Reviewer #1

General comments

This paper comprehensible evaluates the performance of different ensembles based on LSTM and HBV models, and three different forcing datasets, across CAMELS catchments. The ensembles are evaluated in terms of a temporal test, a prediction in ungauged basins test (PUB), and a prediction in ungauged regions test (PUR). The main conclusion is that the data-driven LSTM and process-based HBV ensemble improves NSE, particularly for PUB and PUR tests.

Overall, the manuscript is well-structured and clearly conveys its main point. Please find below some comments and suggestions.

Thanks for the positive comments

Specific comments

1. L150 and L240-L244: Please explicitly indicate which features (static and dynamic) are used by the LSTM model or at least refer to Appendix C. Are static characteristics of the catchment also used during the PUB and PUR tests? Is it the case that for PUB the model does not use previous streamflow observations to generate the predictions? Or does PUB only refer to the model being tested at basins not used during training?

Regarding the question "*Please explicitly indicate which features (static and dynamic) are used by the LSTM model or at least refer to Appendix C*", thanks for your suggestions. The static and dynamic attributes for LSTM are utilized as the same as Kratzert's studies (Kratzert et al., 2022), as shown in Table R1. We considered revising Table C1 of the original manuscript to specify the inputs for LSTM and δ HBV, respectively.

Regarding the question "*Are static characteristics of the catchment also used during the PUB and PUR tests?*", yes. These static characteristics of the catchment are also used during the PUB and PUR tests.

Regarding the question "*Is it the case that for PUB the model does not use previous streamflow observations to generate the predictions?*", yes. For all three kinds of tests, streamflow observations are not included in the inputs and are only used to calibrate the model.

Regarding the question "*does PUB only refer to the model being tested at basins not used during training?*" -- yes, exactly. As described in Section 2.5 of the manuscript, we first divided all basins into 10 subsets. The model was then trained and evaluated over 10 rounds, each time holding out one subset for testing while using the remaining basins for training. After completing all rounds, the test results from all basins were concatenated to evaluate overall performance. Therefore, in each round of evaluation, the test basins were strictly excluded from the training process.

Table R1. Full names for the abbreviations of dynamic data (all but streamflow are "forcings") and static basin attributes used as the LSTM model inputs and outputs. All variables and their values are provided in the CAMELS dataset (Addor et al., 2017) except for the NLDAS and Maurer daily temperature extrema, which are from Kratzert et al. (2021).

Type	Abbreviation	Full name	Unit

Dynamic forcings	prcp	Precipitation	mm/day
	tmax	Maximum air temperature	°C
	tmin	Minimum air temperature	°C
	srad	Shortwave radiation	W/m^2
	vp	Water vapor pressure	pa
	q	Streamflow normalized by basin area (q_vol / area_gages2)	mm/day
Static basin attributes	p_mean	Mean daily precipitation	mm/day
	pet_mean	Mean daily potential evapotranspiration	mm/day
	frac_snow	Fraction of precipitation falling as snow	-
	aridity	Rate of mean values of potential evapotranspiration and precipitation	-
	high_prec_freq	Frequency of high precipitation days	days/year
	high_prec_dur	Average duration of high precipitation events	days
	low_prec_freq	Frequency of dry days	days/year
	low_prec_dur	Average duration of dry periods	days
	elev_mean	Catchment mean elevation	m
	slope_mean	Catchment mean slope	m/km
	area_gages2	Catchment area (GAGES-II estimate)	km^2
	frac_forest	Fraction of catchment area having land cover identified as forest	-
	lai_max	Maximum monthly mean of the leaf area index	-
	lai_diff	Difference between the maximum and minimum monthly mean of the leaf area index	-
	gvf_max	Maximum monthly mean of the green vegetation	-
	gvf_diff	Difference between the maximum and minimum monthly mean of the green vegetation fraction	-
	soil_depth_pelletier	Depth to bedrock	m
	soil_depth_statsgso	Soil depth	m
	soil_porosity	Volumetric soil porosity	-

	soil_conductivity	Saturated hydraulic conductivity	cm/hr
	max_water_content	Maximum water content	m
	sand_frac	Fraction of soil which is sand	-
	silt_frac	Fraction of soil which is silt	-
	clay_frac	Fraction of soil which is clay	-
	carbonate_rocks_frac	Fraction of the catchment area as carbonate sedimentary rocks	-
	geol_permeability	Subsurface permeability	m^2

2. Table 2: Is it possible to draw any conclusions about the skill of the models to extrapolate to a warmer climate based on the temporal test? I assume that the period 1995-2010 is warmer than 1980-1995.

Thanks for the suggestions. We have compared the temperature changes from the training to the test periods, as shown in Figure R1. The results show that most basins are getting warmer, and these models can also get satisfactory performances, which indicates that these models can extrapolate at least modestly their performances under a warmer climate. Note that it is an often asked question for purely data-driven models like LSTM to lose accuracy if the drift is stronger. The issue is that there was a lack of a "substantially warmed" real dataset to assess this behavior. Previous benchmarks suggest the 15-year-scale trend to be accurately captured in temporal test by LSTM, and more poorly captured for PUR (Feng et al., 2023). We will add some sentences in the revised manuscript to describe this finding.

