processing, we hoped to extract daily runoff data for hundreds of basins based on a relatively high-quality, relatively long time span daily runoff data set and a unified basin boundary demarcation standard.

Of course, we also fully agree with what you said that using runoff data products will lead to biased conclusions, so we originally planned to use the real hydrological station data of several basins to calibrate the runoff data products. However, the problem we face is that we only have real daily runoff data for 15 basins (the time span of each basin is about one year). And the above basins cannot cover different climate zones and water systems in space. If we only use these basins for data correction, it may lead to greater deviations. We think about it and make a cautious decision. Since it is impossible to fully guarantee that the data of each basin in the dataset is consistent with the actual observation under the existing objective conditions, we will try to simplify the data acquisition method on the basis of ensuring data consistency. This can ensure that the dataset has a

certain degree of availability and provide a reference for establishing a large-sample dataset in other areas that lack observation sites.

Regarding the issue of conclusion bias, we plan to add appropriate explanations in the Discussion section to ensure that readers can view our conclusions critically while understanding our goals. The relevant discussion added is as follows:

"…

A notable limitation of this study lies in the use of VIC-CN05.1 simulated streamflow as the reference for model evaluation. While this approach ensures nationwide spatial coverage and consistent hydrological boundaries across all 544 catchments, it may introduce systematic biases—particularly when evaluating models that are forced with the same meteorological dataset (CN05.1). The use of simulated rather than observed streamflow data could potentially favor certain models and compromise the neutrality of comparative performance assessments.

The primary rationale for adopting a runoff product instead of observational data stems from the limited accessibility and temporal inconsistency of observed streamflow records in China. Existing observed datasets often cover only specific regions, vary in spatial resolution and delineation standards, and exhibit non-overlapping time periods. These issues hinder the construction of a coherent, large-sample hydrological dataset with sufficient temporal depth and spatial uniformity. Given these constraints, the VIC-CN05.1 product was selected due to its relatively high simulation quality, long-term continuity, and compatibility with the CN05.1 precipitation product and unified basin boundaries.

Although this choice is methodologically justifiable, it is important to acknowledge the limitations it imposes on the interpretation of model performance. Comparative results may partially reflect consistency between inputs and evaluation targets rather than absolute predictive skill. Therefore, the findings should be interpreted with caution, particularly regarding the apparent superiority of models forced by CN05.1. Future work may consider integrating sparse but high-quality observed streamflow data for calibration or validation, thereby enhancing the robustness of the benchmark and supporting broader applicability in ungauged or data-scarce regions.

…"

Comment 2 (About Abstract): *The abstract mentions the use of PBMs, but not what they are used for, neither what we can conclude about them. Line 36: conclusions about the two hybrid models are drawn, but those are not detailed before. L 40-42: This is not a concluding sentence for an abstract, this is the rationale of the study. Here we need you to give us the major guidance resulting from your work.*

Response:

Thank you for your reminder. Your suggestion on the abstract is very detailed. We have made the following changes to the abstract: added an explanation of the purpose of the process-based model; briefly introduced the operation of the two hybrid models (before summarizing the conclusion); added a concluding sentence at the end of the abstract, detailing the recommendations provided by this study. The content of the modified abstract is as follows:

"…

Hydrological modeling plays a key role in water resource management and flood forecasting. However, in China, with diverse geography and complex climate types, a systematic evaluation of different modeling schemes for large-sample hydrological datasets is still lacking. This study preliminarily constructs a dataset of catchment attributes and meteorology covering 544 basins in China, and systematically evaluates the applicability of two process-based models (PBMs: EXP-HYDRO model and Xin'an jiang model), long short-term memory (LSTM) models, and hybrid modeling methods. Among them, four hybrid models are developed: two process-based models are combined with the LSTM model using the alternative hybrid modeling scheme and the differentiable hybrid modeling scheme, respectively. The results demonstrate: (1) The accuracy of meteorological data critically impacts the prediction performance of hydrological models. High-quality precipitation data enables the model to better simulate the runoff generation process in the basin, thereby improving prediction accuracy. (2) The hybrid modeling method possesses regional modeling capabilities comparable to those of LSTM model. It also demonstrates strong generalization capabilities. In predicting ungauged basins, the hybrid model exhibits greater stability than the LSTM model. (3) Among the two hybrid modeling methods, the differentiable hybrid modeling scheme offers a deeper understanding and simulation of hydrological processes, along with the ability to output unobserved intermediate hydrological variables, compared to the alternative hybrid modeling schemes. Its prediction results are more consistent with the water

balance of the basin. The research results provide a systematic analysis for evaluating the applicability of different hydrological modeling methods in 544 basins in China, suggesting that high-quality meteorological data from consistent sources should be selected and considering the use of differentiable hybrid modeling schemes to better understand and simulate hydrological processes. This will help achieve higher prediction accuracy while ensuring the physical consistency of the prediction results.

…"

Comment 3: *Line 52: Why is the complexity of hydrological processes increasing? It seems to me that all this discussion is about natural processes, which do not complexify in time.*

Response:

Thank you for the valuable comment. We agree that natural hydrological processes themselves may not inherently become more complex over time. Our intention was to emphasize that, in recent years, the perception and modeling of hydrological processes have become increasingly complex, due to several factors:

1. Climate variability and change have led to more frequent extreme events (droughts, floods), making it harder to represent hydrological dynamics using traditional process-based models

2. Data limitations and heterogeneity, especially in ungauged basins or regions with sparse observations, increase the challenges in parameterizing and calibrating PBMs.

3. High expectations from stakeholders now require models to perform reliably under novel conditions or for diverse purposes, thus increasing the demand for more robust and generalizable modeling methods.

To avoid misunderstanding, we have revised the sentence as follows:

"…

However, with growing climate variability, and increasing demands on hydrological modeling, the perceived complexity and uncertainty in basin hydrological processes have increased, posing new challenges to the applicability of traditional process-based models (PBMs) in practice, especially under data-scarce and heterogeneous conditions.

…"

Comment 4: *L 86: I do agree for physically-based models, but conceptual/empirical ones only need from 3 variables, namely precipitation, temperature and streamflow. This is not a substantial amount of high-quality data! For example, the EXP-HYDRO used by the authors exactly need these data, plus the day length, and the Xin'an jiang model only needs these data.*

Response:

Thank you for your insightful comment. We agree with your observation that conceptual and empirical hydrological models—such as EXP-HYDRO and Xin'an jiang—typically rely on a limited number of input variables (precipitation, temperature, runoff, and day length), and do not necessarily require a "substantial amount" of input data in terms of variable types.

Our intention, however, was to emphasize that even with a limited number of inputs, the reliability and performance of such models in practical applications often depend on the availability of continuous, high-quality observations, especially for streamflow data and meteorological drivers. Additionally, the calibration of model parameters can still involve subjectivity and nontrivial complexity, particularly when applied to basins with limited or noisy data.

To reflect this point more accurately and avoid confusion, we have revised the sentence as follows:

"…

Even when used to model hydrology for a single basin, such models often rely on the availability of continuous and reliable input data, and their parameterization can involve a degree of subjectivity and complexity, particularly in data-scarce or ungauged conditions.

…"

Comment 5: *L 91: I do not agree, see previous comment*

Response:

Thank you for pointing this out. We understand your concern regarding the comparison between process-based models (PBMs) and data-driven models like LSTM.

We acknowledge that many conceptual models, such as EXP-HYDRO and Xin'an jiang, do not rely on highly detailed or fully physically-based parameters, and that their parameterization often combines empirical knowledge with simplified process representations.

Our intention was not to overstate the advantage of LSTM, but rather to highlight that deep learning models bypass the need for explicit physical parameterization, and instead rely on learning input-

output relationships directly from data. However, we agree that LSTMs come with their own requirements, particularly the need for long, continuous, and high-quality historical datasets for effective training, which can also be a limitation in practice. To better reflect this balance, we have revised the sentence as:

"…

In contrast, deep learning models such as long short-term memory networks (LSTM) can learn the dynamic characteristics of basin hydrological processes from historical data and capture complex nonlinear relationships, without relying on explicit physical parameterizations. However, they typically require long-term, high-quality data for effective training, and their interpretability remains limited compared to PBMs.

