Hydrol. Earth Syst. Sci., 29, 6829–6861, 2025
https://doi.org/10.5194/hess-29-6829-2025

Hydrology and
Earth System
Sciences
Open Access

Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation

Peijun Li[1], **Yalan Song**[1], **Ming Pan**[2], **Kathryn Lawson**[1], **and Chaopeng Shen**[1]

[1]Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, USA
[2]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography,
University of California San Diego, La Jolla, CA, USA

Correspondence: Peijun Li (pql5336@psu.edu) and Chaopeng Shen (cshen@engr.psu.edu)

Received: 3 February 2025 – Discussion started: 10 March 2025
Revised: 16 October 2025 – Accepted: 24 October 2025 – Published: 1 December 2025

Abstract. Streamflow simulations produced by different hydrological models exhibit distinct characteristics and can provide valuable information when ensembled. However, few studies have focused on ensembling simulations from models with significant structural differences and evaluating them under both temporal and spatial tests. Here we systematically evaluated and utilized the simulations from two highly different models with great performances: a purely data-driven long short-term memory (LSTM) network and a physics-informed machine learning ("differentiable") HBV (Hydrologiska Byråns Vattenbalansavdelning) model (δHBV). To effectively display the features of the two models, multiple forcing datasets are employed. The results show that the simulations of LSTM and δHBV have distinct features and complement each other well, leading to better Nash-Sutcliffe model efficiency coefficients (NSE) and improved high-flow and low-flow metrics across all spatiotemporal tests, compared to within-class ensembles. Ensembling models trained on a single forcing outperformed a single model using fused forcings, challenging the paradigm of feeding all available data into a single data-driven model. Most notably, δHBV significantly enhanced spatial interpolation when incorporated into LSTM, and provided even more prominent benefits for spatial extrapolation where the LSTM-only ensembles degraded significantly, attesting to the value of the structural constraints in δHBV. These advances set new benchmark records on the well-known CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) hydrological dataset, reaching median NSE values of ~ 0.83 for the temporal test (densely trained scenario), ~ 0.79 for the ungauged basin test (PUB, Prediction in Ungauged Basins), and ~ 0.70 for the ungauged region test (PUR, Prediction in Ungauged Regions). This study advances our understanding of how various model types, each with distinct mechanisms, can be effectively leveraged alongside multi-source datasets across diverse scenarios.

Highlights.

- Combining LSTM and δHBV with diverse forcings sets new accuracy benchmarks
- Ensembling models with one forcing outperforms merging forcings as an input
- δHBV and LSTM together always increase NSEs, especially spatial generalization
- δHBV provides valuable spatial constraints in the deterministic ensemble simulations
- δHBV and LSTM have different error characteristics that can be offset in an ensemble

1 Introduction

Streamflow, a critical component of the global hydrosphere, profoundly influences both human society and natural ecosystems (Lins and Slack, 1999). Accurate simulation and prediction of streamflow yield numerous benefits, including improved flood prevention strategies (Brunner et al., 2021). Hydrological models serve as indispensable tools for achieving this objective and can be traditionally categorized into two types: data-driven models (Feng et al., 2020; Kratzert et al., 2018; Liu et al., 2024; Nearing et al., 2024) and process-based (or physically-based) models (Newman et al., 2017; Paul et al., 2021). Data-driven models, exemplified by long short-term memory (LSTM) (Feng et al., 2020; Kratzert et al., 2018) and transformer (Liu et al., 2024) networks, excel in learning patterns from multi-source data (Li et al., 2023b, 2024; Liu et al., 2022; Nearing et al., 2024) and generally achieve high performance. However, they often lack interpretability and may not resolve extreme values very well (Li et al., 2020a; Song et al., 2025b). Conversely, process-based models, derived deductively from physical laws or conceptualized views of natural systems, offer insights into internal hydrological processes but may exhibit weaker performance due to structural inadequacies (Li et al., 2020a, 2022; Zhang et al., 2019).

To combine the benefits and counteract the weaknesses of these two kinds of models, many efforts have been made to incorporate physical constraints and structures into data-driven models to align with fundamental physical principles, such as mass and water balances (Bandai and Ghezzehei, 2021; Wang et al., 2020; Xie et al., 2021). The most seamless integration uses neural networks to provide parameterizations or missing process representations for process-based models (Aboelyazeed et al., 2023; Bindas et al., 2024; Feng et al., 2022; Jiang et al., 2020; Kraft et al., 2022; Rahmani et al., 2023; Song et al., 2024; Tsai et al., 2021). These differentiable models (Shen et al., 2023) connect (flexible amounts of) prior physical knowledge to neural networks, and have displayed many advantages, including improved computational efficiency and prediction of untrained variables (Tsai et al., 2021), spatial generalization (Feng et al., 2023b), and representation of extremes (Song et al., 2025b). However, it is also unclear whether current differentiable models, e.g., δHBV, the Hydrologiska Byråns Vattenbalansavdelning (HBV) model implemented within a differentiable framework (Feng et al., 2023b; Ji et al., 2025; Shen et al., 2023; Song et al., 2025b), have unique bias characteristics that are associated with the process-based parts of their structures that cannot be reduced once the equations are prescribed.

Orthogonal to such efforts are ensemble simulations (Yu et al., 2024), which combine many members with different biases and uncertainties to mitigate their respective biases in deterministic predictions. Many previous studies have tried ensemble methods to improve streamflow (Clark et al., 2016; Zounemat-Kermani et al., 2021) based on many factors, like initial conditions (e.g., initial weights and biases in LSTM; Kratzert et al., 2018), data used for parameterization (Feng et al., 2021), and objective functions (Lin et al., 2024). These studies generally use one model to generate the differences among the ensemble members. Furthermore, some studies (Dion et al., 2021; Solanki et al., 2025) have utilized simulations from multiple different models but are limited to process-based models, resulting in ensemble simulations that are better than each individual member. Thus far, however, most studies have focused on simulations from only similar models or model types, and little work has tested an ensemble across the boundary of model types, particularly between data-driven, process-based, and hybrid models, especially on a large number of samples. Presumably, if each model has its own unique bias, data-driven and process-based models are likely to exhibit greater differences due to their inherently distinct characteristics. It remains unclear whether ensembling across model types should bring benefits to deterministic predictions. Furthermore, grounded in the process-based model, the differentiable process-based hydrological model, such as δHBV, significantly enhances performance compared to traditional process-based models, while on the other hand introducing greater uncertainty regarding its potential benefits when ensembled. Moreover, previous studies have primarily focused on evaluating ensemble simulations for temporal predictions. However, streamflow simulation under spatial extrapolation scenarios presents greater challenges, and findings from temporal tests may not be directly applicable in this context.

It is known that the performance of any type of hydrologic model heavily depends on the quality of input data, particularly meteorological forcing data (Bell and Moore, 2000; Yao et al., 2020), and other inputs, like the uncertainties of initial conditions, can be mitigated via warming up (Yu et al., 2019). While independent forcing datasets excel in certain aspects, they each carry different error characteristics (Beck et al., 2017; Behnke et al., 2016; Newman et al., 2019) and accordingly affect the hydrological models in different ways. In order to fully display the different features between LSTM and δHBV, multiple forcing datasets could be considered. Given the utilization of multiple forcing datasets, one could choose to use data fusion to combine them into a single coherent model input (Kratzert et al., 2021; Sawadekar et al., 2025), or to pass each forcing dataset through a model and then afterwards combine the multiple outputs in an ensemble. It is not clear which approach is more beneficial.

Considering the knowledge gaps discussed above, we sought to answer several research questions:

1. Will a cross-model-type ensemble of LSTM and δHBV improve deterministic streamflow prediction more than a within-class ensemble?
2. Is it better to use multiple forcings in one model or to ensemble multiple models, each with a different forcing input?
3. Do process-based equations bring unique value to an ensemble, especially in terms of spatial generalizability?

The remainder of this paper is structured as follows: Sect. 2 outlines the hydrological data and models used in this study, as well as the experimental design. Results and discussions are presented in Sect. 3, with conclusions provided in Sect. 4.

2 Materials and methods

2.1 CAMELS hydrologic dataset

The Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) dataset (Addor et al., 2017) is widely employed for hydrological model evaluation and community benchmarking. The CAMELS dataset encompasses 671 basins distributed across the conterminous United States, with basin sizes ranging from 1 to 25 800 km^2 (median: 335 km^2). This standardized and publicly available dataset serves as a benchmark for evaluating various hydrological models, with LSTM models trained on this dataset often serving as a reference point for comparing other models (Kratzert et al., 2021). CAMELS provides basin-scale data, including streamflow observations and static basin attributes, as well as forcing datasets from three independent sources: Daymet (Thornton et al., 1997), North American Land Data Assimilation System (NLDAS) (Xia et al., 2012), and Maurer (Maurer et al., 2002). Each of the three meteorological forcing datasets operates at a daily temporal resolution, encompassing precipitation, temperature, vapor pressure, and surface radiation variables, with daily temperature extrema of NLDAS and Maurer supplemented from Kratzert et al. (2021). These three meteorological forcing datasets have methodological distinctions in spatial resolution, data generation approaches, and temporal processing (Behnke et al., 2016; Kratzert et al., 2021). Exemplary plots illustrating the differences among the three meteorological forcing datasets are provided in Appendix B. These features can lead to dataset-specific error characteristics and make them valuable for displaying the distinct features of different model types. All model inputs used in this study are detailed in Table C1 in Appendix C.

2.2 Long short-term memory

As one kind of deep learning algorithm, long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) has unique structures like hidden states and gates activated by the tanh and sigmoid functions (Li et al., 2023a), respectively. These features enable LSTM to excel in streamflow simulation tasks (Feng et al., 2020; Kratzert et al., 2018; Nearing et al., 2024). In the current benchmark framework, LSTM models are trained using dynamic atmospheric forcings and static basin attributes as inputs, with streamflow as the target output, making it perform well in both temporal and spatial tests (Fig. 1a). In this work, for cross-group comparability, we used the LSTM model and its hyperparameters as reported in Kratzert et al. (2021).

2.3 Differentiable HBV model (δHBV)

The Hydrologiska Byråns Vattenbalansavdelning (HBV) model is a parsimonious bucket-type hydrologic model that simulates various hydrological variables, including snow water equivalent, soil water, groundwater storage, evapotranspiration, quick flow, baseflow, and total streamflow (Aghakouchak and Habib, 2010; Beck et al., 2020; Bergström, 1976, 1992). Recently demonstrated differentiable HBV (δHBV) model (Feng et al., 2023b; Ji et al., 2025; Shen et al., 2023; Song et al., 2024) incorporates deep neural networks for both regionalized parameterization and missing process representations within a differentiable programming framework that supports “end-to-end” training (Fig. 1b). This innovation enables δHBV to effectively learn from data while obeying physical laws, resulting in high-level performance for streamflow simulations. From the perspective of process-based modeling, LSTM is a regionalized parameter provider that leverages the autocorrelated nature of its inputs to impose an implicit spatial constraint on the generated parameters.

In this study, we used δHBV1.1p (Song et al., 2024, 2025b), which is an updated version of δHBV1.0 (Feng et al., 2022, 2023b). The main improvement is the addition of a capillary rise module, which enhances the characterization of low flows. Three additional modifications are included to address high-flow simulation challenges: the use of three dynamic parameters (γ, β, k_0) (Song et al., 2025b); the removal of log-transform normalization for precipitation; and the adoption of the normalized squared-error loss function (Table C2) (Frame et al., 2022; Kratzert et al., 2021; Song et al., 2025a, b; Wilbrand et al., 2023). We also maintain dynamic parameters during warm-up periods. Although this provides only marginal benefits and increases computational costs, it yields a more realistic representation and reduces uncertainties associated with initial conditions. The basic equations in δHBV are as follows:

$$\theta = \mathrm{LSTM}_\mathrm{w}(\overline{x}, \overline{A_\mathrm{attr}}) \tag{1}$$

$$Q = \mathrm{HBV}(x, \theta) \tag{2}$$

$$W_\mathrm{opt} = \mathrm{argmin}_\mathrm{w}(L(Q, Q^*)) \tag{3}$$

where θ are the dynamic or static physical parameters, w denotes the weights and biases of LSTM, x includes the basin-averaged meteorological forcings, such as precipitation, mean temperature, and potential evapotranspiration, with $\overline{x}$ representing their normalized versions. Similarly, $\overline{A_\mathrm{attr}}$ consists of normalized observable basin-averaged attributes, encompassing basin area, topography, climate, soil texture, land cover, and geology (Table C1). Precipitation and mean temperature are from CAMELS, while potential evapotranspiration is calculated using the Hargreaves (1994) method based on maximum and minimum temperatures along with basin latitudes, all from data described in Sect. 2.1. Q and Q^* are the streamflow simulations (model

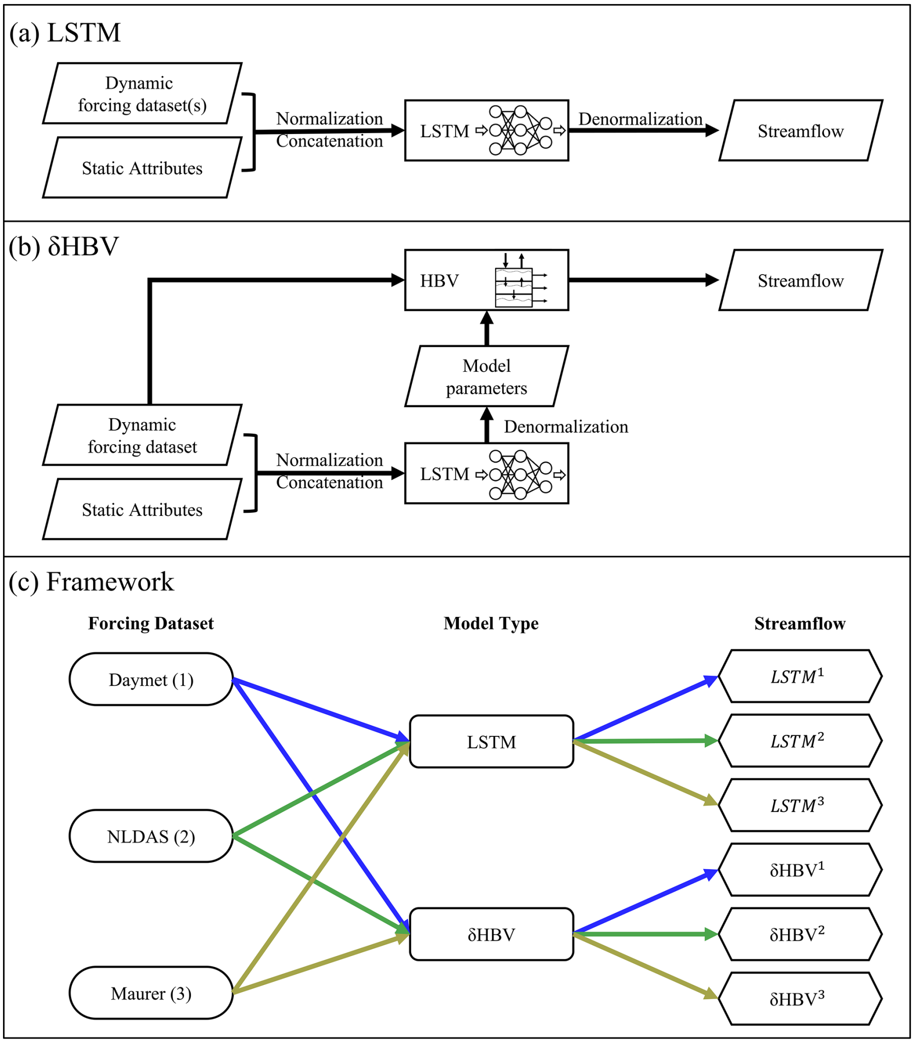

Figure 1. (a) The LSTM structure, **(b)** the δHBV structure, and **(c)** the framework to generate the six individual ensemble members of the streamflow simulations, in which different colors of arrow lines denote the different meteorological forcing datasets (also denoted as 1, 2, 3), respectively.

outputs) and observations (as provided in CAMELS), respectively. HBV is implemented on PyTorch so it is programmatically differentiable: all steps store information related to gradient calculations during backpropagation, allowing this model to be trained together with neural networks in an end-to-end fashion. More details about differentiable HBV can be found in previous studies (Feng et al., 2022; Song et al., 2024). The details of some particularly relevant HBV processes are described in Appendix A.

2.4 Experimental design

In this study, we trained the two models of very different types (LSTM and δHBV), each with one of three meteorological forcing datasets (Daymet, NLDAS, and Maurer), resulting in six corresponding streamflow simulations (Fig. 1c) for each different test scenario (see Sect. 2.5 for additional information). The training processes of LSTM and δHBV followed Kratzert et al. (2021) and Feng et al. (2023b), respectively. Test results and performance metrics for all models are reported for the 531-basin subset that excludes those with areas larger than 2000 km^2 or with more than a 10 % discrepancy between different basin area calculation methods (Newman et al., 2017).