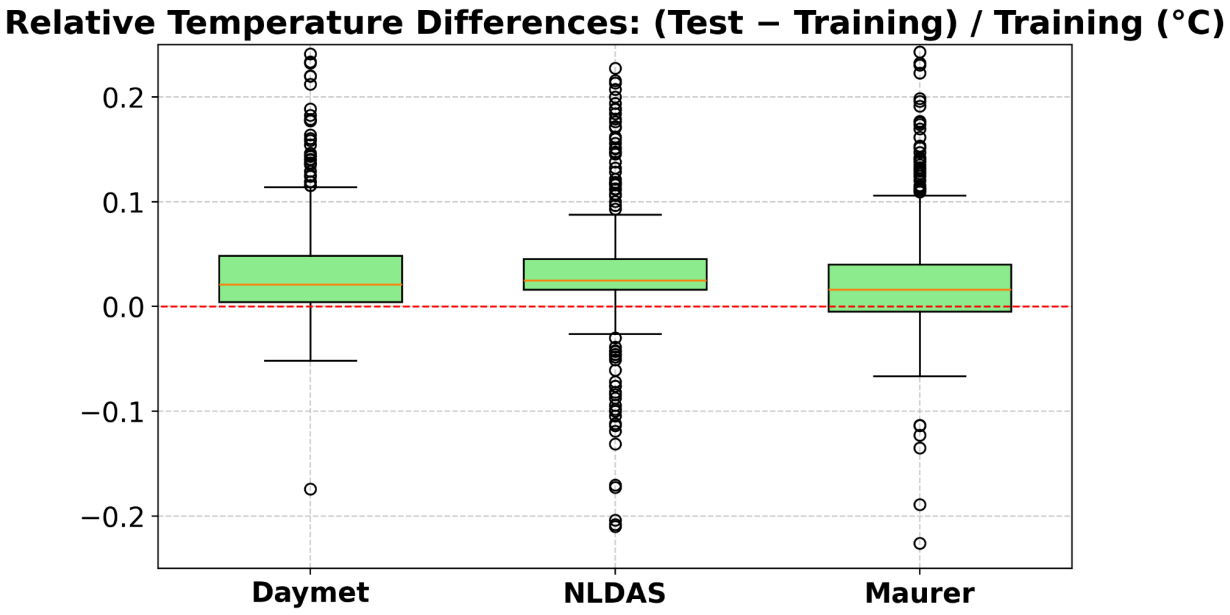

Figure R1. Boxplot of relative temperature differences between the test and training periods, calculated as (Test − Training) / Training. Each box represents the distribution of normalized temperature changes across basins for a specific meteorological forcing dataset: Daymet, NLDAS, and Maurer. Positive values indicate warming in the test period relative to the training period

3. Table 2: What would happen if you were to train HBV using only the same 531-basin subset as for the LSTM instead of the 671 basins?

Thanks for the question. We tested this by conducting the temporal experiments using only the 531-basin subset for the δHBV model. The results were largely similar to those obtained using all 671 basins shown in Figures R2-R3, indicating that the impact of reducing the training set size is limited in this context. We will incorporate some discussion about this in the revised manuscript.

We believe that training the δHBV model on the full 671-basin dataset is still beneficial, as the physical constraints inherent in the model allow it to make more effective use of available data, even when data quality is somewhat limited. That said, the added value from including the additional 140 basins appears to be marginal, and the choice of training on 531 versus 671 basins does not substantially affect the overall model performance.