…"

Comment 6: *L 168: From now onwards, I wonder if most elements should rather appear in the material and methods section of the manuscript*

Response:

Thank you for your constructive suggestion. We agree that much of the content in this paragraph—such as the dataset structure, variable types, model descriptions, and evaluation settings—would be more appropriately placed in the Data section. Our original intention was to briefly highlight the necessity and contribution of building a large-sample hydrological dataset covering diverse Chinese basins, as a key motivation for this study. However, we recognize that the inclusion of detailed technical information in the Introduction may affect the logical flow and clarity.

To address your comment, we have reorganized the manuscript structure:

The technical description of the dataset (number of variables, data sources, processing methods) has been moved to the Data section.

In the Introduction, we now briefly summarize the dataset's role and relevance in the study, without delving into implementation details.

Comment 7: *L 190: Why do accurate daily runoff observation data often need to be kept confidential?*

Response:

Thank you for your reminder. We acknowledge that the reasons for limited accessibility of daily

runoff observation data may not be immediately clear. In some countries, including China, hydrological data, particularly high-resolution runoff observations, are managed under strict institutional frameworks. Access to this data is often restricted for several reasons, including national regulations on water resource management, the perceived strategic importance of water data for flood control, water security, and infrastructure planning, as well as legacy data-sharing policies that allow agencies to maintain ownership and control over observation networks.

As a result, while some aggregated or monthly flow data may be available, high-quality daily discharge records are often difficult to obtain for research purposes or international sharing. Similar challenges have also been reported in other regions (e.g., South Asia, Africa and so on).

This situation underscores the importance of constructing a curated, consistent, and research-accessible dataset, as we have done in this study, to support comparative hydrological modeling and promote reproducibility.

We have clarified this point in the revised manuscript to provide better context. The specific contents are as follows:

"…

In China, obtaining datasets for large-scale hydrological studies is challenging for two main reasons. Firstly, access to accurate daily runoff observations is often restricted due to institutional regulations and data management policies.

…"

Comment 8 (About Figure 1):

Figure 1: Please make the different maps more uniform. Panel b uses a different color for foreign countries. In addition, please do not use the same color for China and seas (panel a). I also suggest removing the bottom right islands, as there as no basins there and they are originally not on the map. Imagine if French researchers put all French territories on all maps!!

Caption of Figure 1: In a I see the areas, in b the DEM, in c the catchments and in d the climates (only this ones correct). Please modify

Response:

Thank you for the reminder. The use of different map backgrounds and inconsistent coloring may affect the visual coherence of the figure. In the revised manuscript, we have unified the background

color schemes across all sub-panels to improve clarity and comparability, and we have modified the colors of surrounding countries and seas to avoid potential confusion.

Regarding the inclusion of islands in the bottom-right corner of the map: we fully understand your concern. However, as this study is conducted using officially released national geographic data (from https://www.tianditu.gov.cn/), we are required to follow the standardized map representation guidelines mandated by relevant authorities. The inclusion of such elements is to comply with formal map-use conventions in Chian. We hope for your understanding in this matter.

In fact, several recent papers published in the journal focusing on China's hydrological research also used similar national base maps including the South China Sea region (as shown in Figs 1, 2, and 3).

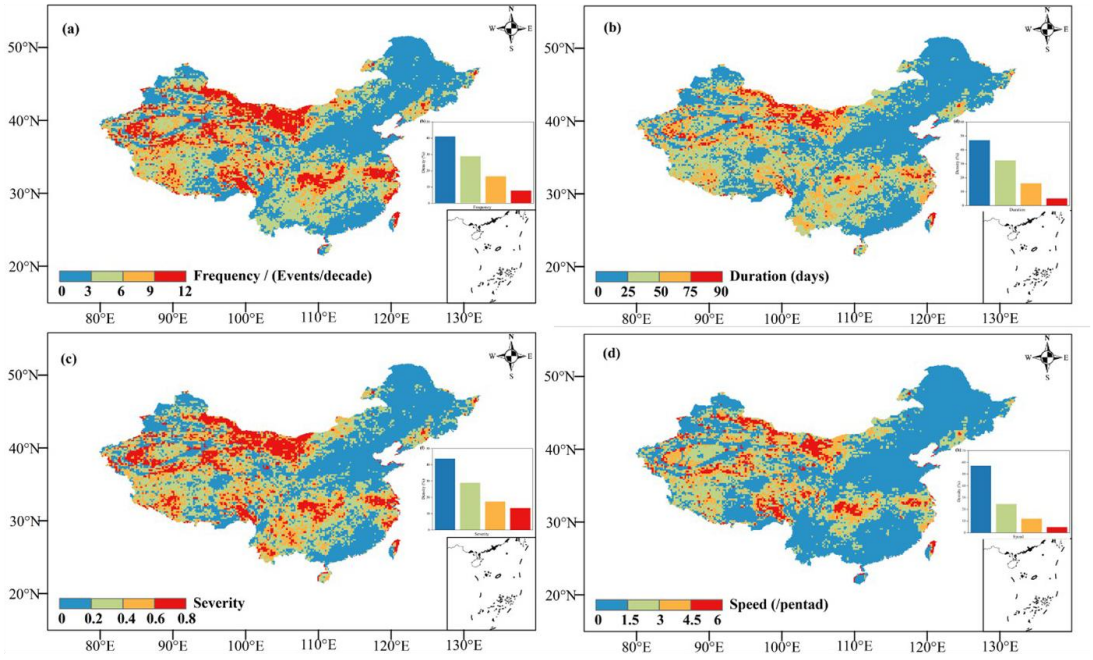

Fig 1. The Figure 2 from *Assessing recovery time of ecosystems in China: insights into flash drought impacts on gross primary productivity* https://doi.org/10.5194/hess-29-613-2025

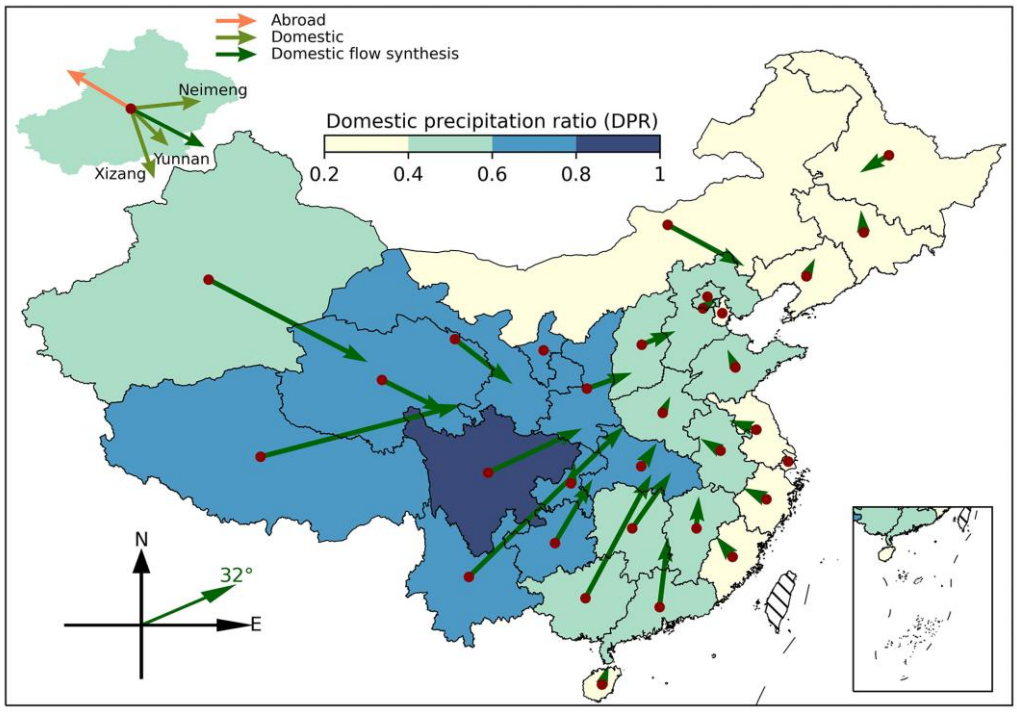

Fig 2. The Figure 3 from *The interprovincial green water flow in China and its teleconnected effects on the social economy* https://doi.org/10.5194/hess-29-67-2025

