Journal of Hydrology 664 (2026) 134477

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Hybrid modeling for daily streamflow forecasting: A study over the contiguous United States☆

Fanzhang Zeng [a], Zhao Zhao [b], Natalie P. Memarsadeghi [c,f], Charles J. McKnight [c], Xudong Wang [b], Sarah Miele [d], Mayank Chadha [e], Dania Ammar [d], Yichao Zeng [b], Magdalena Asborno [c], Kenneth Mitchell [c], Guga Gugaratshan [d], Michael D. Todd [e], Zhen Hu [b,*], Dingbao Wang [a,*]

[a] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA
[b] Department of Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn, Dearborn, MI 48128, USA
[c] Coastal and Hydraulics Laboratory, Engineering Research and Development Center, US Army Corps of Engineers, Vicksburg, MS 39180, USA
[d] Hottinger Bruel & Kjaer Solutions LLC, Southfield, MI 48076, USA
[e] Department of Structural Engineering, University of California San Diego, La Jolla, CA 92093-0085, USA
[f] Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Streamflow forecasting
Hybrid modeling
Delta learning
Data augmentation
LSTM

ABSTRACT

Streamflow forecasting plays a significant role in flood predictions, reservoir operations, and planning for navigation channel dredging. In this paper, the Long Short-Term Memory (LSTM) networks are used for developing the surrogate model, $G_m(\bullet)$, of a hydrologic model, which simulates saturation and infiltration excess runoff in a coherent framework. Two hybrid modelling approaches (delta learning and data augmentation) are applied to 600 watersheds in the contiguous United States for streamflow forecasting. In the delta learning (DL) approach, an LSTM-based surrogate model, $G_\delta(\bullet)$, which is driven by climate data and observed streamflow, is developed. This model aims to capture the discrepancy between the observed streamflow and the streamflow predicted by another model, $G_m(\bullet)$, which also uses climate data and observed streamflow as inputs. In contrast, the data augmentation (DA) approach involves an LSTM-based surrogate model, $G_e(\bullet)$. This model is driven by climate data, observed streamflow, and the outputs from $G_m(\bullet)$, which is similarly driven by climate data and observed streamflow. The findings highlight that both hybrid modeling approaches consistently outperform the hydrologic model in forecasting 30-day streamflow across various watersheds and seasons. The spatial analysis reveals that both DL and DA models consistently deliver strong 30-day streamflow forecasting performance across the eastern U.S., Pacific Northwest, and Northern Rockies, with seasonal variations. Compared to the hydrologic model, hybrid models significantly improve the forecast accuracy, especially in the Pacific Northwest, the Rocky Mountains, and Eastern States. The results show that the performance of the hydrologic model is sensitive to the forecast horizon length, with the performance generally improving at longer lead times, especially in spring. In contrast, the hybrid models maintain consistent and superior accuracy across all forecast horizon lengths. Forecast improvements by DL are positively correlated with latitude but negatively correlated with seasonality and timing of precipitation, frequency of high precipitation days, and streamflow elasticity to precipitation in summer; they are also positively correlated with leaf area index in winter and baseflow index in spring. On the other hand, forecast improvements by DA are negatively correlated with the frequency of high precipitation days in summer and positively correlated with baseflow index in spring. These findings serve as a prospective guide to improve the representation of relevant hydrologic processes for individual watersheds in the hydrologic model.

☆ This article is part of a special issue entitled: 'River Basin Ecohydrological Processes' published in Journal of Hydrology.

* Corresponding authors.
E-mail addresses: zhennhu@umich.edu (Z. Hu), Dingbao.Wang@ucf.edu (D. Wang).

https://doi.org/10.1016/j.jhydrol.2025.134477

Available online 2 November 2025

1. Introduction

Streamflow forecasting is a major area of research within the hydrology and water resources management communities, playing a critical role in supporting decision-making across various time scales. Short-term streamflow forecasting, typically with lead times ranging from a few hours to several days, is essential for real-time decision-making. It supports timely flood warnings and emergency response efforts (Alfieri et al., 2013) and effective multi-year planning for optimal navigation channel dredging (Asborno et al., 2024); it is also vital for urban drainage system operations, helping to prevent urban flooding and infrastructure overload (Piadeh et al., 2022). Medium-range forecasting, which spans from weeks to a few months, is particularly important for proactive water resources management. It enables more effective reservoir operations to ensure reliable water supply and reduce the risk of shortages during dry periods (Zhao et al., 2011). Additionally, it informs agricultural irrigation planning (Zhang et al., 2017), allowing farmers to optimize water use and crop yields, and supports drought preparedness and mitigation strategies (Luo et al., 2023). Long-range streamflow forecasting, with lead times extending from several months to years, plays a strategic role in long-term water resources planning (Troin et al., 2021). These forecasts are crucial for infrastructure development, policy-making, and sustainable management of water systems under changing climatic and socio-economic conditions (Devineni et al., 2008).

Physics-based hydrologic models and data-driven models have previously been used for streamflow forecasting. Physics-based hydrologic models are categorized into lumped, spatially distributed, and semi-distributed models based on representation of spatial variability (e.g., Beven and Kirkby, 1979; Moore, 1985; Liang et al., 1994). Physics-based models capture hydrologic processes, but they are constrained by challenges related to explicitly characterizing or prescribing landscape heterogeneity, capturing process complexity (McDonnell et al., 2007), and handling spatial scale issues when applying governing equations such as Richards' equation in space (Richards, 1931; Blöschl and Sivapalan, 1995). Moreover, physically based distributed hydrological models are often computationally intensive and demand substantial hydrological expertise from both developers and users (Fatichi et al., 2016). Data-driven models include conventional methods such as autoregressive integrated moving average, machine learning methods such as support vector machines (Kisi and Cimen, 2011) and artificial neural networks (e.g., Zealand et al., 1999), and deep learning methods (e.g., Granata et al., 2022; Xu et al., 2022). Particularly in recent years, deep learning techniques—such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs)—have gained significant traction within the field of hydrology (Feng et al., 2020; Xie et al., 2021; Tripathy and Mishra, 2024; Yu et al., 2024). However, data-driven models are limited by physical interpretability and their inability to predict untrained hydrological variables, yet they can effectively capture hydrologic dynamics at the watershed scale (Ng et al., 2023).

Hybrid models integrate physics-based and data-driven models, overcoming the disadvantages of both models and enhancing the overall model performance (Nourani et al. 2014; Ghaith et al., 2020; Yang et al., 2020). For example, Li et al. (2023) integrated neural networks into the conceptual EXP-Hydro model by replacing its internal components, preserving hydrological principles. This hybrid approach enables the prediction of previously untrained hydrological variables without the need for pre-training or post-processing. Li et al. (2024) further implanted this conceptual hydrological model into a recurrent neural network (RNN) cell as a process driver for providing multi-sub-process variables related to runoff process, with an Entity-Aware cell being incorporated as a post-processor layer for simulating daily runoff. Hybrid deep learning models have stronger feature extraction capabilities and more dominant performance (Ng et al., 2023). For example, Yu et al. (2023) explored the synergy between process-based hydrological model (HBV) and LSTM to improve the predictive capability for semi-arid basins by developing three hybrid models (the outputs of the HBV model are used as inputs of the LTSM model to simulate streamflow or the residual of HBV simulated streamflow; the outputs of the LSTM models include simulated streamflow as well as parameters of the HBV model) and found distinct improvements in the three hybrid models when compared with the HBV model and the standalone LSTM model. Mohanty et al. (2024) developed a hybrid model for real-time streamflow forecasting with up to 10-days lead-time, such that the error of simulated streamflow by SWAT model (Arnold et al., 1998) are updated by hierarchical data-driven models including LSTM. Xu et al. (2024) added data-driven models (e.g., random forests, support vector regression, and multilayer perceptron) as the post-processing procedure for residual correction to the results of process-driven XAJ model (Zhao, 1992), enhancing the performance of real-time flood forecasting. Xu et al. (2025) developed a hybrid model using data augmentation to integrate XAJ model and a deep learning model (combination of CNN and Gated Recurrent Unit) for monthly streamflow forecasts in a basin with humid subtropical climate. Zhao et al. (2024) developed two hybrid modeling approaches (delta learning and data augmentation) for forecasting river discharge by integrating LSTM and the physics-based VIC model (Liang et al., 1994) for runoff simulation with the river routing model RAPID (David et al., 2011); the hybrid models yield promising results for a basin located in Colorado. Building upon, this work aims to generalize the approach presented in Zhao et al. (2024) by applying the hybrid modeling methods to watersheds with varying geographies and spatial scales.

This research aims to investigate the performance of two hybrid modeling techniques—delta learning and data augmentation—in forecasting daily streamflow across watersheds in the contiguous United States. Focusing on variations in climate and watershed properties, we conduct a comprehensive analysis to evaluate the applicability and limitations of these techniques. Our findings are intended to offer insights for adopting these hybrid modeling approaches in diverse regions across the United States. The remaining part of the paper proceeds as follows. The study watersheds and data used are presented in Section 2. Section 3 describes the details of the process-based model, deep learning model, and the two hybrid modeling approaches. Results are presented in Section 4 and discussed in Section 5. Section 6 provides the conclusions of this study.

2. Study watersheds and data used

The watersheds used in this study were sourced from the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) dataset, which was developed and expanded by the United States National Center for Atmospheric Research (NCAR; Addor et al., 2017; Newman et al., 2015; Fowler et al., 2021; Coxon et al., 2020) and covers the continental United States. The data extracted from the dataset for this study encompasses daily meteorological forcing inputs, catchment attributes, and streamflow observations for 671 watersheds, making it suitable for large sample hydrology such as comparative studies of hybrid model performance. The CAMELS dataset includes daily data on precipitation, potential evapotranspiration, air temperature, and streamflow from 1985 to 2014 for each watershed. The watersheds, ranging in size from 4 to 25,000 km^2, were selected due to their relatively low anthropogenic impacts. Deep learning studies utilizing the CAMELS dataset have demonstrated state-of-the-art performance, surpassed several calibrated lumped conceptual models and distributed hydrological models (e.g., Ma et al., 2021).

3. Methodology

3.1. Hydrologic model

This study employs the process-based hydrologic model with a unified runoff scheme for saturation and infiltration excess as shown in

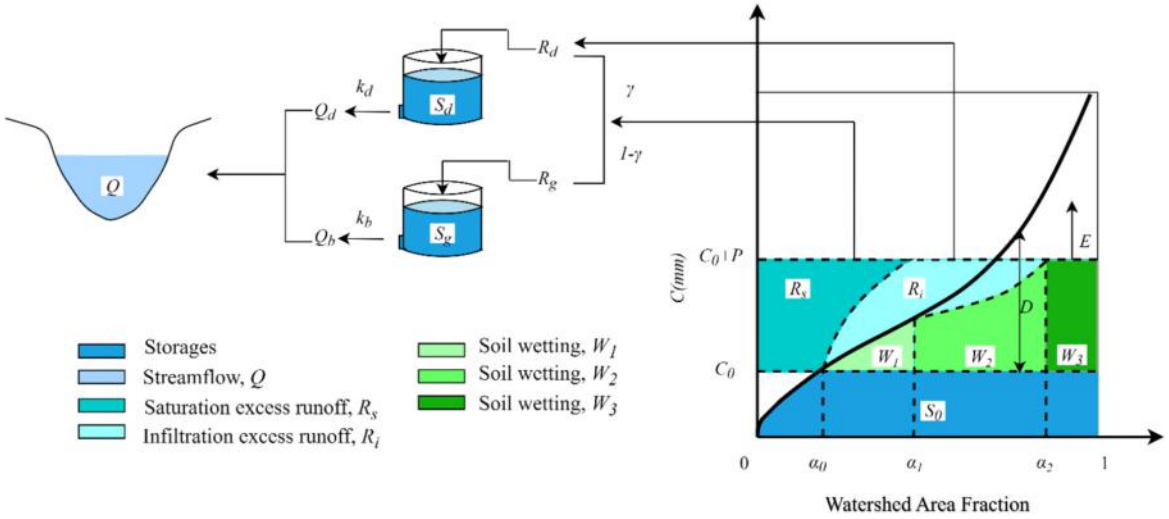

Fig. 1. The structure of hydrologic model with a unified runoff scheme for saturation excess and infiltration excess (Hong et al., 2025).

Fig. 1, as developed by Hong et al. (2025). The model provides a coherent representation of both saturation and infiltration excess runoff, capturing the spatial and temporal transitions between these two runoff generation mechanisms. The snow-related processes are accounted by the degree-day factor method with one parameter (m_s) using mean daily air temperature (Eder et al., 2003; Ye et al., 2012).