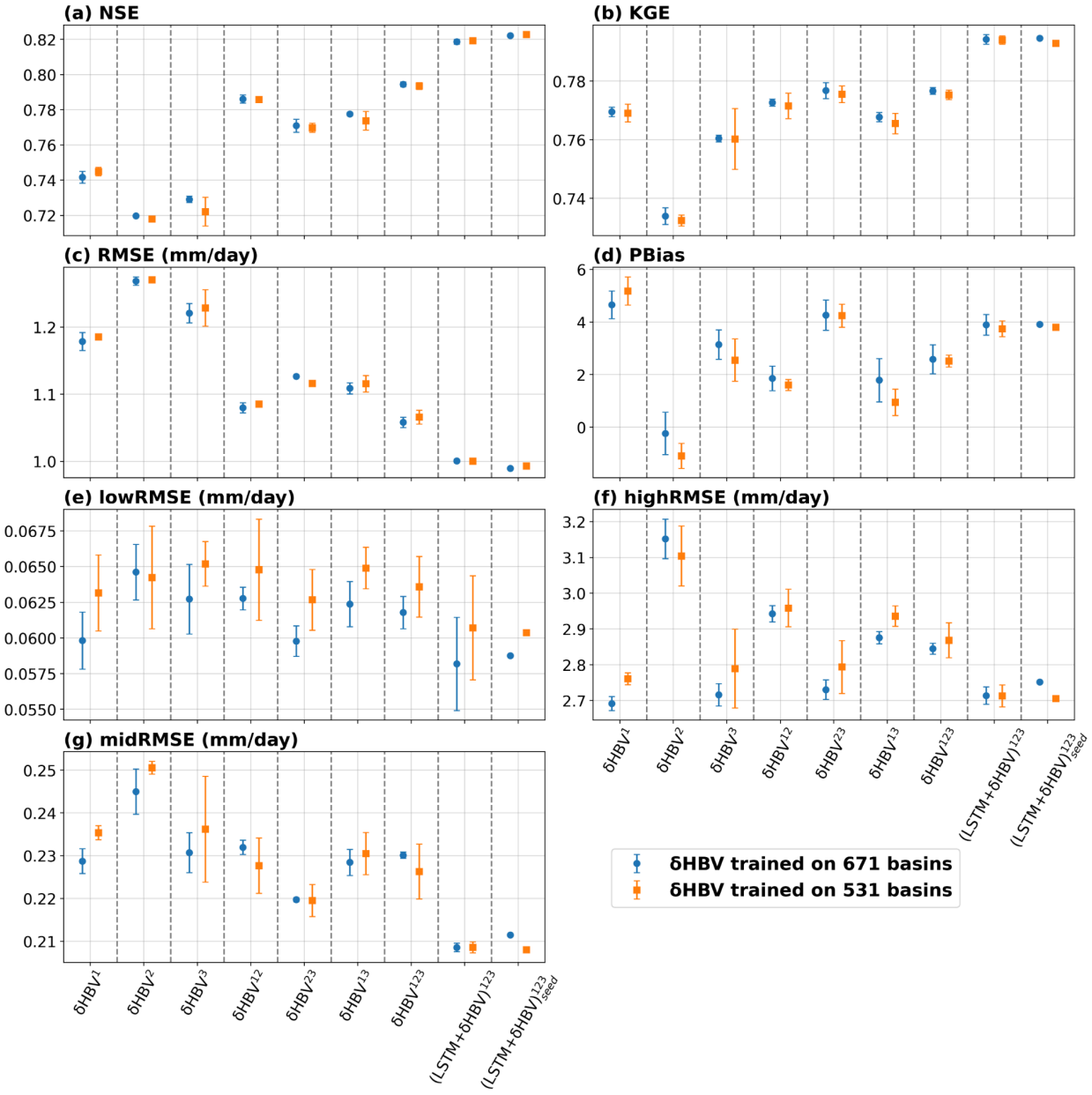

Figure R2. Comparison of δHBV simulations trained on 671 versus 531 basins across performance metrics (a)–(g), with test cases distinguished by varying x-axis labels

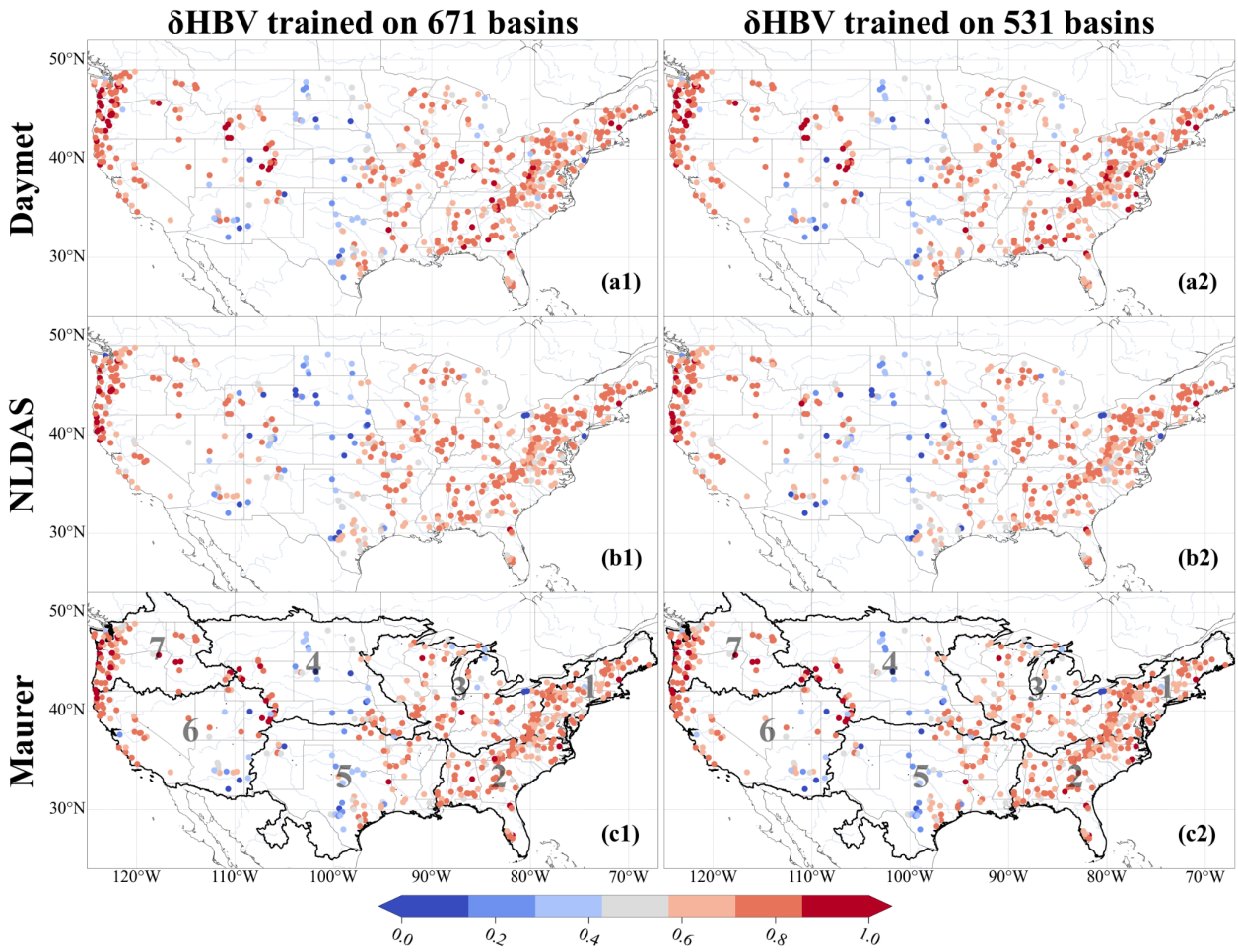

Figure R3. Comparison of NSE spatial distributions for δHBV models trained on 671 versus 531 basins across different meteorological forcing datasets

4. It could be useful to also provide similar plots to Fig. 3 in the appendix where HBV2 and HBV3 are used instead of HBV1.

Thanks for the suggestions. We have plotted Figures R4 and R5 based on δHBV2 and δHBV3, and they show similar results, consistent with the conclusions. We will revise the manuscript accordingly.