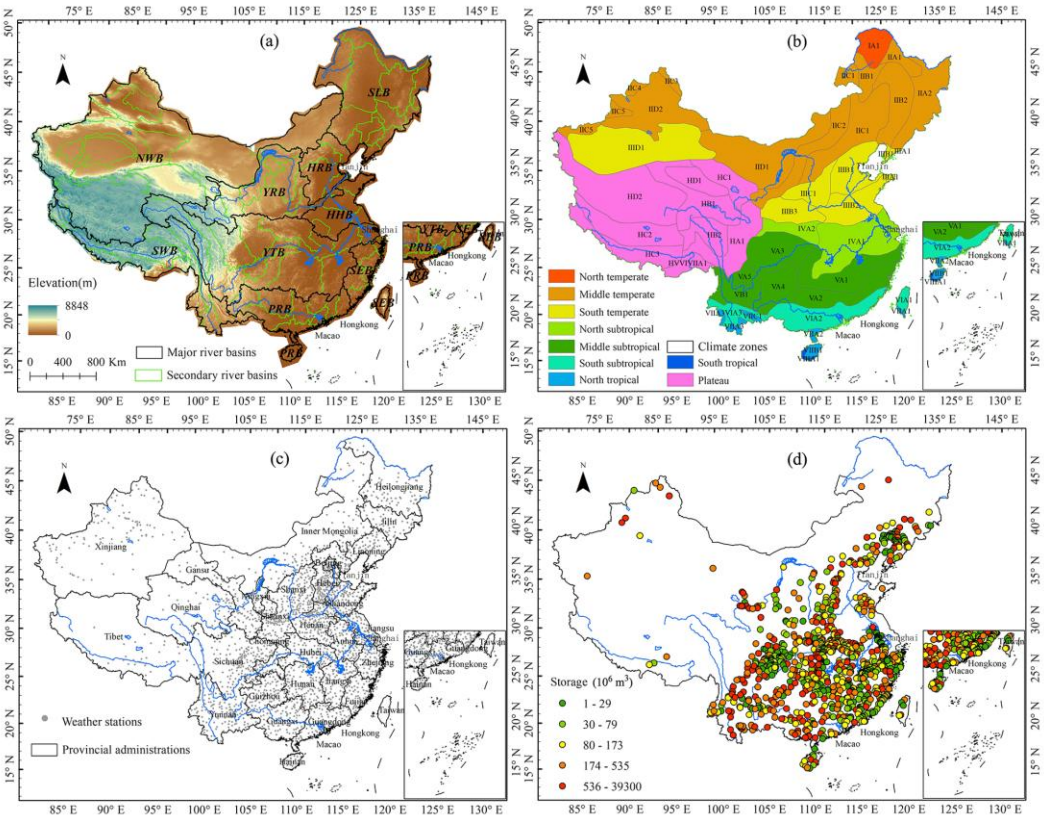

Fig 3. The Figure 1 from *Variation and attribution of probable maximum precipitation of China using a high-resolution dataset in a changing climate* https://doi.org/10.5194/hess-28-1873-2024

We have revised the figure caption accordingly to correctly describe the content of each sub-panel and avoid any previous confusion. The modified Figure 1 and its caption are as follows:

"….

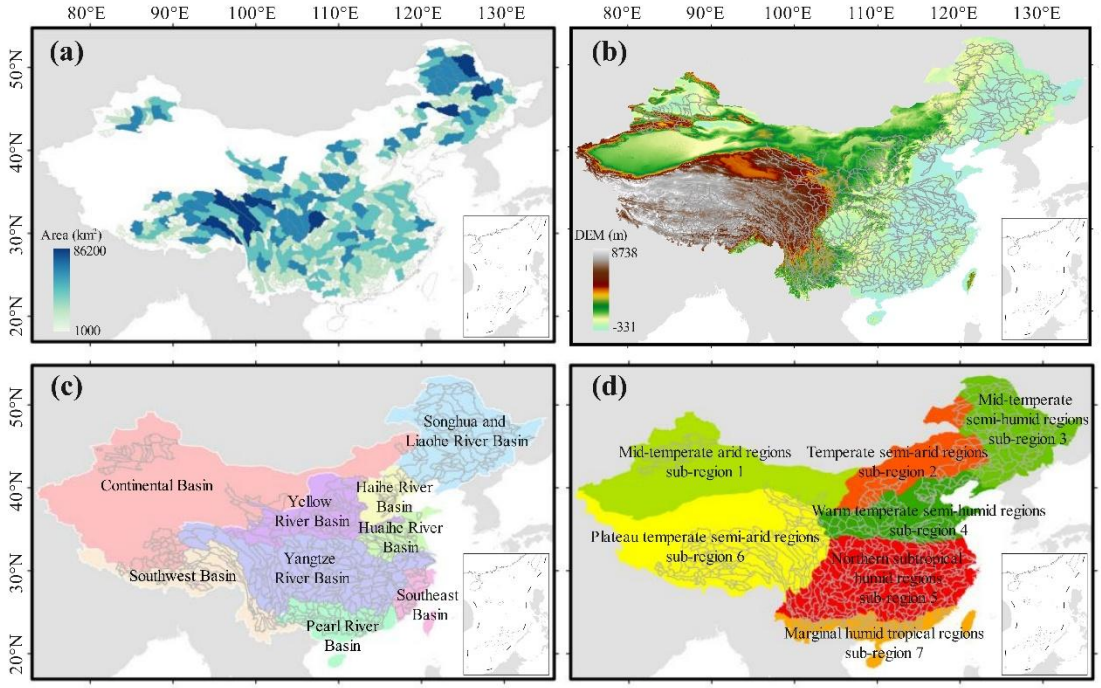

Figure 1. Spatial distribution of the 544 basins used in this study. (a) Basin boundaries and areas. (b) Elevation distribution based on DEM data. (c) Divisions of China's nine major river systems and (d) seven climate regions (The map of China used in this study is from https://www.tianditu.gov.cn/.)

…"

Comment 9: *L 234: I was completely lost here. There must be a nuance between the different terms (observation, runoff, runoff hydrograph), but I initially didn't get it. Only later on, while reading the results, I understood that the VIC-CN05.1 dataset is simulations from the VIC model forced by CN05.1. That was not clear at all.*

Response:

Thank you for your reminder. You are absolutely right to point out the importance of clearly distinguishing between observed and simulated runoff data. In this study, we used a national gridded runoff dataset—VIC-CN05.1, which was produced by driving the VIC (Version 4.2.d) hydrological model with the CN05.1 meteorological forcing. The CN05.1 dataset itself is based on interpolated

observations from more than 2400 stations across China, and the resulting VIC-CN05.1 runoff dataset provides 0.25° × 0.25° daily runoff estimates for the period 1961–2017.

We fully acknowledge that this runoff dataset is simulated, not directly observed. However, due to institutional constraints and the limited public availability of high-resolution daily runoff observations in China, VIC-CN05.1 has been widely adopted in large-scale hydrological studies as a proxy or substitute for runoff observations. If possible, in the future, we will definitely attempt to update the dataset with all runoff and meteorological data using observational data.