At the watershed scale, the spatial distribution of soil water storage capacity is modeled by the distribution function proposed by Wang (2018) and is represented by the black bold curve in Fig. 1. This distribution function has two parameters including average storage capacity (S_b) and shape parameter (α). C is soil water storage capacity at a point and C_0 is corresponding to the average initial soil water storage (S_0) in Fig. 1. The saturation excess runoff (R_s) from this distribution leads to the Soil Conservation Service curve number method (SCS, 1972) at the event scale and Budyko equation at the long-term scale (Wang and Tang, 2014; Yao and Wang, 2022). Yao et al. (2020) applied the distribution function to replace the generalized Pareto distribution of HyMOD model (Moore, 1985) to demonstrate the unified runoff model across time scales.

More recently, Hong et al. (2025) incorporated infiltration excess runoff (R_i) into this framework. In this formulation, infiltration capacity (f_c) at a point is expressed as a power function of degree of soil water deficit, defined as the ratio of soil moisture deficit (D) to S_b. The parameters for the power function include coefficient (m_k) and exponent (n). Therefore, the spatial distribution of infiltration capacity (controlling infiltration excess runoff) is coherently coupled with storage capacity (controlling saturation excess runoff).

As shown in Fig. 1, at the beginning of the time interval (e.g., day), the antecedent saturation area fraction (α_0) corresponds to the initial average storage (S_0). With precipitation (P) during the time interval, saturation area fraction increases to α_1. Over the area fraction of $\alpha_1 - \alpha_0$, runoff generation is switched from infiltration excess to saturation excess, and infiltration (soil wetting) over this area is denoted as W_1. Infiltration excess occurs over the area fraction of $\alpha_2 - \alpha_1$, and infiltration over this area is denoted as W_2. All the rainfall infiltrates into the soil over the area fraction of $1 - \alpha_2$, and infiltration over this area is denoted as W_3. The average infiltration over the entire watershed is the sum of W_1, W_2, and W_3. E is actual evapotranspiration.

Direct runoff (R_d), comprising infiltration excess and a fraction (γ) of saturation excess, flows into the quick storage tank. The discharge from this tank (Q_d) is proportional to its storage (S_d), which is governed by the runoff coefficient k_d. Meanwhile, Groundwater recharge (R_g), represented by the remaining fraction ($1 - \gamma$) of saturation excess, enters the slow storage tank. Baseflow from this tank (Q_b) is similarly proportional to its storage (S_g), with the runoff coefficient k_b. The total streamflow (Q) at the watershed outlet is the sum of Q_d and Q_b. Comprehensive details of the hydrologic model are available in Hong et al. (2025).

3.2. Data-driven model

3.2.1. Nonlinear autoregressive exogenous model

Nonlinear autoregressive models with exogenous inputs (NARX) link the current value of a time series to its past values and to the current and past values of other external time series (Takens, 1981). A nonlinear dynamic system, such as a hydrologic system with external inputs like precipitation and potential evapotranspiration, can be modeled using a NARX model as follows:

$$y_i = F\left(U_i, U_{i-1}, \cdots, U_{i-(q-1)}, y_{i-1}, \cdots y_{i-p}\right) + \varepsilon_i. \quad (1)$$

Here, y_i represents the system output at the i-th time step; $U_i = [u_{1,i}, \cdots, u_{n,i}]$ denotes the exogenous inputs with a dimension of n at the i-th time step; $F(\bullet)$ is a nonlinear mapping function that transforms recent inputs and outputs into the current output; q and p are the number of lags in the delayed input and output vectors, respectively; and ε_i is the residual of the NARX model, typically modeled as Gaussian white noise with zero mean and standard deviation estimated from the residual data. NARX models have been applied in various water resources research, such as establishing multi-step-ahead flood forecast models for the next hour at a 10-minute scale (Chang et al., 2022), groundwater level prediction (Gharehbaghi et al., 2022), daily suspended sediment forecast (Li et al., 2022), predicting the outcomes of ecological restoration from water diversion (Liu et al., 2022), and estimating the enduring effects of long-term drought on water fluxes and storage, as well as projecting future short-term groundwater levels and recovery potential under various precipitation scenarios (Luo et al., 2024).

The nonlinear mapping function $F(\bullet)$ can be approximated using various machine learning models. Shamseldin and O'Connor (2001) developed a multi-layer feedforward neural network-based NARX model for river flow forecasting. Nanda et al (2016) combined artificial neural network (ANN), wavelet analysis, and NARX models for real-time flood forecasting. Wunsch et al. (2021) implemented NARX as a RNN and LSTM networks. In this paper, LSTMs are used to implement the function $F(\bullet)$ in the NARX model, due to their demonstrated effectiveness and widespread adoption by researchers in this field, as detailed in the following section.

3.2.2. Long short-term memory network

RNNs are a type of neural architecture specifically designed for handling sequence data, such as text, speech, and time series. In RNNs, the output from a neuron at one time step is fed back as input to the neuron at the next timestep, allowing RNNs to capture temporal dependencies. However, traditional RNNs struggle with the vanishing gradient problem, where gradient magnitudes diminish across layers during backpropagation, hindering their capacity to capture long-term dependencies. LSTM networks, an advanced variant of RNNs, were introduced to overcome this issue (Hochreiter and Schmidhuber, 1997). LSTMs modify the hidden layer of RNNs to better capture longer dependencies. An LSTM network comprises multiple LSTM layers, each containing a suite of LSTM cells. LSTMs are well-suited for processing and predicting large time spans in time series data by incorporating special gating mechanisms (input gate, output gate, and forget gate). These three gates within the cell unit regulate the flow of information between cell units. The forget gate decides which information from the previous cell unit should be discarded. The input gate determines which new information (both from the forget gate and the current input) should be updated. The output gate decides which information is finally output. Fig. 2 illustrates the architecture of a single LSTM cell, where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$ are the gate signals for the input, forget and output gates, respectively. $\boldsymbol{x}_t$, $\boldsymbol{a}_t$, and $\boldsymbol{C}_t$ are the input, hidden, and cell states, respectively. The relationships among these variables are expressed as follows:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \bullet [\boldsymbol{a}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i), \tag{2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \bullet [\boldsymbol{a}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f), \tag{3}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \bullet [\boldsymbol{a}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o), \tag{4}$$

$$\widetilde{\boldsymbol{C}}_t = tanh(\boldsymbol{W}_c \bullet [\boldsymbol{a}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c), \tag{5}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \otimes \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \otimes \widetilde{\boldsymbol{C}}_t \tag{6}$$

$$\boldsymbol{a}_t = \boldsymbol{o}_t \otimes tanh(\boldsymbol{C}_t). \tag{7}$$

In these equations, $tanh(\bullet)$ is the hyperbolic tangent function, and $\sigma(\bullet)$ is the sigmoid function. Both $tanh(\bullet)$ and $\sigma(\bullet)$ enhance the model's nonlinear expression capabilities. $\otimes$ denotes the element-wise product. $\widetilde{\boldsymbol{C}}_t$ represents the intermediate cell state created by a *tanh* layer. $\boldsymbol{W}$ and $\boldsymbol{b}$ with different subscripts are the weight ($\boldsymbol{W}_c$, $\boldsymbol{W}_i$, $\boldsymbol{W}_f$, $\boldsymbol{W}_o$) and bias ($\boldsymbol{b}_c$, $\boldsymbol{b}_i$, $\boldsymbol{b}_f$, and $\boldsymbol{b}_o$) vectors associated with different gates, and they are the trainable parameters of the LSTM network. $\boldsymbol{a}_t$ is the final output of the unit cell and provides the hidden state for the next timestep. The LSTM network have been combined with the NARX model to predict streamflow (e.g., Hunt et al., 2022).

3.3. Hybrid modeling

Hybrid deep learning models combine one or more modeling techniques with deep learning models (e.g., LSTM) to strengthen the capability of streamflow forecasts. Ng et al. (2023) provided a comprehensive review on hybrid deep learning applications for streamflow forecasts. The modeling techniques include data decomposition (Zuo et al., 2020; Zhao et al., 2021), data convolution (Xu et al., 2022; Wunsch et al., 2022), encoder-decoder (Kao et al., 2020; Ni et al., 2020), attention mechanism (Wang et al., 2023), ensemble modeling (Ma et al., 2023), and physically based models (Cho and Kim, 2022; Han and Morrison, 2022). Despite the performance and applicability of deep learning models being much better than the physics-based model, one significant drawback is the lack of consideration of physical mechanisms in deep learning models (Jiang et al., 2022). Hybrid models can thus be developed by integrating the process-based hydrologic model and the data-driven model (LSTM-NARX) to create a more robust model for accurate streamflow forecasting. The primary benefit of a hybrid model compared to standalone data-driven or physics/process-based models is its ability to merge the strengths of both process-based and data-driven models. For example, Cui et al. (2022) showed that using the forecasted streamflow of the XAJ model as the exogenous input variable of the LSTM decoder enhanced the prediction performance.

3.3.1. Delta learning

This method first trains a surrogate model $G_m(\bullet)$ based on the low-fidelity data from the hydrologic model and then trains another surrogate model $G_\delta(\bullet)$ based on the discrepancy between the physics-based prediction from $G_m(\bullet)$ and the high-fidelity data from the observed streamflow at the gauge station (Fig. 3). After training the two surrogate models, the discharge forecasting in the future period can be made by adding the prediction of $G_\delta(\bullet)$ to that of $G_m(\bullet)$ (Zhao et al., 2024). This method will be introduced through two parts: training stage and prediction stage.

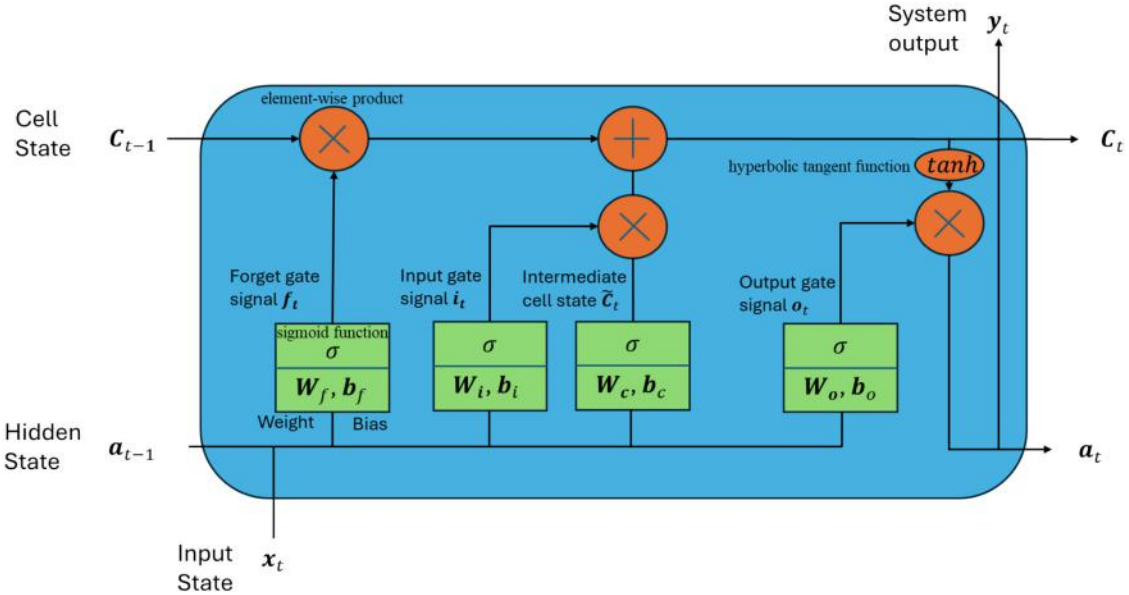

Fig. 2. The architecture of single Long Short-Term Memory (LSTM) cell.

