Author comments on RC3- egusphere-2025-1161

Comment 1 (Major comment 1):

The first major point is the fact that the target variable is not observed data, but an output from VIC. While I understand that this might be required for confidentiality reasons, it also means that the results cannot be trusted or generalized elsewhere. This is because results are compared to the outputs of VIC and as such, the models are rewarded if they emulate VIC, rather than simulate real streamflow. And since VIC has its own biases, strengths and weaknesses, we are simply evaluating the ability of these new modeling techniques to emulate the same biases. To make this point, I contend that we could obtain NSE values of 1 if we simply used VIC as one of the models in this study. Would it mean that VIC is much better than other models? Of course not, and the same is true with the relationship between these new models and the VIC outputs. The same goes for internal variable analysis: Models that are more similar to VIC will perform better. The problem with this whole approach is that we cannot learn from results vs application in the real world, because the response surface of the optimization problem is much much smoother and easier to navigate than one using real observations which are uncertain and error-prone. Models will never perform as well on real data than on these synthetic data. Therefore I think this study has a very limited reach and usefulness while using synthetic streamflow data.

Response:

Thank you very much for your critical and insightful comment. We fully agree that using simulated streamflow (from VIC) instead of observed discharge as the modeling target introduces limitations regarding the realism, uncertainty, and generalizability of the results. This is indeed a non-negligible concern, and we would like to clarify the context, motivations, and limitations of our study in this regard.

First, as the reviewer correctly pointed out, the target variable in this study is derived from a process-based hydrological model (VIC). This choice was made due to the unavailability or confidentiality of observed streamflow data across many basins in the study region. However, the goal of this study was not to evaluate the absolute predictive accuracy of various models against ground truth, but rather to assess the relative behavior and compatibility of different modeling frameworks (process-

based, data-driven, and hybrid models) when exposed to consistent meteorological forcings and boundary conditions.

Second, we fully acknowledge that VIC has its own set of biases and limitations. To mitigate the risk of models merely learning the idiosyncrasies of VIC, we conducted additional experiments using two distinct meteorological forcings (ERA5-Land and CN05.1), and different hydrological models (EXP-HYDRO and XAJ) in hybrid configurations. These strategies were intended to reduce overfitting to a single model structure and better evaluate model robustness across conditions. In other words, we can actually choose other runoff data products, but the time span and resolution of VIC-CN05.1 can meet the research requirements. At the same time, ERA5-Land and CN05.1 are also meteorological data sets that have been used in many studies.

Third, we agree that models trained on synthetic data tend to achieve better scores due to the smoother, less noisy response surface, as noted by the reviewer. However, we emphasize that the methodological contribution of this study lies in exploring the behavior and potential of hybrid model architectures under a controlled environment. We see this study as a stepping stone—a preliminary effort to benchmark different hybridization strategies under reproducible settings before applying them to real-world observations.

Finally, in light of your helpful suggestion, we have added a paragraph in the Conclusion section to explicitly acknowledge this limitation and clarify the scope of applicability of our results. We also emphasize that the findings should not be interpreted as absolute model rankings but as comparative behaviors under VIC-forced hydrological settings.

The new specific contents are as follows:

"…

It is important to note that the runoff data product used in this study is the runoff data simulated by the VIC model, rather than the runoff data observed by real hydrological instruments. While this enables large-scale, consistent benchmarking across data-scarce basins, it also introduces model-induced biases and a potentially smoother response surface. Therefore, the reported predictive performances should be interpreted with caution and not taken as absolute measures of model accuracy. Future work will focus on validating the models against observed streamflow where available, and further investigating their transferability to real-world applications.

…"

Comment 2 (Major comment 2): *The second major issue is linked to the previous one, and that is the use of CN05.1 as the input data to VIC, which is then used again as an input in the other models. This means that any model using CN05.1 will most likely perform better than another using the ERA5 dataset, simply because the processes are artificial and conditioned on the use of CN05.1. In the study, there are a few sections commenting on how CN05.1 performs better than ERA5 (ex lines 467-469: "When using CN05.1 precipitation data, the median NSE for LSTM in regional modeling and PUB reached an impressive 0.95 and 0.93, respectively."). These results, while contextualized by following that this stems from its use in VIC, are still give the impression that CN05.1 is better than ERA5, which is not true given the evaluation method presented here. This issue would not arise if VIC was not used at all (as per my point #1 above), but if the authors decide to continue using VIC for a revised version of this paper, they need to simply remove CN05.1 as one of the datasets in the comparison to be fully independent. Doing so would at least allow simplifying the paper enough that they could then delve into the analysis of internal variables of the hybrid models, which seems to me as a key advantage but that is not discussed or evaluated in the present paper.*