To generate ensembles, we tested various weighting strategies and ultimately employed averaging to combine the six single-forcing, single-model-type simulations, as it yielded the best performance. To better describe various combinations including cross-model ensembles, these simulations were categorized into six groups (Table 1). A shorthand notation is used throughout the remainder of this work to describe the forcing datasets and ensembles. Daymet, NLDAS, and Maurer are abbreviated as superscripts 1, 2, and 3, respectively. The + symbol is used to group model types being ensembled, while superscript clustering (e.g., ^{12} or ^{123}) is used to group the meteorological forcing types being ensembled, with parentheses indicating that the superscripts apply to all model types within. For example, $(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}$ could be explicitly written as $\mathrm{LSTM}^1+\mathrm{LSTM}^2+\mathrm{LSTM}^3+\delta\mathrm{HBV}^1+\delta\mathrm{HBV}^2+\delta\mathrm{HBV}^3$. To compare two different strategies to utilize the multiple meteorological forcing datasets and to benchmark against the previously highest performance, we additionally trained a single LSTM model using all three forcing datasets as simultaneous inputs as done by Kratzert et al. (2021), referred to as $\mathrm{LSTM}^{\mathrm{multi}}$ (the last row in Table 1).

2.5 Evaluation scenarios and criteria

The above cases were comprehensively evaluated for performance in temporal extrapolation (Feng et al., 2022; Kratzert et al., 2018), as well as two types of spatial generalization: prediction in ungauged basins (PUB) (Feng et al., 2023b; Kratzert et al., 2019), and prediction in ungauged regions (PUR) (Feng et al., 2021, 2023b):

- *Temporal Test*: Models were trained using data from all basins and tested across different periods.
- *PUB Test*: Models were trained on randomly selected subsets from all basins and tested on the remaining basins during the same time period.
- *PUR Test*: Different from the PUB test, basins were grouped into continuous regions, one of which was selected to comprise the group of testing basins while the others were used for training.

Temporal generalization is generally considered to be the easiest of these tests. In terms of spatial generalization, which approximates data-sparse scenarios, the PUB test is an example of spatial interpolation, whereas the PUR test involves spatial extrapolation. The PUR test is widely regarded as the most challenging and may therefore produce findings that differ significantly from those in other scenarios. In this study, all basins were divided into 10 spatially stratified groups for the PUB test and 7 fully disjoint regional groups for the PUR test (Table 2) in the same way as Feng et al. (2023b). The spatial extent of the 7 regions for the PUR test is also shown in Fig. 3c1–c2. Therefore, we conducted 10 rounds for the PUB test and 7 rounds for the PUR test, with a different group held out for testing in each round. Model performance was evaluated after concatenating the test results for all basins.

We repeated all the simulations with three different random seeds. Therefore, all the simulations come from a total of $(2\times3+1)\times(1+10+7)\times3$ trained models. The first factor represents the models: two model types (LSTM and δHBV) trained separately with each of the three forcing datasets, along with $\mathrm{LSTM}^{\mathrm{multi}}$, a single model instance trained using all three forcing datasets simultaneously. The second factor accounts for the three types of tests (temporal, PUB, and PUR tests), and the last for the three random seeds. With respect to random seeds, we present two variations in the results, which are visually depicted in Fig. C1 in Appendix C. The results without "seed" as a subscript represent the average metric values from multiple streamflow simulations, each generated from a single model implementation, along with the corresponding uncertainties, visualized using error bars. The results marked with "seed" as a subscript are based on the average of multiple streamflow simulations conducted with different random seeds. In terms of computational cost, training LSTM (30 epochs) and δHBV (50 epochs) for temporal testing under a single meteorological forcing dataset takes approximately 5 and 21 h, respectively, using a single NVIDIA Tesla V100 GPU.

We calculated several well-established performance metrics: Nash-Sutcliffe model efficiency coefficient (NSE) (Nash and Sutcliffe, 1970), Kling-Gupta model efficiency coefficient (KGE) (Kling et al., 2012), percent bias (PBIAS), and root-mean-square error (RMSE). We also considered RMSE values for high (top 2 % "peak" flow, highRMSE), low (bottom 30 % "low" flow, lowRMSE), and mid-range (the remaining flow, midRMSE) flow conditions (Yilmaz et al., 2008). These metrics were computed for each basin and aggregated into error bars and cumulative density functions (CDFs). For brevity, the main text primarily reports NSE values, and other metric values are provided in Appendices D and E. Furthermore, we use the spread values (Li et al., 2021; Reichle and Koster, 2003) to investigate ensemble variability and explore model complementarity. Detailed descriptions of these metrics and their calculations are available in Table C2.

3 Results and discussion

3.1 Temporal extrapolation

For the temporal test, in which models were trained and tested on the same basins but in different time periods, we found that cross-model-type ensembles noticeably surpassed the within-class ensembles when other conditions were the same, with small uncertainties as shown by the error bars in Fig. 2. With a single forcing dataset, the median NSE was elevated from ~ 0.735 for LSTM to ~ 0.79 with δHBV added, though δHBV performance was similar to LSTM (~ 0.74 under Daymet). Even after LSTM achieved very high performance when its simulations, each derived separately from different meteorological forcing datasets, were ensembled (ef = 123, ~ 0.808), adding δHBV still improved the results to ~ 0.818. This finding was robust for all different combinations of the tested meteorological forcing datasets. Conversely, adding LSTM also helped to improve δHBV ensembles. These results highlight the benefits of the cross-

Table 1. **(a)** The six groups of streamflow simulations, and **(b)** the streamflow simulation via LSTM based on a different strategy, in which three meteorological forcing datasets were combined as a single set of inputs (Kratzert et al., 2021). Superscripts 1, 2, and 3 denote Daymet, NLDAS, and Maurer, respectively. The ensemble across forcings ("ef") superscript indicates an ensemble of model simulations, each of which uses a different single meteorological forcing, e.g., LSTM^{12} means the average of LSTM^{1} and LSTM^{2}.

(a) Six Groups of Streamflow Simulations		
Group Name	Group Members	
LSTM	LSTM^1, LSTM^2, LSTM^3	
δHBV	$\delta\mathrm{HBV}^1$, $\delta\mathrm{HBV}^2$, $\delta\mathrm{HBV}^3$	
LSTM + δHBV	$(\mathrm{LSTM}+\delta\mathrm{HBV})^1$, $(\mathrm{LSTM}+\delta\mathrm{HBV})^2$, $(\mathrm{LSTM}+\delta\mathrm{HBV})^3$	
$\mathrm{LSTM}^{\mathrm{ef}}$	LSTM^{12}, LSTM^{13}, LSTM^{23}, LSTM^{123}	
$\delta\mathrm{HBV}^{\mathrm{ef}}$	$\delta\mathrm{HBV}^{12}$, $\delta\mathrm{HBV}^{13}$, $\delta\mathrm{HBV}^{23}$, $\delta\mathrm{HBV}^{123}$,	
$(\mathrm{LSTM}+\delta\mathrm{HBV})^{\mathrm{ef}}$	$(\mathrm{LSTM}+\delta\mathrm{HBV})^{12}$, $(\mathrm{LSTM}+\delta\mathrm{HBV})^{13}$, $(\mathrm{LSTM}+\delta\mathrm{HBV})^{23}$, $(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}$	
(b) Using forcing datasets as simultaneous inputs to an LSTM		
Streamflow Simulation	Model Type	Meteorological Forcing Dataset
$\mathrm{LSTM}^{\mathrm{multi}}$	LSTM	Daymet, NLDAS, Maurer

Table 2. Differences of temporal, PUB, and PUR tests.

Test Scenario	Training		Testing	
	Basin	Time	Basin	Time
Temporal	All[a]	1980–1995[b]	All	1995–2010
PUB	Random nine-tenths	1980–1999	Holdout[c]	1995–1999
PUR	Random six of seven regions	1980–1999	Holdout	1995–1999

[a] δHBV training followed Feng et al. (2023b) using all 671 CAMELS basins, while LSTM training followed Kratzert et al. (2021) using the selected 531-basin subset. Test results and performance metrics for all models are reported for the 531 basins. [b] Each hydrological year spans from 1 October to 30 September of the following year. [c] In the PUB and PUR tests, models are run for 10 and 7 rounds, respectively, with the group held out for testing changed in each round. The simulation performance was evaluated after concatenating the test results for all basins.

model-type ensemble framework and indicate distinct simulation features for each model type. LSTM is a data-driven method that has low bias and large variance. Data errors (Li et al., 2020b), different sampling strategies (Nai et al., 2024), or even different weight initializations (Narkhede et al., 2022) can lead to substantively different outcomes. Conversely, δHBV may have a smaller variance but a larger bias due to the fixed HBV formulation (Moges et al., 2016) for some scenarios like low flows (Feng et al., 2023b; Song et al., 2024) or in basins with significant water uses (Song et al., 2025a). These errors with varying characteristics from different model classes can partially offset each other in an ensemble. On a side note, δHBV models seem more reliant on the quality of the forcing data, as shown in Fig. 2. δHBV with the Maurer and NLDAS forcing datasets generally performs worse than it does with Daymet, which has lower biases. However, even in those cases, the combination of LSTM and δHBV was still better than LSTM alone, attesting to the robustness of these benefits.

In the lower-performing basins where LSTM^1 had advantages over $\delta\mathrm{HBV}^1$, the ensemble of meteorological forcings $\delta\mathrm{HBV}^{123}$ also tended to be higher than $\delta\mathrm{HBV}^1$ (Fig. 3), suggesting that forcing quality was a significant reason behind the underperformance of $\delta\mathrm{HBV}^1$ in these basins. Similar patterns were also observed when analyzing $\delta\mathrm{HBV}^2$ and $\delta\mathrm{HBV}^3$ values (Figs. D1 and D2 in Appendix D). These basins previously contributed to LSTM's cumulative distribution function of NSE diverging from that of $\delta\mathrm{HBV}^1$ at the low end (Feng et al., 2022). Forcing errors can exist in the form of systematic timing errors, low or high bias for larger events, etc., which can be difficult for the mass-balanced conceptual HBV^1 structure to adapt to these errors. Because the ensemble of forcings tends to suppress the errors in each forcing source, part of the advantages of $\delta\mathrm{HBV}^{123}$ over $\delta\mathrm{HBV}^1$ can be attributed to reducing forcing bias or timing errors. Since the advantages of LSTM^1 over $\delta\mathrm{HBV}^1$ also tend to occur with these same basins, this also explains how LSTM^1 surpasses $\delta\mathrm{HBV}^1$ in some basins with poorer-quality forcings. In contrast to δHBV, LSTM has the innate ability to shift information in time and moderately adjust the input scale. Moving from temporal validation to PUB to PUR scenarios, the advantages of diverse forcing datasets appear to diminish,

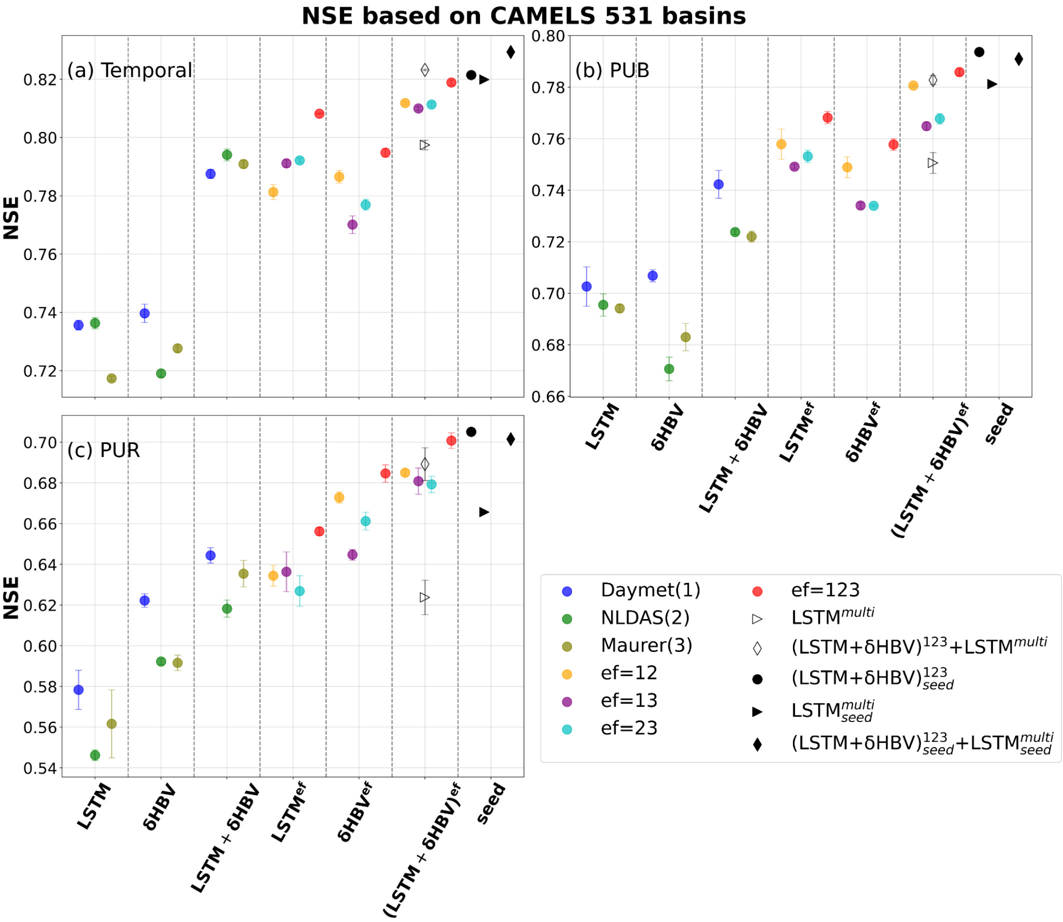

Figure 2. Median NSE values for 531 CAMELS basins, indicating model and ensemble performances for **(a)** temporal, **(b)** prediction in ungauged basin (PUB), and **(c)** prediction in ungauged region (PUR) tests. Different simulations are represented by variously-shaped and -colored points, and are organized by ensemble group, listed along the x-axis: LSTM, δHBV, LSTM + δHBV, and their "ensemble forcing" counterparts, $\mathrm{LSTM}^{\mathrm{ef}}$, $\delta\mathrm{HBV}^{\mathrm{ef}}$, and $(\mathrm{LSTM} + \delta\mathrm{HBV})^{\mathrm{ef}}$. $\mathrm{LSTM}^{\mathrm{multi}}$ is a single LSTM model trained directly on all three forcing datasets at once. The superscript "ef" denotes the forcing datasets involved in each ensemble (choices of 1 for Daymet, 2 for NLDAS, and 3 for Maurer), while the "+" connects the model types used within an ensemble. The x-axis group and subscript "seed" indicate that simulation results were averaged based on three different random seeds (see Fig. C1). Other points without "seed", along with their corresponding error bars, are derived from the averages of metrics computed over repeated runs with three different random seeds. The error bar indicates one standard deviation above and below the average value for each simulation.

as evidenced by the decreasing ratio of points above versus below the diagonal line, since the forcing error patterns remembered by LSTM may not generalize well in space (discussed in more detail in Sect. 3.2).

Ensembling streamflow simulations from different meteorological forcing datasets demonstrates certain advantages over the previous approach of simultaneously sending multiple forcings into a data-driven model like LSTM (Kratzert et al., 2021). Ensembling LSTM simulations each using a single forcing dataset (LSTM^{123}) resulted in an NSE value of 0.8082, higher than that of 0.7974 from feeding multiple forcing datasets into a single LSTM ($\mathrm{LSTM}^{\mathrm{multi}}$). This difference was more pronounced in the cross-model-type ensemble, after including δHBV, compared to the previous within-class ensemble, and particularly notable for the spatial generalization tests (to be discussed in more detail in Sect. 3.2). The corresponding specific performance metrics are summarized in Tables D1–D5 in Appendix D, with seasonal evaluations provided in Fig. D3. These results indicate that the trained LSTM in $\mathrm{LSTM}^{\mathrm{multi}}$ may be overfit to the significant redundant information in these three forcing datasets, and that LSTM models alone cannot fully exploit the information hidden in the multiple forcing datasets. Training separate ensemble members via different nonlinear hydrological processes, on the other hand, seems to allow different bias features to emerge with separate forcing datasets, accordingly mitigating them during the subsequent ensembling process.