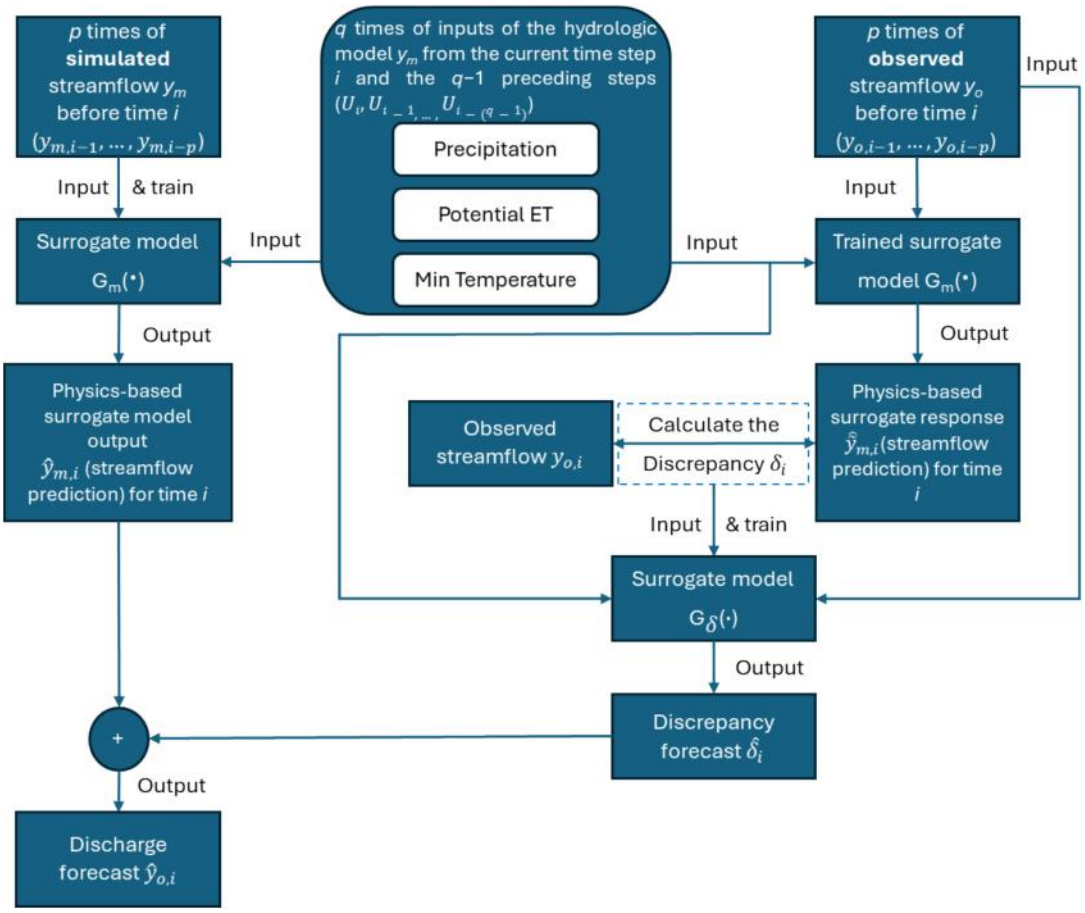

Fig. 3. Hybrid forecasting approach for delta learning.

3.3.1.1. Model training. The first step aims to build an underlying physics-based surrogate model based on the low-fidelity data from the hydrologic model. Following NARX architecture described in Section 3.2, the underlying physics-based surrogate model can be constructed as follows:

$$\hat{y}_{m,i} = G_m\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{m,i-1}, \cdots, y_{m,i-p}\right), \forall i = \omega, \cdots, N_t, \tag{8}$$

where $\omega = max\{q, p+1\}$; $y_{m,i}$ is the simulated streamflow by the hydrologic model; $\boldsymbol{U}_i = [u_{1,i}, u_{2,i}, u_{3,i}]$ denotes the inputs of the hydrologic model (i.e., $u_{1,i}$ is precipitation, $u_{2,i}$ is potential evapotranspiration, and $u_{3,i}$ is minimum temperature at time i), obtained from the CAMELS dataset; and N_t is the number of time step for training. $G_m(\bullet)$ utilizes the output of the process-based model as an input feature, and it enables the hybrid model to capture the physical process of watersheds. Existing studies have enforced that the utilization of these inputs proves successful in estimating daily discharge (e.g., Yang et al., 2019).

The main goal of the 2nd step is to compute the discrepancy between physics-based prediction and observations. To ensure that another NARX model can then be established for this discrepancy term, we use the observation $y_{o,i}$ instead of simulated streamflow by the hydrologic model ($y_{m,i}$) in the delayed input–output vector to yield the physics-based surrogate response:

$$\hat{\hat{y}}_{m,i} = G_m\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}\right), \forall i = \omega, \cdots, N_t \tag{9}$$

With the physics-based surrogate response $\hat{\hat{y}}_{m,i}$, the discrepancy term can be computed by:

$$\delta_i = y_{o,i} - \hat{\hat{y}}_{m,i}, \forall i = \omega, \cdots, N_t. \tag{10}$$

Based on the δ_i values obtained in the previous step, we can continue to construct another surrogate model $G_\delta(\bullet)$ for the discrepancy term:

$$\hat{\delta}_i = G_\delta\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}\right), \forall i = \omega, \cdots, N_t. \tag{11}$$

Substituting Eq. (9) and Eq. (11) into Eq. (10), one obtains:

$$\hat{y}_{o,i} = \hat{\hat{y}}_{m,i} + \hat{\delta}_i. \tag{12}$$

From Eq. (12), it can be found that the basic idea of delta learning is to recover the missing physics by adding a discrepancy term into the underlying physics. The additive form facilities the full use of the underlying physics information from hydrologic models and meanwhile learns the missing physics information based on limited observations.

3.3.1.2. Model forecast. With the two surrogate models $G_m(\bullet)$ and $G_\delta(\bullet)$ trained based on the historical data, the discharge $\hat{y}_{o,i}$, $i = N_t + 1, \cdots, N$, at the future time steps can be forecasted in the following recursive manner. With the known delayed input–output vector $\left[\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}\right]$ where $\boldsymbol{U}_{N_t+1}$ is the forecasted input, we can obtain the surrogate responses from $G_m(\bullet)$ and $G_\delta(\bullet)$ at the (N_t+1)-th time step:

$$\hat{\hat{y}}_{m,N_t+1} = G_m\left(\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}\right), \tag{13}$$

$$\hat{\delta}_{N_t+1} = G_\delta\left(\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}\right). \tag{14}$$

Adding the two surrogate response yields the corrected prediction $\hat{y}_{o,N_t+1}$ at the (N_t+1)-th time step:

$$\widehat{y}_{o,N_t+1} = \widehat{\widehat{y}}_{m,N_t+1} + \widehat{\delta}_{N_t+1}. \tag{15}$$

Let $y_{o,N_t+1} \approx \widehat{y}_{o,N_t+1}$. The delayed input–output vector can move forward one time step, i.e., $\left[\boldsymbol{U}_{N_t+2}, \boldsymbol{U}_{N_t+1}, \cdots, \boldsymbol{U}_{N_t+2-(q-1)}; y_{o,N_t+1}, \cdots, y_{o,N_t+2-p}\right]$, and then $\widehat{y}_{o,N_t+2}$ can be predicted by the above operations. And so on until the discharge at the N-th time step is predicted.

3.3.2. Data augmentation

Like delta learning approach, this method first trains an underlying physics-based surrogate model $G_m(\bullet)$ based on the low-fidelity data from the hydrologic model (Fig. 4). Then, the physics-based surrogate prediction is integrated with the streamflow observations into an augmented dataset to train another surrogate model $G_e(\bullet)$ as shown in Fig. 4. Finally, the trained $G_e(\bullet)$ will provide the corrected predictions (Zhao et al., 2024). The model training and prediction for this approach are described in the following sections.

3.3.2.1. Model training. The first step is the same as delta learning as discussed in Section 3.3.1, and its aim is to build an underlying physics-based surrogate model based on the low-fidelity data from the hydrologic model, denoted as $G_m(\bullet)$ and defined in Eq. (8). Like delta learning, this step also requires physics-based surrogate response $\widehat{\widehat{y}}_{m,i}$ by Eq. (9). The physics-based surrogate response $\widehat{\widehat{y}}_{m,i}$ is then integrated with the delayed input–output vector $\left[\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}\right]$ to train another surrogate model $G_e(\bullet)$, which can directly output the final corrected prediction:

$$\widehat{y}_{o,i} = G_e\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}; \widehat{\widehat{y}}_{m,i}\right), \forall i = \omega, \cdots, N_t. \tag{16}$$

The following equation is obtained by substituting Eq. (9) into Eq. (16):

$$\widehat{y}_{o,i} = G_e\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}; G_m\left(\boldsymbol{U}_i, \boldsymbol{U}_{i-1}, \cdots, \boldsymbol{U}_{i-(q-1)}; y_{o,i-1}, \cdots, y_{o,i-p}\right)\right), \tag{17}$$

where i is between ω and N_t.

From Eq. (17), it can be found that the basic idea of data augmentation is like multi-layer neural networks that embed one layer into another layer. The embedding physics information from the hydrologic model will help to improve the prediction ability of the machine learning model, which allows for more accurate predictions based on limited observed data.

3.3.2.2. Model prediction. With the known delayed input–output vector $\left[\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}\right]$ where $\boldsymbol{U}_{N_t+1}$ is the forecasted input, the surrogate response $\widehat{\widehat{y}}_{m,N_t+1}$ is generated from $G_m(\bullet)$ at

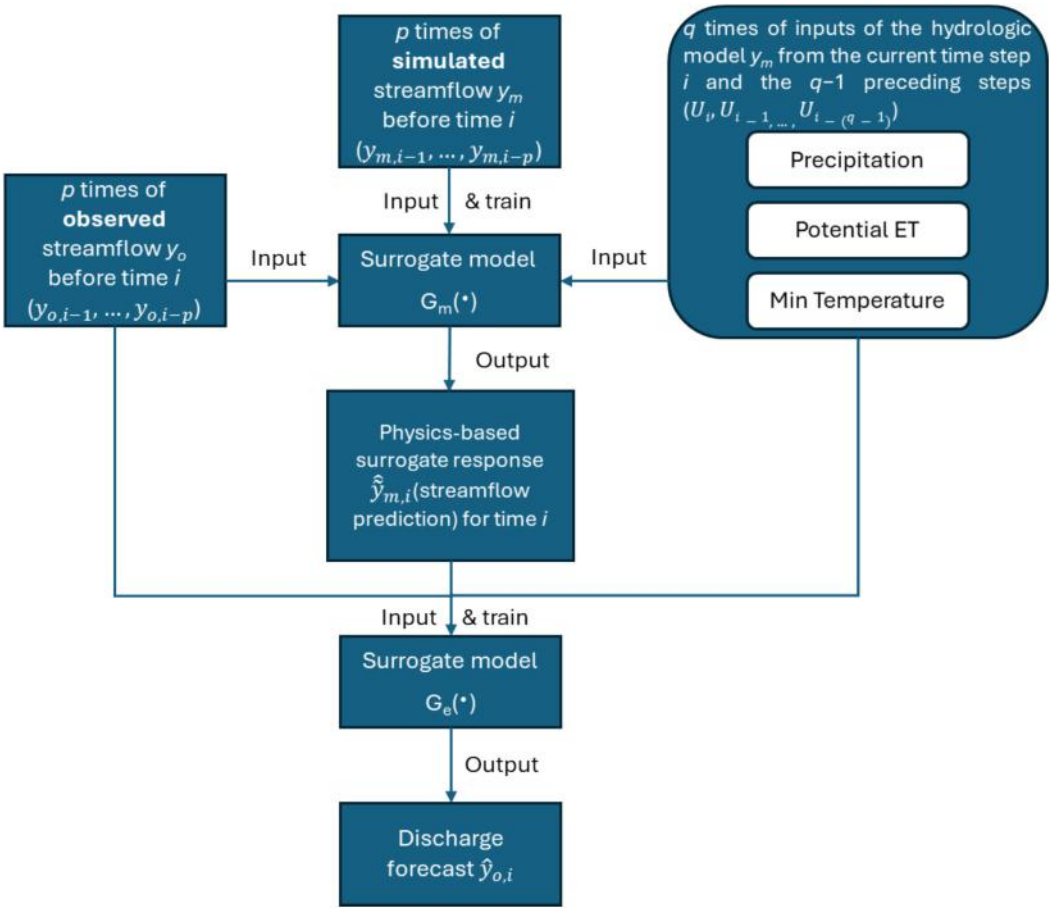

Fig. 4. Hybrid forecasting approach for data augmentation.

the (N_t+1)-th time step, as shown in Eq. (13). Plugging the delayed input–output vector $\left[\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}\right]$ along with the surrogate response $\widehat{\widehat{y}}_{m,N_t+1}$ into the surrogate model $G_e(\bullet)$ yields the corrected prediction at the (N_t+1)-th time step:

$$\widehat{y}_{o,N_t+1} = G_e\left(\boldsymbol{U}_{N_t+1}, \boldsymbol{U}_{N_t}, \cdots, \boldsymbol{U}_{N_t+1-(q-1)}; y_{o,N_t}, \cdots, y_{o,N_t+1-p}; \widehat{\widehat{y}}_{m,N_t+1}\right). \quad (18)$$

Let $y_{o,N_t+1} \approx \widehat{y}_{o,N_t+1}$. Taking the delayed input–output vector one time step further, i.e., $\left[\boldsymbol{U}_{N_t+2}, \boldsymbol{U}_{N_t+1}, \cdots, \boldsymbol{U}_{N_t+2-(q-1)}; y_{o,N_t+1}, \cdots, y_{o,N_t+2-p}\right]$, $\widehat{y}_{o,N_t+2}$ can be predicted by the above operations. And so on until the discharge forecasting at the N-th time step is made.

