Reviewer #2

The manuscript presents an innovative ensemble strategy that combines a differentiable process-based model (δHBV) with a data-driven Long Short-Term Memory (LSTM) model, further diversified through the application of multiple meteorological forcing datasets. The approach is evaluated across a wide range of generalization scenarios (temporal extrapolation, PUB, and PUR) using the CAMELS dataset. Although the paper is well-written and the main ideas are clearly communicated, it would benefit from additional details in the methods and a deeper discussion of model complementarity and limitations.

Thanks for the positive comments and constructive suggestions. We will revise the manuscript accordingly. Please find our point-by-point responses below.

Strength

The proposed ensemble framework is conceptually strong and offers a well-justified combination of complementary data and algorithmic modeling paradigms.
The study is well evaluated across well-defined training protocols and temporal-spatial splits, which improves confidence in its generalizability.
The use of multiple data sources for meteorological forcings addresses input uncertainty better than traditional single-source modeling.
The results, specifically the finding that δHBV improves spatial generalization, have clear implications for prediction in ungauged regions.

Thanks for the positive comments.

Weakness

Interpretability: While the δHBV model's performance is shown to be beneficial in spatial generalization, the underlying reasons for this complementarity (e.g., structural constraints, parameter smoothness) are not deeply explored. A discussion of how each model contributes to ensemble diversity would strengthen the scientific value of the work.

We thank the reviewer for this insightful comment regarding model interpretability and the complementarity between LSTM and δHBV within the ensemble framework. We fully agree that a deeper understanding of the relative contributions of each model would enhance the scientific value of our study. Besides this, we also plan to dig deeper and examine cases where the errors of LSTM and dHBV cancel each other.

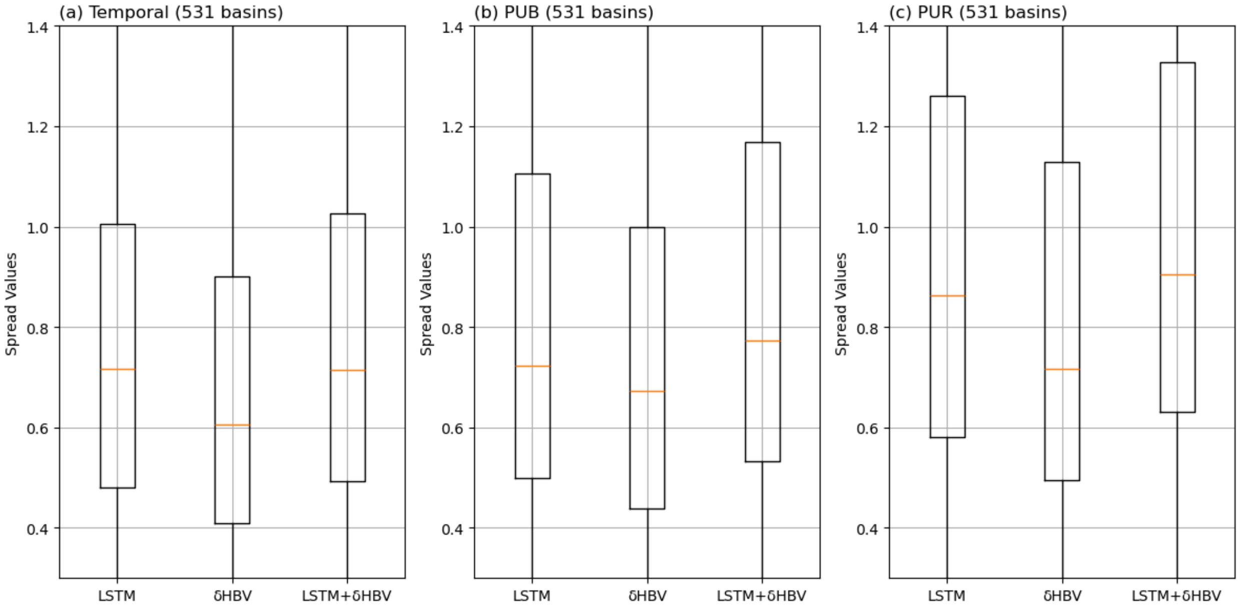

Figure R1. Boxplots of the spread values of simulations based on LSTM, δHBV, and LSTM + δHBV with different meteorological forcings and random seeds across temporal, PUB, and PUR tests.