Response:

Thank you for your thoughtful critique. We fully agree that using the CN05.1 product both as a forcing input to VIC and as an input to other models in the current study introduces a potential bias. Specifically, this setup may favor models driven by CN05.1 data, as they are inherently more consistent with the streamflow outputs generated by VIC, thus compromising the independence between the training data and the target variable.

Your observation is highly relevant and correctly points out that any apparent performance advantage of CN05.1 over ERA5-Land in this context does not necessarily reflect its superiority as a meteorological product, but rather its compatibility with the VIC-generated target streamflow. We have revised the manuscript to clarify this critical point and to ensure that no misleading interpretation is conveyed.

In particular:

In related section, we now explicitly state that the observed performance advantage of CN05.1-driven models stems in large part from the fact that CN05.1 was also used to force VIC in

generating the target runoff data.

We have softened or removed language suggesting that CN05.1 is categorically better than ERA5-Land and have reframed our results in terms of consistency with the VIC simulation rather than absolute model accuracy.

A cautionary statement has been added in the Discussion section emphasizing that these results do not translate directly to real-world predictive skill and must be interpreted within the context of a VIC-forced synthetic experiment.

As for your suggestion to remove CN05.1 entirely in order to preserve independence between input and output, we acknowledge that this would yield a cleaner comparison. However, we also believe that retaining CN05.1 offers an opportunity to test model robustness under two distinct meteorological inputs—one highly consistent with the target and one more independent (ERA5-Land). By presenting both cases and transparently communicating the differences, we aim to provide a fuller picture of the model behaviors under realistic data conditions. Still, we understand the trade-off and plan to conduct further experiments in future work using observational data and independently generated targets to address this concern more completely.

We are also grateful for your suggestion to simplify the comparative analysis and expand the focus on internal hydrological variables, which is a unique advantage of differentiable hybrid models. In the revised manuscript, we have added Section 4.6, which explores intermediate outputs such as soil moisture, snowpack, and evaporation, and compares them with ERA5-Land datasets. This addition directly supports the interpretability and process fidelity of the hybrid models, in line with your suggestion.

The details of the newly added discussion are as follows:

"…

Due to the lack of long-term large-scale observed runoff records with consistent quality and coverage over China, this study uses runoff data products generated by the VIC model driven by the CN05.1 forcing as a proxy for observed runoff. This approach inevitably leads to a tight coupling between the input meteorological dataset (CN05.1) and the target runoff data, which may influence model evaluation results. Specifically, models using CN05.1 as input tend to achieve higher simulation accuracy compared to those driven by ERA5-Land, not necessarily due to the intrinsic superiority of the CN05.1 dataset, but because of its consistency with the target runoff data source.

This methodological limitation does not aim to suggest that CN05.1 is inherently better than ERA5-Land in hydrological prediction tasks. Rather, it highlights the challenge of conducting large-sample modeling research under data-scarce conditions. Future studies based on observed runoff data or runoff derived from independent forcing–target pairings will be essential to further evaluate the generalization and robustness of the models across different input conditions.

…"

Comment 3 (Major comment 3): *This issue is related to the way the deep learning and hybrid models are trained, and has two distinct sub-issues. We often see this from hydrologists that work with deep learning models for the first time and is a common mistake that is easy to make but has important consequences. The first is the fact that the authors have only 2 periods of data: Training (or calibration; 1975-1995) and a testing period (1995-2015). While adequate for PBM calibration, this is simply unacceptable for Deep-learning models. These models need 3 periods of data: Training, Validation, and Testing. Training is used on the forward pass and backpropagation steps to tune the weights and parameters according to the chosen gradient descent method. The model is then evaluated on the Validation period after each epoch, and the objective function score is computed on that period specifically. The model training is then stopped when the validation period loss stops improving and starts regressing. Finally, when the model has stopped improving and the training is stopped, then the model performance is evaluated on the third, independent Testing period. Failing to stop training will inevitably lead to overfitting and unreliable results, which is the case here. Therefore, a revised study should follow best practices and add an independent testing period for the deep learning models and also the PBM which should share the same testing period. I also note that there are no details on the selected objective function for training the LSTMs nor do we know how this was computed for the regional models? How are error/loss metrics calculated on multiple basins at the same time? A few studies proposed some methods to do this, for example just in HESS see Kratzert et al (2019) and Arsenault et al. (2023) listed below. The second sub-point is that the authors state: "All input data are normalized before training" without further details on how this was done. This is critical, because the data need to preserve independence between the [Training] and [Validation; Testing] periods. Data need to be normalized using a scaler of some sort (which one was used?) using the training period data, and then the scaler is applied to the*