4. Results

4.1. Hydrologic model calibration

The period 1985–1986 are used for model warm-up, 1987–2005 for parameter determination, and 2006–2014 for model evaluation. The hydrologic model has 8 parameters including m_s, S_b, a, γ, k_b, k_d, m_k, and n (Hong et al., 2025). The range for m_s (snow melting parameter) is between 0 and 1 (Martinez and Gupta, 2010). The range for average storage capacity (S_b) is set to be between 0 mm and 1500 mm (Yao et al., 2020), along with a larger upper bound to be used if needed. The shape parameter of spatial distribution of storage capacity (a) is between 0 and 2 (Wang, 2018). Based on its definition, the range of γ is designated to be between 0 and 1. The value of k_d (coefficient for quick storage tank) should be higher than that of k_b (coefficient for slow storage tank) since the residence time of fast flow should be shorter than that of slow flow; therefore, the range of k_b is set to (0, 0.14) day^{-1} and that of k_d is set to [0.14, 1] day^{-1} (Kollat et al., 2012). The range of m_k (coefficient of power function for infiltration capacity) is assigned to be between 0 and 2000 $mm \bullet day^{-1}$, and that of exponent n is assigned to be between 0 and 1.

The model parameters are determined following a two-stage strategy proposed by Abeshu et al. (2023). In the first stage, runoff parameter sets are generated through stratified sampling, and the best-performing sets are identified by comparing simulated and observed runoff at annual and monthly scales. In the second stage, routing parameters are systematically varied using the top-performing runoff sets as inputs, and the final optimal parameter set is selected based on model performance metrics. For each of the study watersheds, samples from the entire parameter space are selected to obtain one million sets of parameter values and run the model using each parameter set. The parameter estimation process is carried out in three sequential steps to ensure robust calibration of the hydrologic model. The first step is parameter sampling. Initially, one million parameter sets are generated using Latin Hypercube Sampling, a statistical method that ensures a comprehensive and stratified exploration of the parameter space. This approach enhances the diversity and representativeness of the sampled parameter combinations. In the second step, each of the one million parameter sets is used to run the hydrologic model, resulting in a corresponding set of simulation outputs. These outputs are evaluated against observed hydrologic data to assess model performance. The third step involves a multi-stage filtering process based on hydrologic signatures at various temporal scales, as described by Hong et al. (2025). The filtering proceeds as follows: The normalized root mean square error (NRMSE) between simulated and observed annual streamflow is computed for all one million simulations. The 100,000 parameter sets with the lowest NRMSE values are retained for further evaluation. For these 100,000 retained sets, the NRMSE is calculated between the simulated and observed regime curves (i.e., mean monthly streamflow). The top 1,000 parameter sets with the smallest NRMSE values are selected for the next stage. The NRMSE between simulated and observed annual flood peaks is then computed for the 1,000 selected parameter sets. From these, the 100 best-performing sets are chosen based on their flood peak accuracy. Finally, the Kling-Gupta Efficiency (KGE; Gupta et al., 2009) is used to evaluate the agreement between simulated and observed daily streamflow for the remaining 100 parameter sets. The parameter set that achieves the highest KGE value is identified as the optimal set and is selected as the final estimate.

Of the 671 watersheds analyzed, 600 with positive KGE values during the calibration period are selected for hybrid modeling. This selection is conservative, as even a KGE score above −0.41 still outperforms using mean flow as a predictor (Knoben et al., 2019). Fig. S1 in the Supporting Information shows the exceedance probability distribution of KGE values during the calibration and validation periods for the selected watersheds. During this period, 98 % of the watersheds are observed to achieve KGE values above 0.3, 90 % to exceed 0.5, and 40 % to surpass 0.7. In the validation period, 91 % of the watersheds are depicted to maintain KGE values above 0.3, 66 % to remain above 0.5, and 23 % to exceed 0.7, suggesting acceptable model performance. These results indicate a decline in model performance during the validation phase as compared to the calibration phase. However, the use of hybrid models has the potential to enhance the performance of streamflow forecasts.

4.2. Hybrid models

4.2.1. Surrogate model $\boldsymbol{G_m}(\bullet)$

By comparing the performance of LSTM-NARX model $G_m(\bullet)$ with different time lags, it is identified that thirty days for the number of lags in the delayed input (q) and output (p) vectors produce the best prediction accuracy. The LSTM model structure has a total of three layers, including an LSTM layer with 80 units, a dropout layer (applying 1 % dropout) and a dense layer with linear activation for regression. The model is trained using the Adam optimizer with a learning rate of 1×10^{-3} and the Mean Squared Error (MSE) function is used as the loss function for the optimizer. It undergoes 500 epochs with a batch size of 2048. During training, 20 % of the training data is reserved for validation, and the data is shuffled at the start of each epoch to reduce order bias. The input of LSTM layer has a shape of (2048, 30, 4), corresponding to the batch size, time step size, and the feature size respectively. The data (simulated streamflow by the hydrologic model $y_{m,i}$, precipitation $u_{1,i}$, potential evapotranspiration $u_{2,i}$, minimum temperature $u_{3,i}$) during 1985 and 2012 is used for training of $G_m(\bullet)$.

4.2.2. Surrogate model $\boldsymbol{G_\delta}(\bullet)$

The architecture and parameters of LSTM-NARX model $G_\delta(\bullet)$ are consistent with those of $G_m(\bullet)$. The data (discrepancy term δ_i, precipitation $u_{1,i}$, potential evapotranspiration $u_{2,i}$, minimum temperature $u_{3,i}$) during 1985 and 2012 is used for training of $G_\delta(\bullet)$.

4.2.3. Surrogate model $\boldsymbol{G_e}(\bullet)$

The architecture and parameters of LSTM-NARX model $G_e(\bullet)$ are also similar to those of $G_m(\bullet)$. The only difference lies in that the input shape of the LSTM layer is (2048, 30, 5), where an additional item $\widehat{\widehat{y}}_{m,i}$ is added to the input feature, as described in Sec. 3.3.2. As a result, the data (simulated streamflow by the hydrologic model $y_{m,i}$, precipitation $u_{1,i}$, potential evapotranspiration $u_{2,i}$, minimum temperature $u_{3,i}$, physics-based surrogate response $\widehat{\widehat{y}}_{m,i}$) during 1985 and 2012 is used for training of $G_e(\bullet)$.

4.3. Performance of hybrid modeling

The streamflow forecasts by the hydrologic model (HM), DL, and DA are presented in this section. Seasonal streamflow forecasts with a 30-day lead time are issued on July 4, 2013 (summer), October 2, 2013 (fall), December 31, 2013 (winter), and March 31, 2014 (spring).

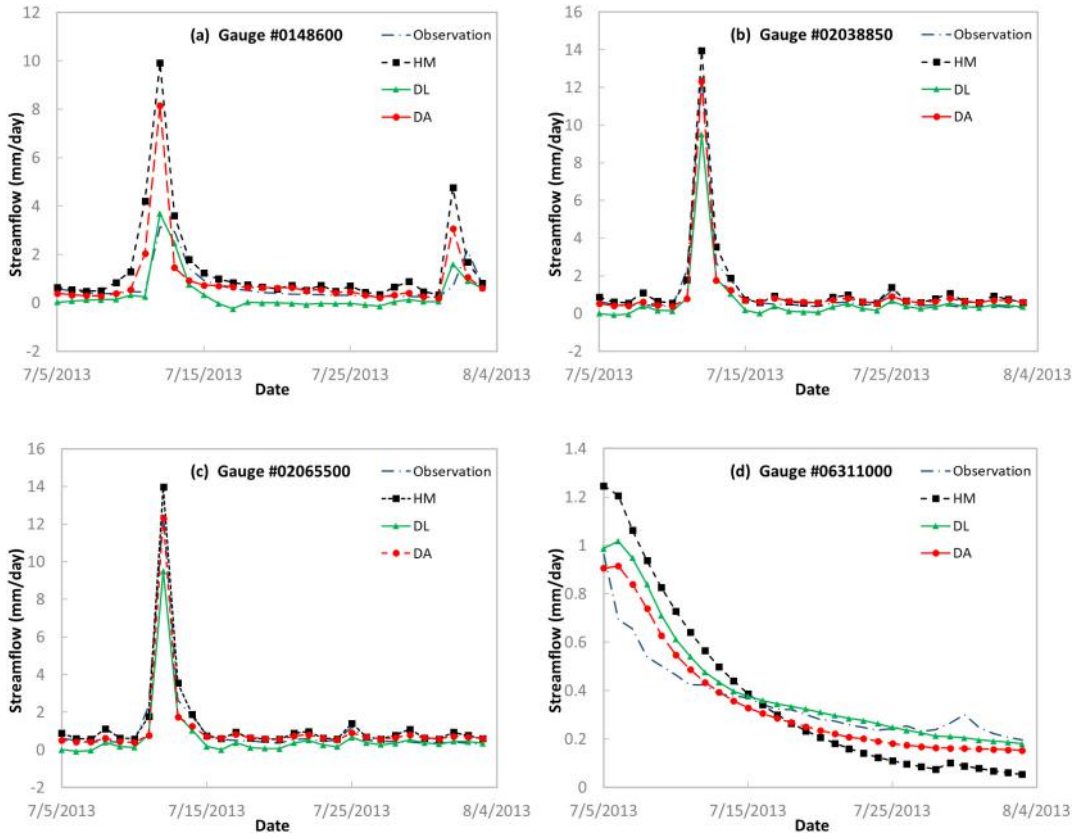

Fig. 5. Observed and 30-day lead time forecasted hydrographs for July 4, 2014, generated using the hydrologic model, Delta Learning, and Data Augmentation approaches across four representative watersheds: (a) Manokin Branch, Maryland; (b) Holiday Creek, Virginia; (c) Cub Creek, Virginia; and (d) North Fork Powder River, Wyoming.

4.3.1. *Hydrographs*

For demonstration purposes, Fig. 5 presents the forecasted daily streamflow for July 4, 2013, with a 30-day lead time across four selected watersheds. Starting with the first watershed, Fig. 5a depicts the observed and forecasted streamflow for Manokin Branch in Maryland, using HM, DL, and DA. The watershed has a climate aridity index of 0.82 and a drainage area of 12.4 km^2 (USGS gauge #01486000). During the 30-day period, two peaks were evident. Compared to observations, HM overestimates high flows; DA improves high flow predictions but still overestimates them; DL significantly enhances high flow predictions, closely matching real observations. HM generally overestimates low flows; DA accurately matches low flows; and DL underestimates low flows. The KGE values are −0.91 for HM, 0.52 for DL, and −0.08 for DA. Fig. 5b shows the observed and forecasted streamflow for the second watershed, Holiday Creek in Virginia, with a drainage area of 22.1 km^2 (USGS gauge #02038850) and a climate aridity index of 0.92. One event is identified during the 30-day period. HM slightly overestimates streamflow, while DL slightly underestimates it and DA matches both high and low flows well. The KGE values are 0.72 for HM, 0.55 for DL, and 0.95 for DA. Moving to the third watershed, Fig. 5c illustrates the observed and forecasted streamflow for Cub Creek in Virginia, with a drainage area of 252.8 km^2 (USGS gauge #02065500) and a climate aridity index of 1.18. Both HM and DA overestimate the peak flow, whereas DL accurately matches the peak flow. The KGE values are 0.51 for HM, 0.97 for DL, and 0.83 for DA. Lastly, Fig. 5d displays the observed and forecasted streamflow for the fourth watershed, North Fork Powder River in Wyoming, with a drainage area of 61.9 km^2 (USGS gauge #06311000) and a climate aridity index of 1.68. The hydrograph during the 30-day period shows a recession event. HM overestimates streamflow when it is higher (July 5, 2013 – July 15, 2013) but underestimates when it is lower (July 15, 2013 – August 4, 2013), resulting in a faster recession than observed. The recession curves for both DL and DA are flatter compared to HM. The KGE values are −0.13 for HM, 0.50 for DL, and 0.61 for DA.

4.3.2. *Grouping by KGE*

The KGE values for the 30-day lead time streamflow predictions by the HM, DL, and DA models are calculated for each of the four forecast dates representing seasons. Watersheds with negative KGE values for all three models (HM, DL, and DA) are excluded. The remaining watersheds, which have positive KGE values for at least one of the models, are categorized into seven groups based on the signs of their KGE values. It should be noted that KGE values greater than −0.41 indicate that a model improves upon the mean flow benchmark, generally considered acceptable (Knoben et al., 2019). As shown in Table 1, Group 1 includes watersheds with positive KGE values for HM, DL, and DA. The numbers of watersheds in this group are 232 for July 4, 2013, 165 for October 2, 2013, 156 for December 31, 2013, and 217 for March 31, 2014. Group 2 includes watersheds with negative KGE values for HM but positive KGE

Table 1
Number of watersheds in each of the seven groups with positive or negative KGE values for 30-day lead time streamflow predictions.