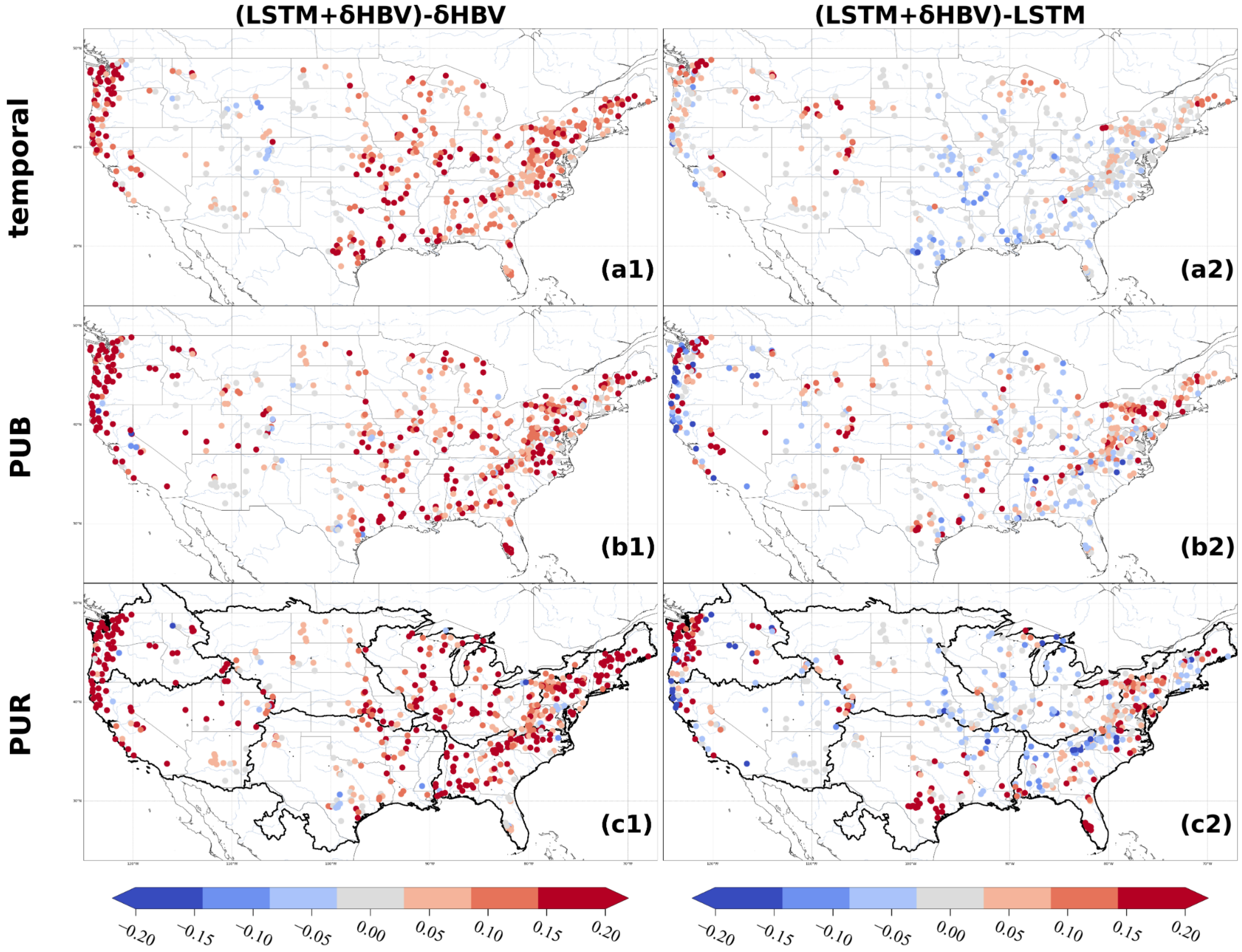

Figure R2. Spatial distributions of spread increase from δHBV and LSTM to the LSTM+δHBV ensemble across temporal, PUB, and PUR tests.