Our most diverse ensemble, $(\mathrm{LSTM} + \delta\mathrm{HBV})^{123}_{\mathrm{seed}} + \mathrm{LSTM}^{\mathrm{multi}}_{\mathrm{seed}}$, achieved a median NSE value of ∼ 0.83, surpassing the ∼ 0.82 benchmark set by $\mathrm{LSTM}^{\mathrm{multi}}_{\mathrm{seed}}$ (Table D4).

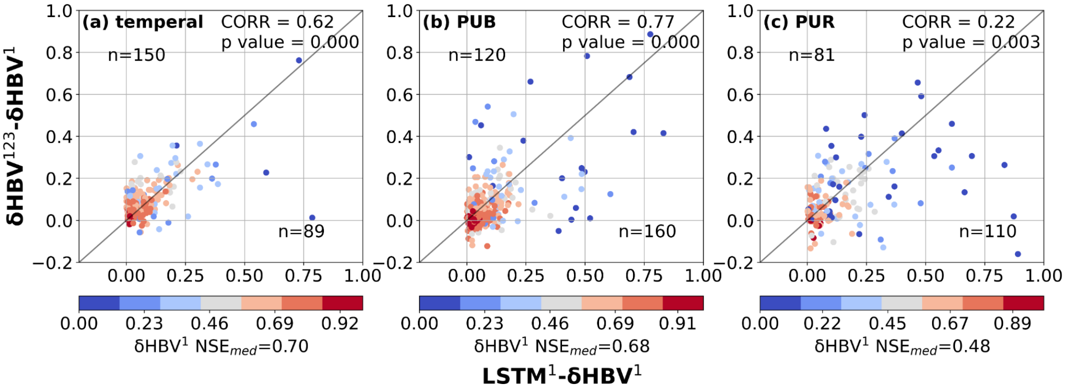

Figure 3. Scatter plots comparing the performance differences between hydrological models for the basins where LSTM outperformed δHBV (the basins where δHBV outperformed are not shown in this plot). The x-axis represents the NSE differences between LSTM^1 and δHBV^1 (LSTM^1 – δHBV^1), while the y-axis shows the NSE differences between δHBV^{123} and δHBV^1 (δHBV^{123} – δHBV^1). Points are color-coded according to the NSE values of δHBV^1. The correlation coefficient (CORR) and p values between the x-axis values and the y-axis values, along with the median NSE value of δHBV^1 (NSE$_{\text{med}}$) on these basins, are also noted. We note that NSE is not additive and should generally not be subtracted. Here the purpose is only to confirm that basins where LSTM outperforms δHBV also tend to be those that benefit from the ensemble of forcings.

This advancement was achieved through random seed variation and cross-model-type ensembling. The performance of (LSTM + δHBV)^{123} ensemble proved more robust than LSTM$^{\text{multi}}$, with only a slight boost when we incorporated random seeds, i.e., (LSTM+δHBV)$^{123}_{\text{seed}}$. Notably, the derived (LSTM + δHBV)$^{123}_{\text{seed}}$ ensemble outperformed LSTM^1 across almost all basins (Fig. 4). Further incorporation of LSTM$^{\text{multi}}$ into this framework, especially when using multiple random seeds, (LSTM + δHBV)$^{123}_{\text{seed}}$ + LSTM$^{\text{multi}}_{\text{seed}}$, yielded the best overall performance. Here, the margin over the previous benchmark was small in the temporal test. However, as we will show in Sect. 3.2, the previous benchmark, LSTM$^{\text{multi}}_{\text{seed}}$, lacked robustness, exhibited greater deficiencies in spatial generalization, and negatively impacted ensemble simulations.

When we changed the number of random seeds from 3 to 10, we found that although all model and ensemble performances slightly improved, the gaps between them did not change much (Fig. 5; Table D5 for 10 seeds, Table D4 for 3 seeds). In particular, the gap between (LSTM + δHBV)$^{123}_{\text{seed}}$ + LSTM$^{\text{multi}}_{\text{seed}}$ and (LSTM + δHBV)$^{123}_{\text{seed}}$ or LSTM$^{\text{multi}}_{\text{seed}}$ remained unchanged. This indicates that the benefits from more random seeds rapidly become marginal, and our results based on 3 random seeds were sufficiently robust. For LSTMs alone, different random seeds displayed higher variation, and ensembling them led to greater improvement than ensembling (LSTM + δHBV)^{123} with additional random seeds. It was noteworthy that while the (LSTM+δHBV)^{123} ensemble generally showed the lowest RMSE values, it did not always show the best high flow performance, as indicated by highRMSE (Tables D1–D4). After incorporating the LSTM$^{\text{multi}}_{\text{seed}}$ variant into (LSTM+δHBV)$^{123}_{\text{seed}}$+LSTM$^{\text{multi}}_{\text{seed}}$, overall RMSE and highRMSE both improved. Nevertheless, this ensemble did not always obtain the best values in other metrics like low flow (lowRMSE) and requires further improvement.

3.2 Spatial generalization

It is clear that cross-model-type ensembling and the incorporation of δHBV significantly improved prediction in ungauged basins (PUB) or regions (PUR), mitigating the difficulty of spatial generalization (Fig. 2b–c). In particular, the previous record-holder for temporal test performance, LSTM$^{\text{multi}}_{\text{seed}}$, incurred large drops in the PUB and PUR tests, once again reminding us of the limitations of LSTM in spatial generalization. Given the same forcings, δHBV-only individual simulations or ensembles consistently outperformed LSTM-only counterparts in the PUR test. Furthermore, adding δHBV to the same-model-type LSTM ensembles improved median NSE by 0.02–0.03 for PUB. The role of δHBV became even more prominent in the harder PUR tests, with an increased gap (0.04–0.07), e.g., LSTM^{123} (median NSE $\sim$0.656) and (LSTM + δHBV)^{123} (median NSE $\sim$0.701). The increased significance of δHBV is also illustrated by the optimized weights shown in Fig. E1 in Appendix E, which were estimated using a genetic algorithm with streamflow observations from the test periods. These weights are presented solely to illustrate the relative contributions of the different ensemble components. The significantly different spatial distribution patterns of these weights among different test scenarios also indicate the differences among temporal, PUB, and PUR tests (Figs. E2–E3). The performance of (LSTM+δHBV)^{123} improved compared to LSTM$^{\text{multi}}$ regardless of whether multiple random seeds were employed to form an ensemble. As such, we can conclude that the inclusion of a differentiable process-based

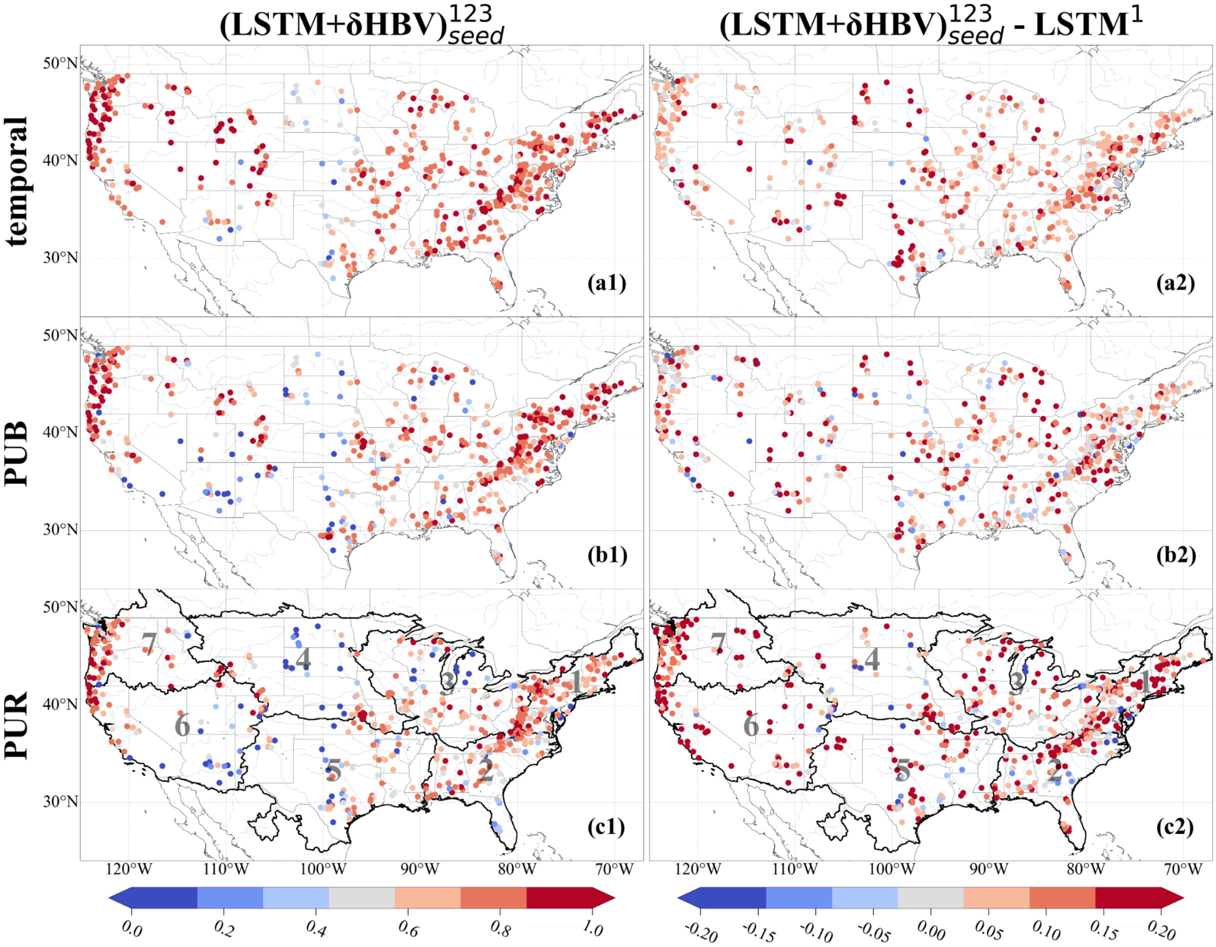

Figure 4. Spatial distributions of NSE values over 531 basins. Subplots are arranged in rows, indicating **(a)** temporal, **(b)** PUB, and **(c)** PUR test results, and columns, denoting (1) NSE values from $(\text{LSTM}+\delta\text{HBV})^{123}_{\text{seed}}$ and (2) the differences between these NSE values and those of LSTM^1 (models using only forcing 1, Daymet). For LSTM^1 each NSE value reported was the average of three NSE values from three simulations using three different random seeds. The seven continuous regions used to divide up basins for the PUR test are outlined and numbered in the PUR test maps.

model like δHBV in an ensemble is a systematic way to reduce the risks of failed generalizations of LSTM.

Utilizing a cross-model-type ensemble led to widespread improvements over LSTM-only ensembles, with the exception of a few scattered basins for each temporal (Fig. 4a2), PUB (Fig. 4b2), and PUR (Fig. 4c2) test. The most significant improvements due to the ensemble were concentrated on the center of the Great Plains along with the midwestern US, while the eastern US was moderately improved, suggesting data uncertainty is a larger issue in the central and midwestern US. The Great Plains have historically had poor performance for all kinds of models (Mai et al., 2022) and even the ensemble model had NSE values of only 0.3–0.4 for many of the basins there, although this still marked significant improvements over LSTM^1 (Fig. 4a2, b2, c2). Some western basin NSE values were elevated by more than 0.15 for the temporal test (Fig. 4a2) and even more for PUB and PUR. Meteorological stations are generally sparse on the Great Plains, and an ensemble seems to be an effective way to leverage the different forcing datasets that are available. The poor performances in some basins highlight some remaining deficiencies in current models, which clearly cannot fully consider the heterogeneities of different basins; thus, multiscale formulations that resolve such heterogeneities may have advantages (Song et al., 2025a).

To investigate why ensembles outperformed single-model, single-forcing approaches, we compared their temporal, PUB, and PUR test simulation time series against observations for 531 basins (Fig. 6). Analysis of averaged hydrological year data revealed that while individual ensemble members using single-source forcing datasets performed similarly for easily simulated periods, they showed significant divergence during challenging periods, particularly peak flows. This divergence stems from distinct systematic errors inher-

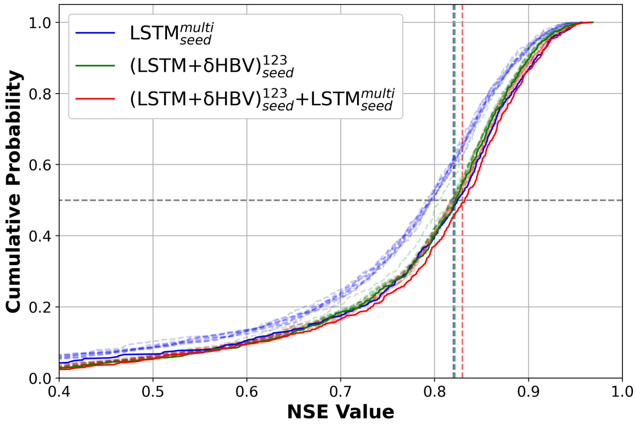

Figure 5. Cumulative distribution function (CDF) curves based on temporal test results for LSTM$^{\text{multi}}$, (LSTM + δHBV)^{123}, and [(LSTM + δHBV)^{123} + LSTM$^{\text{multi}}$]. The solid lines (with "seed") denote the results with 10 random seeds while the corresponding dashed and translucent lines denote the performances of their individual members each based on one random seed. The median NSE values computed with 3 random seeds are also indicated by vertical dashed and translucent lines in the corresponding colors.

ent to different model types and forcing datasets. Notably, LSTM-based simulations alone proved insufficient in generating adequate spread to capture these divergent points. By averaging individual model outputs and stabilizing uncertainties, ensemble simulations achieved effective and robust performance across all conditions, which can be shown via the metric highRMSE and lowRMSE values in Tables D1–D4. 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.

3.3 Ensemble variability and robustness analysis

Although δHBV (median spread 0.61) exhibits lower spreads than LSTM (mean spread 0.72), their combination increases the ensemble spreads, thereby enhancing diversity (Fig. 7). This pattern holds across the temporal, PUB, and PUR tests. Ensemble effectiveness depends on the diversity of model behaviors and their distinct error characteristics. Consequently, larger spreads are generally associated with greater ensemble benefits. Figure D4 further demonstrates that δHBV + LSTM exhibits larger spreads than LSTM in most basins.

As the warming signal is already clear across most basins under any forcing across the periods of simulation (Fig. D5), the models' strong performance in the temporal test suggests decent extrapolation capability under warming scenarios. It is often questioned whether data-driven models like LSTM lose accuracy under stronger climate drift, but no substantially warmed dataset is available to test this. Benchmarks suggest LSTM captures 15-year trends well in temporal tests, but less so in data-sparse scenarios (Feng et al., 2023b). Introducing a 10 % precipitation perturbation (multiplying precipitation by 1.1) slightly reduced performance for both models as expected (Fig. D6a and b), but ensemble benefits remained robust across models despite the perturbation.

Training sample size, dynamic parameter choices, and lookback windows exert only a limited impact on our conclusions. δHBV shows limited sensitivity to sample size, with similar results when trained on 531 versus 671 basins (Fig. D6c). Regarding parameter uncertainties, fixing one δHBV parameter (k_0) as static increased structural errors and reduced performance (Fig. D6d), yet ensemble benefits remained robust. For LSTM, alternative window sizes of 182 and 730 d were tested, with the default 365 d window yielding optimal performance (Fig. D6e). Importantly, variations in the lookback window had only minor effects on model performance, underscoring the robustness of ensemble benefits.

3.4 Further discussion

Based on our results, we identified several avenues for future research. First, while we have explored various weighting strategies and found that averaging yields the best performance yet, we believe that dynamic or adaptive weighting schemes could further enhance performance in future studies. It is also demonstrated by Table E1 that estimated uneven weights can significantly improve simulation performance. Moreover, within specific basins, the estimated weights of different components are often highly imbalanced, as evidenced by the spatial distribution of optimized weights (Figs. E2–E3). Some potential feasible ways include using the simulations from these individually-trained models as inputs of a data-driven model (Solanki et al., 2025), and making the weight estimation and the ensemble member training simultaneously.