Group	KGE Value			Number of Watersheds			
	HM	DL	DA	Summer	Fall	Winter	Spring
1	+	+	+	232	165	156	217
2	−	+	+	95	84	118	158
3	−	−	+	46	48	37	38
4	−	+	−	19	39	35	20
5	+	−	+	13	24	12	8
6	+	+	−	16	17	10	12
7	+	−	−	34	25	27	15

values for both DL and DA. The number of watersheds in this group is 95 for summer, 84 for fall, 118 for winter, and 158 for spring. Group 3 includes watersheds with negative KGE values for HM and DL but positive KGE values for DA. Group 4 includes watersheds with negative KGE values for HM and DA but positive KGE values for DL. Group 5 includes watersheds with positive KGE values for HM and DA but negative KGE values for DL. Group 6 includes watersheds with positive KGE values for HM and DL but negative KGE values for DA. Group 7 includes watersheds with positive KGE values for HM but negative KGE values for both DL and DA. Compared to the other six groups, Group 1 has the largest number of watersheds. Combining Groups 2, 3, and 4—which are characterized by negative KGE value for HM but positive value for DL and/or DA—results in a total number of 160 watersheds for summer, 171 for fall, 190 for winter, and 216 for spring.

4.3.3. Performance of hybrid modeling

To evaluate and compare the performance of hybrid modeling approaches against the traditional hydrologic model, Table 2 presents the number of watersheds in which each method has achieved the highest KGE values for 30-day streamflow forecasts. The results are broken down by season and by watershed group. Focusing on Group 1 as an illustrative example, HM outperforms the hybrid methods in 59 watersheds during summer, 35 in fall and winter, and 50 in spring. Whereas DL has the highest KGE values in 80 watersheds in summer, 58 in fall and winter, and 94 in spring; and DA leads in 93 watersheds during summer, 72 in fall, 63 in winter, and 73 in spring. When aggregating results across all seven watershed groups, the hybrid modeling approaches (DL or DA) demonstrates superior performance over the hydrologic model in a substantial majority of cases: 75 % of watersheds in summer, 81 % in fall and winter, and 84 % in spring. These findings highlight the consistent advantage of hybrid modeling techniques in enhancing streamflow forecast accuracy across different seasons and watershed conditions.

To further assess the effectiveness of hybrid modeling approaches, the exceedance probability distributions of KGE values for the HM, DL, and DA are illustrated in Fig. 6. These distributions are derived from all watersheds across the seven groups listed in Table 1. For consistency and better visualization of lower-performing cases, the y-axis in Fig. 6 is truncated at a minimum value of −2. In summer (Fig. 6a), both hybrid models outperform HM. Notably, DA exhibits slightly better performance than DL. Specifically, KGE values exceed 0 in 64 % of watersheds for HM, 79 % for DL, and 84 % for DA. When considering a higher threshold (KGE > 0.5), 23 % of watersheds for HM, 40 % for DL, and 45 % for DA are identified. In fall (Fig. 6b), a similar trend is observed. KGE values are greater than 0 in 57 % of watersheds for HM, 75 % for DL, and 79 % for DA. For KGE values above 0.5, the proportions are 23 % for HM, 39 % for DL, and 41 % for DA. During winter (Fig. 6c), the performance gap between the models becomes more pronounced. KGE values exceed 0 in 52 % of watersheds for HM, compared to 80 % for DL and 81 % for DA. For KGE values above 0.5, the percentages are 18 % for HM, 36 % for DL, and 41 % for DA. In spring (Fig. 6d), the hybrid models show their strongest relative performance. KGE values are greater than 0 in 54 % of watersheds for HM, 87 % for DL, and 90 % for DA. For KGE values exceeding 0.5, the results are 22 % for HM, 54 % for DL, and 48 % for DA. Overall, these seasonal exceedance probability distributions clearly demonstrate the superior performance of hybrid modeling approaches over HM in forecasting 30-day streamflow across a wide range of watersheds.

4.3.4. Spatial distribution of model performance

Fig. 7 illustrates the spatial distribution of KGE values for 30-day streamflow forecasts generated using DL and DA across the contiguous United States. In summer (Fig. 7a for DL and Fig. 7e for DA), medium to large KGE values—specifically those represented in the green (0.41–0.60), blue (0.61–0.80), and dark blue (0.81–0.97) ranges—are primarily concentrated in parts of the eastern United States, including areas in the Midwest, Southeast, and Northeast, where many forecast points show good to excellent streamflow prediction performance. Additionally, some isolated pockets in the Pacific Northwest and Northern Rockies also exhibit high KGE values, indicating strong model skill in those locations. These regions contrast with the central and southwestern U.S., where lower KGE values are more prevalent, highlighting spatial variability in the effectiveness of DL for summer streamflow forecasting. In the fall season (Fig. 7b for DL and Fig. 7f for DA), medium to high KGE values are predominantly located in the eastern half of the United States, particularly in the Midwest, Southeast, and parts of the Northeast, where DL shows strong streamflow forecasting performance. Additionally, scattered high-performing locations appear in the Pacific Northwest and Northern Rockies, similar to the summer pattern. These areas contrast with the central and southwestern U.S., where lower KGE values are more common. During the winter season (Fig. 7c for DL and Fig. 7g for DA), the spatial distribution of KGE values for 30-day streamflow forecasts generated by DL reveals strong model performance across much of the eastern United States, particularly in the Appalachian Mountains and surrounding regions. This suggests that DL effectively captures the hydrological behavior of the Appalachian region during winter, likely benefiting from more stable snowmelt and precipitation patterns. Beyond the Appalachians, other areas with notable concentrations of high KGE values include Pacific Northwest and parts of the Midwest. In the spring season (Fig. 7d for DL and Fig. 7h for DA), medium to high KGE values are distributed across several key areas of the United States. Notable concentrations of medium KGE values appear in parts of the Midwest, Northeast, Southeast, West Coast, and Northern Rockies. These patterns suggest that DL maintains a solid predictive skill in diverse hydrological and climatic regions during the spring, with particularly strong performance in select northern and coastal areas. Overall, both DL and DA demonstrate robust and geographically consistent forecasting capabilities, particularly in hydrologically active and topographically complex regions such as the eastern U.S. and the Pacific Northwest. These results underscore the models' ability to adapt to seasonal and regional variability in streamflow dynamics.

Fig. 8 illustrates the spatial distribution of KGE improvements for 30-day streamflow forecasts by hybrid modeling compared with the baseline hydrologic model. In summer (Fig. 8a for DL and 7e for DA), significant improvements, indicated by blue and dark blue dots, are

Table 2
Number of watersheds with the best KGE performance at a 30-day forecast horizon, categorized by hydrologic, Delta learning, and data augmentation models.

Group	Summer			Fall			Winter			Spring		
	HM	DL	DA	HM	DL	DA	HM	DL	DA	HM	DL	DA
1	59	80	93	35	58	72	35	58	63	50	94	73
2	0	39	56	0	36	48	0	64	54	0	93	65
3	0	0	46	0	0	48	0	0	37	0	0	38
4	0	19	0	0	39	0	0	35	0	0	20	0
5	9	0	4	11	0	13	6	0	6	5	0	3
6	10	6	0	6	11	0	6	4	0	4	8	0
7	34	0	0	25	0	0	27	0	0	15	0	0

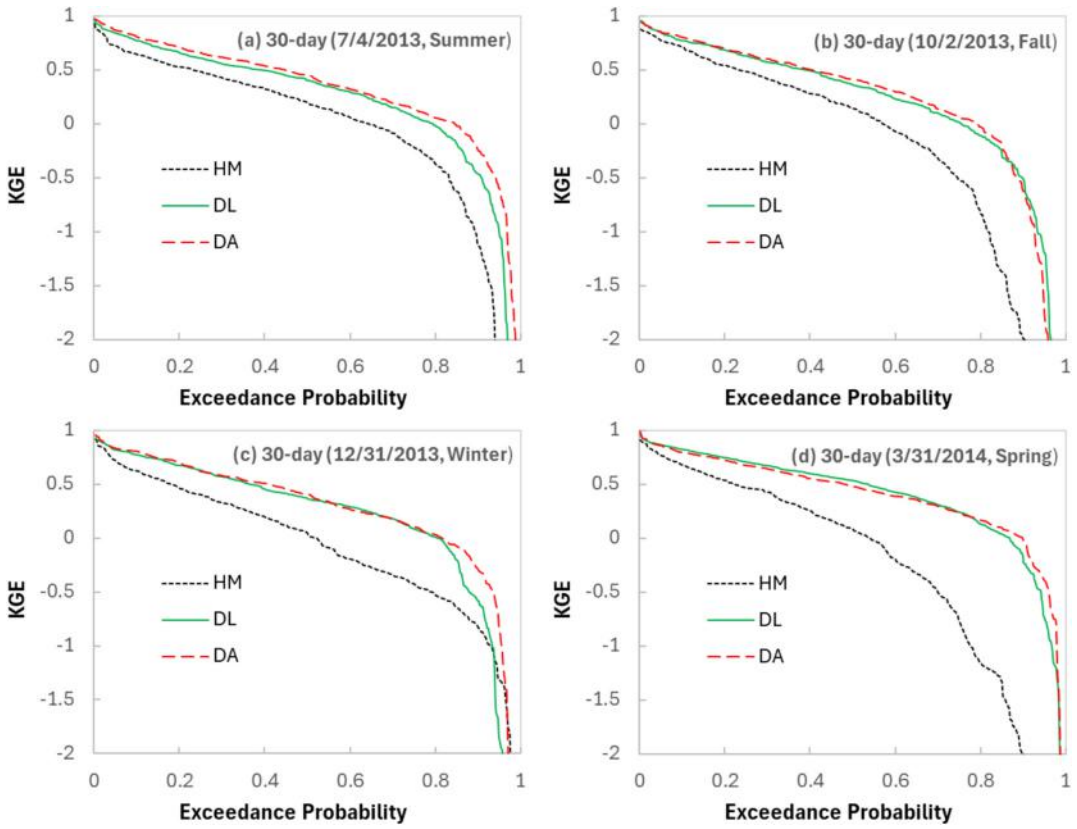

Fig. 6. Comparison of exceedance probability distributions of KGE values for 30-day streamflow forecasts generated by the hydrologic model, Delta Learning, and Data Augmentation methods, shown separately for summer (a), fall (b), winter (c), and spring (d).

concentrated in the Pacific Northwest (with the average value of 0.49 for DL, 0.55 for DA) and the Rocky Mountains (0.62 for DL, 0.67 for DA), as well as scattered locations in the Midwest (1.00 for DL, 1.56 for DA) and along the East Coast (0.39 for DL, 0.40 for DA). These regions exhibit the largest positive differences in KGE values, highlighting areas where DL and DA substantially outperform HM. In fall (Fig. 8b for DL and 7f for DA), regions with substantial KGE improvement are primarily located in the western U.S. (e.g., parts of California (0.99 for DL, 1.26 for DA), the Pacific Northwest(0.29 for DL, 0.11 for DA), and Northern Rockies (0.90 for DL, 0.99 for DA)), as well as in scattered areas in the Midwest (0.80 for DL, 0.53 for DA) and along the East Coast (0.42 for DL, 0.40 for DA). In winter (Fig. 8c for DL and 7 g for DA), strong performance gains are located Northern Rockies (0.10 for DL, 0.75 for DA), Midwest (0.42 for DL, 0.58 for DA), Northeastern (0.15 for DL, 1.24 for DA), Alabama (0.81 for DL, 0.68 for DA), and Florida (0.75 for DL, 0.71 for DA). In spring (Fig. 8d for DL ad 7 h), substantial improvements are prominently observed in the western U.S., particularly in California (0.70 for DL, 0.78 for DA), the Pacific Northwest (0.63 for DL, 0.64 for DA), and key areas of the Rocky Mountains (1.50 for DL, 1.47 for DA), including northern Colorado (2.64 for DL, 2.55 for DA), southern Wyoming (3.74 for DL, 3.83 for DA), western Montana (1.27 for DL, 1.32 for DA), and eastern Idaho (1.56 for DL, 1.48 for DA). Additional clusters of strong performance appear in parts of the Midwest (0.65 for DL, 0.52 for DA) and along the East Coast (0.83 for DL, 0.75 for DA) (e.g., Florida (0.54 for DL, 0.30 for DA)).