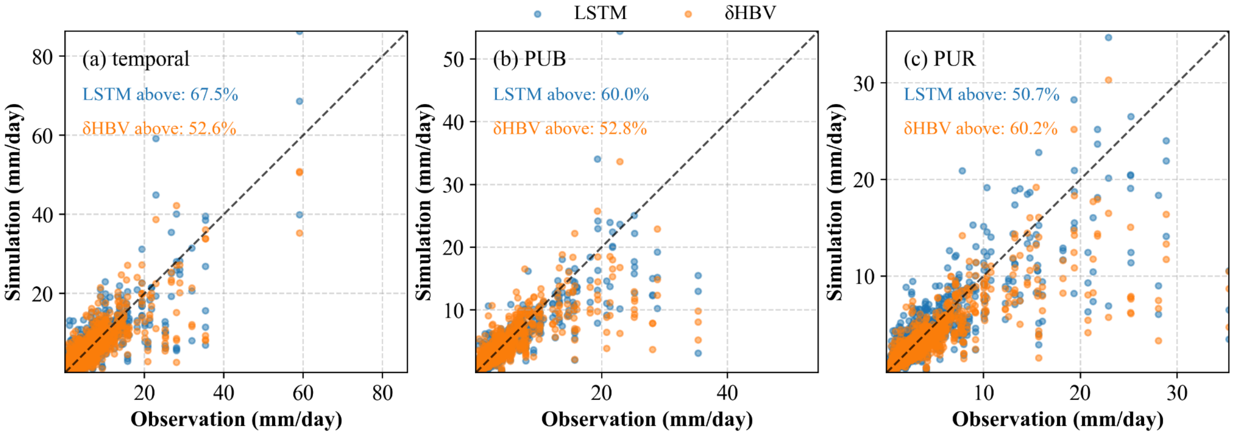

Figure R3. Distributions of observation–simulation pairs from LSTM and δHBV models along the 1:1 line across temporal, PUB, and PUR tests. Percentages of pairs lying above the 1:1 line for both models are also indicated.

Since the benefits of the different ensemble members to the deterministic precision have been displayed in the original manuscript, we have conducted additional analyses in terms of ensemble variability as suggested. Specifically, we use the spread values (Li et al., 2021; Reichle and Koster, 2003), which are widely adopted to quantify ensemble variability, to further explore model complementarity. The spread value is calculated as follows,

$$Spread = \sqrt{\frac{1}{n}\frac{1}{r}\sum_{i=1}^{n}\sum_{j=1}^{r}(S_{i,j}-\overline{S}_{i})^{2}}$$

Where *n* is the number of simulated days, r is the number of ensemble members, and S is the simulations of each ensemble member, $\overline{()}$ indicates the average of values.

Figure R1 presents the boxplots of spread values for ensemble simulations using random seed variations with LSTM, δHBV, and the combined LSTM + δHBV, across the temporal, PUB, and PUR test settings. We observe that the overall spread increases from temporal to PUB and PUR tests, reflecting growing uncertainty. Notably, δHBV consistently exhibits lower spread values than LSTM across all tests, indicating its higher stability. This aligns with our prior discussion: δHBV tends to constrain the learnable function space, thus having lower variability and potentially higher bias. This difference stems from their structural characteristics—δHBV is governed by more rigid physical constraints, which limit unrealistic dynamics and enhance stability, while LSTM is more flexible and capable of capturing patterns that may not be explicitly represented in physical models, such as human influences or unmodeled processes. The combination of both models (LSTM + δHBV) yields greater spread values, indicating enhanced ensemble diversity. This suggests that the two models offer complementary strengths—LSTM contributes flexibility and capacity to represent data-driven nuances, while δHBV anchors the ensemble with physically constrained behavior.

Figure R2 illustrates the spatial distributions of spread increase resulting from incorporating LSTM and δHBV, respectively, and further supports our previous analysis. Incorporating LSTM leads to an increase in spread values across all basins, reflecting its higher variability. In contrast, the δHBV model, characterized by stronger physical constraints and generally lower variability, results in a decrease in spread values for many basins. However, δHBV still contributes to a spread increase in most northern basins and gradually leads to spread increases in a larger number of basins across the CONUS. This suggests notable differences in simulated streamflow behavior between LSTM and δHBV, largely attributable to their distinct model structures. Figure R3 reveals relatively limited differences between the streamflow behaviors simulated by LSTM and δHBV, with LSTM generally producing higher streamflow estimates than δHBV. A more systematic investigation of these differences would be valuable in future studies.

Following the reviewer's suggestion, we will incorporate these analyses and discussions about the ensemble spread in the revised manuscript.

Robustness and Sensitivity Analysis: The paper lacks an explicit assessment of how ensemble performance responds to errors or biases in the forcing datasets or uncertainty in model parameters. Including even a limited robustness analysis would improve confidence in the ensemble's reliability. Additionally, the authors should consider running one or two experiments to understand whether changing the size of the lookback window (i.e., the number of historical timesteps) for the LSTMs impacted the overall performance of the ensemble.

Thank you for the suggestions. Based on them, we conducted several experiments using temporal tests to demonstrate the robustness of ensemble benefits under various factors, including precipitation errors, parameter uncertainties in the δHBV model, and hyperparameter uncertainties in the LSTM model.