validation and testing data. Failing to do so means that the testing period data are included in the scaler and as such the training will benefit from knowing the scale of data it can expect to get. This is called data contamination and needs to be avoided. Nothing in the paper at this stage seems to suggest this was performed. As such, I believe the results are flawed and performance is overestimated in this study.

Response:

Thank you very much for pointing out the importance of data splitting and normalization strategies in deep learning model training. We sincerely acknowledge your concern regarding the potential risk of overfitting and data contamination. Although our manuscript originally stated that the modeling period was divided into a training period (1975–1995) and a testing period (1995–2015), we would like to clarify the following points:

1. Normalization Strategy: All input features were normalized using the mean and standard deviation computed solely from the training period (1975–1995). This ensures that no statistical information from the testing period was introduced during model training or data preprocessing. Therefore, the issue of data leakage or contamination is avoided.
2. Training and Early Stopping: While we did not explicitly set aside a separate validation period, an early stopping mechanism was implemented based on the loss trend during training. Specifically, training was stopped if the loss did not improve for a specified number of epochs, which helps prevent overfitting. Although a formal validation split was not used, this strategy has been shown effective in previous large-sample hydrology studies. We will clarify this point in the revised manuscript.
3. Model Generalization: The testing period (1995–2015) remained completely independent throughout model development and was only used for final performance evaluation. This helps ensure a fair assessment of the model's generalization capability.
4. Loss Function and Regional Training: The loss function used for LSTM training was the average Nash efficiency coefficient (NSE). In regional training, the loss was computed by averaging NSE across all basins in each batch, following a similar strategy as outlined in Kratzert et al. (2019) and Arsenault et al. (2023). We will provide additional details in the revised version to enhance clarity and reproducibility.

The added explanations are as follows:

"…

To ensure proper training of deep learning models while avoiding data leakage, all input features were normalized using the mean and standard deviation calculated exclusively from the training period (1975–1995). This normalization strategy was consistently applied to both the training and testing periods, thereby preventing any use of future information and ensuring the independence of the testing dataset. Although the dataset was divided into two main temporal segments—training (1975–1995) and testing (1995–2015). An early stopping mechanism was implemented during training to mitigate the risk of overfitting. Specifically, the training process was halted when the model loss failed to improve for a predefined number of epochs, based on monitoring the training loss trend. This approach has been adopted in prior large-sample hydrology studies where formal validation sets are not always feasible. The objective function used for training the LSTM-based models was the average Nash efficiency coefficient (NSE) between predicted and target runoff values. In the case of regional modeling, the loss was computed by averaging the NSE across all basins within each training batch, allowing the model to generalize across catchments with heterogeneous hydrological behaviors.

…"

Comment 4: *Line 32-33: This seems trivial that better inputs will lead to better modelling. I would remove.*

Response:

Thank you for your comment. We agree that the statement as originally written may appear too self-evident or generic. Our original intention was to briefly highlight that differences in precipitation data quality can influence the reliability of model outputs, especially in regions where precipitation variability is a dominant driver of runoff processes.

To address your concern, we have removed the sentence from the abstract.

Comment 5: *Line 33: At this stage, readers don't know what hybrid modelling refers to. Please add a few key details to set the table, maybe a 1-sentence description.*

Response:

Thank you for your reminder. We agree that the term "hybrid modeling" may be unclear to readers

at this early stage of the manuscript. To improve clarity, we have revised the sentence in the abstract by briefly explaining the core idea of the hybrid model. The modified contents are as follows:

"…

By combining process-based model with data-driven deep learning methods, the hybrid modeling approach achieves regional modeling capabilities comparable to those of the standalone LSTM model.