Collectively, these results demonstrate that hybrid modeling approaches consistently enhance streamflow forecast accuracy across a range of geographic and climatic regions, with particularly strong gains in the western U.S. and other hydrologically dynamic areas.

5. Discussions

5.1. Impact of lead time

Fig. 9 presents a comparative analysis of the three streamflow forecasting models evaluated across 30-, 60-, and 90-day forecast horizons using the KGE metric, plotted against the percentage of watersheds (exceedance probability). The results reveal that the HM is notably sensitive to forecast horizons, with seasonal variations influencing this sensitivity. In summer (Fig. 9a), HM performance improves with longer lead times, as the 90-day forecast consistently yields higher KGE values than the 30- and 60-day forecasts across most watersheds. This trend becomes even more pronounced in spring (Fig. 9d), where KGE values increase substantially with forecast horizon, suggesting that HM benefits from extended lead times during periods of dynamic hydrologic activity, such as snowmelt and spring runoff. It should be noted that true values are used for climatic forecasts. In practice, climate forecasts may degrade with longer forecast horizons. For example, Shukla et al. (2012) and Schepen et al. (2016) reported that skill at seasonal horizons is strongly conditioned by basin memory and the quality of meteorological forcing. Physics-based hydrologic models are typically sensitive to forecast horizon because their performance strongly depends on the quality of climatic inputs, which often degrade with longer lead times (Shukla et al., 2012; Schepen et al., 2016). In contrast, winter (Fig. 9c) shows minimal sensitivity to the forecast horizon for HM, with KGE values remaining relatively consistent across 30-, 60-, and 90-day forecasts. This indicates that extending the forecast horizon in winter neither significantly enhances nor degrades HM performance, possibly due to more stable hydrologic conditions which aligns with the research by Robertson (2012). The fall season (Fig. 9b) presents a more complex pattern: while 90-day forecasts generally outperform shorter horizons, the 30-day forecast shows higher KGE values than the 60-day forecast

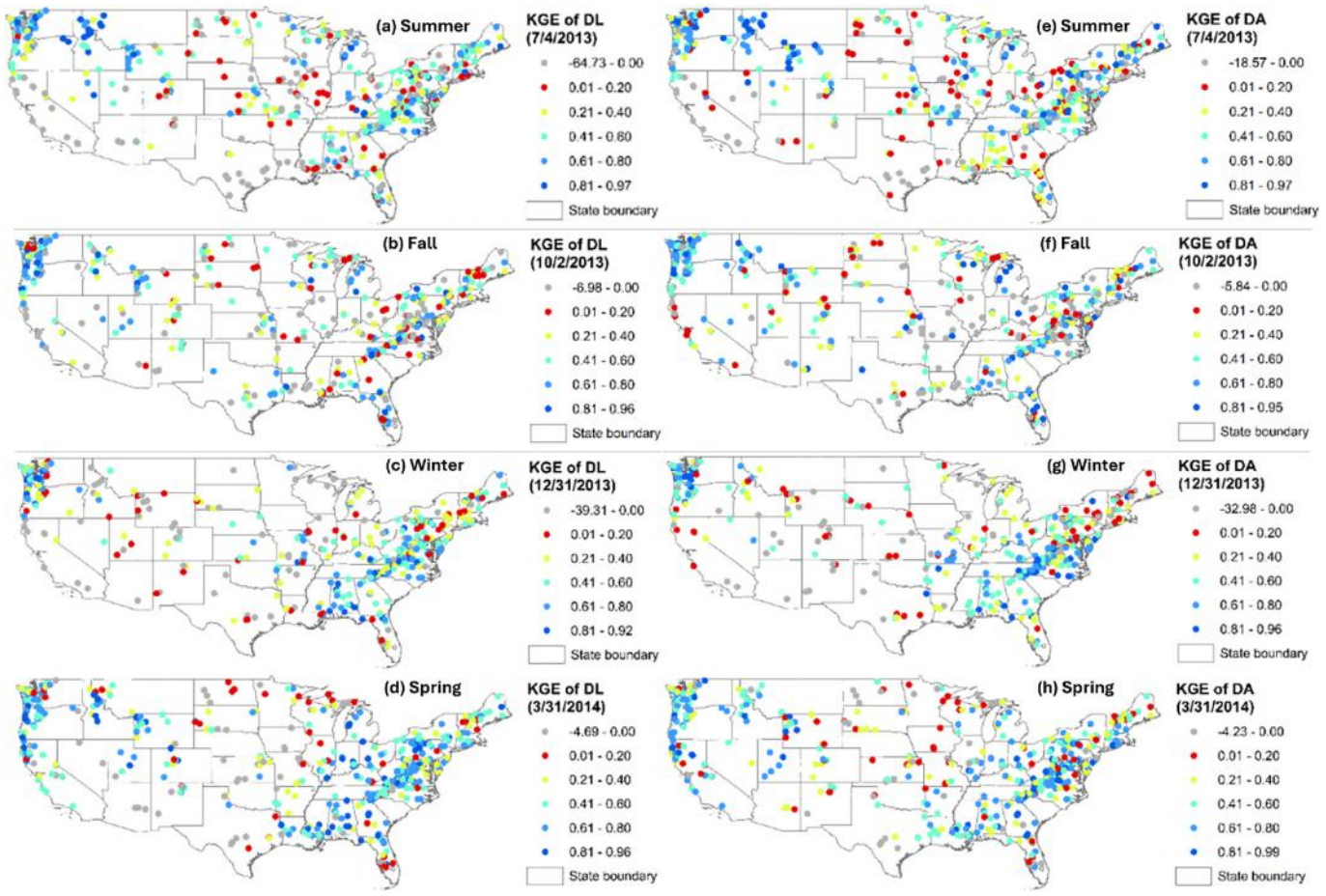

Fig. 7. Spatial distribution of KGE values for 30-day streamflow forecasts generated by Delta Learning (DL) for (a) summer, (b) fall, (c) winter, and (d) spring, and by Data Augmentation (DA) for (e) summer, (f) fall, (g) winter, and (h) spring.

for lower exceedance probabilities (<0.7) but lower values for higher exceedance probabilities (>0.7). This intricate behavior was also noted in Luo's study (2007). This suggests that HM's performance in fall may be influenced by transitional hydrologic conditions and watershed-specific variability.

In contrast to HM, the DL and DA models exhibit relatively stable KGE values across all forecast horizons and seasons, indicating that their performance is largely insensitive to lead time. This finding is consistent with recent machine learning research in hydrology, where residual learning approaches (similar to DL) have effectively captured systematic model errors (Kratzert et al., 2019), and data augmentation strategies have enhanced skill for extreme events by leveraging additional data sources (Nearing et al., 2021). This consistency highlights the robustness of hybrid modeling approaches, which leverage data-driven techniques to maintain accuracy over varying temporal scales. Notably, DL excels at capturing residual patterns not modeled by HM, while DA enhances performance by incorporating additional data, particularly for extreme events.

Across all four seasons, the results (Fig. 9) collectively demonstrate that forecast performance generally improves with longer horizons for HM, especially in spring and summer when hydrologic variability is high. However, DL and DA consistently outperform HM, particularly at shorter horizons (30 and 60 days), underscoring their effectiveness in capturing complex, nonlinear dynamics. These findings emphasize the importance of tailoring forecasting strategies to both seasonal hydrologic conditions and forecast lead times, particularly when selecting between a process-based model and a hybrid model.

5.2. *Performance improvements and watershed characteristics*

To gain physical insights into the performance of hybrid modeling approaches, a correlation analysis is conducted between the improvements in 30-day streamflow forecast skill—measured by the difference in KGE values between the hybrid model and the HM—and various watershed characteristics across different seasons. For the DL approach, six watershed properties, obtained from the CAMELS dataset (Addor et al., 2017), are found to be significantly associated with forecast improvements: four in summer, one in winter, and one in spring (Fig. 10). In summer, Fig. 10a suggests a weak positive trend between gauge latitude and KGE improvement, indicating that higher-latitude basins may benefit slightly more from DL, which aligns with previous finding that hydroclimatic variability with latitude influences model predictability (Berghuijs et al., 2014). Conversely, negative correlations are observed between KGE improvement and (i) seasonality and timing of precipitation (Fig. 10b), (ii) frequency of high precipitation days (Fig. 10c), and (iii) streamflow-precipitation elasticity (Fig. 10d). The seasonality metric is derived from fitting sine curves to annual temperature and precipitation cycles, where positive values indicate summer-peaking precipitation and values near zero reflect uniform precipitation throughout the year. High precipitation days are defined as those exceeding five times the mean daily precipitation. Streamflow-precipitation elasticity quantifies the sensitivity of annual streamflow to changes in annual precipitation, with higher values indicating more responsive basins. This metric has been associated with hydrologic model transferability and forecast skill (Sankarasubramanian et al., 2001; Troch et al., 2013).

In winter, DL-based forecast improvements are found to be positively

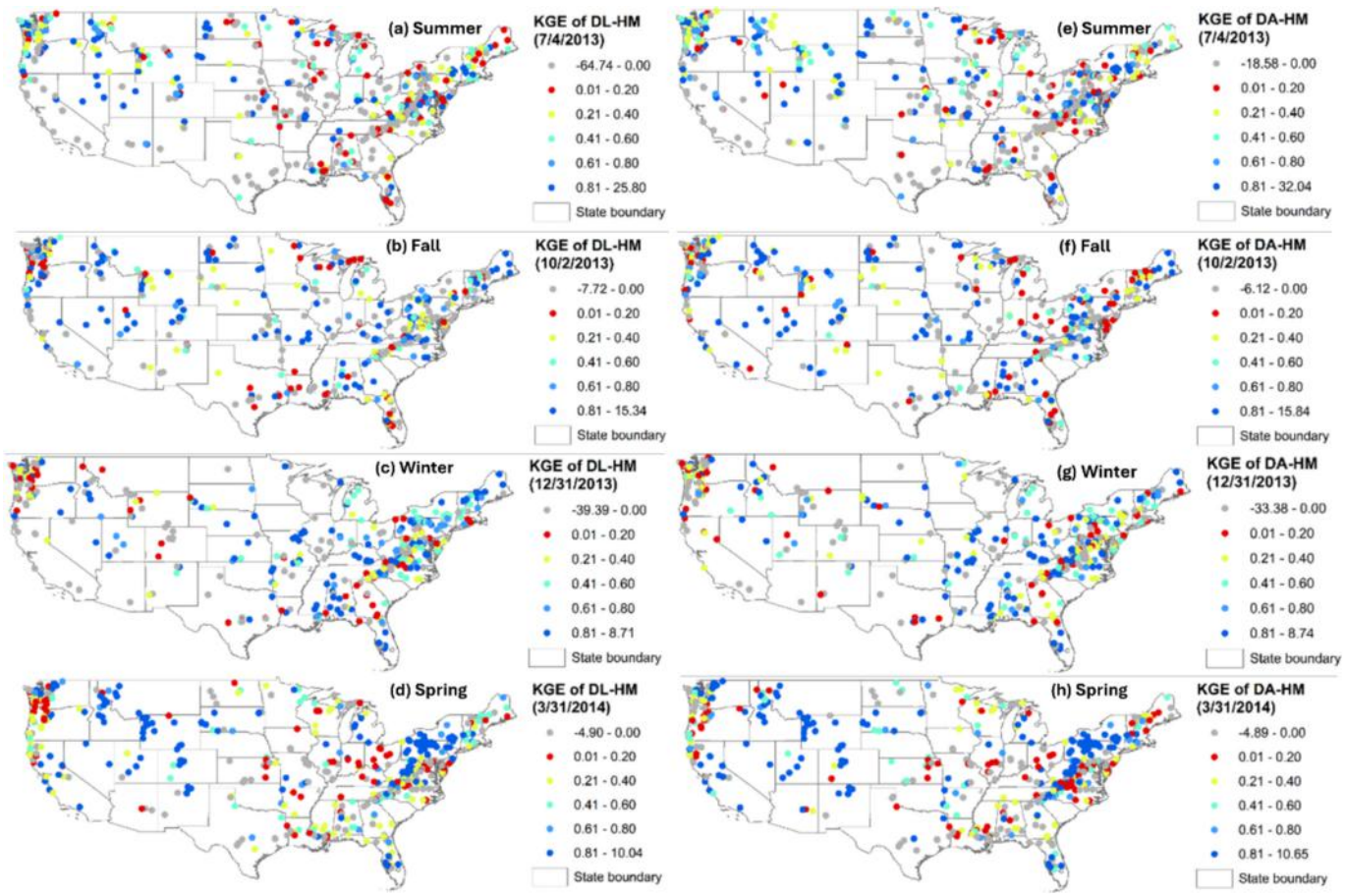

Fig. 8. Spatial distribution of KGE differences between hybrid modeling and the hydrologic model for 30-day streamflow by DL for (a) summer, (b) fall, (c) winter, and (d) spring, and by DA for (e) summer, (f) fall, (g) winter, and (h) spring.

correlated with the maximum monthly mean of leaf area index (LAI), suggesting that vegetation dynamics—such as snow interception or evapotranspiration—may influence model performance during this season. This observation is consistent with earlier study showing that vegetation plays a critical role in modulating water and energy fluxes in snow-dominated basins (Mao & Cherkauer, 2009). In spring, a positive correlation is observed between KGE improvement and the baseflow index, which represents the ratio of mean daily baseflow to total streamflow, calculated using the digital filter method of Ladson et al. (2013). This indicates that basins with stronger groundwater contributions tend to benefit more from DL in spring.