Regarding sensitivity to the forcing datasets, we ran the δHBV and LSTM models under a temporal test, both without and with a precipitation error introduced by multiplying precipitation by 0.1, to examine differences across ensemble groups. The results, shown in Figures R4 and R5, indicate that although the performance of both LSTM and δHBV decreases when the precipitation error is introduced, the decrease is not substantial, demonstrating a certain degree of robustness to precipitation errors and some capacity of both models to adapt to such errors. Interestingly, LSTM and δHBV respond differently to this type of precipitation error: for LSTM, the error tends to reduce ensemble performance mainly under low and high flow regimes, whereas for δHBV, the reduction is more pronounced under low and middle flow regimes. These differences reflect the fact that LSTM does not need to respect mass balance and can adjust precipitation up or down internally, but has trouble learning the contrast, while δHBV needs to distort the low flow to capture the high flows. Despite these differences, the ensemble benefits remain significant and robust when comparing different ensemble groups and assessing the impact of precipitation errors.

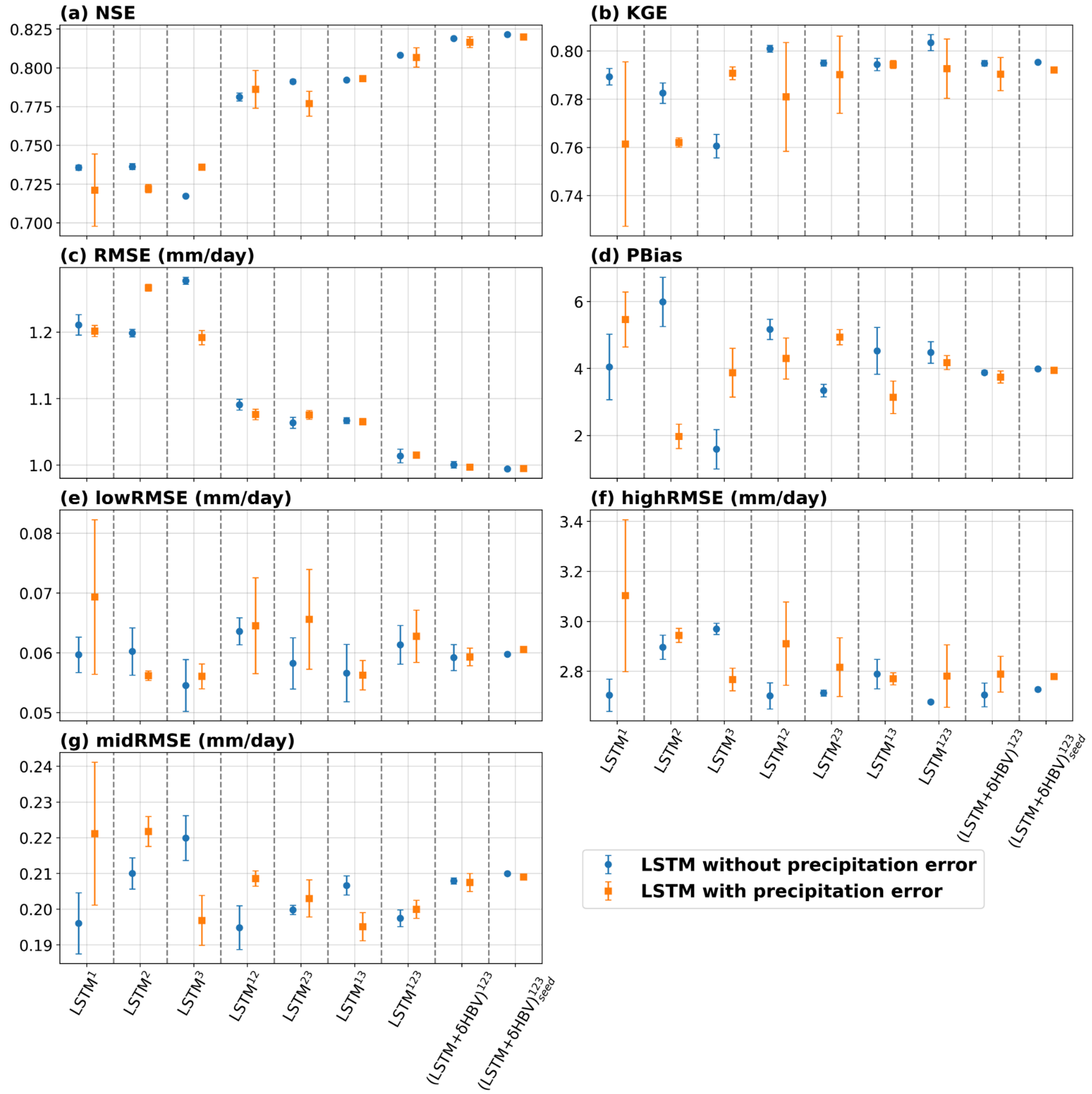

Figure R4. Simulation performance under the temporal test using the LSTM model with and without a precipitation error equal to 0.1 times the precipitation, compared across metrics (a)–(g).