…"

Comment 6: *Line 76: "solve" is a strong word. Perhaps "help address"?.*

Response:

Thank you for pointing this out. We agree that "solve" may overstate the role of our benchmark in addressing the runoff prediction problem. Following your suggestion, we have revised the sentence to use a more appropriate and balanced expression. The updated sentence now reads:

"…

This benchmark can assist in improving relevant hydrological models to help address the runoff prediction problem in China and in catchments with similar basin conditions.

…"

Comment 7: *Line 86-88: This is also true for deep learning models (perhaps even more so than PBMs!)*

Response:

Thank you for your thoughtful reminder. We fully agree that deep learning models also require large volumes of data and careful configuration. However, the intention of our original sentence was to highlight a commonly recognized limitation of process-based models: their dependency on high-quality forcing data and the often subjective or complex nature of parameter calibration, which may hinder their large-scale or cross-regional applicability. To avoid misunderstanding while preserving this point, we have rephrased the sentence to better reflect our intent. The updated sentence now reads:

"…

When applied to a single basin, process-based hydrological models often rely on high-quality

forcing data and require subjective and complex parameterization procedures, which can limit their applicability and accuracy, particularly in data-sparse regions.

…"

Comment 8: *Lines 90-91: "LSTMs... can effectively capture nonlinear relationships...": so do PBMs, depending on the structure. The difference really lies in the model learning the relationships between weather and flows without humans providing any physical sense.*

Response:

Thank you for your insightful comment. We agree that process-based models can also represent nonlinear hydrological processes depending on their structure. Our intention was not to suggest otherwise, but rather to emphasize that LSTM models can automatically learn the relationship between inputs and outputs from data, without the need for explicit physical formulations or manual parameterization. To clarify this point, we have revised the sentence accordingly.

The updated sentence now reads:

"…

In contrast, deep learning models, such as long short-term memory networks (LSTM), can automatically learn the relationship between meteorological inputs and streamflow responses from data, without relying on explicit physical equations or manually calibrated parameters.

…"

Comment 9: *Lines 97-98: redundant sentence with the previous.*

Response:

Thank you for pointing this out. We agree that the sentence is redundant. We have removed the repetitive statement.

Comment 10: *Lines 127-130: Would need a bit more details. Are the equations of the model kept as-is? Is it an emulator of the PBM? Do the parameters preserve the same meaning?*

Response:

Thanks for your comment. We agree that additional clarification is necessary regarding the structure and characteristics of the differentiable hybrid model. In our approach, the core hydrological

equations of the process-based models are retained in their original form and embedded within a differentiable framework. Therefore, the model is not an emulator, but a fully integrated differentiable physical model. The key innovation lies in the neural parameterization, where a neural network learns to generate model parameters from static basin attributes, and these parameters are optimized via backpropagation based on daily runoff prediction errors. Importantly, the physical meaning of the parameters is preserved. The neural network serves only as a mapping function to infer parameter values rather than changing their definitions or functions within the governing equations.

We have revised the corresponding paragraph in the manuscript to make these points clearer. The modified contents are as follows:

"…

Specifically, this hybrid modeling approach retains the original physical equations of the process-based model and embeds them within a differentiable architecture. A neural network serves as a parameter generator, mapping static catchment attributes to the model parameters, which retain their physical meanings. During training, model parameters are optimized through gradient-based backpropagation using daily runoff prediction errors. This approach allows the model to preserve the interpretability and physical constraints of process-based modeling while improving performance through data-driven learning.

…"

Comment 11: *Line 214: Any reason why ERA5-Land is not used? Should be better/more precise especially in mountainous areas?*

Response:

Thank you very much for your sharp observation. We appreciate your suggestion regarding the use of ERA5-Land, and we apologize for the lack of clarity in the original manuscript. In fact, we did use the ERA5-Land dataset in our experiments. The meteorological variables were extracted and clipped using the ERA5-Land daily aggregated data products available on the Google Earth Engine platform (https://developers.google.com/earthengine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR).

However, due to a terminological oversight, the manuscript referred to this data source simply as "ERA5" throughout, which may have caused confusion.