In contrast, the DA approach demonstrates fewer and more selective relationships with watershed characteristics (Fig. 11). Only two correlations are identified: a negative correlation with the frequency of high precipitation days in summer (Fig. 11a), and a positive correlation with the baseflow index in spring (Fig. 11b). This suggests that DA may be more robust to a wide range of watershed conditions but less sensitive to specific hydrologic features compared to DL. Similar observations have been reported in prior hybrid modeling study, where simpler augmentation schemes yielded robust yet less condition-dependent improvements (Nearing et al., 2021). While DL appears to leverage a broader set of physical and climatic signals to enhance forecast skill, DA's improvements are more narrowly tied to hydrologic stability (as indicated by baseflow) and are potentially hindered by extreme precipitation variability. This contrast highlights the complementary nature of the two hybrid approaches and underscores the importance of tailoring modeling strategies to watershed-specific characteristics and seasonal dynamics.

6. Conclusion

Streamflow forecasting is a critical component of water resources engineering, yet it remains a complex and challenging endeavor, whether approached through physics-based or data-driven models. Hybrid models, integrating both process-based and data-driven models, could enhance the performance of streamflow forecasting. In this study, hybrid modelling is applied to a large number of watersheds in the contiguous United States for streamflow forecasting. The recently developed hydrologic model (Hong et al., 2025), which simulates saturation and infiltration excess runoff in a coherent framework due to the dependence of the spatial distribution of infiltration capacity on the spatial distribution of storage capacity and soil water storage, is used as the process-based model. LSTM, a deep learning model, is then utilized for developing data-driven models. An LSTM-based surrogate model, $G_m(\bullet)$, which is driven by hydroclimatic observations and simulated streamflow by HM, is developed for capturing the underlying processes of the hydrologic model. Two hybrid modeling approaches, including DL and DA, are applied for forecasting streamflow. For the DL approach, an LSTM-based surrogate model, $G_\delta(\bullet)$, which is driven by hydroclimatic observations and observed streamflow, is developed to capture the discrepancy between observed streamflow and predicted streamflow by $G_m(\bullet)$, which in turn is driven by hydroclimatic observations and observed streamflow. For the DA approach, the LSTM-based surrogate model, $G_e(\bullet)$, is driven by hydroclimatic observations, observed streamflow, and outputs of $G_m(\bullet)$, which in turn is driven by hydroclimatic observations and observed streamflow.

The findings highlight that both hybrid modeling approaches consistently outperform the hydrologic model in forecasting 30-day

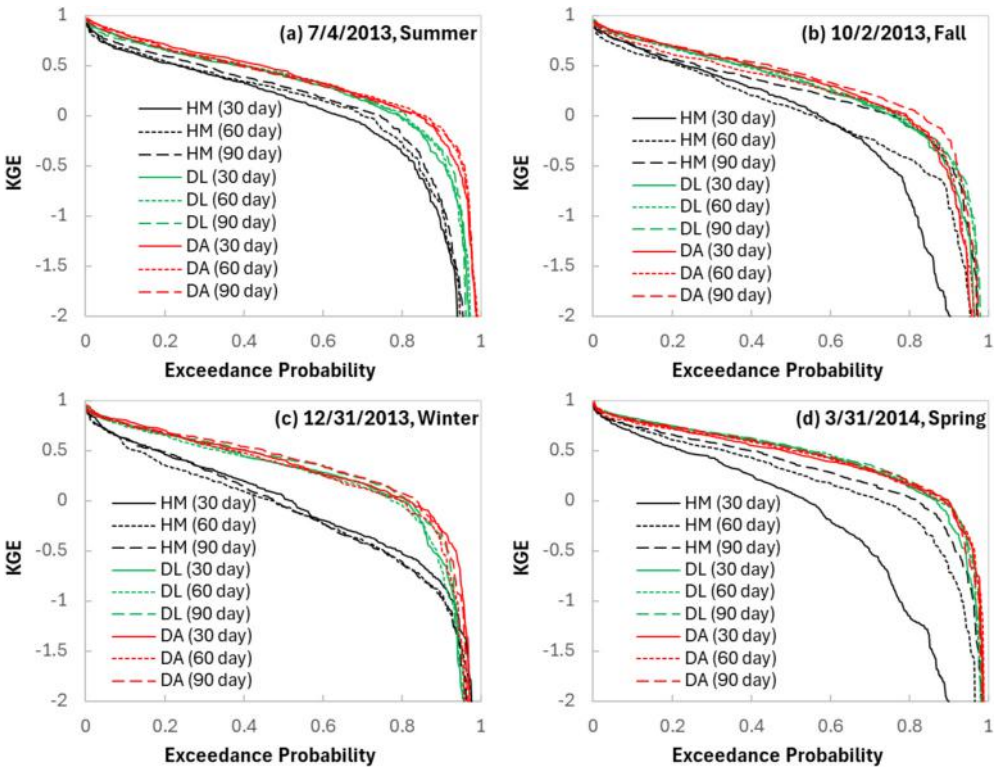

Fig. 9. Exceedance probability distributions of KGE values for 30-day, 60-day, and 90-day streamflow forecasts generated by the hydrologic model, Delta Learning, and Data Augmentation methods, shown separately for summer (a), fall (b), winter (c), and spring (d).

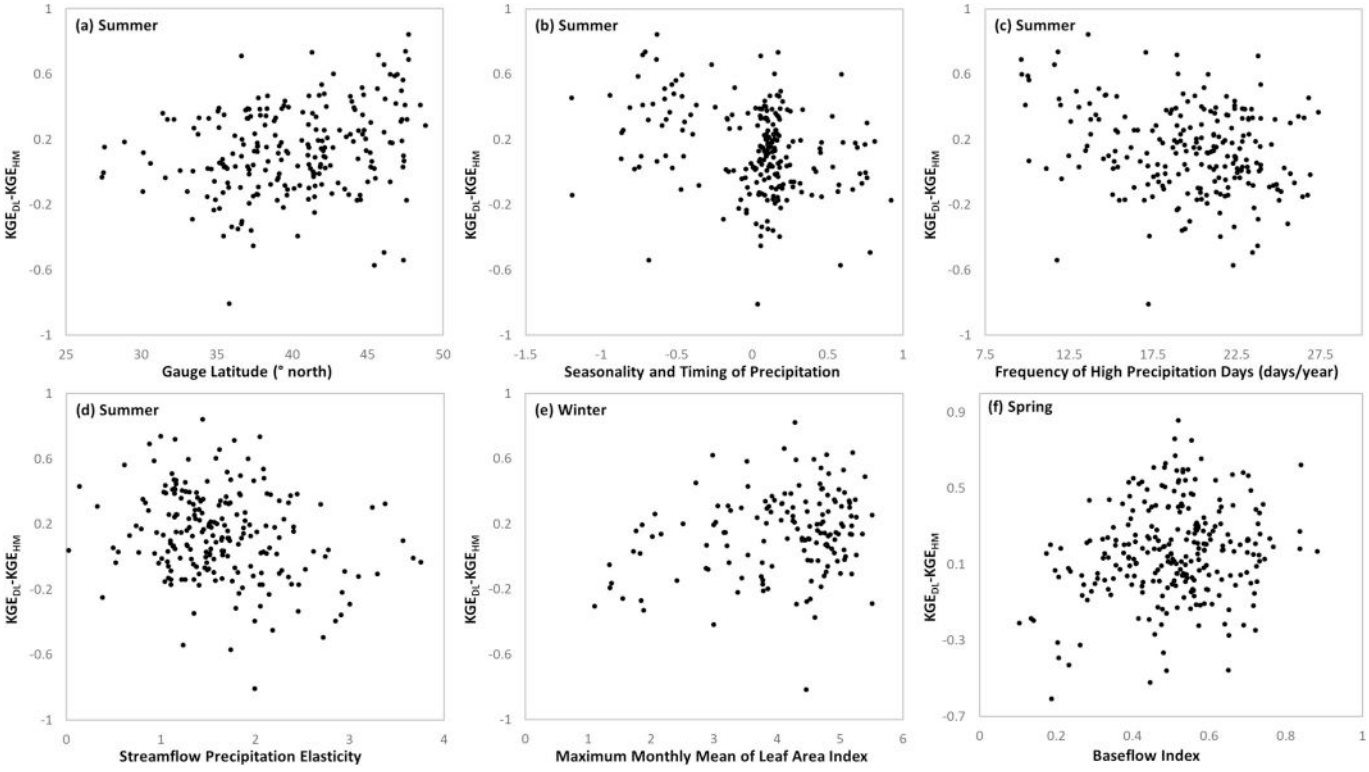

Fig. 10. The correlation of KGE difference for 30-day streamflow forecast by delta learning and the hydrologic model and watershed characteristics: streamflow forecasts improvement in summer versus latitude of gauge station (a), seasonality and timing of precipitation (b), frequency of high precipitation days (c), and streamflow precipitation elasticity (d); streamflow forecasts improvement in winter versus maximum monthly mean of leaf area index (e); and streamflow forecasts improvement in spring versus baseflow index (f).

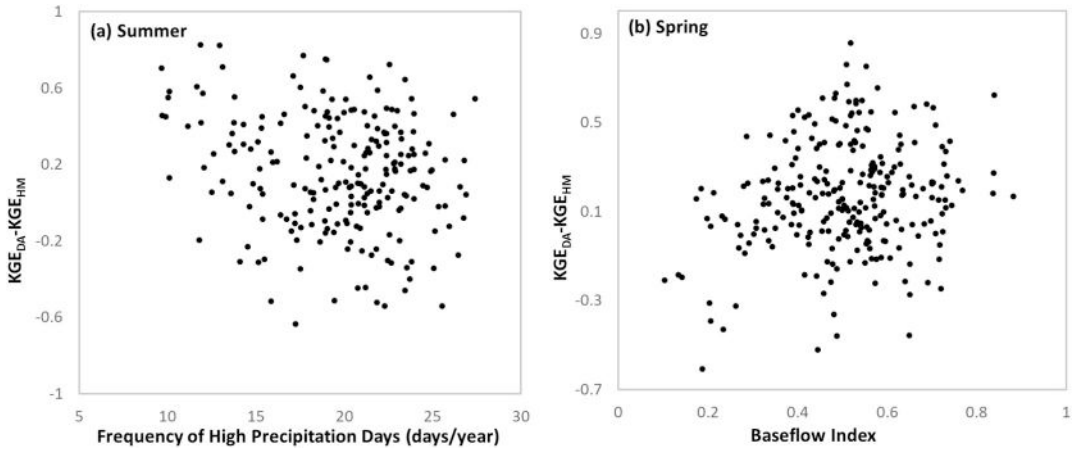

Fig. 11. The correlation of KGE difference for 30-day streamflow forecast by data augmentation and the hydrologic model and watershed characteristics: streamflow forecasts improvement in summer versus frequency of high precipitation days (a), and streamflow forecasts improvement in spring versus baseflow index (b).

streamflow across various seasons and watersheds. DA exhibits slightly better performance than DL. During winter, the performance gap between the models becomes more pronounced. Seasonal analysis reveals that fall provides the most stable forecasting conditions, whereas summer, winter, and spring present challenges due to low flows and snow-related processes. The spatial analysis reveals that both hybrid models consistently deliver strong 30-day streamflow forecasting performance across the eastern U.S., Pacific Northwest, and Northern Rockies, with seasonal variations. Compared to the HM, hybrid models significantly improve forecast accuracy, especially in the western U.S., Midwest, and along the East Coast. Graph neural network architectures can be used to explore temporal and spatial patterns in hydrological signatures for improving streamflow forecasting (Sun et al., 2021; Sun et al., 2022). The results show that the performance of the HM is sensitive to forecast horizon length. In contrast, the hybrid models maintain consistent and superior accuracy across all forecast horizons and seasons. These findings highlight the hybrid models' adaptability and robustness to diverse hydrologic and climatic conditions across seasons and regions as well as forecast horizon lengths.