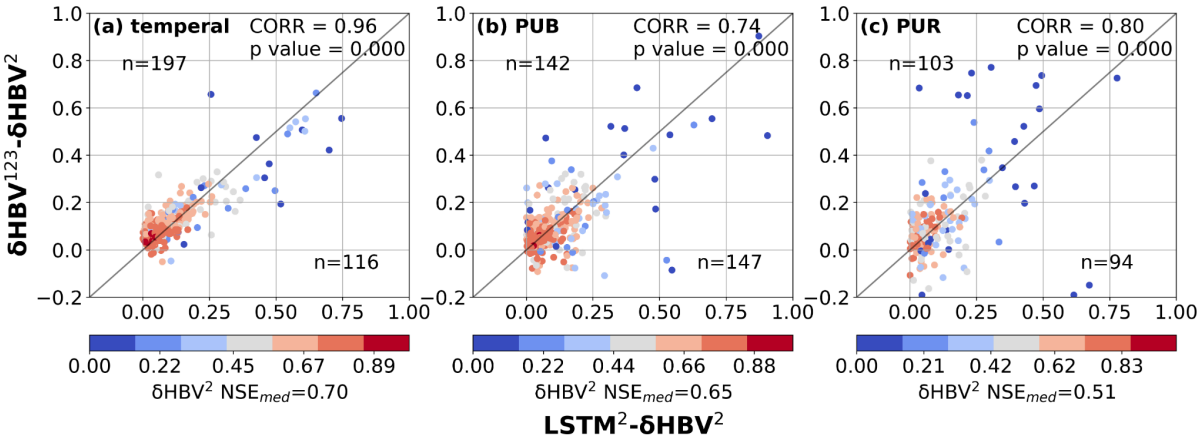

Figure R4. Scatter plots comparing the performance differences between hydrological models for the basins where LSTM outperformed δHBV (the basins where δHBV outperformed are not shown in this plot). The x-axis represents the NSE differences between $LSTM^2$ and δHBV^2 ($LSTM^2$ - δHBV^2), while the y-axis shows the NSE differences between δHBV^{123} and δHBV^2 (δHBV^{123}- δHBV^2). Points are color-coded according to the NSE values of δHBV^2. The correlation coefficient (CORR) and p values between x-axis values and y-axis values, along with the median NSE value of δHBV^2 (NSE_{med}) on these basins are also noted.

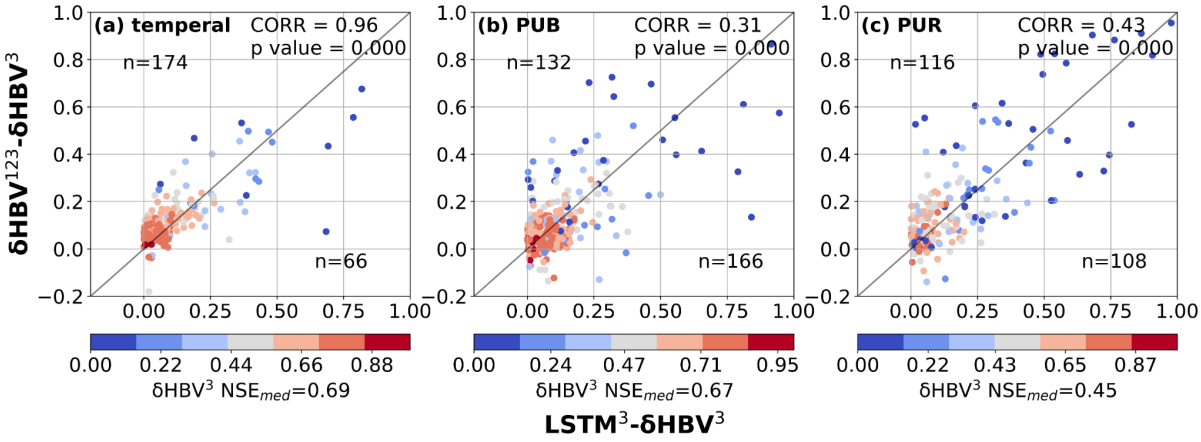

Figure R5. Scatter plots comparing the performance differences between hydrological models for the basins where LSTM outperformed δHBV (the basins where δHBV outperformed are not shown in this plot). The x-axis represents the NSE differences between $LSTM^3$ and δHBV^3 ($LSTM^3$ - δHBV^3), while the y-axis shows the NSE differences between δHBV^{123} and δHBV^3 (δHBV^{123}- δHBV^3). Points are color-coded according to the NSE values of δHBV^3. The correlation coefficient (CORR) and p values between x-axis values and y-axis values, along with the median NSE value of δHBV^3 (NSE_{med}) on these basins are also noted.

5. Fig. 4 and L439-L454: Can you expand on potential reasons for the lower model-skill in midwestern and western basins? Is human management of streamflow an important factor here, despite being CAMELS basins?

Thanks for the suggestions. We have directly compared the spatial patterns of performance with the spatial distributions of evaporation and the dam number. We found that over the midwestern and western CONUS, there are also high evaporation climate conditions (Figure 2 of (Heidari et al., 2020), shown here as Figure R6) and a large number of basins (Figure 1 of (Ryan Bellmore et al., 2017), shown here as Figure R7), which tend to have complex water use processes that cannot be simulated via the models (Figure 11 of (Wada et al., 2016), shown here as Figure R8). All these factors indicate that anthropogenetic influence can be an important factor that causes the model to perform poorly. And these factors have been implicitly expressed in lines 486-490 in section 3.3 of the original manuscript. Based on your suggestions, these sentences will be revised to be clearer.