Both LSTM and δHBV models exhibit limitations in regions with significant anthropogenic impacts, such as dam presence, as well as arid climatic and highly heterogeneous geological conditions. These regions are mainly located in the midwestern and western CONUS, where high evaporation conditions (Heidari et al., 2020) and numerous dams (Bellmore et al., 2017) coincide with complex water use processes (Wada et al., 2016) that current models cannot simulate well. Together, these factors suggest that anthropogenic influence is likely an important driver of poor model performance. Further improvements may include incorporating additional data that capture these factors like capacity-to-runoff

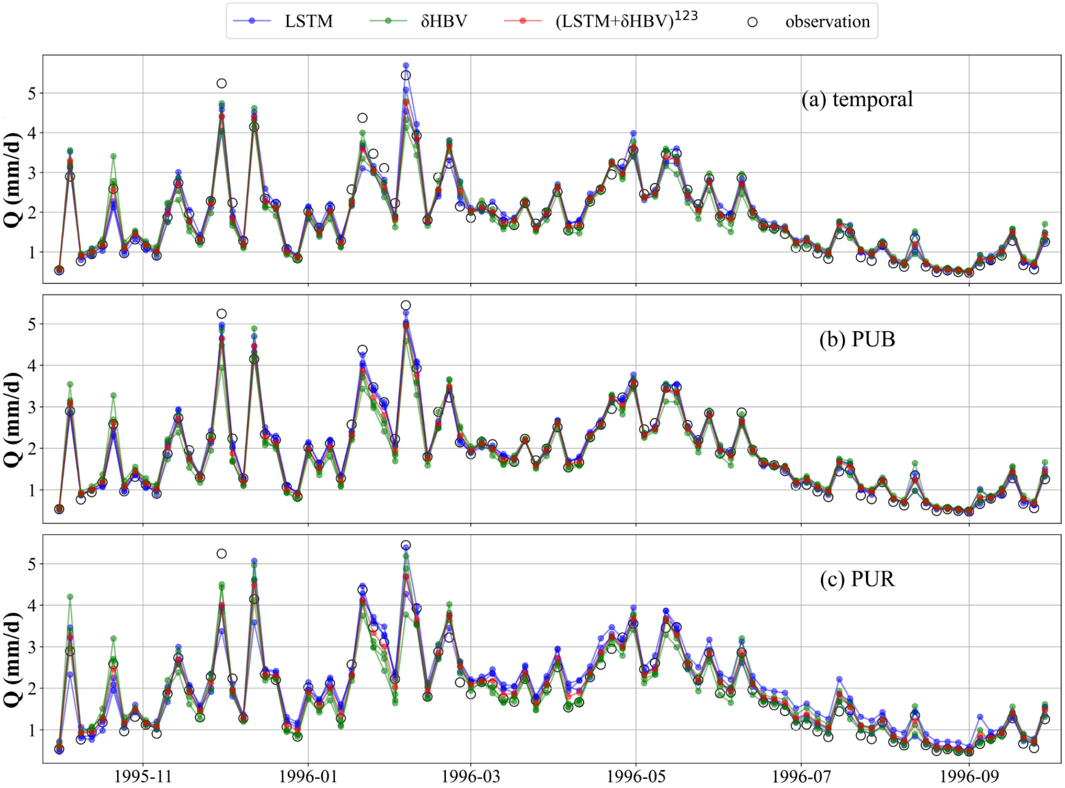

Figure 6. Comparisons between multi-basin-averaged streamflow observations and simulations across 531 basins. The time series points are displayed at four-day intervals for clarity and conciseness. Ensemble members based on the same model (LSTM or δHBV) but driven by different forcing datasets are shown in the same color to highlight the differences between models more clearly.

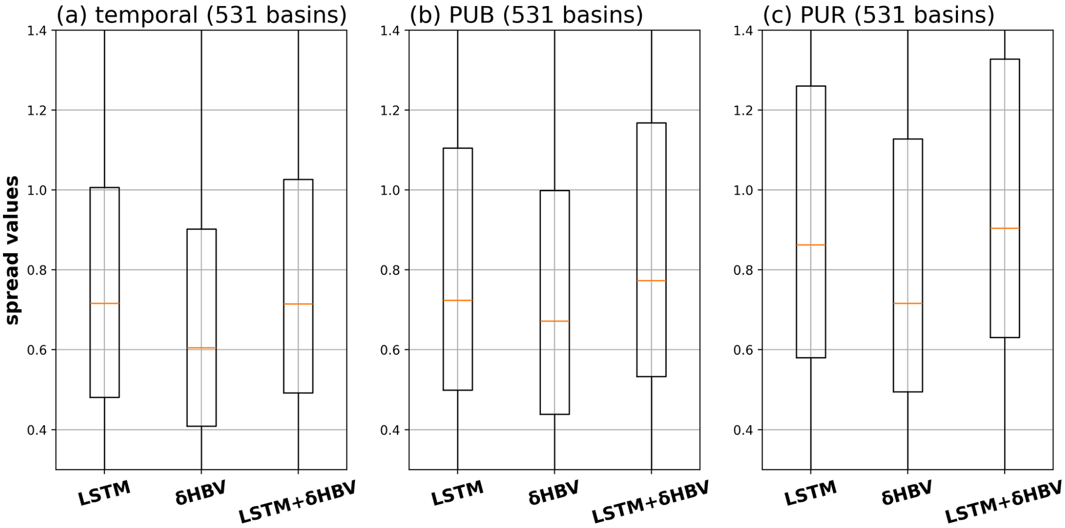

Figure 7. Spread values (Table C2) of each model for LSTM, δHBV, and LSTM + δHBV due to different meteorological forcings and random seeds across temporal, PUB, and PUR tests.

ratio (Ouyang et al., 2021) or integrating specialized modules, such as reservoirs (Hanazaki et al., 2022; West et al., 2025). Compared with LSTM, δHBV is more sensitive to precipitation biases. For example, the differences between δHBV simulations under different forcing datasets were generally larger than those for LSTM, and δHBV using the Daymet forcing dataset showed largely better performance than with the other two forcing datasets, which indicates that δHBV may not be able to fit different forcing datasets well. Therefore, many potential structural optimizations can be implemented to improve δHBV. Our analysis provided corroborating evidence that forcing error is an important reason why LSTM can outperform δHBV in the temporal test for some basins, although such patterns may not generalize well in space. A meteorological forcing data correction module can be developed in the future to account for timing and magnitude errors in precipitation. 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 pretrained 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.

For this work, we did not create a $\delta\text{HBV}^{\text{multi}}$ model (in the same vein as $\text{LSTM}^{\text{multi}}$) using all forcings as an input to a single model, since a similar experiment has already been conducted by Sawadekar et al. (2025). We also did not examine "seed" combinations of a $\delta\text{HBV}^{\text{multi}}$ as we believed they would not result in a significant performance boost (unlike that seen with $\text{LSTM}^{\text{multi}}$), because LSTM has high variability and low bias, while δHBV has lower variance and potentially higher bias. As a result, random seeds would likely not create large enough perturbations for δHBV and wouldn't bring the benefits seen with $\text{LSTM}^{\text{multi}}_{\text{seed}}$. To achieve an equivalent perturbation level for δHBV, it may be necessary to incorporate multiple distinct hydrological models, such as SAC-SMA, PRMS, and GR4J, similar to the approach implemented in the Framework for Understanding Structural Errors (FUSE) (Clark et al., 2008). Work is ongoing to create a combination of a series of differentiable process-based models, which is expected to produce a further improved ensemble with great interpretability. Given the success of cross-model-type ensembles shown in this work, we also encourage further exploration of ensemble simulations involving models with other distinct mechanisms.

4 Summary and conclusions

This study comprehensively analyzes ensemble combinations of two advanced model types (LSTM and δHBV), each with distinct mechanisms, for streamflow simulation across 531 basins in the US. 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. The performance of ensemble simulations was evaluated under three distinct testing scenarios (temporal, PUB, and PUR tests), surpassing the previous highest performances. Our findings enhance the understanding of how to effectively utilize diverse model types and multi-source datasets to improve streamflow simulations. The principal conclusions are:

1. Cross-model-type ensembles (LSTM + δHBV) consistently outperformed single-model approaches across all test scenarios, setting new performance benchmarks on the CAMELS dataset. These ensembles demonstrated the complementarity of data-driven (LSTM) and physics-informed (δHBV) approaches in capturing diverse hydrological behaviors.

2. Ensembling models trained on different forcing datasets proved more effective than using multiple forcing datasets as simultaneous inputs to a single model. This suggests that separate training allows each model to capture unique features contained in each forcing dataset, which can then be effectively leveraged in the ensemble.

3. δHBV provided significant benefits to ensemble simulations on spatial generalization. Ensembling LSTM with δHBV showed increasing benefits as generalization challenges increased, from temporal to spatial interpolation (PUB) to spatial extrapolation (PUR) tests. This underscores the value of physics-informed constraints in improving model transferability to ungauged basins and regions.

4. While ensemble methods significantly improved overall performance, they did not fully mitigate consistent deficiencies in certain challenging areas (e.g., regions with high dam density or heterogeneous hydrogeological conditions). This indicates areas for future model development.

These findings have important implications for hydrological modeling and water resources management. The improved accuracy and spatial generalization of our ensemble approach can enhance streamflow predictions, benefiting water resources planning and management, particularly in data-scarce regions. Our results also suggest that future hydrological model development should focus on combining data-driven and physics-based approaches to improve model generalizability across diverse conditions. The superior performance of ensembling models with different forcing datasets over using merged forcings as a single input highlights the risk of indiscriminately feeding all available data into one data-driven model. While computational demands certainly require consideration, the potential improvements in prediction accuracy offer significant value for both research and operational applications. Future work should focus on refining these ensemble techniques, addressing model limitations in challenging regions, and exploring ensemble implementation in operational settings.

Appendix A: Detailed processes of HBV employed in this study

The Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Aghakouchak and Habib, 2010; Beck et al., 2020; Bergström, 1976, 1992) is a simple yet effective bucket-type hydrologic model that simulates hydrologic components including snow water equivalent, soil moisture, groundwater storage, evapotranspiration, quick flow, baseflow, and total streamflow.

In the following, we describe these processes in detail with their corresponding equations. Uppercase letters denote state variables, while lowercase letters denote parameters. The overall water balance is expressed as Eq. (A1).

$$\mathrm{EP}-\mathrm{AE}-Q_{\mathrm{t}}=\mathrm{SN}+\mathrm{SM}+\mathrm{SUZ}+\mathrm{SLZ}+\mathrm{LAKE} \tag{A1}$$

where EP is effective precipitation, AE is actual evapotranspiration, Q_{t} is total simulated runoff, SN is snow storage, SM is soil moisture storage, SUZ and SLZ are the upper and lower groundwater storages, respectively, and LAKE represents lake storage (omitted in this study).

First, effective precipitation (EP) is partitioned into rain (RN) and snow (SN) components based on the air temperature (T) relative to a threshold temperature (tt):

$$\mathrm{RN}=\mathrm{EP}\ \text{ if }\ T\geq \mathrm{tt} \tag{A2}$$

$$\mathrm{SN}=\mathrm{EP}\ \text{ if }\ T<\mathrm{tt} \tag{A3}$$

Snow (SN) accumulates in the snowpack (SNP), while snowmelt (SNM) happens when $T\geq$ tt, which is calculated based on a melt factor (cfm) and the temperature difference ($T-$tt). The computed snowmelt (SNM) is constrained by the available snowpack (SNP).

$$\mathrm{SNM}=\min[\max\left(\mathrm{cfm}\cdot(T-\mathrm{tt}),0\right)\mathrm{SNP}] \tag{A4}$$

The snowmelt (SNM) contributes to meltwater (MW), while the snowpack (SNP) is updated as:

$$\mathrm{MW}=\mathrm{MW}+\mathrm{SNM} \tag{A5}$$

$$\mathrm{SNP}=\mathrm{SNP}+\mathrm{SN}-\mathrm{SNM} \tag{A6}$$

A portion of the meltwater (MW) may refreeze when $T<$ tt, controlled by the refreezing parameter (cfr):

$$\mathrm{RFZ}=\min\left[\max\left(\mathrm{cfr}\cdot\mathrm{cfm}\cdot(\mathrm{tt}-T),0\right),\mathrm{MW}\right] \tag{A7}$$

$$\mathrm{SNP}=\mathrm{SNP}+\mathrm{RFZ} \tag{A8}$$

$$\mathrm{MW}=\mathrm{MW}-\mathrm{RFZ} \tag{A9}$$

The remaining meltwater (MW) exceeding the snowpack's liquid water holding capacity (cwh · SNP) infiltrates into the soil (IF), with the remainder retained in MW:

$$\mathrm{IF}=\max\left(\mathrm{MW}-\mathrm{cwh}\cdot\mathrm{SNP},0\right) \tag{A10}$$

$$\mathrm{MW}=\mathrm{MW}-\mathrm{IF} \tag{A11}$$

The fraction of soil moisture (SM) relative to the field capacity (fc), raised to the power index β, modulates shallow seepage (SP) according to the available water (IF + RN):

$$\mathrm{SP}=\left(\frac{\mathrm{SM}}{\mathrm{fc}}\right)^{\beta}(\mathrm{IF}+\mathrm{RN}) \tag{A12}$$

$$\mathrm{SM}=\mathrm{SM}+\mathrm{IF}+\mathrm{RN}-\mathrm{SP} \tag{A13}$$

Excess soil water above the field capacity contributes to direct infiltration ($\mathrm{IF}_{\mathrm{dir}}$):

$$\mathrm{IF}_{\mathrm{dir}}=\max\left(\mathrm{SM}-\mathrm{fc},0\right) \tag{A14}$$

$$\mathrm{SM}=\mathrm{SM}-\mathrm{IF}_{\mathrm{dir}} \tag{A15}$$

Actual evapotranspiration (AE) is estimated as the product of potential evapotranspiration (PE) and an evapotranspiration coefficient (PEC). The PEC depends on soil moisture storage (SM), field capacity (fc), a shape parameter (λ), and a threshold parameter (lp).

$$\mathrm{PEC}=\min\left[1,\max\left(0,\left(\frac{\mathrm{SM}}{\mathrm{lp}\cdot\mathrm{fc}}\right)^{\lambda}\right)\right] \tag{A16}$$

$$\mathrm{AE}=\min\left(\mathrm{PE}\cdot\mathrm{PEC},\mathrm{SM}\right) \tag{A17}$$

$$\mathrm{SM}=\mathrm{SM}-\mathrm{AE} \tag{A18}$$

Capillary rise (CP) from the lower zone (SLZ) replenishes SM, controlled by a coefficient (c) and constrained by the soil moisture deficit:

$$\mathrm{CP}=\min\left[c\cdot\mathrm{SLZ}\cdot\left(1-\frac{\mathrm{SM}}{\mathrm{fc}}\right),\mathrm{SLZ}\right] \tag{A19}$$

$$\mathrm{SM}=\mathrm{SM}+\mathrm{CP} \tag{A20}$$

$$\mathrm{SLZ}=\mathrm{SLZ}-\mathrm{CP} \tag{A21}$$

Recharge from the soil, consisting of shallow seepage (SP) and direct infiltration ($\mathrm{IF}_{\mathrm{dir}}$), enters the upper groundwater zone (SUZ). Water in the upper zone either percolates to the lower groundwater zone (SLZ) at a constant percolation rate (prc) or contributes to direct runoff (Q_0) when the upper zone (SUZ) exceeds a threshold (uzl). Flow from the upper and lower zones is computed using linear reservoir formulations, with parameters k_0, k_1, k_2 controlling the respective runoff components Q_0, Q_1, Q_2. The total simulated streamflow (Q_t) is then computed as the sum of these components.

$$\mathrm{SUZ} = \mathrm{SUZ} + \mathrm{SP} + \mathrm{IF}_{\mathrm{dir}} \tag{A22}$$

$$\mathrm{PERC} = \min(\mathrm{prc}, \mathrm{SUZ}) \tag{A23}$$

$$\mathrm{SUZ} = \mathrm{SUZ} - \mathrm{PERC} \tag{A24}$$

$$Q_0 = \max\left[k_0 \cdot (\mathrm{SUZ} - \mathrm{uzl}), 0\right] \tag{A25}$$

$$\mathrm{SUZ} = \mathrm{SUZ} - Q_0 \tag{A26}$$

$$Q_1 = \mathrm{SUZ} \cdot k_1 \tag{A27}$$

$$\mathrm{SUZ} = \mathrm{SUZ} - Q_1 \tag{A28}$$

$$\mathrm{SLZ} = \mathrm{SLZ} + \mathrm{PERC} \tag{A29}$$

$$Q_2 = \mathrm{SLZ} \cdot k_2 \tag{A30}$$

$$\mathrm{SLZ} = \mathrm{SLZ} - Q_2 \tag{A31}$$

$$Q_\mathrm{t} = Q_0 + Q_1 + Q_2 \tag{A32}$$

Finally, a routing module (Feng et al., 2022) is used to process Q_t to produce the final streamflow output (Q_t^*). This module with two parameters ($\theta_\alpha, \theta_\tau$) assumes a gamma function for the unit hydrograph and convolves the unit hydrograph with the runoff as,

$$Q_\mathrm{t}^* = \int_0^{\mathrm{tmax}} \xi(s : \theta_\alpha, \theta_\tau) \cdot Q(t - s)\mathrm{d}s \tag{A33}$$

$$\xi(s : \theta_\alpha, \theta_\tau) = \frac{1}{\Gamma(\theta_\alpha)\theta_\tau^{\theta_\alpha}} s^{\theta_\alpha - 1} e^{-\frac{s}{\theta_\tau}} \tag{A34}$$