Recognizing the limitation of using simulated data, we conducted a comparison between the VIC-CN05.1 runoff series and actual observed streamflow records from 15 gauged basins with similar boundaries (see Supplementary Figure S2). The results show that, although not strictly calibrated, the simulated runoff series closely capture the overall temporal variation and seasonal trends of observed runoff, suggesting that the dataset is reasonably suitable for large-sample hydrological modeling.

We also acknowledge that constructing a fully observation-based hydrological dataset across hundreds of basins with consistent boundary delineation and time coverage is currently infeasible in China, due to limited public access to daily streamflow data. Thus, our study aims not to deliver a high-precision observational dataset, but rather to build a relatively comprehensive and internally consistent dataset that enables comparative and reproducible evaluation of hydrological models.

In light of your comment, we have revised the manuscript to avoid mislabeling simulated data as "observed." Instead, we now refer to the VIC-CN05.1 runoff data as a proxy for observations, and clearly acknowledge that it is model-simulated data used as an observation substitute due to the unavailability of direct measurements. We hope this clarification preserves the intent of our study while improving transparency and accuracy.

In the revised manuscript, we clarified the origin and nature of the VIC-CN05.1 dataset in the Data section and explicitly stated the purpose and findings of the comparison with observed data in Supplementary Figure S2. The specific contents are as follows:

"…

Due to the limited accessibility of daily observed runoff records in China, this study uses the VIC-CN05.1 dataset as a proxy runoff dataset. This dataset was generated using the VIC model driven by the CN05.1 meteorological forcing, and provides daily runoff estimates at 0.25° × 0.25°

resolution. Although this dataset is not based on direct streamflow observations, it has been widely adopted in previous studies and offers a physically consistent, nationwide runoff product. In this study, we treat it as a substitute for observed runoff in basins where actual measurements are unavailable. To validate its feasibility, runoff time series from 15 gauged basins were compared with actual streamflow records from hydrological stations (see Supplementary Figure S2). The results suggest that VIC-CN05.1 data can reasonably capture seasonal patterns and interannual variability, making it a suitable alternative for large-sample hydrological model evaluation in data-scarce regions.

…"

Comment 10: *L 284: Do you mean 4? There are 5 clusters*

Response:

Thank you for catching this typographical error. The manuscript should refer to four clusters, not nine. We have corrected the sentence to accurately describe the 5-fold cross-validation procedure. The specific contents are as follows:

"… the model is trained using the training period data from the basins in four of the clusters, and…"

Comment 11: *Figure 4: While a is understandable, I do not get b at all. What is FCNN? It is never defined in the text. Please improve or develop the caption.*

Response:

Thank you for pointing out that "FCNN" was not defined in the figure or the main text. In fact, FCNN stands for Fully Connected Neural Network, which serves as the parameterization channel: it takes static basin attributes (e.g., soil, terrain metrics) as inputs and maps them to the hydrological model parameters θ. We have now (1) added the definition of FCNN in the caption and main text, and (2) enriched the caption to make panels (a) and (b) fully self-contained. The specific content after the modification is as follows:

"…

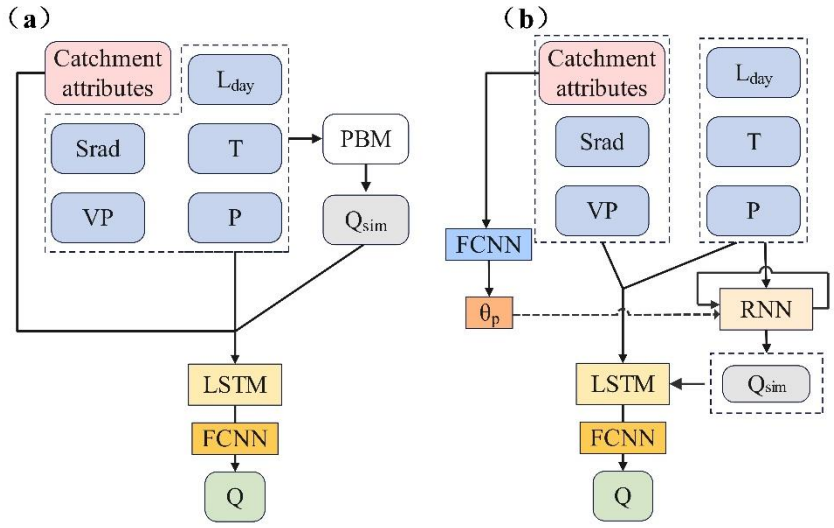

Figure 4. Structure of the hybrid hydrological models. (a) In the conventional hybrid scheme, a process-based hydrological model (PBM) produces simulated runoff Q_{sim}, which—together with meteorological forcings (P: precipitation; T: temperature; Srad: solar radiation; VP: vapor pressure; L_{nay}: day length) and static catchment attributes—is fed into a data-driven network (LSTM + FCNN) to predict runoff Q. (b) In the differentiable hybrid scheme, the PBM's discrete equations are embedded into RNN units, while a Fully Connected Neural Network (FCNN) parameterization channel maps catchment attributes to the model parameters θ_p. The entire architecture (parameters within both RNN and FCNN) is then optimized jointly via back-propagation, allowing hydrological parameters to vary adaptively across basins and climatic regimes.

…

In the differentiable hybrid modeling scheme (Figure 4b), standard recurrent neural network (RNN) units encode the discrete ordinary differential equations of the process-based hydrological model, ensuring mass balance and fundamental process representation. At the same time, we introduce a parameterization channel implemented as a Fully Connected Neural Network (FCNN), which ingests static basin attributes and produces the spatially varying parameter vector θ_p. By jointly optimizing both the RNN weights and the FCNN parameters via back-propagation, the model can dynamically adjust its physical parameters conditioned on basin characteristics, overcoming the

fixed-parameter limitation of traditional PBMs and enabling cross-basin generalization.

…”

Comment 12 (About Figure 5):

Figure 5: Please use the same range for the distribution of P values for the two products over the diverse basins. Also make sure to use the same categories, it seems that there are many more categories for CN05.1 than for ERA5. I guess this is basin-averaged P and T? Please specify.

Figure 5: The scale indicates a gradual color scale for P and T, but the maps only display categorical values, with only 5 colors. Please correct. What is the period? Is it the total period or the evaluation period (1995-2015)? These two comments are valid for most figures that follow

Response:

Thank you for pointing out the potential confusion between the histogram scales and the map legend. To clarify our presentation—and without altering the original figure content—we have made the following changes:

1. Clarified the time period: In the revised caption we now explicitly state that all basin-averaged values are calculated over the full study period (from October 1, 1975, to September 30, 2015).

2. Explained the categorical coloring: Although the histograms use continuous bins to show the full distribution, the maps intentionally use five discrete categories to highlight broad hydro-climatic classes across China. These categories were chosen based on natural breaks in the combined ERA5-Land & CN05.1 distribution (<500 mm, 500–1000 mm, 1000–1500 mm, 1500–2000 mm, >2000 mm for precipitation), and similarly spaced for temperature. We have revised the figure legend and added a sentence to the caption to make this explicit.

3. Unified legend ranges: We confirmed that both the ERA5-Land and CN05.1 datasets share the same class boundaries: red consistently represents the highest precipitation category, while blue indicates the lowest. For temperature, green represents the lowest values, and red represents the highest. This alignment ensures direct comparability, even though the underlying datasets have slightly different numerical ranges.