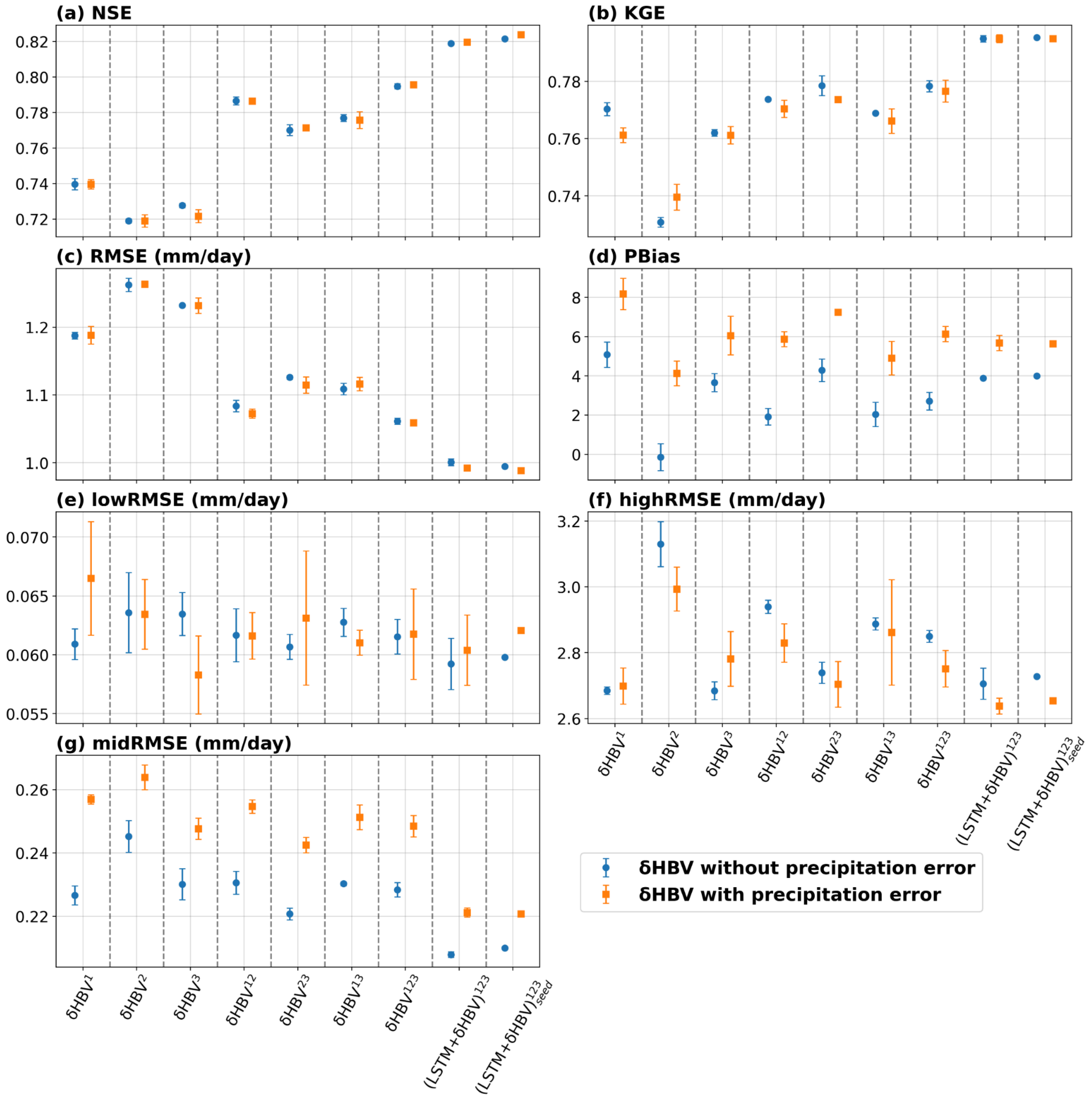

Figure R5. Simulation performance under the temporal test using the δHBV model with and without a precipitation error equal to 0.1 times the precipitation, compared across metrics (a)–(g).