Appendix B: Illustrated differences among the three meteorological forcing datasets

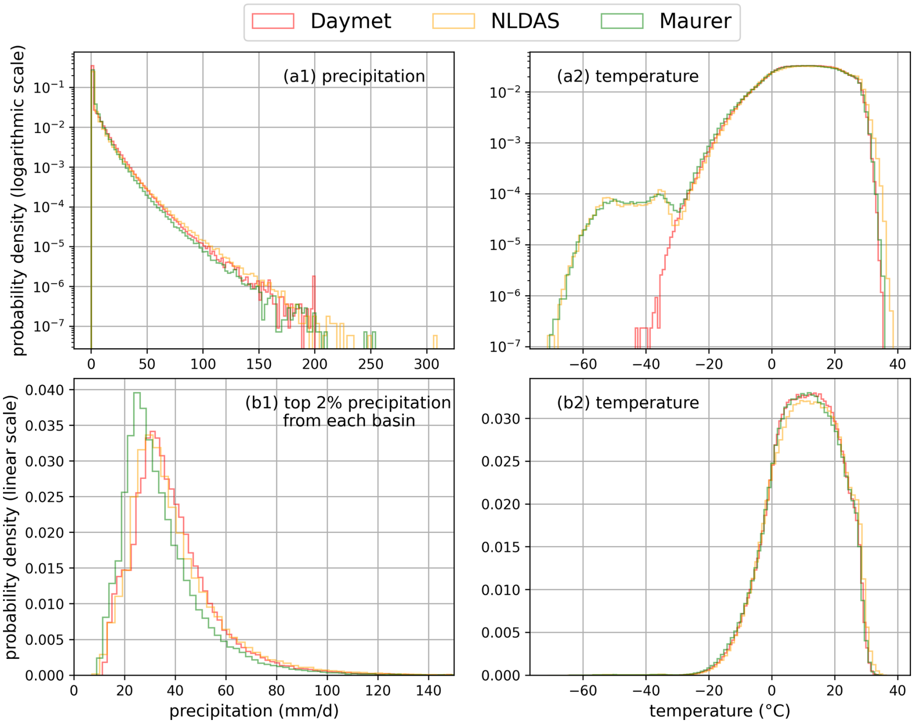

Figure B1. Probability density distributions (top panel in logarithmic scale, bottom panel in linear scale) of precipitation and temperature across three meteorological forcing datasets.

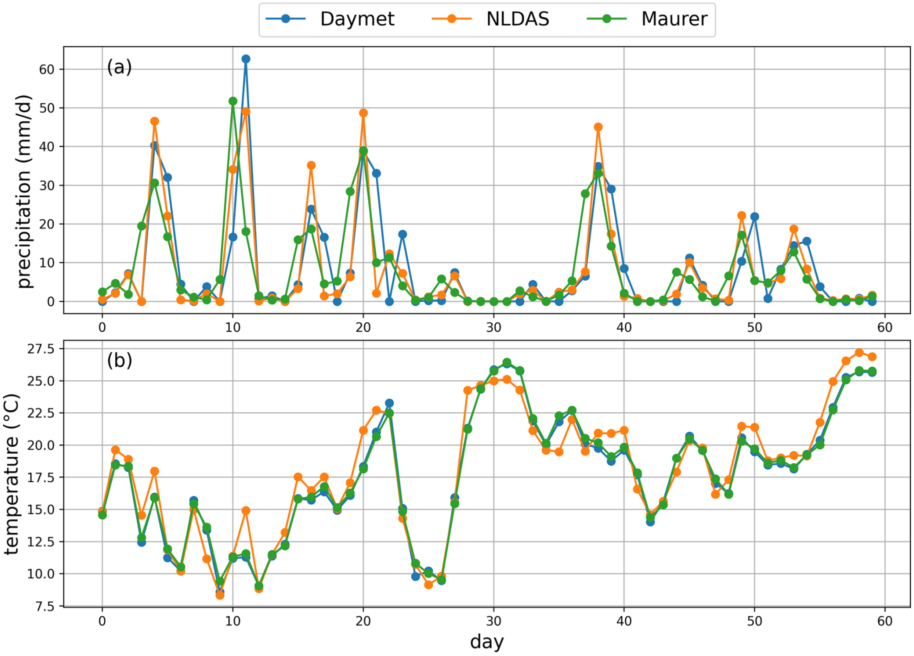

Figure B2. Illustrated temporal variations of precipitation and temperature in a basin across three meteorological forcing datasets.

Appendix C: Details of model inputs, ensemble frameworks, and evaluations

Table C1. Full names for the abbreviations of dynamic data (all but streamflow are "forcings") and static basin attributes used as 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). Potential evapotranspiration and normalized streamflow were calculated in this work, using CAMELS data. The number in parentheses specifies model usage: 1 denotes use in the LSTM model, and 2 denotes use in the δHBV model.

Type	Abbreviation	Full name	Unit
Dynamic data	prcp (1,2)	Precipitation	$\mathrm{mm\,d^{-1}}$
	pet (2)	Potential evapotranspiration (calculated in this work using the Hargreaves equation and CAMELS data)	$\mathrm{mm\,d^{-1}}$
	tmean (2)	Mean air temperature	°C
	tmax (1)	Maximum air temperature	°C
	tmin (1)	Minimum air temperature	°C
	srad (1)	Shortwave radiation	$\mathrm{W\,m^{-2}}$
	vp (1)	Water vapor pressure	pa
	q_vol	Volumetric streamflow	$\mathrm{ft^3\,s^{-1}}$
	q (1,2)	Streamflow normalized by basin area (q_vol/area_gages2)	$\mathrm{mm\,d^{-1}}$
Static basin attributes	p_mean (1,2)	Mean daily precipitation	$\mathrm{mm\,d^{-1}}$
	pet_mean (1,2)	Mean daily potential evapotranspiration	$\mathrm{mm\,d^{-1}}$
	p_seasonality (2)	Seasonality and timing of precipitation	–
	frac_snow (1,2)	Fraction of precipitation falling as snow	–
	aridity (1,2)	Rate of mean values of potential evapotranspiration and precipitation	–
	high_prec_freq (1,2)	Frequency of high precipitation days	$\mathrm{d\,yr^{-1}}$
	high_prec_dur (1,2)	Average duration of high precipitation events	days
	low_prec_freq (1,2)	Frequency of dry days	$\mathrm{d\,yr^{-1}}$
	low_prec_dur (1,2)	Average duration of dry periods	days
	elev_mean (1,2)	Catchment mean elevation	m
	slope_mean (1,2)	Catchment mean slope	$\mathrm{m\,km^{-1}}$
	area_gages2 (1,2)	Catchment area (GAGES-II estimate)	$\mathrm{km^2}$
	frac_forest (1,2)	Fraction of catchment area having land cover identified as forest	–
	lai_max (1,2)	Maximum monthly mean of the leaf area index	–
	lai_diff (1,2)	Difference between the maximum and minimum monthly mean of the leaf area index	–
	gvf_max (1,2)	Maximum monthly mean of the green vegetation	–

Table C1. Continued.

Type	Abbreviation	Full name	Unit
Static basin attributes	gvf_diff (1,2)	Difference between the maximum and minimum monthly mean of the green vegetation fraction	–
	dom_land_cover_frac (2)	Fraction of the catchment area associated with the dominant land cover	–
	dom_land_cover (2)	Dominant land cover type	–
	root_depth_50 (2)	Root depth at 50th percentile, extracted from a root depth distribution based on the International Geosphere-Biosphere Programme (IGBP) land cover	m
	soil_depth_pelletier (1,2)	Depth to bedrock	m
	soil_depth_statsgso (1,2)	Soil depth	m
	soil_porosity (1,2)	Volumetric soil porosity	–
	soil_conductivity (1,2)	Saturated hydraulic conductivity	$\mathrm{cm\,h^{-1}}$
	max_water_content (1,2)	Maximum water content	m
	sand_frac (1,2)	Fraction of soil which is sand	–
	silt_frac (1,2)	Fraction of soil which is silt	–
	clay_frac (1,2)	Fraction of soil which is clay	–
	geol_class_1st (2)	Most common geologic class in the catchment basin	–
	geol_class_1st_frac (2)	Fraction of the catchment area associated with its most common geologic class	–
	geol_class_2nd (2)	Second most common geologic class in the catchment basin	–
	geol_class_2nd_frac (2)	Fraction of the catchment area associated with its 2nd most common geologic class	–
	carbonate_rocks_frac (1,2)	Fraction of the catchment area as carbonate sedimentary rocks	–
	geol_porosity (2)	Subsurface porosity	–
	geol_permeability (1,2)	Subsurface permeability	$\mathrm{m^2}$

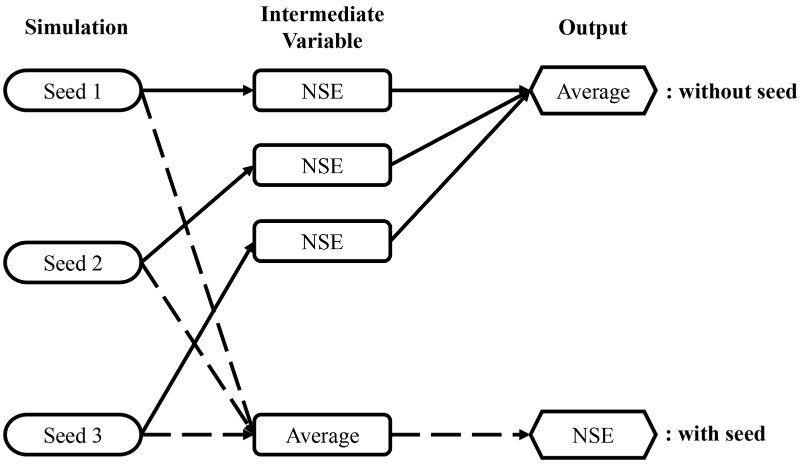

Figure C1. Ensemble frameworks to generate metrics for ensembles named without (solid arrows) and with (dashed arrows) "seed" as a subscript.

Table C2. Loss function and evaluation metrics.

Statistic	Equation*	Range	Optimal Value
Loss	$\frac{1}{n}\sum_{i=1}^{n}\frac{(O_i-S_i)^2}{(\sigma_O+\epsilon)^2}$	0.0 to ∞	0.0
NSE	$1-\frac{\sum_{i=1}^{n}(O_i-S_i)^2}{\sum_{i=1}^{n}(O_i-\mu_O)^2}$	$-\infty$ to 1.0	1.0
KGE	$1-\sqrt{(r-1)^2+(\beta-1)^2+(\gamma-1)^2}$, $\beta=\frac{\mu_S}{\mu_O}, \gamma=\frac{CV_S}{CV_O}=\frac{\sigma_S/\mu_S}{\sigma_O/\mu_O}$	$-\infty$ to 1.0	1.0
PBIAS	$\frac{\sum_{i=1}^{n}(O_i-S_i)}{\sum_{i=1}^{n}O_i}\times 100$	$-\infty$ to ∞	0.0
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i-S_i)^2}$	0.0 to ∞	0.0
spread	$\sqrt{\frac{1}{n}\frac{1}{e}\sum_{i=1}^{n}\sum_{j=1}^{e}(S_{i,j}-\mu_{S,i})^2}$	0.0 to ∞	None

* S is a streamflow simulation; O is the corresponding observation; n is the number of total S or O; ϵ is a numerical stabilizer, with a default value of 0.1; e is the number of ensemble members; r is the linear Pearson correlation between S and O; β is the mean bias; and γ is the variability bias. The mean and standard deviation of simulations are denoted as μ_S and σ_S, respectively, and μ_O and σ_O are the mean and standard deviation of the observations.

Appendix D: Additional details on model performance

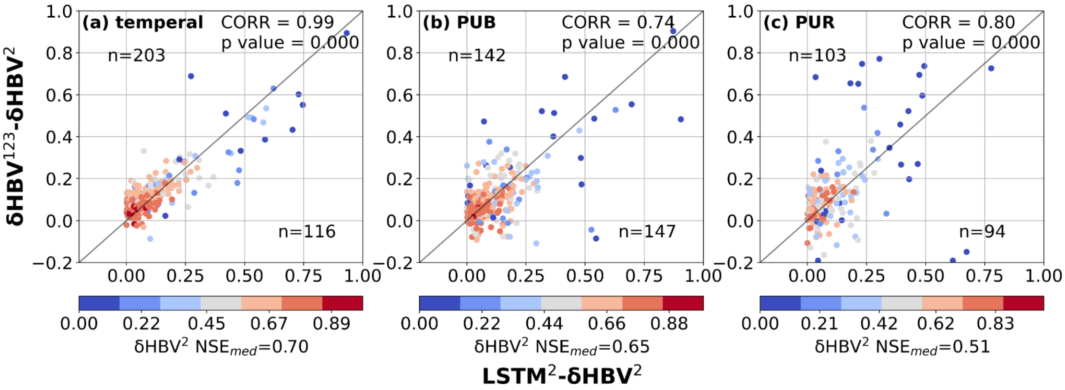

Figure D1. Scatter plots comparing the performance differences between hydrological models for the basins where LSTM outperformed δHBV (the basins where δHBV outperformed are not shown in this plot). The x-axis represents the NSE differences between LSTM^2 and δHBV^2 (LSTM^2 – δHBV^2), while the y-axis 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 the x-axis values and the y-axis values, along with the median NSE value of δHBV^2 (NSE$_{\text{med}}$) on these basins, are also noted.

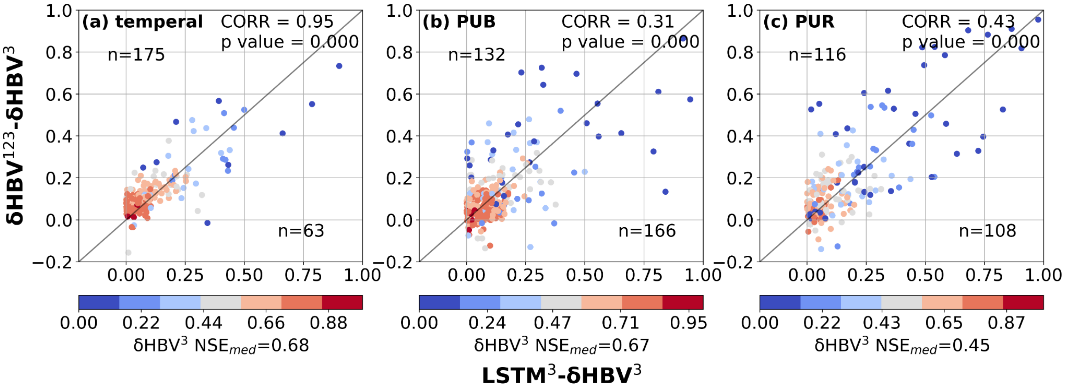

Figure D2. Scatter plots comparing the performance differences between hydrological models for the basins where LSTM outperformed δHBV (the basins where δHBV outperformed are not shown in this plot). The x-axis represents the NSE differences between LSTM^3 and δHBV^3 (LSTM^3 – δHBV^3), while the y-axis 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 the x-axis values and the y-axis values, along with the median NSE value of δHBV^3 (NSE$_{\text{med}}$) on these basins, are also noted.