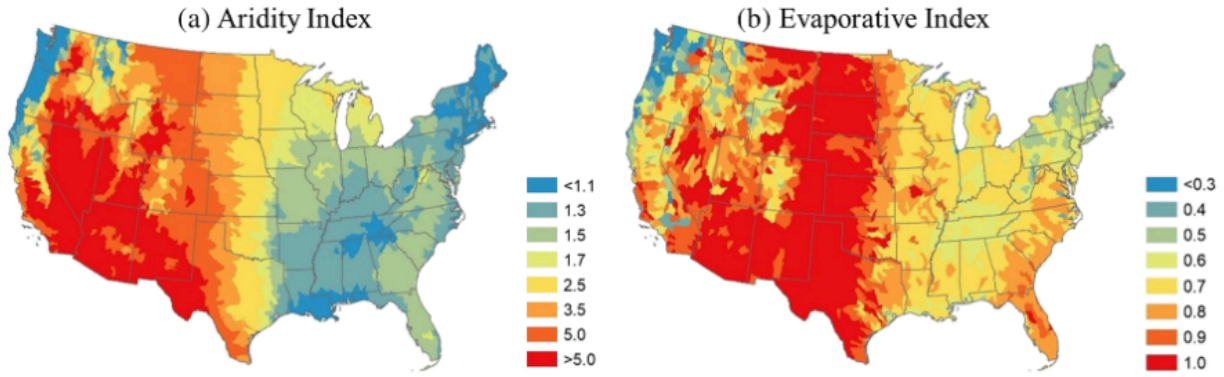

Figure R6. Maps of current (a) aridity index and (b) evaporative index for the baseline period (1986–2015) from (Heidari et al., 2020).

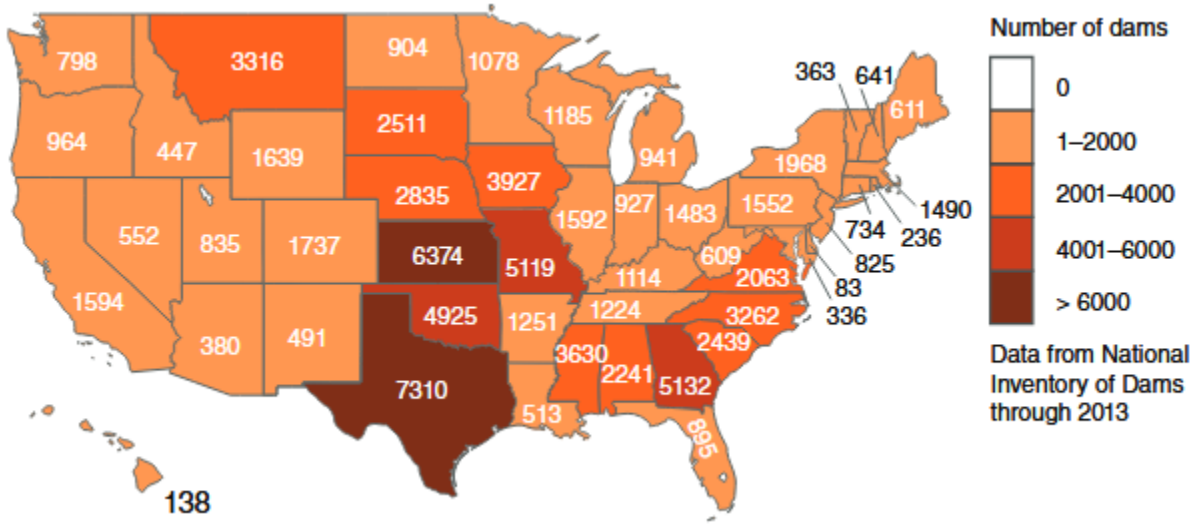

Figure R7. Distribution of dams in the contiguous U.S. from the corrected figure of (Ryan Bellmore et al., 2017).