Similar results are observed in cases investigating the effects of parameter uncertainties in δHBV (Figure R6) and hyperparameter uncertainties in LSTM (Figure R7). Regarding parameter uncertainties, we additionally ran a case using the δHBV model with fewer dynamic parameters—reducing the number from three in the benchmark case to two—by fixing the infiltration rate parameter K0 as static to assess the resulting performance changes, which may reduce δHBV's ability to represent dynamic water release processes influenced by changing groundwater levels, bank and wetland storages, and other factors (Song et al., 2025b). This leads to increased structural errors and decreased model performance. Nevertheless, the contribution of δHBV to ensemble simulations remains robust, with ensemble benefits substantially outweighing the negative effects of parameter uncertainties.

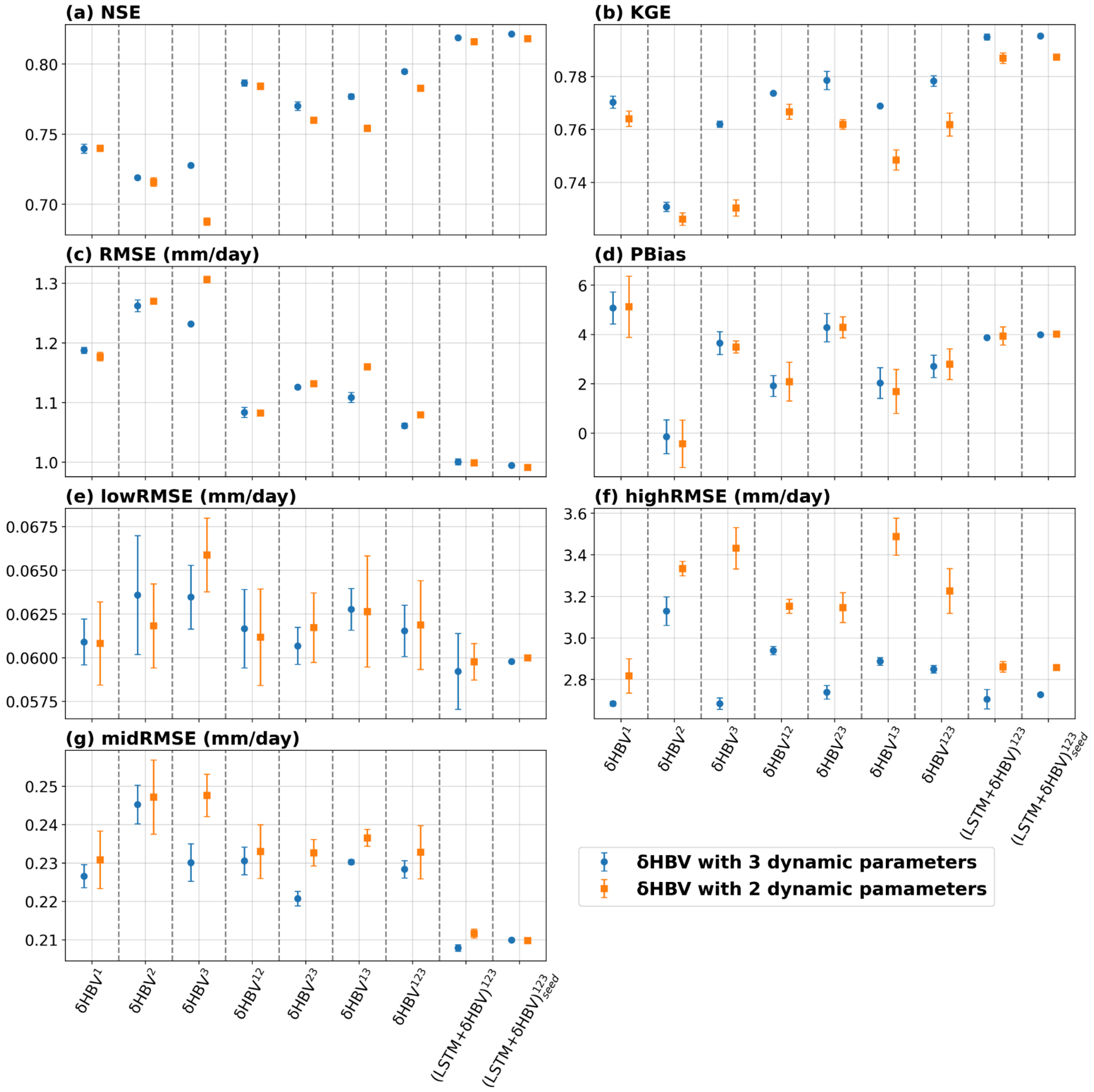

Figure R6. Simulation performance under the temporal test using the δHBV model with 3 and 2 dynamic parameters, compared across metrics (a)–(g).