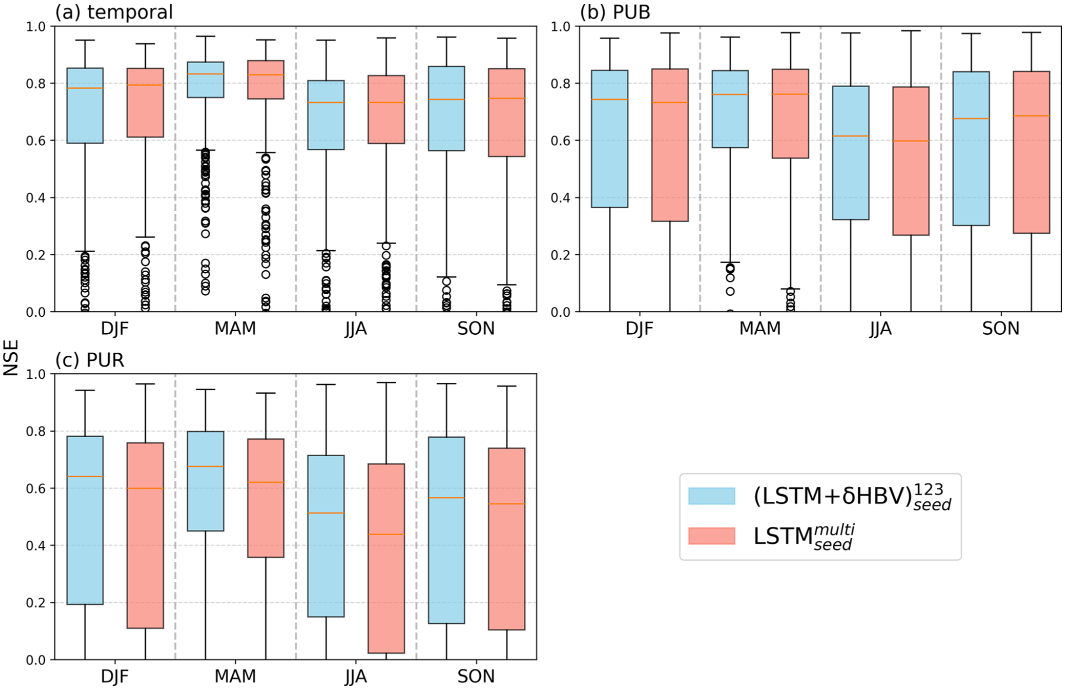

Figure D3. Seasonal comparison of Nash–Sutcliffe efficiency (NSE) values for $(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}_{\mathrm{seed}}$ (blue) and $\mathrm{LSTM}^{\mathrm{multi}}_{\mathrm{seed}}$ (red) in **(a)** temporal, **(b)** PUB, and **(c)** PUR tests. Each box represents the distribution of NSE values across 531 basins for a given season (DJF: December–February, MAM: March–May, JJA: June–August, SON: September–November). Vertical dashed lines separate different seasons. $(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}_{\mathrm{seed}}$ performs better than $\mathrm{LSTM}^{\mathrm{multi}}_{\mathrm{seed}}$ in most cases, especially during MAM, likely due to differences in snowmelt representation.

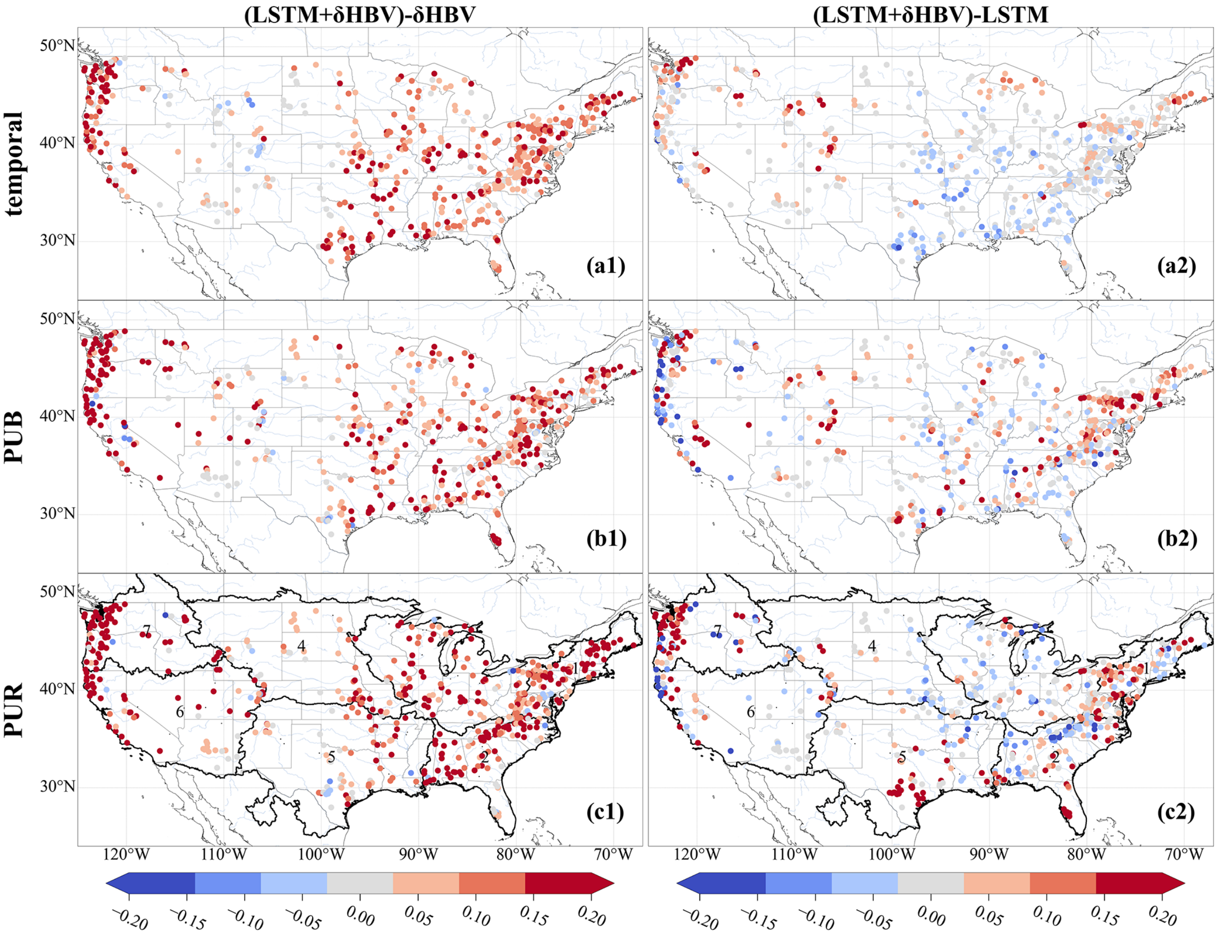

Figure D4. Spatial distributions of model spread values increase from δHBV and LSTM to the LSTM + δHBV ensemble across temporal, PUB, and PUR tests.

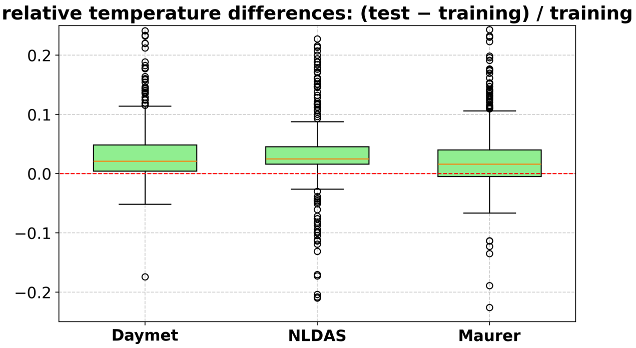

Figure D5. Boxplot of relative temperature differences between the test and training periods, calculated as (test − training)/training. Each box represents the distribution of normalized temperature changes across basins for a specific meteorological forcing dataset: Daymet, NLDAS, and Maurer. Positive values indicate warming in the test period relative to the training period.

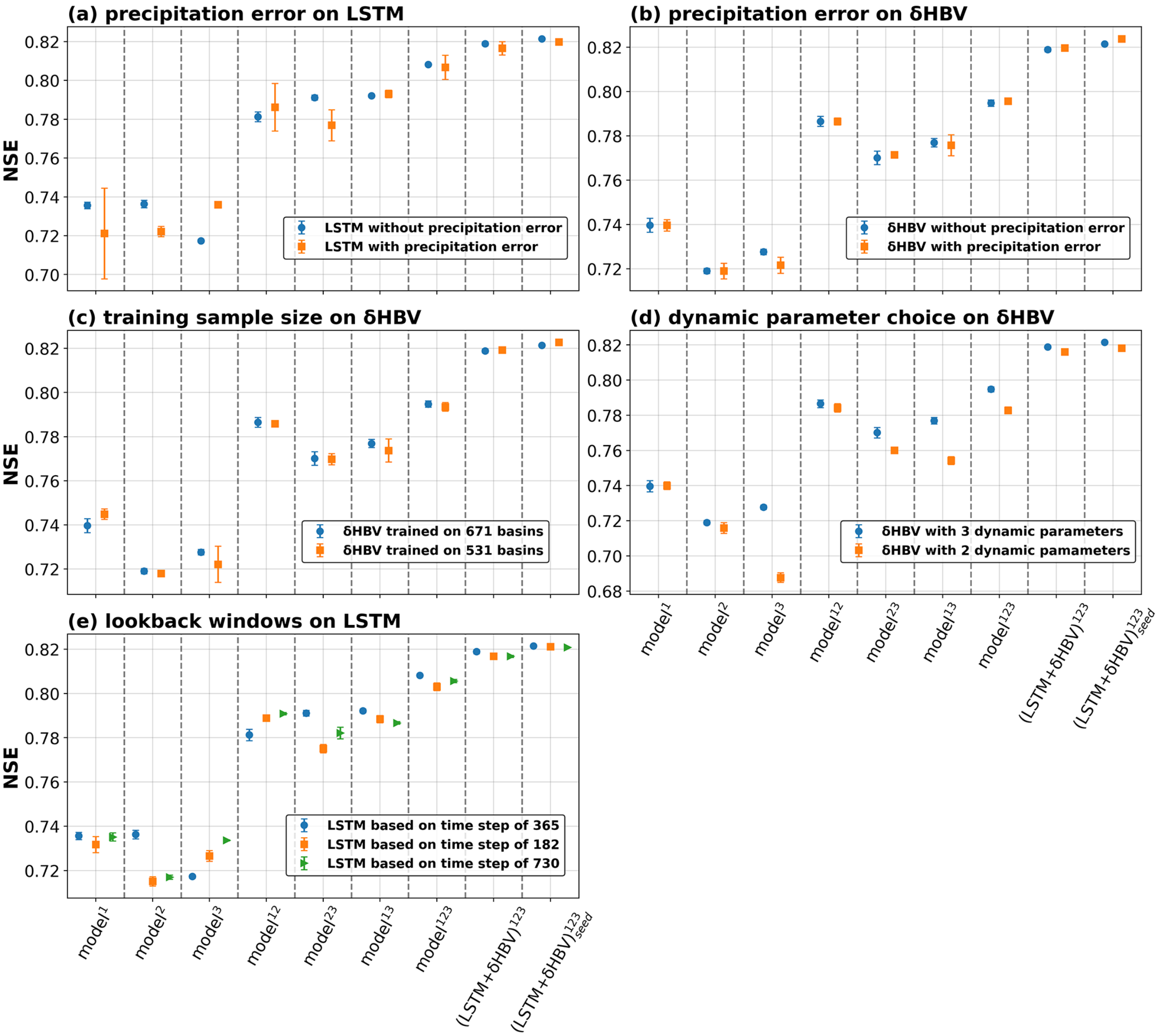

Figure D6. Simulation performance (NSE) under the temporal test: **(a)** LSTM model with and without a 10 % precipitation error (precipitation × 1.1); **(b)** δHBV model with and without a 10 % precipitation error; **(c)** δHBV model trained on 671 versus 531 basins; **(d)** δHBV model with 3 versus 2 dynamic parameters; **(e)** δHBV model using time steps of 365, 182, and 730 d. Individual and ensemble groups are distinguished along the x-axis. Ensemble benefits are indicated by the gap between columns of the same color within each panel–columns 1–7 correspond to individual LSTM or δHBV groups, and the last two columns correspond to LSTM + δHBV ensembles.

Table D1. Median NSE, KGE, RMSE, PBIAS, and RMSE values under low (lowRMSE), high (highRMSE), and middle (midRMSE) flows based on 531 basins under the temporal test. The values are the mean of three simulations run with different random seeds.

Temporal	Number	Daymet	NLDAS	Maurer
LSTM	NSE	0.735639	0.736301	0.717337
	KGE	0.789375	0.782555	0.760575
	RMSE	1.21088	1.19847	1.27723
	PBIAS	4.04818	5.99486	1.58911
	lowRMSE	0.0596913	0.0602381	0.0545577
	highRMSE	2.70508	2.89684	2.97028
	midRMSE	0.196039	0.210022	0.219922
δHBV	NSE	0.739688	0.71903	0.727669
	KGE	0.77033	0.730753	0.762022
	RMSE	1.18752	1.26239	1.23193
	PBIAS	5.07898	−0.14449	3.65263
	lowRMSE	0.060906	0.063581	0.063466
	highRMSE	2.68479	3.13011	2.6845
	midRMSE	0.226595	0.245242	0.230125
LSTM + δHBV	NSE	0.787545	0.794053	0.790903
	KGE	0.794412	0.78383	0.786854
	RMSE	1.0777	1.0716	1.07141
	PBIAS	4.59065	3.33053	3.45501
	lowRMSE	0.059955	0.059565	0.054838
	highRMSE	2.70216	2.88511	2.69633
	midRMSE	0.20394	0.214726	0.212514

Temporal	Number	Daymet + NLDAS	Daymet + Maurer	NLDAS + Maurer	All
LSTM	NSE	0.781275	0.791158	0.792144	0.808176
	KGE	0.800955	0.795026	0.794441	0.803476
	RMSE	1.09103	1.06374	1.06701	1.01395
	PBIAS	5.17159	3.34362	4.5305	4.48263
	lowRMSE	0.0636155	0.0582563	0.0566306	0.0613625
	highRMSE	2.70218	2.71366	2.78962	2.67803
	midRMSE	0.194849	0.199809	0.206653	0.197469
δHBV	NSE	0.786562	0.77012	0.776938	0.794796
	KGE	0.773732	0.778557	0.768854	0.77834
	RMSE	1.08362	1.12584	1.10875	1.06118
	PBIAS	1.91507	4.28194	2.03584	2.71021
	lowRMSE	0.061667	0.060679	0.062765	0.061539
	highRMSE	2.93961	2.7394	2.88758	2.84994
	midRMSE	0.230576	0.220743	0.230272	0.228375
LSTM + δHBV	NSE	0.811825	0.809964	0.811316	0.818907
	KGE	0.797564	0.797635	0.78735	0.794936
	RMSE	1.01938	1.01755	1.0314	1.00067
	PBIAS	4.14594	4.23333	3.19652	3.88096
	lowRMSE	0.0603	0.058022	0.057882	0.059221
	highRMSE	2.75275	2.67122	2.81393	2.70606
	midRMSE	0.207637	0.205965	0.213191	0.207905

Table D2. Median NSE, KGE, RMSE, PBIAS, and RMSE values under low (lowRMSE), high (highRMSE), and middle (midRMSE) flows based on 531 basins under the PUB test. The values are the mean of three simulations run with different random seeds.

PUB	Number	Daymet	NLDAS	Maurer
LSTM	NSE	0.702636	0.695496	0.694156
	KGE	0.693998	0.677438	0.6909
	RMSE	1.31714	1.3394	1.34233
	PBIAS	0.669018	0.283106	0.936582
	lowRMSE	0.087648	0.088393	0.086873
	highRMSE	4.2852	4.49292	4.16042
	midRMSE	0.354458	0.364921	0.368124
δHBV	NSE	0.706809	0.670636	0.682998
	KGE	0.703137	0.66566	0.686912
	RMSE	1.35541	1.41185	1.37942
	PBIAS	1.49234	−2.43395	0.291966
	lowRMSE	0.0798196	0.0808967	0.0846775
	highRMSE	4.21648	4.49582	4.18003
	midRMSE	0.335159	0.351271	0.356903
LSTM + δHBV	NSE	0.74227	0.723778	0.72202
	KGE	0.715931	0.690154	0.707292
	RMSE	1.24887	1.278	1.26697
	PBIAS	1.27863	−0.599778	0.903464
	lowRMSE	0.0816748	0.0795686	0.0825691
	highRMSE	4.08432	4.23483	3.94929
	midRMSE	0.327459	0.33851	0.347169

PUB	Number	Daymet + NLDAS	Daymet + Maurer	NLDAS + Maurer	All
LSTM	NSE	0.757853	0.749151	0.753136	0.768181
	KGE	0.713319	0.720099	0.716497	0.727143
	RMSE	1.18251	1.22254	1.19718	1.15026
	PBIAS	0.320396	0.931656	0.766216	0.970047
	lowRMSE	0.0875191	0.0864129	0.0835341	0.0874717
	highRMSE	4.1296	4.06602	4.17217	4.0061
	midRMSE	0.334683	0.349856	0.342819	0.333534
δHBV	NSE	0.748916	0.734052	0.733955	0.757749
	KGE	0.699768	0.714323	0.69436	0.714048
	RMSE	1.26852	1.27637	1.27244	1.23229
	PBIAS	0.0446112	1.212	−1.04135	0.201809
	lowRMSE	0.0808293	0.0792486	0.0814476	0.0808359
	highRMSE	4.19575	3.97788	4.21623	4.07419
	midRMSE	0.311826	0.33668	0.339257	0.318165
LSTM + δHBV	NSE	0.780625	0.764866	0.767761	0.785833
	KGE	0.719781	0.725373	0.715982	0.723972
	RMSE	1.14924	1.17659	1.16881	1.13591
	PBIAS	0.186062	0.881644	0.405548	0.565489
	lowRMSE	0.0805946	0.0814251	0.0817114	0.0826379
	highRMSE	3.97373	3.86834	3.88	3.91692
	midRMSE	0.313708	0.324777	0.324089	0.323671

Table D3. Median NSE, KGE, RMSE, PBIAS, and RMSE values under low (lowRMSE), high (highRMSE), and middle (midRMSE) flows based on 531 basins under the PUR test. The values are the mean of three simulations run with different random seeds.