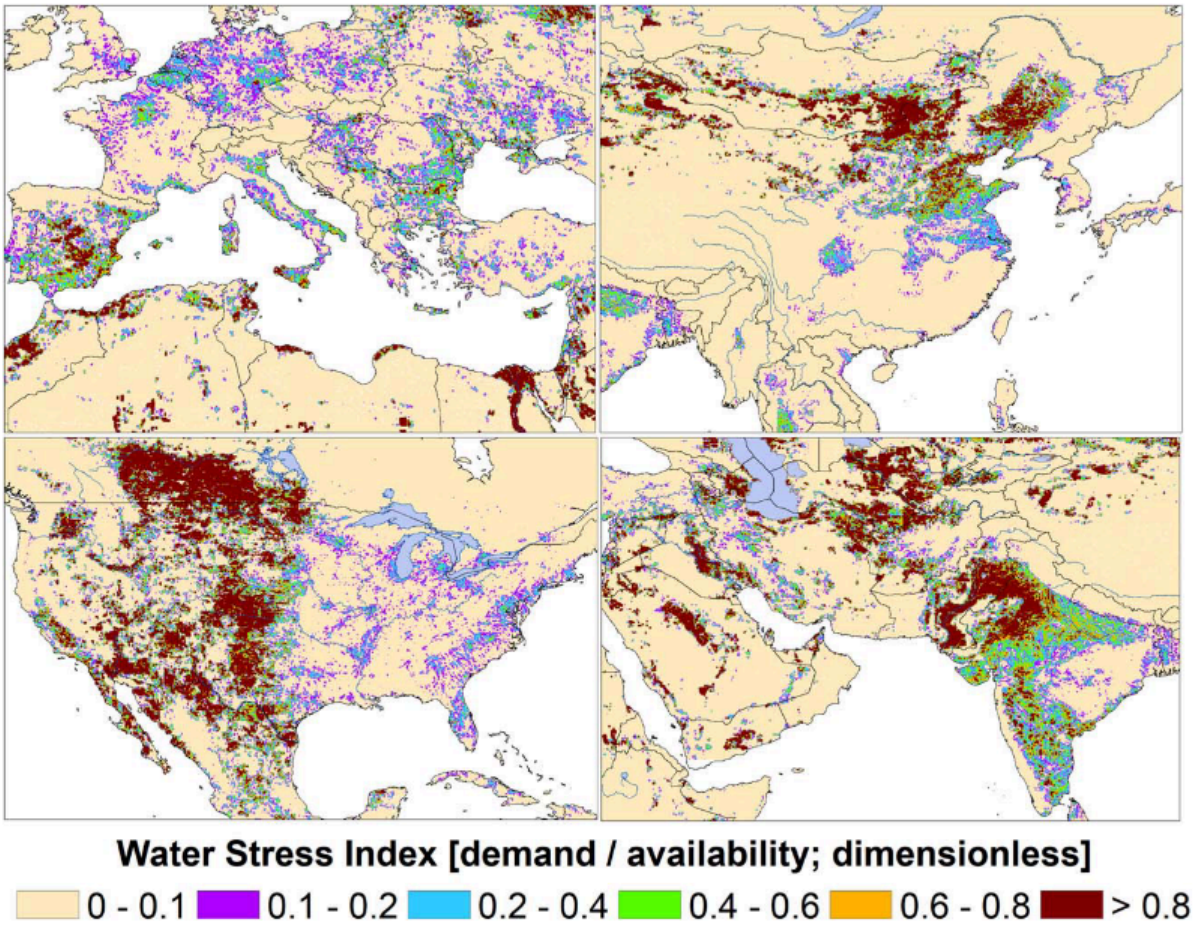

Figure R8. Water Stress Index for 2010 calculated at a 6 min spatial resolution from (Wada et al., 2016)

6. Fig. 5: Please clarify. L399-L401 says there is a small difference when using 3 or 10 seeds, but Fig. 5 shows the difference between individual seeds and using 10 seeds. It is interesting to note in Fig. 5 that $LSTM^{multi}$ with 10 seeds achieves a similar skill as $(LSTM + HBV)^{123}$ at least for the temporal test. This could also be an important conclusion.

Thanks for the comments. It is true that LSTM, after ensembling different random seeds, increases much higher compared with the individual component and other ensemble strategies. But here, we want to show that there are smaller differences between ensembling 3 random seeds and ensembling 10 seeds. Specific metric values of the LSTM model with different random seeds can be found in Table D4 and Table D5 (0.81992 (3 seeds) v.s. 0.824 (10 seeds)). We also implicitly describe the boost performance improvement of $LSTM_{seed}^{multi}$ and owe it to the instability of LSTM simulations in the original manuscript as, "The

performance of $(LSTM + \delta HBV)^{123}$ ensemble proved more robust than $LSTM^{multi}$, with only a slight boost when we incorporated random seeds, i.e., $(LSTM + \delta HBV)^{123}_{seed}$. " in lines 390-391. Based on your suggestions, we will add some sentences to make it clearer in the revised manuscript.

7. Fig. 6: Suggestion to have LSTM models with shades of one color and HBV models with shades of a different color to better highlight the differences between LSTM and HBV mentioned in L461-L464.

Thanks for the suggestions. We replotted the figure accordingly as shown in Figure R9 here.