We trust these changes resolve the inconsistencies and improve the clarity of our data presentation. At the same time, we have thoroughly checked all the figures in the manuscript and made

modifications to the legends and figure captions to ensure that readers can clearly understand the meaning of each figure. The modified Figure 5 and its caption are as follows:

"…

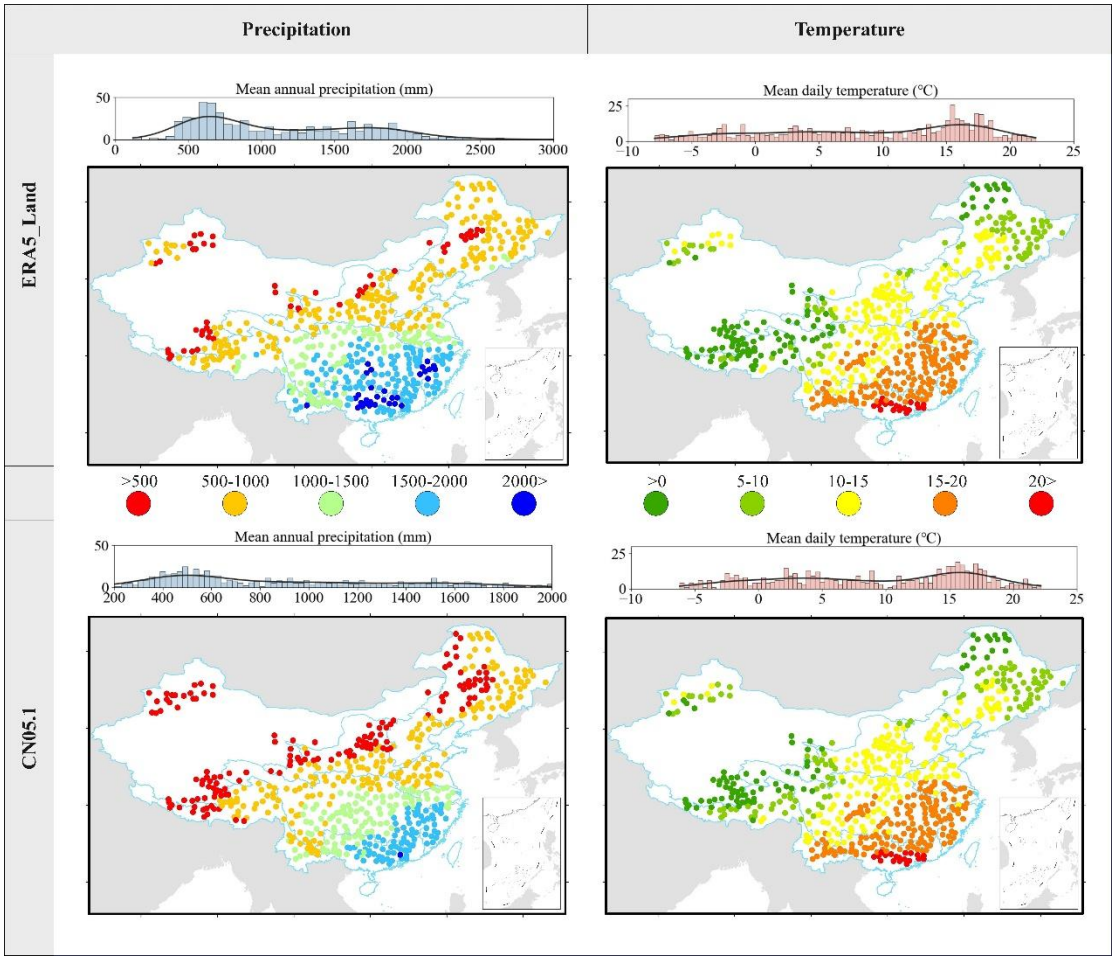

Figure 5. Spatial distribution of five hydro-climatic categories of basin-averaged precipitation and temperature (1 Oct 1975–30 Sep 2015). (a) Mean annual precipitation (mm): Top histograms display the continuous distribution of basin means. Maps use five discrete classes: <500, 500 to 1000, 1000 to 1500, 1500 to 2000, and >2000 mm, applied uniformly to both ERA5-Land (top row) and CN05.1 (bottom row). (b) Mean daily temperature (°C): Top histograms display the continuous distribution of basin means. Maps use five discrete classes: <5, 5 to 10, 10 to 15, 15 to 20, and >20 °C, applied identically to both datasets.

…"

Comment 13: *L 407: How is the drought index calculated?*

Response:

Thank you for your reminder. In the revised manuscript, we have added a clear definition of the drought index and its calculation. The details are as follows:

"…

Figure 6 shows a scatter plot of the evaporation index (EI, the ratio of annual average evapotranspiration to annual average precipitation) and the drought index (Aridity, the ratio of annual average potential evaporation to annual average precipitation).

…"

Comment 14: *L 415: That definitely induces a bias! It is easier to reproduce streamflow obtained from a model forced by a dataset, when you use the same dataset…*

Response:

Thank you for your reminder. You are absolutely right that using the same meteorological forcing (CN05.1) both to generate our "proxy" runoff via VIC and then to drive other models can introduce a positive bias in cross-model comparisons. Our primary objective, however, is not to report absolute predictive skill but to perform a relative evaluation of different modeling approaches under identical forcing conditions. By holding the input data constant, we ensure that differences in performance arise from model structure rather than from differences in meteorological inputs. We added relevant explanation in the Methodology to clarify this point:

"…

While the use of the same CN05.1 forcing to generate VIC-simulated runoff and to drive all subsequent models may inflate apparent performance, this design was chosen to isolate the effect of model formulation by eliminating variability in meteorological inputs.

…"

Comment 15: *L 416: This is methods, not results*

Response:

Thank you for your correction. We agree that the description of the VIC-CN05.1 product belongs in

the Methods section rather than in Results. Accordingly, we have moved the sentence "…the runoff data product used in this study was simulated by the VIC model, which uses CN05.1 meteorological data. …" into the Methods section.

Comment 16: *L 420-425: This is discussions, not results*

Response:

Thank you for your correction. These statements should indeed be placed in the Discussion section rather than the Results section. In the revised manuscript, we have made several corresponding adjustments.

We have separated the Results and Discussion into distinct sections. The Results section now focuses solely on descriptive findings, such as the number of basins violating balance under each forcing. In contrast, interpretative content including the implications for water–energy closure and recommendations for future dataset construction has been relocated to the new Discussion section. Additionally, we have expanded the Discussion to provide deeper insights, addressing your earlier observation that our interpretation was too brief. Specifically, we now discuss the critical importance of ensuring mass–energy closure at the watershed scale across diverse climatic regimes. We also explore how calibration strategies, such as multi-objective optimization of flow and energy fluxes, can help reduce balance violations. Furthermore, we try to discuss the specific challenges and solutions for high-altitude, humid basins where snow processes and significant latent heat fluxes play a dominant role. The specific discussion content corresponding to this point is as follows:

"…

The marked decrease in balance-violating basins under CN05.1 forcing highlights the critical role of accurate precipitation and energy inputs in large-sample hydrological datasets. Ensuring mass and energy closure is especially challenging in humid, high-altitude basins where snow accumulation, melt dynamics, and evapotranspiration interact strongly.

To mitigate balance errors, future dataset-building efforts could incorporate multi-objective calibration routines that jointly optimize streamflow, snowmelt timing, and energy fluxes. For high-altitude watersheds, integrating remote-sensing snow cover, station-based radiation corrections, and physically based snowpack models may further improve closure and data fidelity.

…"

Comment 17: *Figure 6: what is the blue shaded area?*

Response:

Thank you for your reminder. The translucent blue band in both panels is intended to highlight the relatively "humid" basins. Specifically, basins where the aridity index (mean PET / mean P) is less than 1.5. In our original caption this feature was not described, which understandably caused confusion. The revised Figure 6 caption as follows:

"…

Figure 6. Water balance for 544 basins, illustrated in a Budyko scheme for ERA5-Land (a) and CN05.1 (b). Markers are coloured by the basin mean elevation. The translucent blue band marks the relatively humid regime (aridity index < 1.5).

…."

Comment 18: *L 433: This is a somehow unfair comparison, as the reference data used to calculate NSE comes from VIC forced by CN05.1. Then, when you compare models forced by ERA5 to these data, you include the error coming from the PBM and the error coming from the input data set.*

Response:

Thank you for your timely correction. You are correct that, by using VIC-CN05.1 as the reference "observed" hydrograph, the NSE computed for models driven by ERA5-Land reflects both the structural error of each model and the mismatch between ERA5-Land and CN05.1 forcings.