Regarding hyperparameter uncertainties in the LSTM model, we focus on a key hyperparameter: the lookback window size, as suggested. We treat this parameter as having physical significance related to the temporal period rather than a typical hyperparameter. Therefore, we fix the window size to one year (365 days) to capture a full annual cycle while accounting for interannual variability. To evaluate the impact of different window lengths, we include two additional scenarios with 182 and 730 timesteps. As shown in Figure R7, the LSTM model with a 365-day window generally achieves better performance across most scenarios. However, compared to the overall benefits of the ensemble, this difference is not substantial, indicating the robustness of ensemble simulations to variations in this LSTM hyperparameter.

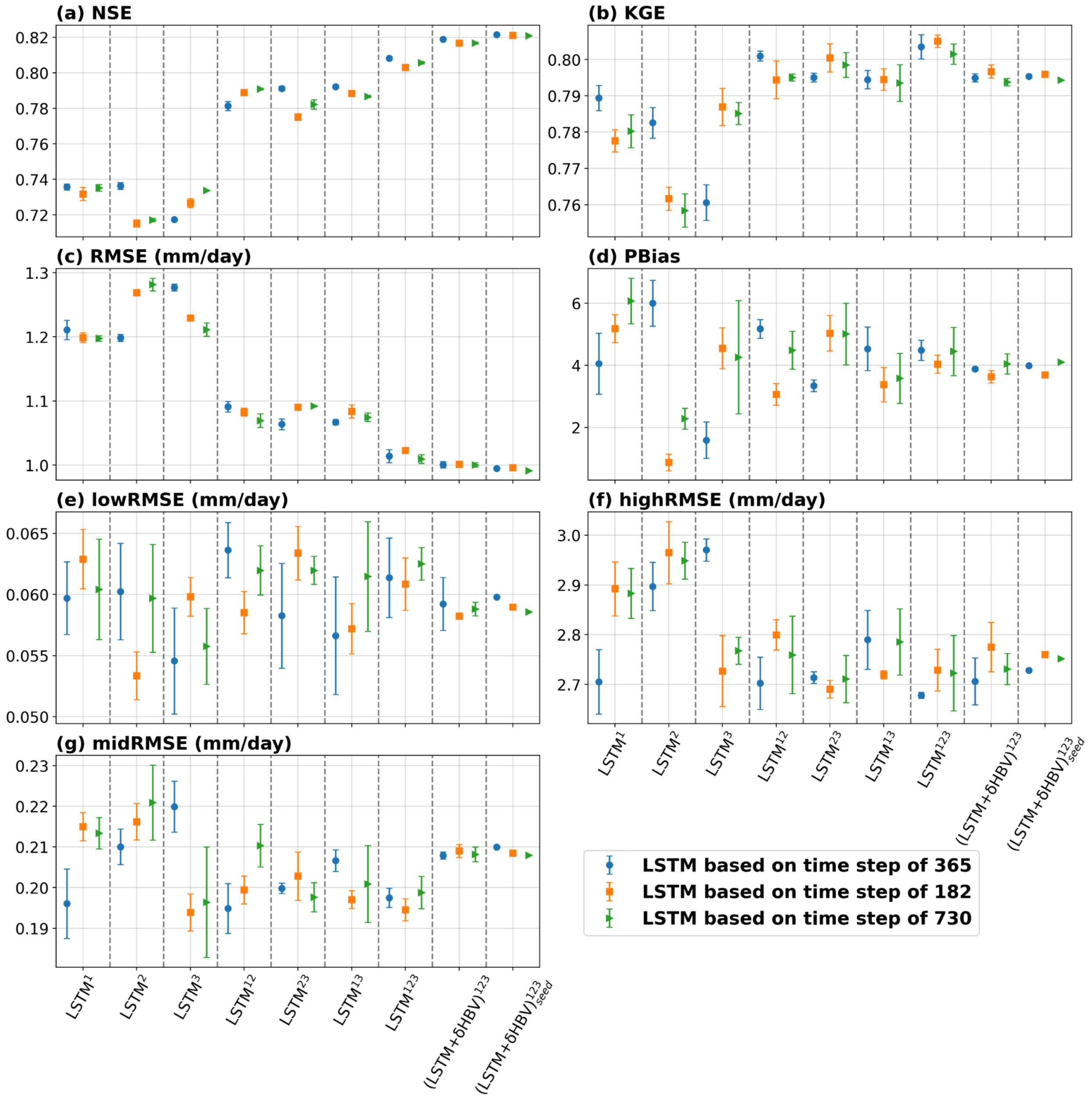

Figure R7. Simulation performance under the temporal test using the δHBV model on the time steps of 365, 182, 730, compared across metrics (a)–(g).