PUR	Number	Daymet	NLDAS	Maurer
LSTM	NSE	0.578365	0.546217	0.56164
	KGE	0.557788	0.559986	0.567231
	RMSE	1.59111	1.63626	1.5833
	PBIAS	−0.575328	−2.77709	−0.623183
	lowRMSE	0.124837	0.118971	0.118695
	highRMSE	5.42346	5.38886	5.05212
	midRMSE	0.498133	0.498442	0.471744
δHBV	NSE	0.622278	0.592306	0.59161
	KGE	0.638818	0.601338	0.620877
	RMSE	1.57189	1.61191	1.63628
	PBIAS	1.27223	−1.60075	1.62709
	lowRMSE	0.10142	0.102975	0.101075
	highRMSE	5.07706	5.16093	4.99602
	midRMSE	0.447879	0.474516	0.439697
LSTM + δHBV	NSE	0.644398	0.618255	0.635444
	KGE	0.627481	0.605237	0.615883
	RMSE	1.46185	1.5153	1.48393
	PBIAS	−0.269697	−0.719505	0.197859
	lowRMSE	0.105146	0.100944	0.106272
	highRMSE	4.95749	4.99478	4.78638
	midRMSE	0.431456	0.4575	0.426126

PUR	Number	Daymet + NLDAS	Daymet + Maurer	NLDAS + Maurer	All
LSTM	NSE	0.634398	0.636369	0.626939	0.656228
	KGE	0.59844	0.600371	0.605007	0.612858
	RMSE	1.4434	1.43416	1.43009	1.38042
	PBIAS	−0.547128	−0.687947	−0.865748	−0.543918
	lowRMSE	0.118989	0.120228	0.115004	0.117728
	highRMSE	5.03277	5.02434	4.84415	4.74281
	midRMSE	0.462923	0.455257	0.453912	0.449598
δHBV	NSE	0.672839	0.644732	0.661231	0.684685
	KGE	0.653841	0.65646	0.6515	0.66205
	RMSE	1.43224	1.50803	1.48604	1.43376
	PBIAS	0.564363	1.55134	−0.156553	0.956961
	lowRMSE	0.0975783	0.0984076	0.100773	0.100807
	highRMSE	4.83843	4.81176	4.72529	4.71255
	midRMSE	0.447828	0.431252	0.433688	0.432018
LSTM + δHBV	NSE	0.685032	0.680872	0.679321	0.700814
	KGE	0.638788	0.647826	0.646782	0.649999
	RMSE	1.35303	1.3873	1.36795	1.3185
	PBIAS	−0.0150729	0.406127	−0.135091	−0.0232668
	lowRMSE	0.103284	0.101814	0.104528	0.102916
	highRMSE	4.80178	4.72583	4.70024	4.70713
	midRMSE	0.426819	0.411727	0.41573	0.41081

Table D4. Median NSE, KGE, RMSE, PBIAS, and RMSE values under low (lowRMSE), high (highRMSE), and middle (midRMSE) flows based on 531 basins under the temporal, PUB, and PUR tests of $\text{LSTM}^{\text{multi}}$, $(\text{LSTM}+\delta\text{HBV})^{123}+\text{LSTM}^{\text{multi}}$, their "seed" version, and $(\text{LSTM}+\delta\text{HBV})^{123}_{\text{seed}}$.

Test	Metric	$\text{LSTM}^{\text{multi}}$	$(\text{LSTM}+\delta\text{HBV})^{123}+\text{LSTM}^{\text{multi}}$
Temporal	NSE	0.797448	0.82321
	KGE	0.811064	0.810248
	RMSE	1.05987	0.983168
	PBIAS	3.95241	4.08594
	lowRMSE	0.056221	0.05702
	highRMSE	2.7089	2.58881
	midRMSE	0.183526	0.192442
PUB	NSE	0.750605	0.782727
	KGE	0.71469	0.734731
	RMSE	1.20586	1.11509
	PBIAS	0.475674	0.706777
	lowRMSE	0.0861127	0.0836
	highRMSE	4.13615	3.83009
	midRMSE	0.347562	0.326814
PUR	NSE	0.623755	0.68923
	KGE	0.593757	0.633971
	RMSE	1.47379	1.31221
	PBIAS	−2.6737	−1.38119
	lowRMSE	0.112434	0.107646
	highRMSE	4.98202	4.59232
	midRMSE	0.501807	0.436811

Test	Metric	$(\text{LSTM}+\delta\text{HBV})^{123}_{\text{seed}}$	$\text{LSTM}^{\text{multi}}_{\text{seed}}$	$(\text{LSTM}+\delta\text{HBV})^{123}_{\text{seed}}+\text{LSTM}^{\text{multi}}_{\text{seed}}$
Temporal	NSE	0.821444	0.81992	0.829385
	KGE	0.795317	0.82078	0.812581
	RMSE	0.99455	1.00908	0.967779
	PBIAS	3.99009	4.09469	4.08882
	lowRMSE	0.059782	0.057346	0.057015
	highRMSE	2.7279	2.62815	2.58384
	midRMSE	0.209943	0.183656	0.195557
PUB	NSE	0.793673	0.781175	0.790921
	KGE	0.726188	0.736191	0.739284
	RMSE	1.12957	1.13079	1.09176
	PBIAS	0.370674	1.13671	0.869057
	lowRMSE	0.083423	0.084038	0.085728
	highRMSE	3.89363	3.93473	3.79505
	midRMSE	0.323045	0.329772	0.325627
PUR	NSE	0.705154	0.665723	0.701504
	KGE	0.651538	0.614649	0.64373
	RMSE	1.30377	1.3727	1.2851
	PBIAS	−0.283645	−2.74069	−1.39149
	lowRMSE	0.100525	0.111229	0.108121
	highRMSE	4.74889	4.88127	4.58344
	midRMSE	0.406797	0.473783	0.432447

Table D5. Median NSE values based on ten different random seeds during the temporal test. Each number (1 through 10) represents metric values calculated for an individual simulation based on only one random seed. "Seed" indicates metric values calculated by averages of these ten simulations based on different random seeds, while "mean" denotes the average of metrics from 1–10 individual simulations (visualized in Fig. C1).

Number	$\mathrm{LSTM}^{\mathrm{multi}}$	$(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}$	$(\mathrm{LSTM}+\delta\mathrm{HBV})^{123}+\mathrm{LSTM}^{\mathrm{multi}}$
1	0.797742	0.818436	0.82315
2	0.795312	0.820188	0.823559
3	0.799291	0.818097	0.822922
4	0.796388	0.818251	0.821791
5	0.791192	0.818285	0.820132
6	0.795691	0.81966	0.823268
7	0.795912	0.821511	0.82352
8	0.796625	0.81831	0.825204
9	0.794062	0.804959	0.816497
10	0.796066	0.817122	0.82169
Seed	0.82425	0.822528	0.832197
Mean	0.795828	0.817482	0.822173

Appendix E: Intuitive visualization of the relative contributions of ensemble members based on optimized weights

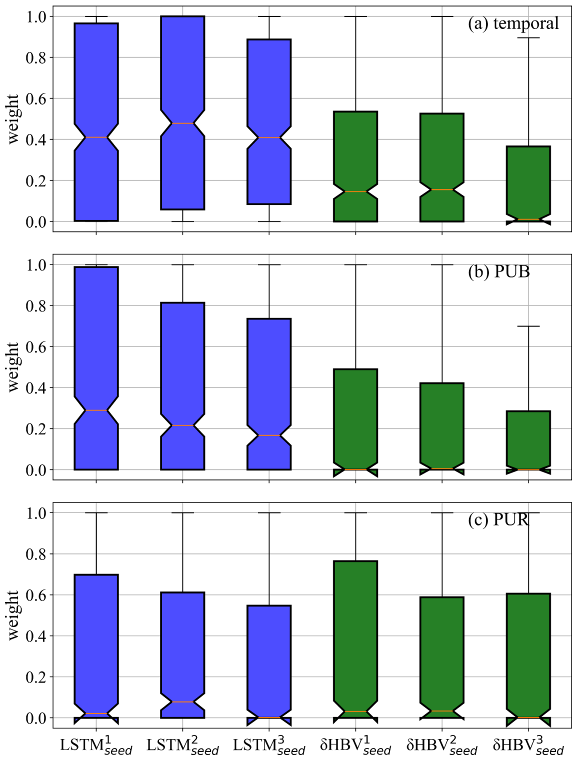

Figure E1. Weights of six components across 531 basins, estimated basin-by-basin using a genetic algorithm based on streamflow observations during the test periods. The weights are normalized by the maximum weight within each ensemble group. These weights are used exclusively for qualitatively analyzing the relative contributions of different ensemble members, with higher values indicating larger relative contributions.

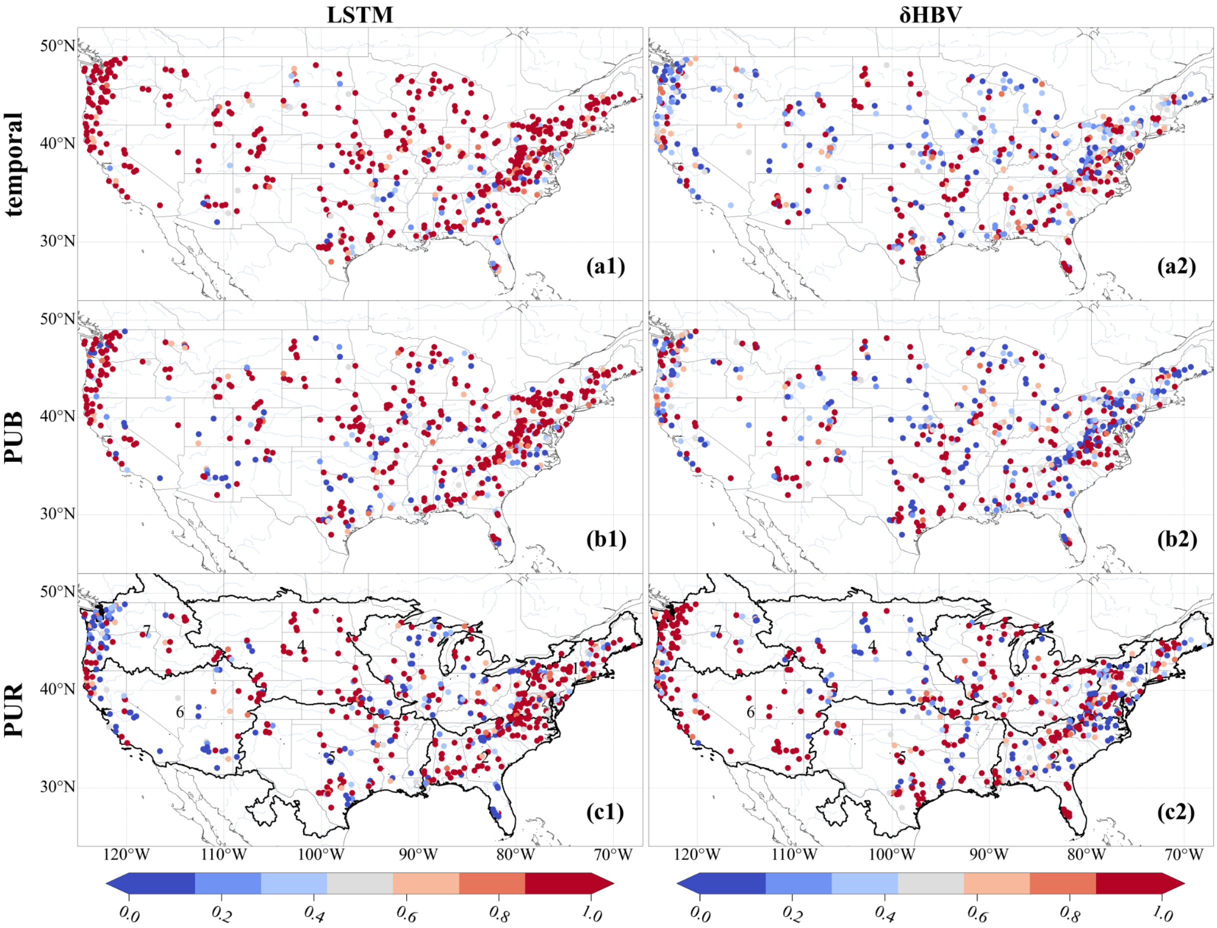

Figure E2. Spatial distributions of weights of the LSTM and δHBV models, estimated by a genetic algorithm based on streamflow observations during the test periods. The weights are normalized by the maximum weight within each ensemble group. These weights are used exclusively for qualitatively analyzing the relative contributions of different ensemble members, with higher values indicating larger relative contributions.

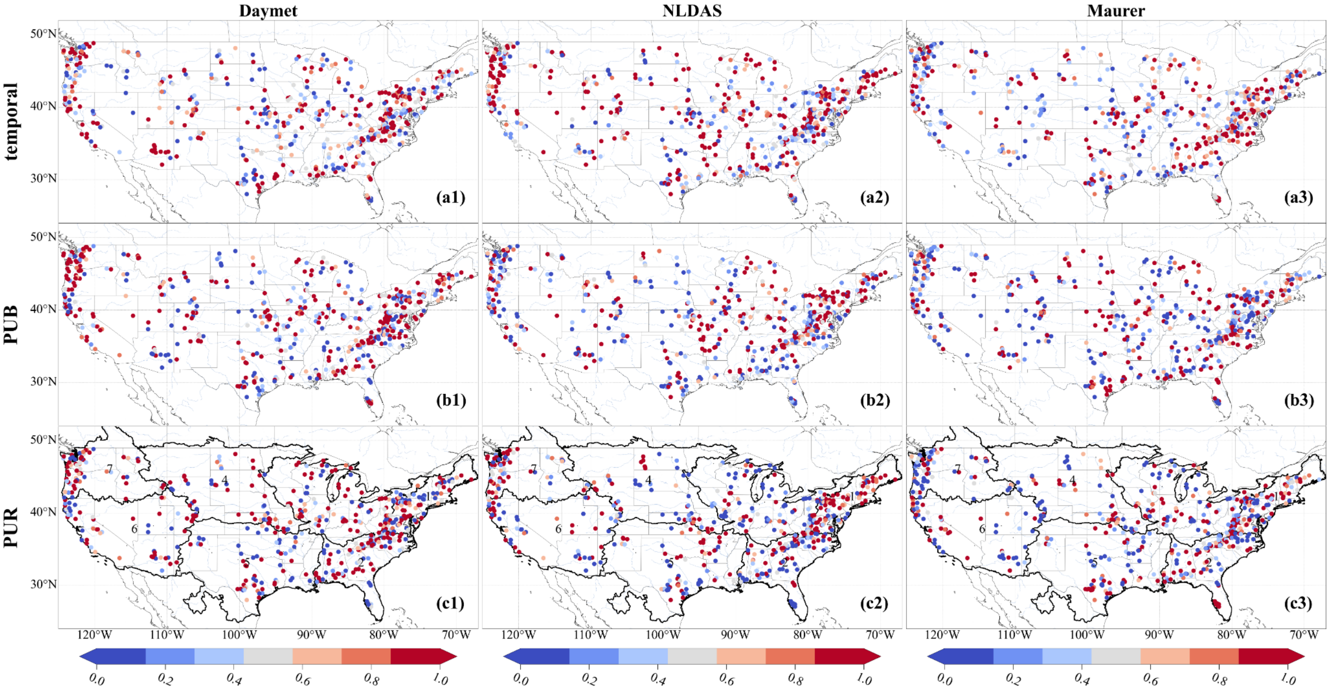

Figure E3. Spatial distributions of weights of the Daymet, NLDAS, and Maurer meteorological forcing datasets, estimated by a genetic algorithm based on streamflow observations during the test periods. The weights are normalized by the maximum weight within each ensemble group. These weights are used exclusively for qualitatively analyzing the relative contributions of different ensemble members, with higher values indicating larger relative contributions.

Table E1. Comparisons of metric values between averaged ensemble simulations and optimized weighted simulations, estimated using a genetic algorithm based on streamflow observations during the test periods. The results highlight the potential for further improvements in ensemble simulations.