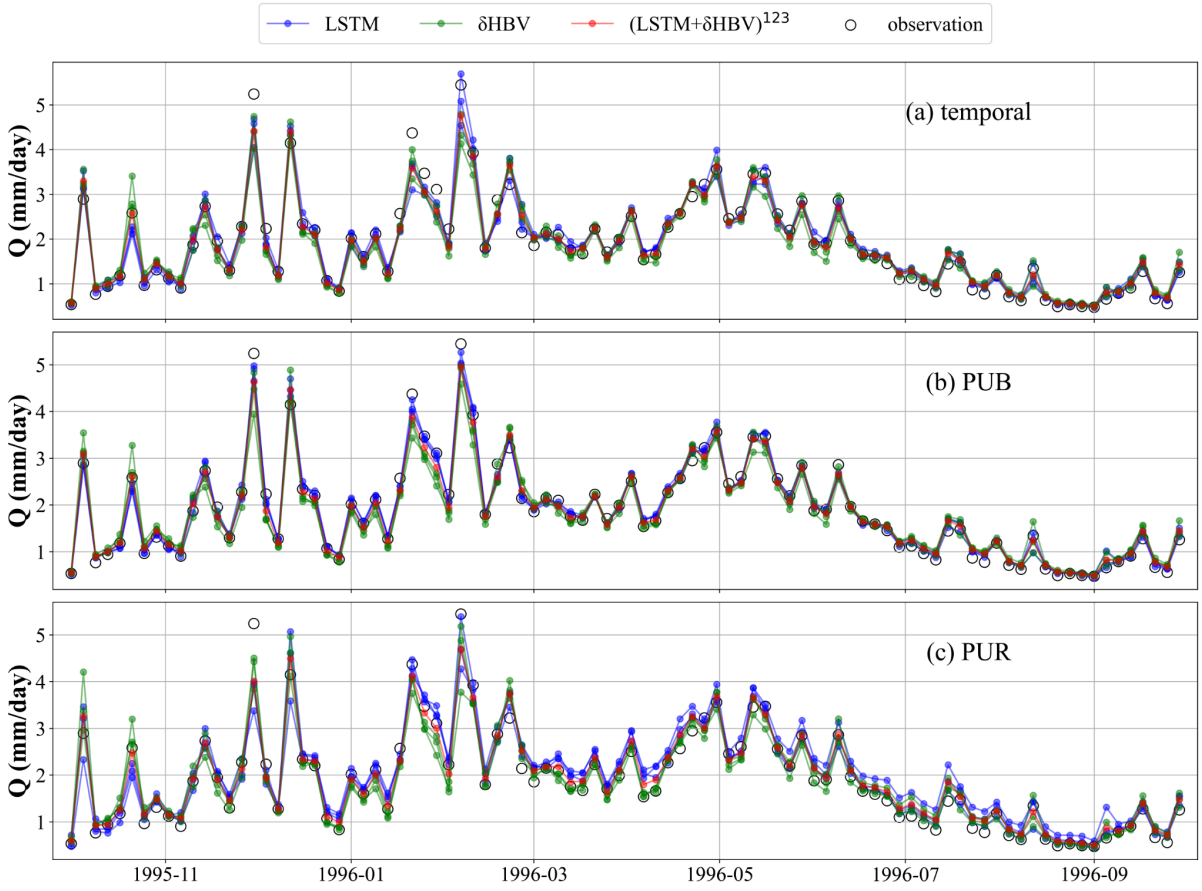

Figure R9. Comparisons between multi-basin-averaged streamflow observations and simulations across 531 basins. The time series points are displayed at four-day intervals for clarity and conciseness. Ensemble members based on the same model (LSTM or δHBV) but driven by different forcing datasets are shown in the same color to highlight the differences between models more clearly.

8. L467-L469: Suggestion to extend the sentence to clarify what is meant here.

Thanks for the suggestion. We will revise the original sentence

"This highlights the critical importance of comprehensive training for each ensemble member to enable the development of distinct characteristics in their streamflow simulations, ultimately enhancing ensemble performance."
as,
"This highlights the critical importance of comprehensive training for each ensemble member, including diverse forcing inputs, full-period model calibration, and rigorous hyperparameter tuning, to ensure that each member develops distinct simulation behaviors. These differences allow the ensemble to better represent a range of hydrological responses, particularly under extreme or uncertain conditions. By capturing complementary strengths and compensating for individual weaknesses, such well-trained ensemble members collectively enhance the robustness and accuracy of streamflow simulations."

Minor comments and technical corrections

1. L13: Replace "while" for "however".

Thanks, it will be revised.

2. L20: Suggest deleting "utilized in two ways". Here it just raises the question which two ways? Also in L526 it would be good to explicitly mention the "two ways".

Thanks for the suggestions. We will delete the "utilized in two ways" in the Abstract section.
As for "*Also in L526 it would be good to explicitly mention the "two ways".*" in the conclusion section, we will revise the original words as,
"Three meteorological forcing datasets (Daymet, NLDAS, and Maurer) are employed to fully capture the characteristics of the two models. Their applications are also tested in two distinct ways: (1) by feeding all diverse forcing datasets simultaneously into a single LSTM model, and (2) by ensembling the outputs of multiple LSTM models, each trained separately using a single forcing dataset."

3. L177-L179: Are all modifications to HBV1.0 of similar importance? Or can it be said which of them are more important?

Thanks for the insightful comment. We are very cautious to make modifications, and to determine these modifications, we have evaluated various structural changes across multiple studies using diverse datasets. Each modification targets specific aspects of model improvement, and most contribute significantly to overall performance.

To address high-flow simulation challenges, we implemented three key modifications: the use of three dynamic parameters (γ, β, θ_{K0}) during training and testing periods; the removal of log-transform normalization for precipitation; and the adoption of the normalized squared-error loss function.

Our recent study (Song et al., 2024) shows that δHBV1.1p with three dynamic parameters (γ, β, θ_{K0}) outperforms the two-parameter version (γ, β). The dynamic shape coefficient (β) and evapotranspiration coefficient (γ) capture the nonlinear relationships between surface soil moisture and effective rainfall, as

well as evapotranspiration. The dynamic θ_{K0} parameter reflects variable water release rates influenced by changing groundwater levels, bank and wetland storage, and other factors. By remaining small during low-flow periods and increasing during peak-flow events, dynamic θ_{K0} allows the upper soil layer to retain more moisture before extreme events, thereby enhancing peak-flow contributions..

The elimination of log-transform normalization for precipitation, paired with the adoption of the normalized squared-error (NSE) loss function, synergistically enhances model performance. By removing log-transform normalization, the model becomes more sensitive to high precipitation events, thus better capturing high-flow conditions. Simultaneously, the NSE loss function amplifies the impact of significant deviations in peak flows, further improving the model's ability to predict high-flow events effectively (Frame et al., 2022; Kratzert et al., 2021; Song et al., 2025a, b; Wilbrand et al., 2023).

In contrast, maintaining dynamic parameters during warm-up periods offers marginal benefits while increasing computational costs. However, it provides a more realistic representation and mitigates potential uncertainties from initial conditions.

Based on your feedback, we will revise the relevant sections with more details for clarity and precision.

4. Table 2: Isn't there more recent data for PUB and PUR? Why are they trained only until 1999?

Thanks for the question. All three tests (temporal, PUB, and PUR) are conducted using the same underlying dataset. However, due to differences in testing strategies, the computational cost for PUB and PUR is significantly higher than for the temporal test. Specifically, each complete evaluation requires 10 runs for PUB and 7 runs for PUR. Based on prior studies (Feng et al., 2021, 2023; Kratzert et al., 2019) and to balance computational efficiency with the objectives of our analysis, we limited the training period to data up to 1999. This choice allows us to preserve the core evaluation goals while keeping the computational demand manageable.

5. Figure B1. Xlabel on right panels should be temperature (C), correct?

Thanks for pointing it out. It has been replotted (Figure R10) and will be added in the revised manuscript.