Although it is practically impossible to test the effects of all possible configurations on ensemble benefits, we expect these benefits to remain robust against other factors to some extent, based on the representative results presented. Following the suggestions, we will include these additional cases in the revised manuscript to further demonstrate the reliability of our ensemble simulations.

Scalability and Practical Deployment: The manuscript does not address the computational or operational feasibility of deploying this ensemble framework in practice, especially over large domains or in real-time forecasting contexts. A short discussion (1-2 sentences) on this topic would add practical relevance.

We appreciate the reviewer's suggestion to further discuss the computational and operational feasibility of deploying the ensemble framework. This point is partially addressed in Section 3.3 of the original manuscript, where we note:
"Moreover, ensemble simulations may face challenges when computational resources are limited and calculations are performed sequentially. However, we remain optimistic about these challenges, as the processes can be addressed by leveraging parallel computing with multiple GPUs, benefiting from ongoing advancements in computational power."
In response to the reviewer's comment, we plan to expand this discussion as follows:
"Ensemble simulations may face challenges when computational resources are constrained, particularly for large-scale or real-time applications. Nevertheless, we remain optimistic about overcoming these challenges due to several promising solutions. These include tailoring the hydrological model by simplifying less relevant components to specific simulation objectives (Clark et al., 2015; Kraft et al., 2022) and cloud-based computing infrastructures that offer scalable, on-demand resource allocation (He et al., 2024; Leube et al., 2013). Importantly, the majority of computational costs are incurred during model training. In practice, ensemble members are typically pre-trained by different research or application groups (Bodnar et al., 2025; Nearing et al., 2024; Song et al., 2025a), enabling direct reuse of these well-trained models and significantly improving computational efficiency."

References:

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth system, Nature, 641, 1180–1187, https://doi.org/10.1038/s41586-025-09005-y, 2025.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, https://doi.org/10/f7db99, 2015.

He, Y., Chen, M., Wen, Y., Duan, Q., Yue, S., Zhang, J., Li, W., Sun, R., Zhang, Z., Tao, R., Tang, W., and Lü, G.: An open online simulation strategy for hydrological ensemble forecasting, Environmental Modelling & Software, 174, 105975, https://doi.org/10.1016/j.envsoft.2024.105975, 2024.

Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Towards hybrid modeling of the global hydrological cycle, Hydrology and Earth System Sciences, 26, 1579–1614, https://doi.org/10.5194/hess-26-1579-2022, 2022.

Leube, P. C., de Barros, F. P. J., Nowak, W., and Rajagopal, R.: Towards optimal allocation of computer resources: Trade-offs between uncertainty quantification, discretization and model reduction, Environmental Modelling & Software, 50, 97–107, https://doi.org/10.1016/j.envsoft.2013.08.008, 2013.

Li, P., Zha, Y., Shi, L., and Zhong, H.: Identification of the terrestrial water storage change features in the North China Plain via independent component analysis, Journal of Hydrology: Regional Studies, 38, 100955, https://doi.org/10.1016/j.ejrh.2021.100955, 2021.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Reichle, R. H. and Koster, R. D.: Assessing the Impact of Horizontal Error Correlations in Background Fields on Soil Moisture Estimation, Journal of Hydrometeorology, 4, 1229–1242, https://doi.org/10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2, 2003.

Song, Y., Bindas, T., Shen, C., Ji, H., Knoben, W. J. M., Lonzarich, L., Clark, M. P., Liu, J., van Werkhoven, K., Lamont, S., Denno, M., Pan, M., Yang, Y., Rapp, J., Kumar, M., Rahmani, F., Thébault, C., Adkins, R., Halgren, J., Patel, T., Patel, A., Sawadekar, K. A., and Lawson, K.: High-resolution national-scale water modeling is enhanced by multiscale differentiable physics-informed machine learning, Water Resour. Res., 61, e2024WR038928, https://doi.org/10.1029/2024WR038928, 2025a.

Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M., Knoben, W. J. M., Wood, A. W., Lawson, K. E., Patel, T., and Shen, C.: Physics-informed, differentiable hydrologic models for capturing unseen extreme events, https://doi.org/10.22541/essoar.172304428.82707157/v2, 2025b.