To make this explicit, we have corrected this sentence in the revised manuscript and added relevant discussion. The specific content is as follows:

"…

Prediction performance (Nash–Sutcliffe efficiency) of both PBMs was higher when using CN05.1 precipitation than when using ERA5-Land precipitation; however, because CN05.1 was also used to generate the reference VIC-simulated runoff, this apparent improvement includes both model structural error and reduced forcing-mismatch error.

Relevant discussion

It should be noted that, by using VIC-CN05.1 simulations as the reference hydrograph, models

forced with ERA5-Land incur not only structural discrepancies relative to VIC, but also additional error arising from differences between ERA5-Land and CN05.1 inputs. In contrast, models driven by CN05.1 avoid the latter source of error. Consequently, the higher NSE observed under CN05.1 should be interpreted as the combined effect of more consistent meteorological inputs and model structural performance, rather than as a pure indicator of intrinsic model skill. Future work employing independent observed streamflow records will be required to disentangle these two components.

…."

Comment 19 (About Figure 7):

Figure 7, left: what is this scale? It does not include regular intervals between values

Figure 7, caption: the authors state that the colormap include vales from 0 to 1. That would be great, to compare the four maps together. Unfortunately, the left maps do not use the same range as the right maps

Response:

Thank you for your correction on the standardization of Figure 7. In the ERA5-Land-forced runs (left column), several basins yield negative NSE, whereas most CN05.1-forced runs (right column) have NSE ≥ 0. To preserve visibility of poor performance under ERA5-Land, the left maps are plotted over the range [–0.5, 1], while the right maps are restricted to [0, 1].

We have redrawn the legend and revised the caption to make this explicit. The specific contents are as follows:

"…

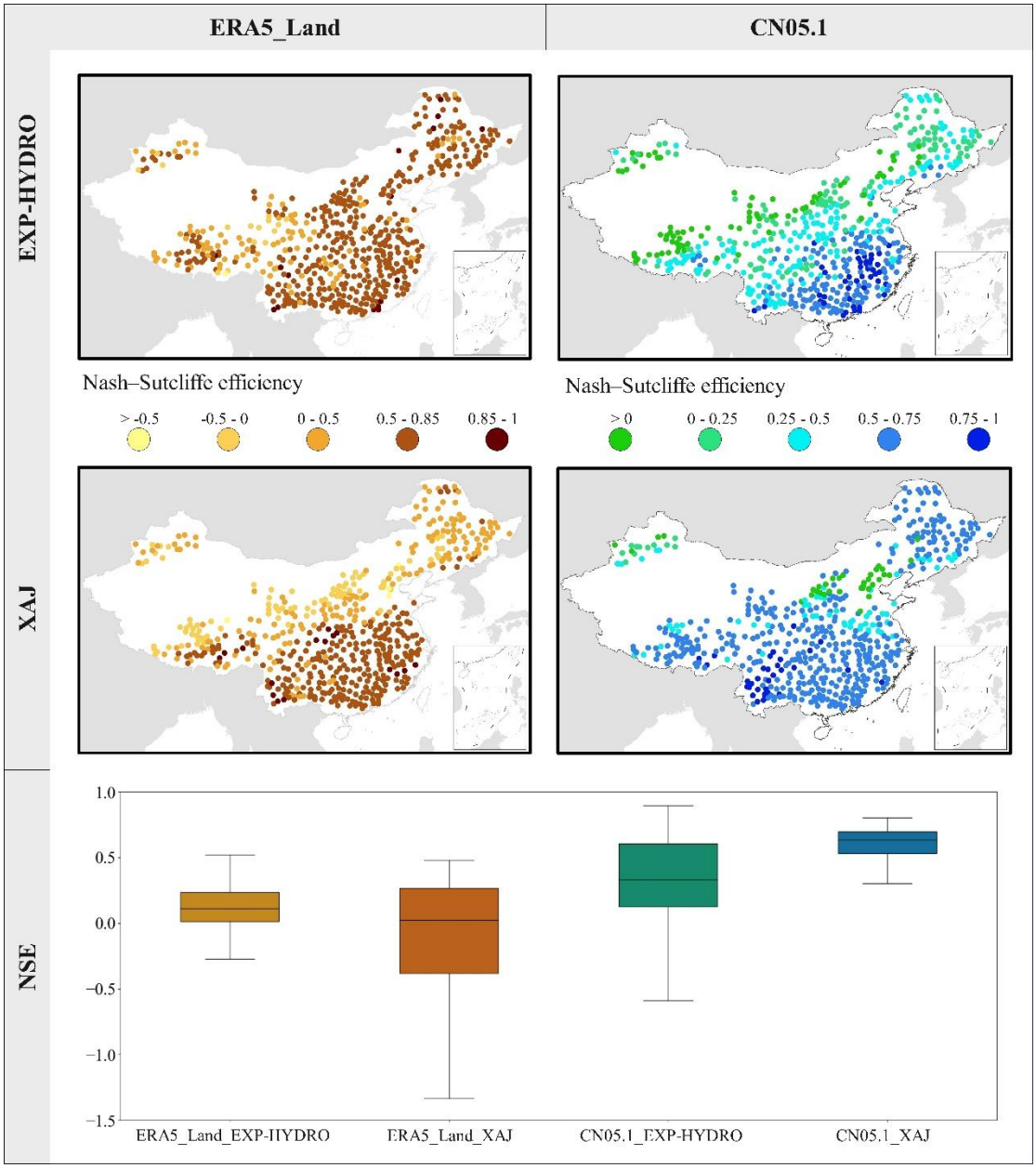

Figure 7. Performance of process-based hydrological models during the testing period (1995.10.1–2015.9.30). Spatial distributions of Nash–Sutcliffe efficiency (NSE) for (top row) EXP-HYDRO and (middle row) Xin'anjiang (XAJ) models under two precipitation forcings: ERA5-Land (left) and CN05.1 (right).

…."

Comment 20: *L 411 and following: The differences should be discussed in terms of what processes are important for these basins and what is the link with the processes present in the PBMs. We need interpretation!*

Response:

Thank you for this insightful suggestion.

It seems there might have been a slight misunderstanding regarding the line number. Based on the order in which you suggested, we guess that you meant line 441. We will take line 441 as an example for the following modification. If this is not correct, please feel free to let us know.

We have expanded the Discussion to link the observed model-selection patterns to key hydrological processes in each basin and to the mechanistic formulations within EXP-HYDRO (EXP) and Xin'an jiang (XAJ). Specifically, we added a new subsection titled "5.x Process-based Drivers of Model Preference" has been added, incorporating the rewritten content from lines 441 and following into this subsection, thereby integrating the new interpretation. The specific contents are as follows:

"…

5.x Process-based drivers of model preference

The differing strengths of EXP-HYDRO (EXP) and Xin'an jiang (XAJ) across China reflect their handling of soil moisture dynamics and runoff generation under distinct climate regimes. In the humid lowlands of the mid–lower Yangtze and southeastern rivers, deep soil moisture storage and vegetation-controlled evapotranspiration govern streamflow seasonality. XAJ's multi-layer soil moisture accounting and temperature-index evapotranspiration routine effectively reproduce the gradual release of baseflow and ET peaks, yielding consistently higher skill than EXP regardless of precipitation source. Conversely, in more arid northern basins such as the Haihe River system, rapid runoff via infiltration-excess and shallow subsurface flow is dominant. Here, EXP's explicit representation of infiltration, percolation, and evaporation processes aligns more closely with observed quick-flow responses, so EXP outperforms XAJ even when the forcing dataset changes.