In the previous version of this manuscript, we referred to this data source as "ERA5" for brevity, but to avoid confusion, we now clarify that ERA5-Land is the precise dataset used in this study.

We have thoroughly revised the manuscript accordingly to explicitly clarify that ERA5-Land was used, and corrected all terminology throughout the relevant sections.

The sources of meteorological data are described as follows:

"…

The meteorological data used in this study were sourced from the ERA5-Land dataset and the CN05.1 dataset (Gao et al., 2013). ERA5-Land provides daily aggregated surface meteorological variables at high spatial resolution, and the data were obtained and clipped using the Google Earth Engine platform (https://developers.google.com/earthengine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR).

…"

Comment 12: *Line 230: "provided by the originates" : missing word here.*

Response:

Thank you for pointing this out. We acknowledge the grammatical error in this sentence. The phrase "provided by the originates" was the result of an editing oversight. We have corrected the sentence in the revised manuscript for clarity. The updated sentence now reads:

"…

The runoff data used in this study originate from the VIC-CN05.1 dataset

…"

Comment 13: *Line 258 (and multiple others): many times the word "relatively" is used to tone down some element. I suggest rephrasing to say that they are accessible or some other word that would be more precise. Same goes everywhere.*

Response:

Thank you for pointing out the frequent use of the word "relatively" in our manuscript. We agree

that its repeated usage may weaken the clarity and precision of our statements. Following your suggestion, we have revised these expressions throughout the manuscript to adopt more accurate and assertive terms such as “accessible,” “available,” or “widely used,” depending on the context. The sentence in Line 258 has been revised to:

“…

Instead, we aim to utilize accessible datasets to evaluate the performance of different models in China.

…”

Comment 14: *Line 295: all process based models follow this law of water balance, I would remove.*

Response:

Thank you for your valuable comment. We agree that stating a model follows the law of water balance is indeed common for all process-based models and may be redundant in manuscript. In response to your suggestion, we have removed the phrase to avoid unnecessary repetition and improve the conciseness of the description. The revised sentence now reads:

“…

EXP-HYDRO is a conceptual hydrological model that operates on a daily time step

…”

Comment 15: *Line 300-320: Xin'anjiang does not model snow processes? How is it used in mountainous and other basins where snow is present?*

Response:

Thank you for Thank you for your thoughtful question. You are correct in pointing out that the original Xin’an jiang model does not include explicit representation of snow processes such as snow accumulation or melt. In our study, although the 544 selected basins span a wide range of climatic and topographic conditions—including some high-elevation regions where snowfall may occur—we used a simplified version of the Xin’an jiang model without snow modules for the following reasons:

Focus on model structure comparison: The simplified Xin’anjiang model serves as a consistent and interpretable baseline to compare with other process-based and hybrid models. We aimed

to examine how different modeling structures perform under the same forcing and evaluation settings.

Scope of model generalization: While the absence of an explicit snow module may reduce simulation accuracy in snow-dominated basins, our aim was not to develop a specialized snow model, but rather to test the general applicability and flexibility of the hybrid framework.

We have added a clarification in the revised manuscript as follows:

"…

It should be noted that the simplified Xin'anjiang model used in this study does not include a dedicated snow module. While some high-altitude basins in the dataset may be affected by seasonal snow, the hybrid modeling scheme allows the data-driven components to implicitly capture snow-related dynamics from meteorological inputs.

…"

Comment 16: *Line 349: This is quite high initial learning rate. What is the learning rate decay rate or function? Also, what is the objective function used? What is the model training patience for the stopping criterion? is there a stopping criterion or are all runs leading to 150 iterations? If not, at 150 training iterations, the model will definitely be overfitting and providing poor results compared to a well-tuned model.*

Response:

Thank you for your detailed comment. We used the Nash–Sutcliffe Efficiency (NSE) as the loss function to directly optimize the performance metric relevant to hydrological modeling. Model parameters were optimized using the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.01. To ensure stable convergence and avoid overfitting, a convergence-based stopping rule was adopted. Specifically, training was terminated early when the absolute difference in NSE between two consecutive epochs was less than 0.001. This served as an effective early stopping criterion. Although the maximum number of training iterations was set to 150, the majority of model runs converged earlier based on this rule.