	Temporal	Averaged	Optimized weighted
Temporal	NSE	0.821444	0.844303212
	KGE	0.795317	0.829996445
	RMSE	0.99455	0.920954559
	PBIAS	3.99009	3.252278013
	lowRMSE	0.059782	0.057137161
	highRMSE	2.7279	2.451194907
	midRMSE	0.209943	0.183127162
PUB	NSE	0.793673	0.842396015
	KGE	0.726188	0.79571295
	RMSE	1.12957	0.987170488
	PBIAS	0.370674	1.023040859
	lowRMSE	0.0834234	0.079807878
	highRMSE	3.89363	3.030715903
	midRMSE	0.323045	0.285110115
PUR	NSE	0.705154	0.790796063
	KGE	0.651538	0.746396324
	RMSE	1.30377	1.13058149
	PBIAS	−0.283645	0.273698787
	lowRMSE	0.100525	0.093595304
	highRMSE	4.74889	3.665495069
	midRMSE	0.406797	0.351694421

Code and data availability. The source codes and datasets utilized in this study are publicly accessible through the following repositories: The δHBV modeling framework, including all computational scripts and documentation, is hosted on Zenodo (https://doi.org/10.5281/zenodo.7943626) (Feng et al., 2023a), with an updated version and comprehensive software release scheduled upon manuscript acceptance. The implementation of the LSTM architecture is accessible through Zenodo (https://doi.org/10.5281/zenodo.6326394) (Kratzert et al., 2022). The CAMELS hydrometeorological dataset, which provides the foundational basin characteristics and time series data used in our analysis, can be obtained via https://doi.org/10.5065/D6MW2F4D (Newman et al., 2022; Addor et al., 2017). The streamflow simulations produced in this study can be downloaded at https://doi.org/10.5281/zenodo.16895228 (Li et al., 2025).

Author contributions. PL and CS designed the experiments and PL carried them out. YS developed the modified δHBV code. PL prepared the manuscript with contributions from all co-authors.

Competing interests. Chaopeng Shen and Kathryn Lawson have financial interests in HydroSapient, Inc., a company that could potentially benefit from the results of this research. This interest has been reviewed by the Pennsylvania State University in accordance with its individual conflict of interest policy for the purpose of maintaining the objectivity and the integrity of research. The other authors have no competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. PL, CS, and KL were supported by the Office of Biological and Environmental Research of the U.S. Department of Energy (contract no. DESC0016605). PJ and MP were also partially supported by California Department of Water Resources Atmospheric River Program Phase III (Grant 4600014294). YS and CS were partially supported by subaward A23-0252-S002 from the Cooperative Institute for Research to Operations in Hydrology (CIROH) through the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement (Grant no. NA22NWS4320003).

Financial support. Financial support. This research has been supported by the Biological and Environmental Research (grant no. DESC0016605), the Department of Water Resources (grant no. 4600014294), and the National Oceanic and Atmospheric Administration (grant no. A23-0252-S002).

Review statement. This paper was edited by Thomas Kjeldsen and reviewed by Aggrey Muhebwa and one anonymous referee.

References

Aboelyazeed, D., Xu, C., Hoffman, F. M., Liu, J., Jones, A. W., Rackauckas, C., Lawson, K., and Shen, C.: A differentiable, physics-informed ecosystem modeling and learning framework for large-scale inverse problems: demonstration with photosynthesis simulations, Biogeosciences, 20, 2671–2692, https://doi.org/10.5194/bg-20-2671-2023, 2023.

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.

Aghakouchak, A. and Habib, E.: Application of a Conceptual Hydrologic Model in Teaching Hydrologic Processes, International Journal of Engineering Education, 26, 963–973, 2010.

Bandai, T. and Ghezzehei, T. A.: Physics-informed neural networks with monotonicity constraints for Richardson-Richards equation: Estimation of constitutive relationships and soil water flux density from volumetric water content measurements, Water Resources Research, 57, e2020WR027642, https://doi.org/10.1029/2020wr027642, 2021.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881–2903, https://doi.org/10.5194/hess-21-2881-2017, 2017.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., and Wood, E. F.: Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments, Journal of Geophysical Research: Atmospheres, 125, e2019JD031485, https://doi.org/10.1029/2019JD031485, 2020.

Behnke, R., Vavrus, S., Allstadt, A., Albright, T., Thogmartin, W. E., and Radeloff, V. C.: Evaluation of downscaled, gridded climate data for the conterminous United States, Ecological Applications, 26, 1338–1351, https://doi.org/10.1002/15-1061, 2016.

Bell, V. A. and Moore, R. J.: The sensitivity of catchment runoff models to rainfall data at different spatial scales, Hydrol. Earth Syst. Sci., 4, 653–667, https://doi.org/10.5194/hess-4-653-2000, 2000.

Bellmore, J. R., 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, Wiley Interdisciplinary Reviews: Water, 4, e1164, https://doi.org/10.1002/wat2.1164, 2017.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, PhD Thesis, Swedish Meteorological and Hydrological Institute (SMHI), Norköping, Sweden, urn:nbn:se:smhi:diva-5738, 1976.

Bergström, S.: The HBV model – its structure and applications, SMHI, urn:nbn:se:smhi:diva-2672, 1992.

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., and Shen, C.: Improving river routing using a differentiable Muskingum-Cunge model and physics-informed machine learning, Water Resources Research, 60, e2023WR035337, https://doi.org/10.1029/2023WR035337, 2024.

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.

Brunner, M. I., Slater, L., Tallaksen, L. M., and Clark, M.: Challenges in modeling and predicting floods and droughts: A review, WIREs Water, 8, e1520, https://doi.org/10.1002/wat2.1520, 2021.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, https://doi.org/10/chvc6k, 2008.

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.

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., Fowler, H. J., Prudhomme, C., Arnold, J. R., and Brekke, L. D.: Characterizing uncertainty of the hydrologic impacts of climate change, Curr. Clim. Change Rep., 2, 55–64, https://doi.org/10.1007/s40641-016-0034-x, 2016.

Dion, P., Martel, J.-L., and Arsenault, R.: Hydrological ensemble forecasting using a multi-model framework, Journal of Hydrology, 600, 126537, https://doi.org/10.1016/j.jhydrol.2021.126537, 2021.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, Water Resources Research, 56, e2019WR026793, https://doi.org/10.1029/2019WR026793, 2020.

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., Liu, J., Lawson, K., and Shen, C.: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy, Water Resources Research, 58, e2022WR032404, https://doi.org/10.1029/2022WR032404, 2022.

Feng, D., Shen, C., Liu, J., Lawson, K., and Beck, H.: differentiable parameter learning (dPL) + HBV hydrologic model, Zenodo [code, data set], https://doi.org/10.5281/zenodo.7943626, 2023a.

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, Hydrol. Earth Syst. Sci., 27, 2357–2373, https://doi.org/10.5194/hess-27-2357-2023, 2023b.

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, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

Hanazaki, R., Yamazaki, D., and Yoshimura, K.: Development of a reservoir flood control scheme for global flood models, JAMES, 14, e2021MS002944, https://doi.org/10.1029/2021MS002944, 2022.

Hargreaves, G. H.: Defining and using reference evapotranspiration, Journal of Irrigation and Drainage Engineering, 120, 1132–1139, https://doi.org/10.1061/(ASCE)0733-9437(1994)120:6(1132), 1994.

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.

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.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Ji, H., Song, Y., Bindas, T., Shen, C., Yang, Y., Pan, M., Liu, J., Rahmani, F., Abbas, A., Beck, H., Lawson, K., and Wada, Y.: Distinct hydrologic response patterns and trends worldwide revealed by physics-embedded learning, Nat. Commun., 16, 9169, https://doi.org/10.1038/s41467-025-64367-1, 2025.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning, Geophys. Res. Lett., 47, e2020GL088229, https://doi.org/10.1029/2020GL088229, 2020.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

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

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-Runoff modelling using Long-Short-Term-Memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.17605/OSF.IO/QV5JZ, 2018.

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, Hydrol. Earth Syst. Sci., 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, Zenodo [data set], https://doi.org/10.5281/zenodo.6326394, 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., Tso, C. H. M., Zhang, Y., and Zeng, W.: Comparison of the use of a physical-based model with data assimilation and machine learning methods for simulating soil water dynamics, Journal of Hydrology, 584, 124692, https://doi.org/10.1016/j.jhydrol.2020.124692, 2020a.

Li, P., Zha, Y., Tso, C. H. M., Shi, L., Yu, D., Zhang, Y., and Zeng, W.: Data assimilation of uncalibrated soil moisture measurements from frequency-domain reflectometry, Geoderma, 374, 114432, https://doi.org/10.1016/j.geoderma.2020.114432, 2020b.

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.

Li, P., Zha, Y., Shi, L., and Zhong, H.: Assessing the Global Relationships Between Teleconnection Factors and Terrestrial Water Storage Components, Water Resources Management, 36, 119–133, https://doi.org/10.1007/s11269-021-03015-x, 2022.

Li, P., Zha, Y., Zuo, B., and Zhang, Y.: A family of soil water retention models based on sigmoid functions, Water Resources Research, 59, e2022WR033160, https://doi.org/10.1029/2022WR033160, 2023a.

Li, P., Zha, Y., and Tso, C.-H. M.: Reconstructing GRACE-derived terrestrial water storage anomalies with in-situ groundwater level measurements and meteorological forcing data, Journal of Hydrology: Regional Studies, 50, 101528, https://doi.org/10.1016/j.ejrh.2023.101528, 2023b.

Li, P., Zha, Y., Zhang, Y., Michael Tso, C.-H., Attinger, S., Samaniego, L., and Peng, J.: Deep learning integrating scale conversion and pedo-transfer function to avoid potential errors in cross-scale transfer, Water Resources Research, 60, e2023WR035543, https://doi.org/10.1029/2023WR035543, 2024.

Li, P., Song, Y., Pan, M., Lawson, K., and Shen, C.: Streamflow Simulation Data from Differentiable HBV and LSTM Models Using CAMELS Datasets, Zenodo [data set], https://doi.org/10.5281/zenodo.16895228, 2025.

Lin, Y., Wang, D., Zhu, J., Sun, W., Shen, C., and Shangguan, W.: Development of objective function-based ensemble model for streamflow forecasts, Journal of Hydrology, 632, 130861, https://doi.org/10.1016/j.jhydrol.2024.130861, 2024.

Lins, H. F. and Slack, J. R.: Streamflow trends in the United States, Geophysical Research Letters, 26, 227–230, https://doi.org/10/d5zbbd, 1999.

Liu, J., Rahmani, F., Lawson, K., and Shen, C.: A multiscale deep learning model for soil moisture integrating satellite and in situ data, Geophysical Research Letters, 49, e2021GL096847, https://doi.org/10.1029/2021GL096847, 2022.

Liu, J., Bian, Y., Lawson, K., and Shen, C.: Probing the limit of hydrologic predictability with the Transformer network, Journal of Hydrology, 637, 131389, https://doi.org/10.1016/j.jhydrol.2024.131389, 2024.

Mai, J., Craig, J. R., Tolson, B. A., and Arsenault, R.: The sensitivity of simulated streamflow to individual hydrologic processes across North America, Nat. Commun., 13, 455, https://doi.org/10.1038/s41467-022-28010-7, 2022.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, Journal of Climate, 15, 3237–3251, https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2, 2002.

Moges, E., Demissie, Y., and Li, H.-Y.: Hierarchical mixture of experts and diagnostic modeling approach to reduce hydrologic model structural uncertainty, Water Resources Research, 52, 2551–2570, https://doi.org/10.1002/2015WR018266, 2016.

Nai, C., Liu, X., Tang, Q., Liu, L., Sun, S., and Gaffney, P. P. J.: A novel strategy for automatic selection of cross-basin data to improve local machine learning-based runoff models, Water Resources Research, 60, e2023WR035051, https://doi.org/10.1029/2023WR035051, 2024.

Narkhede, M. V., Bartakke, P. P., and Sutaone, M. S.: A review on weight initialization strategies for neural networks, Artificial Intelligence Review, 55, 291–322, https://doi.org/10.1007/s10462-021-10033-z, 2022.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

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.

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., Nearing, G., Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, Journal of Hydrometeorology, 18, 2215–2225, https://doi.org/10/gbwr9s, 2017.

Newman, A. J., Clark, M. P., Longman, R. J., and Giambelluca, T. W.: Methodological intercomparisons of station-based gridded meteorological products: Utility, limitations, and paths forward, Journal of Hydrometeorology, 20, 531–547, https://doi.org/10.1175/JHM-D-18-0114.1, 2019.

Newman, A. J., Sampson, K., Clark, M., Bock, A., Viger, R., Blodgett, D., Addor, N. and Mizukami, M.: CAMELS: Catchment Attributes and MEteorology for Large-sample Studies (1.2), Zenodo [data set], https://doi.org/10.5065/D6MW2F4D, 2022.

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., and Shen, C.: Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy, Journal of Hydrology, 599, 126455, https://doi.org/10.1016/j.jhydrol.2021.126455, 2021.

Paul, P. K., Zhang, Y., Ma, N., Mishra, A., Panigrahy, N., and Singh, R.: Selecting hydrological models for developing countries: Perspective of global, continental, and country scale models over catchment scale models, Journal of Hydrology, 600, 126561, https://doi.org/10.1016/j.jhydrol.2021.126561, 2021.

Rahmani, F., Appling, A., Feng, D., Lawson, K., and Shen, C.: Identifying structural priors in a hybrid differentiable model for stream water temperature modeling, Water Resources Research, 59, e2023WR034420, https://doi.org/10.1029/2023WR034420, 2023.

Reichle, R. H. and Koster, R. D.: Assessing the impact of horizontal error correlations in background fields on soil moisture estimation, Journal of Hydrometeorology,

orology, 4, 1229–1242, https://doi.org/10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2, 2003.

Sawadekar, K., Song, Y., Pan, M., Beck, H., McCrary, R., Ullrich, P., Lawson, K., and Shen, C.: Improving differentiable hydrologic modeling with interpretable forcing fusion, J. Hydrol., 659, 133320, https://doi.org/10.1016/j.jhydrol.2025.133320, 2025.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nat. Rev. Earth Environ., 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Solanki, H., Vegad, U., Kushwaha, A., and Mishra, V.: Improving streamflow prediction using multiple hydrological models and machine learning methods, Water Resources Research, 61, e2024WR038192, https://doi.org/10.1029/2024WR038192, 2025.

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, Hydrol. Earth Syst. Sci., 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, ESS Open Archive, https://doi.org/10.22541/essoar.172304428.82707157/v2, 2025b.

Thornton, P. E., Running, S. W., and White, M. A.: Generating surfaces of daily meteorological variables over large regions of complex terrain, Journal of Hydrology, 190, 214–251, https://doi.org/10.1016/S0022-1694(96)03128-9, 1997.

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

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.

Wang, N., Zhang, D., Chang, H., and Li, H.: Deep learning of subsurface flow via theory-guided neural network, Journal of Hydrology, 584, 124700, https://doi.org/10.1016/j.jhydrol.2020.124700, 2020.

West, B. D., Maxwell, R. M., and Condon, L. E.: A scalable and modular reservoir implementation for large-scale integrated hydrologic simulations, Hydrol. Earth Syst. Sci., 29, 245–259, https://doi.org/10.5194/hess-29-245-2025, 2025.

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.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011JD016048, 2012.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., and Shen, C.: Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships, Journal of Hydrology, 603, 127043, https://doi.org/10.1016/j.jhydrol.2021.127043, 2021.

Yao, L., Libera, D. A., Kheimi, M., Sankarasubramanian, A., and Wang, D.: The roles of climate forcing and its variability on streamflow at daily, monthly, annual, and long-term scales, Water Resources Research, 56, e2020WR027111, https://doi.org/10.1029/2020WR027111, 2020.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, https://doi.org/10/fpvsgb, 2008.

Yu, D., Yang, J., Shi, L., Zhang, Q., Huang, K., Fang, Y., and Zha, Y.: On the uncertainty of initial condition and initialization approaches in variably saturated flow modeling, Hydrol. Earth Syst. Sci., 23, 2897–2914, https://doi.org/10.5194/hess-23-2897-2019, 2019.

Yu, M., Huang, Q., and Li, Z.: Deep learning for spatiotemporal forecasting in Earth system science: a review, International Journal of Digital Earth, 17, 2391952, https://doi.org/10.1080/17538947.2024.2391952, 2024.

Zhang, Q., Shi, L., Holzman, M., Ye, M., Wang, Y., Carmona, F., and Zha, Y.: A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation, Advances in Water Resources, 132, 103407, https://doi.org/10.1016/j.advwatres.2019.103407, 2019.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, Journal of Hydrology, 598, 126266, https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.