Although EXP includes energy-balance routines for snowmelt, XAJ does not perform rain–snow separation but directly routes total precipitation into its soil and runoff modules. As a result, XAJ's robustness in snow-affected catchments stems from its simpler reliance on temperature-driven ET and multilayer storage, rather than on explicit snow physics. These patterns highlight that model preference depends not only on climatic setting—humid versus arid, lowland versus montane—but also on whether a model's structural emphasis (multilayer soil storage versus explicit infiltration-excess) matches the basin's dominant hydrological processes.

…."

Comment 21: *L 469: The fact that the LSTM performs very well with CN05.1 comes from the fact that the authors do not try to reproduce observed streamflow but simulated streamflow. This means that LSTM does not excels in reproducing the processes leading to streamflow from meteorological input, but rather excels in mimicking the behavior of the VIC model. This is highly different and is caused by the experiment setup. In addition, this might indicate that the LSTM cannot cope with input errors*

Response:

Thank you for highlighting this important point. Your suggestion is correct and professional. Because the LSTM is trained to reproduce the VIC-CN05.1 simulated runoff, its high NSE under CN05.1 forcing largely reflects its ability to mimic VIC's behavior rather than to reconstruct the true physical processes leading to streamflow. This experimental setup therefore inflates apparent LSTM performance and masks its sensitivity to input errors.

We have revised this sentence and added a corresponding explanation to the Discussion section to clarify that the LSTM skill reported under CN05.1 is conditional on the simulated target and underscore the need for independent validation against real observations. The specific contents are as follows:

"…

When using CN05.1 precipitation data, the median NSE for LSTM in regional modeling and PUB reached 0.95 and 0.93, respectively. However, because the target hydrographs are themselves VIC-CN05.1 simulations, these high values primarily indicate LSTM's capacity to emulate the VIC model outputs, rather than its standalone process-learning skill.

…

…

Relevant discussion:

5.x Implications of learning from simulated targets

The exceptional NSE achieved by LSTM under CN05.1 forcing arises from training the network on VIC-simulated runoff. While this demonstrates the LSTM's flexibility in capturing the input–output mapping of a given process model, it does not necessarily imply proficiency in learning the underlying physics of runoff generation. Moreover, this setup can obscure the LSTM's

vulnerability to input biases. When driven by ERA5-Land, which differs more substantially from the VIC-CN05.1 climate statistics, the LSTM performance declines markedly, revealing its dependence on consistent forcing. To assess true hydrological generalization, future work should train and evaluate LSTM models against independent observed streamflow records and beyond the bounds of a single process model's behavior.

…."

Comment 22: *L 488-491: these are discussions, not results*

Response:

Thank you for noting that these statements are interpretative rather than strictly results. In the revised manuscript, we have taken your advice to clearly separate descriptive results from interpretative discussions by restructuring the sections. The "Results" and "Discussion" are now presented as two distinct sections. Additionally, all interpretative sentences, such as "This not only influences the … model input data," have been moved to the Discussion section.

In response to earlier feedback regarding the sparse discussion, we have enriched this part by elaborating on how the quality of input data and differences in samples affect both process representation and model transferability, as detailed in the new subsection 5.x:

"…

5.x Influence of Forcing Quality on Model Generalization

The accuracy and spatial consistency of meteorological forcing critically shape hydrological model performance and their ability to generalize. When inputs faithfully represent orographic precipitation patterns and energy fluxes—as in CN05.1—both process-based and data-driven models reproduce runoff dynamics more reliably. Conversely, mismatches or biases in precipitation phase, timing, or intensity introduce systematic errors that propagate through model components, degrading skill and transferability across basins. Future large-sample studies should therefore not only ensure balanced sampling of hydroclimatic regimes but also rigorously assess and, where possible, correct input data quality before model calibration and comparison.

…."

Comment 23: *Figure 9, 10: random scales prevent from comparing the different parts of the figure*

Response:

Thank you for noting that the use of different axis and colorbar limits can make cross-panel comparison difficult. In our original figures, each row's scale was chosen to best display the full spread of NSE values or density peaks for that particular model, but we recognize this hamper direct visual comparison across models and forcings. We have restructured the figure legends and made them clear in the captions so that readers can interpret each panel correctly. The revised Figures 9 and 10 and their captions are as follows:

"…

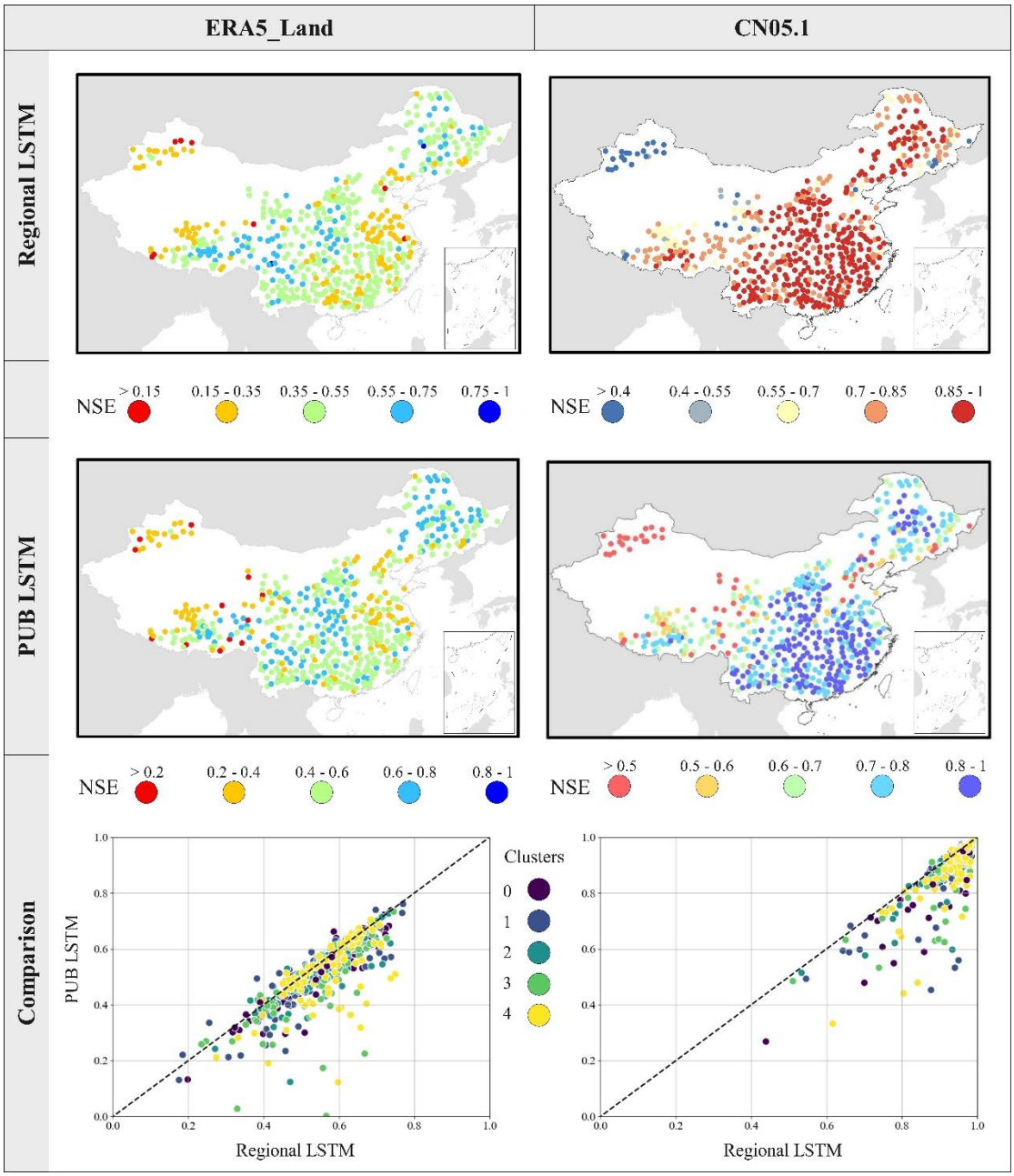

Figure 9. Performance of LSTM models using ERA5 (left column) and CN05.1 (right

column) precipitation during regional modeling and PUB testing. Top row (maps): Spatial distribution of NSE during regional modeling. Middle row (maps): Spatial distribution of NSE during PUB testing. Bottom row (scatter): Basin-by-basin comparison of PUB NSE (vertical axis) vs. regional NSE (horizontal axis). Points are colored by clusters. The axes both span [0, 1].