We have clarified these methodological details in the revised manuscript to avoid misunderstanding and ensure reproducibility. The modified contents are as follows:

"…

The deep neural network adopts the Adam optimizer (Kingma & Ba, 2014) to update both the parameters of the network and hydrological parameters. The initial learning rate is set to 0.01, and the Nash–Sutcliffe Efficiency (NSE) is used as the loss function during training. A convergence-based early stopping criterion is applied: training is terminated when the absolute difference in NSE between two consecutive epochs is less than 0.001. Although the maximum number of training iterations is set to 150, most models converge earlier according to this rule. This strategy ensures efficient training while mitigating the risk of overfitting.

…"

Comment 17: *Line 363: capture nonlinear relationships that evade the physics depicted in the PBMs*

Response:

Thank you for your suggestion. We agree with your point that the nonlinear relationships learned by LSTM in the hybrid framework may capture patterns that are not explicitly represented in the physics of traditional process-based models (PBMs). We have revised the sentence accordingly to better reflect this idea, as shown below:

"…

This approach allows the alternative hybrid modeling method to retain the explanatory power of the physical mechanisms inherent in the process model while leveraging LSTM's capability to effectively capture nonlinear relationships that evade the physics depicted in PBMs, thereby compensating for the limitations of PBMs in large-sample hydrological datasets and complex basins.

…"

Comment 18: *Figure 4: PMB is used throughout the study, I would change PBHM to PBM*

Response:

Thank you for your careful review. We acknowledge the inconsistency in terminology. As you pointed out, the acronym "PBM" (Process-Based Model) is used throughout the manuscript, so we have revised "PBHM" to "PBM" in Figure 4 to maintain consistency. The corrected Figure 4 is as follows:

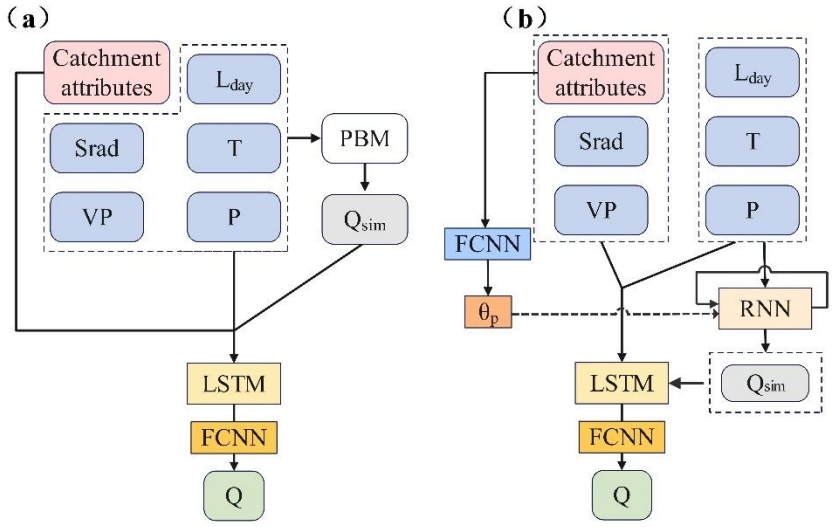

Figure 4. The structure of the hybrid hydrological models.

Comment 19: *Lines 395-397: It seems to me that CN05.1 is more evenly distributed but has more lower extremes, opposed to what is written here.*

Response:

Thank you very much for your reminder. We agree that the original description was not entirely accurate. Upon re-examination of Figure 5, it is evident that CN05.1 precipitation data appear more uniformly distributed spatially but indeed show a concentration of lower precipitation extremes compared to ERA5-Land, which displays a wider range including both wetter and drier basins. We have revised the corresponding sentence in the manuscript to better reflect this observation. The revised sentence reads:

"…

The precipitation data from ERA5-Land exhibit a broader spatial variability, with some basins showing extremely wet or dry conditions, while CN05.1 data appear more spatially uniform but tend to show more frequent lower precipitation values.

…"

Comment 20: *Line 409: water-heat? perhaps mass-energy?*

Response:

Thank you for pointing this out. We agree that the term "water-heat balance" may be imprecise in this context. Since the discussion pertains to the consistency between precipitation, runoff, and evapotranspiration, it is more appropriate to refer to the mass balance (the water balance) rather than energy or heat balance. To avoid confusion, we have revised the sentence accordingly. The revised sentence reads:

"…

When using the same runoff and evapotranspiration data, the precipitation data provided by ERA5-Land resulted in more basins (111) exhibiting significant deviations from the water balance."