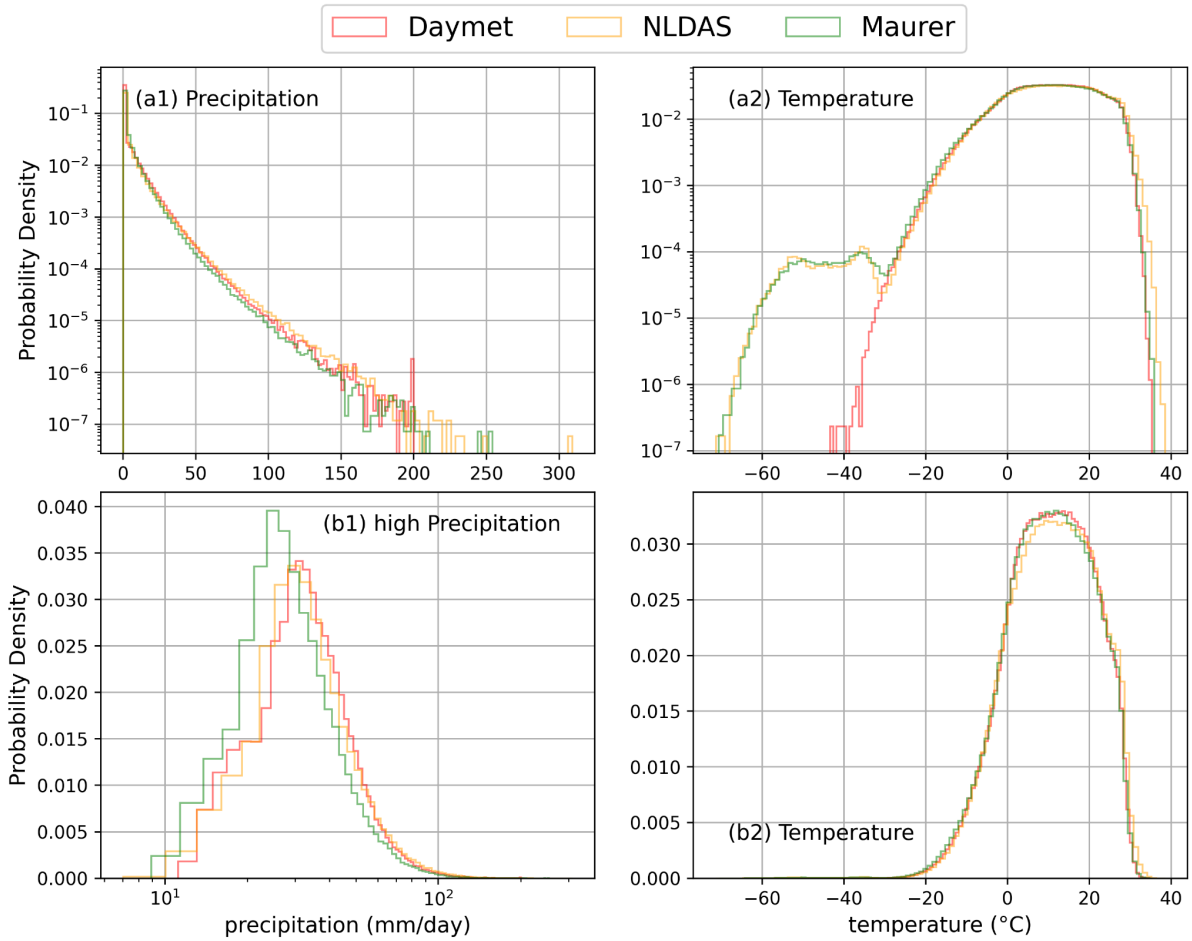

Figure R10. *Probability density distributions of precipitation and temperature across three meteorological forcing datasets.*

References:

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: 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, 2017.

Feng, D., Lawson, K., and Shen, C.: Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data, Geophysical Research Letters, 48, e2021GL092999, https://doi.org/10.1029/2021GL092999, 2021.

Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment, Hydrology and Earth System Sciences, 27, 2357–2373, https://doi.org/10.5194/hess-27-2357-2023, 2023.
Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.
Heidari, H., Arabi, M., Warziniack, T., and Kao, S.-C.: Assessing shifts in regional hydroclimatic conditions of U.S. river basins in response to climate change over the 21st century, Earth's Future, 8, e2020EF001657, https://doi.org/10.1029/2020EF001657, 2020.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11344–11354, https://doi.org/10/gg4ck8, 2019.
Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.
Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, , https://doi.org/10.5281/zenodo.6326394, 2022.
Ryan Bellmore, J., Duda, J. J., Craig, L. S., Greene, S. L., Torgersen, C. E., Collins, M. J., and Vittum, K.: Status and trends of dam removal research in the United States, WIREs Water, 4, e1164, https://doi.org/10.1002/wat2.1164, 2017.
Song, Y., Knoben, W. J. M., Clark, M. P., Feng, D., Lawson, K., Sawadekar, K., and Shen, C.: When ancient numerical demons meet physics-informed machine learning: adjoint-based gradients for implicit differentiable modeling, Hydrology and Earth System Sciences, 28, 3051–3077, https://doi.org/10.5194/hess-28-3051-2024, 2024.
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.
Wada, Y., de Graaf, I. E. M., and van Beek, L. P. H.: High-resolution modeling of human and climate impacts on global water resources, Journal of Advances in Modeling Earth Systems, 8, 735–763, https://doi.org/10/f8wgpv, 2016.
Wilbrand, K., Taormina, R., ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., and Dahm, R.: Predicting streamflow with LSTM networks using global datasets, Front. Water, 5, https://doi.org/10.3389/frwa.2023.1166124, 2023.