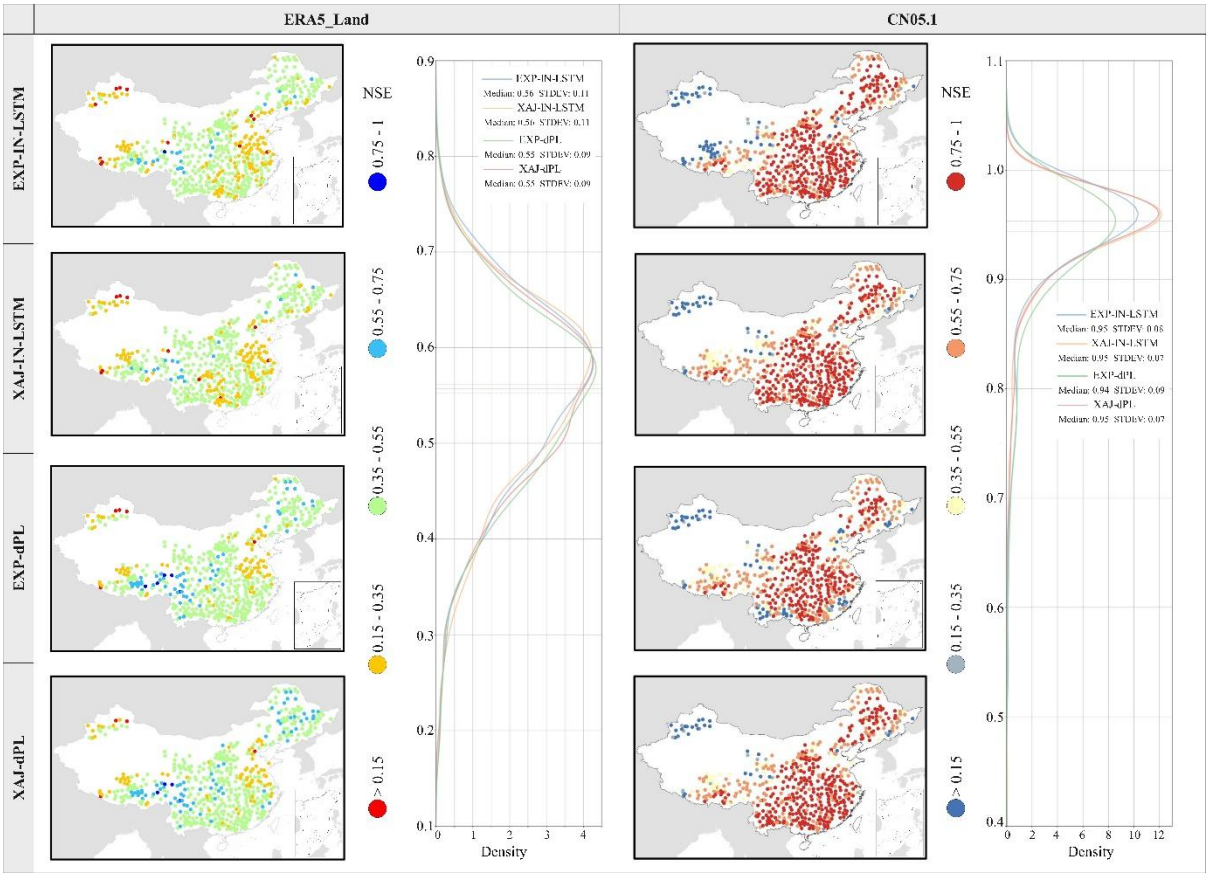

Figure 10. Four hybrid models prediction performances using ERA5-Land (left) and CN05.1 (right) precipitation data. Spatial maps: Colorbars cover ranges of NSE values for different models. Density plots: x-axes display the NSE values; y-axes are density.

…."

Comment 24: *L 528-537: these are discussions, not results*

Response:

Thank you for noting that the explanations of EXP-HYDRO's snow handling and EXP-dPL's suitability in the Qinghai–Tibet Plateau are interpretative rather than strictly "results." In the revised manuscript, we have made adjustments. We have separated the Results and Discussion sections. All descriptive performance metrics, including NSE values and model rankings, now reside in the Results section. On the other hand, interpretative statements such as the reasons behind EXP-

HYDRO's improved rain–snow partitioning and its positive impact on snow-affected runoff, as well as the factors contributing to EXP-dPL's performance in high-altitude regions have been relocated to a new subsection in the Discussion. Furthermore, we have deepened the Discussion by expanding on the interpretation of snow processes. We explain how EXP-HYDRO's energy-balance snowmelt and partitioning routines effectively capture storage and melt dynamics. Additionally, we highlight how the improved accuracy of precipitation phase in CN05.1 enhances these effects. Our discussion also addresses how EXP-dPL's differentiable parameter channel is tailored to the unique energy and precipitation regimes of the Qinghai–Tibet Plateau, enabling it to outperform other models when driven by CN05.1 forcing. The specific content is as follows:

"…

5.x Snow-process representation and high-altitude performance

EXP-HYDRO's explicit rain–snow separation and energy-balance snowmelt modules store winter snowfall and release it based on temperature changes, yielding a more accurate runoff response in snow-dominated basins. CN05.1's station-interpolated precipitation better resolves snowfall events and snow–rain transitions than ERA5-Land, which explains EXP-HYDRO's improved NSE under CN05.1 forcing.

In the high-altitude Qinghai–Tibet Plateau (climate region 6), EXP-dPL further benefits from its differentiable parameterization channel: by learning spatially varying thermal degree-day factor and other relevant hydrological parameters directly from static attributes and CN05.1 inputs, it dynamically tailors its snow and runoff routines to local conditions. This flexibility leads to superior performance in this challenging environment compared to fixed-parameter PBMs.

…."

Comment 25: *Figure 12, 13: fonts are too small, we cannot read*

Response:

Thank you for your reminder. We have increased the font sizes in both Figures 12 and 13 (including axis labels, tick labels, legends, and colorbar annotations) to ensure their readability. All text elements now use at least an 8 pt font. At the same time, we carefully checked the font sizes of other figures in the manuscript and adjusted the fonts that were too small. We hope these adjustments can address your concern.

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:

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21, 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

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

Lu, M., Sun, H., Yang, Y., Xue, J., Ling, H., Zhang, H., and Zhang, W., 2025. Assessing recovery time of ecosystems in China: insights into flash drought impacts on gross primary productivity, Hydrol. Earth Syst. Sci., 29, 613–625, https://doi.org/10.5194/hess-29-613-2025.

Sang, S., Li, Y., Hou, C., Zi, S., and Lin, H., 2025. The interprovincial green water flow in China and its teleconnected effects on the social economy, Hydrol. Earth Syst. Sci., 29, 67–84, https://doi.org/10.5194/hess-29-67-2025.

Xiong, J., Guo, S., Abhishek, Yin, J., Xu, C., Wang, J., and Guo, J., 2024. Variation and attribution of probable maximum precipitation of China using a high-resolution dataset in a changing climate, Hydrol. Earth Syst. Sci., 28, 1873–1895, https://doi.org/10.5194/hess-28-1873-2024.

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.

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.

Ren-Jun, Z., 1992. The Xinanjiang model applied in China. J. Hydrol. 135, 371–381. https://doi.org/10.1016/0022-1694(92)90096-E