…"

Comment 21: *Figure 8: this figure has 2 panel "b". Also this figure needs more details in the legends and caption to fully understand, it is unclear. Add details to what each panel refers to / is presenting.*

Response:

Thank you very much for pointing this out. We acknowledge the error in panel labeling, as there are indeed two sub-figures labeled as "(b)" in the original version of Figure 8. We have corrected this mistake and revised the labels accordingly. Additionally, to address the concern about clarity, we have updated the figure caption and improved the legend descriptions to provide more detailed information about what each panel represents. Specifically, the revised caption now clearly states that:

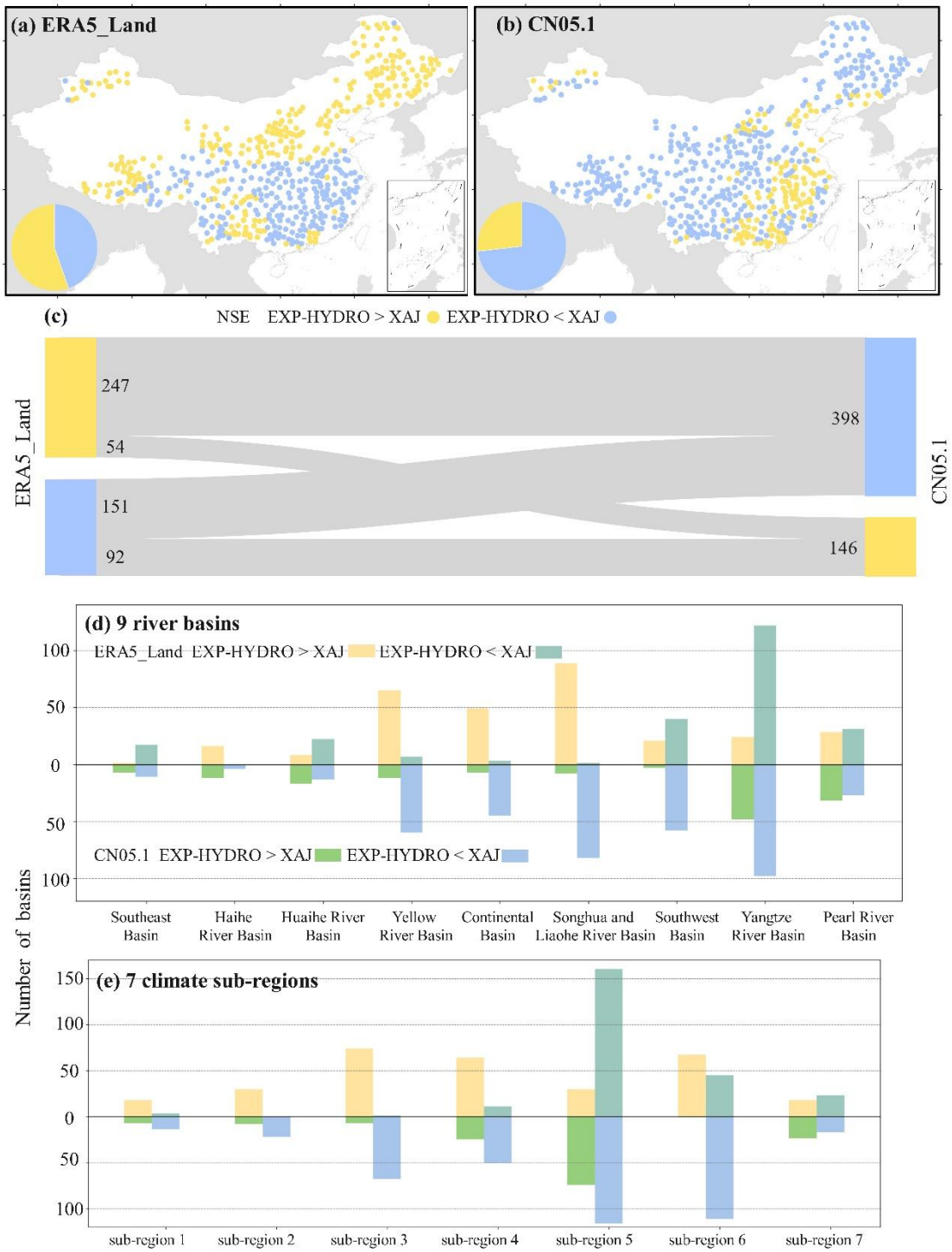

Figure 8. Comparison of PBMs (EXP-HYDRO and Xin'an jiang model) performances in different basins using ERA5-Land and CN05.1 precipitation data. (a–b) Spatial distribution of basins where EXP-HYDRO outperforms (yellow) or underperforms (blue) the Xin'an jiang model under ERA5-Land (a) and CN05.1 (b) forcing. (c) Sankey diagram showing the consistency of best-performing PBMs across datasets. (d) Number of basins where EXP-HYDRO performs better or worse than Xin'anjiang model in each of the 9 major river basins under both forcings. (e)

Number of basins where EXP-HYDRO performs better or worse than Xin'an jiang model in each of the 7 climate sub-regions under both forcings.

…"

Comment 22: *Line 467: These two references do not support he statement that NSE>=0.55 is good. Knoben says 0.50, Newman says it means the model has some skill, and NSE=0.8 shows reasonably good performance. Please clarify.*

Response:

Thank you for your correction. We agree that our previous interpretation of the cited studies was too strong. As clarified in Newman et al. (2015), an NSE of 0.55 is considered to indicate some model skill, whereas NSE = 0.8 reflects reasonably good performance. We have revised the manuscript to more accurately reflect this distinction. The sentence has been modified to:

"…

Specifically, when using ERA5-Land precipitation data, the median NSE of LSTM across 544 basins reaches 0.57, which suggests that the model demonstrates some skill in most basins, according to the evaluation threshold proposed by Newman et al. (2015).

…"

Comment 23: *Line 534-537: Indeed, since XAJ does not have snow process representation?*

Response:

Thank you for your comment. Indeed, the Xin'an jiang (XAJ) model does not explicitly represent snow accumulation and melt processes, which likely contributes to the inferior performance of the XAJ+LSTM hybrid scheme in snow-affected basins. We agree that this limitation should be acknowledged more clearly in the manuscript. The relevant contents after modification are as follows:

"…

This distinction highlights a key difference in the suitability of hybrid model structures across basins with varying snow dynamics. Although both EXP and XAJ are widely used conceptual hydrological models, they differ significantly in terms of process representation. The EXP-HYDRO model includes a simplified snow accumulation and melt component, whereas the Xin'an jiang (XAJ)

model, in its original and simplified forms, does not explicitly account for snow-related processes. As a result, in snow-affected basins, particularly those basins located in high-altitude regions with seasonal snowpack, the hybrid scheme coupling EXP and LSTM generally achieves superior predictive performance compared to the XAJ-LSTM hybrid. This performance gain can be attributed to the EXP model's capacity to represent key aspects of snow dynamics, enabling the hybrid model to capture runoff behavior more realistically in these regions. These findings underscore the importance of selecting appropriate process components within hybrid models to align with the dominant hydrological processes of a region, such as snow accumulation and melt in high altitude mountainous areas.

…"

Comment 24: *Figure 11: there are two B panels. 2nd B panel missing numbers of the overall source/destination bins. Same comments as for Figure 8.*

Response:

Thank you for your correction. We have corrected the subfigure labels to avoid duplication. The Sankey diagram previously labeled as Figure 11(b) has now been relabeled as Figure 11(c), and the subsequent subplots have been updated accordingly. The modified 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.

…."

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:

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., and Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models, Hydrol. Earth Syst. Sci., 27, 139–157, https://doi.org/10.5194/hess-27-139-2023.

Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. Geophysical Research Letters 47, e2020GL088229. https://doi.org/10.1029/2020GL088229

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv Preprint arXiv:1412.6980.

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

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019.

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

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19, 209–223. https://doi.org/10.5194/hess-19-209-2015

Zhan, Y. , Guo, Z. , Yan, B. , Chen, K. , Chang, Z. , & Babovic, V., 2024. Physics-informed identification of pdes with lasso regression, examples of groundwater-related equations. Journal of Hydrology, 638. https://doi.org/10.1016/j.jhydrol.2024.131504