To further explore the processes related to the improvements by hybrid modeling, correlation analysis is conducted between KGE improvement and watershed characteristics including hydroclimatic variables. Forecast improvements by DL are positively correlated with gauge latitude and negatively correlated with seasonality and timing of precipitation, frequency of high precipitation days, and precipitation elasticity in summer; add to that, they are positively correlated with leaf area index in winter and baseflow index in spring. Forecast improvements by DA are negatively correlated with the frequency of high precipitation days in summer and positively correlated with baseflow index in spring. These findings provide potential guidance to improve the representation of relevant hydrologic processes for individual watersheds in the hydrologic model.

Future research will be focused on developing streamflow forecasts utilizing the hybrid models driven by real-time precipitation and temperature forecasts. Moreover, the streamflow forecasts will be utilized as inputs to predict shoaling rate in navigable channels.

CRediT authorship contribution statement

Fanzhang Zeng: Data curation, Formal analysis, Writing – review & editing. **Zhao Zhao:** Data curation, Resources. **Natalie P. Memarsadeghi:** Data curation, Formal analysis, Supervision. **Charles J. McKnight:** Conceptualization, Formal analysis, Validation. **Xudong Wang:** Validation, Writing – review & editing. **Sarah Miele:** Conceptualization, Resources, Writing – review & editing. **Mayank Chadha:** Resources, Validation, Writing – review & editing. **Dania Ammar:** Resources, Validation, Writing – review & editing. **Yichao Zeng:** Data curation, Methodology, Writing – review & editing. **Magdalena Asborno:** Investigation, Supervision, Writing – review & editing. **Kenneth Mitchell:** Formal analysis, Resources, Writing – review & editing. **Guga Gugaratshan:** Resources, Software, Writing – review & editing. **Michael D. Todd:** Methodology, Validation, Writing – review & editing. **Zhen Hu:** Data curation, Methodology, Resources. **Dingbao Wang:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the United States Army Corps of Engineers through the U.S. Army Engineer Research and Development Center Research Cooperative Agreement W912HZ24C0044. The support is gratefully acknowledged.

Appendix A. Summary of notation

Symbol	Description	Appears In
C	Soil water storage capacity at a point	Fig. 1
S_0	Average initial soil water storage	Fig. 1
C_0	Soil water storage capacity corresponding to S_0	Fig. 1
P	Precipitation	Fig. 1
D	Soil water deficit at a point	Fig. 1
R_s	Saturation excess runoff	Fig. 1

(*continued on next page*)

(*continued*)

Symbol	Description	Appears In
R_i	Infiltration excess runoff	Fig. 1
α_0	Area fraction of initial saturation	Fig. 1
α_1	Area fraction of ending saturation	Fig. 1
α_2	Area fraction with runoff generation	Fig. 1
W_1	Infiltration over the area fraction of $\alpha_1 - \alpha_0$	Fig. 1
W_2	Infiltration over the area fraction of $\alpha_2 - \alpha_1$	Fig. 1
W_3	Infiltration over the area fraction of $1 - \alpha_2$	Fig. 1
E	Actual evapotranspiration	Fig. 1
R_d	Direct runoff	Fig. 1
R_g	Groundwater recharge	Fig. 1
γ	Partitioning parameter of saturation excess between direct runoff and groundwater recharge	Fig. 1
S_d	Storage in quick storage tank	Fig. 1
S_g	Storage in slow storage tank	Fig. 1
k_d	Runoff coefficient of direct runoff	Fig. 1
k_b	Runoff coefficient of baseflow	Fig. 1
Q_d	Direct streamflow	Fig. 1
Q_b	Baseflow	Fig. 1
Q	Total streamflow	Fig. 1
$F(\bullet)$	Nonlinear mapping function that transforms recent inputs and outputs into the current output	Eq. (1)
i	Time step	Eq. (1); Fig. 3
y_i	System output	Eq. (1)
$\boldsymbol{U}_i$	Exogenous inputs	Eq. (1); Fig. 3
n	Dimension of exogenous inputs	Eq. (1)
q	Number of lags in the delayed input vector	Eq. (1)
p	Number of lags in the delayed output vector	Eq. (1)
ε	Residual of NARX model	Eq. (1)
$\boldsymbol{i}_t$	Signal for input gate	Eqs. (2), (6); Fig. 2
$\sigma(\bullet)$	Sigmoid function	Eqs. (2), (3), (4); Fig. 2
$\boldsymbol{a}_t$	Hidden state	Eqs. (2), (3), (4), (5), (7); Fig. 2
$\boldsymbol{x}_t$	Input state	Eqs. (2), (3), (4), (5); Fig. 2
$\boldsymbol{W}_i$	Weight vector for input gate	Eq. (2); Fig. 2
$\boldsymbol{b}_i$	Bias vector for input gate	Eq. (2); Fig. 2
$\boldsymbol{f}_t$	Signal for forget gate	Eqs. (3), (6); Fig. 2
$\boldsymbol{W}_f$	Weight vector for forget gate	Eq. (3); Fig. 2
$\boldsymbol{b}_f$	Bias vector for forget gate	Eq. (3); Fig. 2
$\boldsymbol{o}_t$	Signal for output gate	Eqs. (4), (7); Fig. 2
$\boldsymbol{W}_o$	Weight vector for output gate	Eq. (4); Fig. 2
$\boldsymbol{b}_o$	Bias vector for output gate	Eq. (4); Fig. 2
$\widetilde{\boldsymbol{C}}_t$	Intermediate cell state	Eqs. (5), (6); Fig. 2
$tanh(\bullet)$	Hyperbolic tangent function	Eqs. (5), (7); Fig. 2
$\boldsymbol{W}_c$	Weight vector for Intermediate cell	Eq. (5); Fig. 2
$\boldsymbol{b}_c$	Bias vector for Intermediate cell	Eq. (5); Fig. 2
$\boldsymbol{C}_t$	Cell state	Eqs. (6), (7); Fig. 2
$\bigotimes$	Element-wise product	Eqs. (6), (7); Fig. 2
$G_m(\bullet)$	Surrogate model based on the low-fidelity data from the hydrologic model	Eq. (8), (9), (13); Fig. 3
y_m	Simulated streamflow by the hydrologic model	Eq. (8)
$\widehat{y}_m$	Modeled streamflow by $G_m(\bullet)$ using y_m	Eq. (8)
y_o	Observed streamflow	Eq. (9); Fig. 3
$\widehat{\widehat{y}}_m$	Output of $G_m(\bullet)$ using y_o	Eq. (9)
δ	Difference between y_o and $\widehat{\widehat{y}}_m$	Eq. (10)
$G_\delta(\bullet)$	Surrogate model based on the discrepancy between the physics-based prediction from $G_m(\bullet)$ and the high-fidelity data from the observed streamflow	Eq. (11)
$\widehat{\delta}$	Output of $G_\delta(\bullet)$ using y_o	Eq. (11)
$\widehat{y}_o$	Modelled streamflow by delta learning	Eq. (12)

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.134477.

Data availability

Data will be made available on request.

References

Abeshu, G.W., Tian, F., Wild, T., Zhao, M., Turner, S., Chowdhury, A.F.M., Vernon, C.R., Hu, H., Zhuang, Y., Hejazi, M., Li, H.-Y., 2023. Enhancing the representation of water management in global hydrological models. Geosci. Model Dev. 16 (18), 5449–5472. https://doi.org/10.5194/gmd-16-5449-2023.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21 (10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., Pappenberger, F., 2013. Glofas-global ensemble streamflow forecasting and flood early warning. Hydrol. Earth Syst. Sci. 17 (3). https://doi.org/10.5194/hess-17-1161-2013.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large-area hydrologic modeling and assessment. I: Model development. J. Am. Water Resour. Assoc. 34 (1), 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Asborno, M., Broders, J., Mitchell, K.N., Hartman, M.A., Dunkin, L.D., 2024. Forecasting sediment accumulation in the southwest pass with machine-learning models. J. Waterw., Port, Coast. Ocean Eng. 150 (2). https://doi.org/10.1061/JWPED5.WWENG-2009.

Berghuijs, W.R., Sivapalan, M., Woods, R.A., Savenije, H.H.G., 2014. Patterns of similarity of seasonal water balances: a window into streamflow variability over a range of timescales. Water Resour. Res. 50 (7), 5638–5661. https://doi.org/10.1002/2014WR015692.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24 (1), 43–69. https://doi.org/10.1080/02626667909491834.

Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Process. 9 (3–4), 251–290. https://doi.org/10.1002/hyp.3360090305.

Chang, L.-C., Liou, J.-Y., Chang, F.-J., 2022. Spatial-temporal flood inundation nowcasts by fusing machine learning methods and principal component analysis. J. Hydrol. 612, 128086. https://doi.org/10.1016/j.jhydrol.2022.128086.

Cho, K., Kim, Y., 2022. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. J. Hydrol. 605, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297.

Coxon, G., Addor, N., Bloomfield, J.P., Freer, J., Fry, M., Hannaford, J., Howden, N.J.K., Lane, R., Lewis, M., Robinson, E.L., Wagener, T., Woods, R., 2020. CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. Earth Syst. Sci. Data Disc. 12 (4), 2459–2483. https://doi.org/10.5194/essd-12-2459-2020.

Cui, Z., Zhou, Y., Guo, S., Wang, J., Xu, C.-Y., 2022. Effective improvement of multi-step-ahead flood forecasting accuracy through encoder-decoder with an exogenous input structure. J. Hydrol. 609, 127764. https://doi.org/10.1016/j.jhydrol.2022.127764.

David, C.H., Habets, F., Maidment, D.R., Yang, Z.-L., 2011. RAPID applied to the SIM-France model. Hydrol. Process. 25 (22), 3412–3425. https://doi.org/10.1002/hyp.8070.

Devineni, N., Sankarasubramanian, A., Ghosh, S., 2008. Multimodel ensembles of streamflow forecasts: role of predictor state in developing optimal combinations. Water Resour. Res. 44 (9). https://doi.org/10.1029/2006WR005855.

Eder, G., Sivapalan, M., Nachtnebel, H.P., 2003. Modelling water balances in an Alpine catchment through exploitation of emergent properties over changing time scales. Hydrol. Process. 17 (11), 2125–2149. https://doi.org/10.1002/hyp.1325.

Fatichi, S., Vivoni, E.R., Ogden, F.L., Ivanov, V.Y., Mirus, B., Gochis, D., Downer, C.W., Camporese, M., Davison, J.H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., Tarboton, D., 2016. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. J. Hydrol. 537, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026.

Feng, D., Fang, K., Shen, C., 2020. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resour. Res. 56 (9). https://doi.org/10.1029/2019WR026793.

Fowler, K.J., Acharya, S.C., Addor, N., Chou, C., Peel, M.C., 2021. CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia. Earth Syst. Sci. Data Disc. 2021, 1–30. https://doi.org/10.5194/essd-13-3847-2021.

Ghaith, M., Siam, A., Li, Z., El-Dakhakhni, W., 2020. Hybrid hydrological data-driven approach for daily streamflow forecasting. J. Hydrol. Eng. 25 (2), 04019063. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001866.

Gharehbaghi, A., Ghasemlounia, R., Ahmadi, F., Albaji, M., 2022. Groundwater level prediction with meteorologically sensitive Gated Recurrent Unit (GRU) neural networks. J. Hydrol. 612, 128262. https://doi.org/10.1016/j.jhydrol.2022.128262.

Granata, F., Di Nunno, F., de Marinis, G., 2022. Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: a comparative study. J. Hydrol. 613, 128431. https://doi.org/10.1016/j. jhydrol.2022.128431.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Han, H., Morrison, R.R., 2022. Improved runoff forecasting performance through error predictions using a deep-learning approach. J. Hydrol. 608, 127653. https://doi.org/10.1016/j.jhydrol.2022.127653.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, Y., Abeshu, G.W., Li, H.-Y., Wang, D., Zhao, M., Wild, T., Blöschl, G., Leung, L.R., 2025. A unified scheme for modeling saturation and infiltration excess runoff. EGUsphere [preprint]. https://doi.org/10.5194/egusphere-2025-5039.

Hunt, K.M.R., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrol. Earth Syst. Sci. 26 (21), 5449–5472. https://doi.org/10.5194/hess-26-5449-2022.

Jiang, S., Zheng, Y., Wang, C., Babovic, V., 2022. Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. Water Resour. Res. 58, e2021WR030185. https://doi.org/10.1029/2021WR030185.

Kao, I.F., Zhou, Y., Chang, L.-C., Chang, F.-J., 2020. Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting. J. Hydrol. 583. https://doi.org/10.1016/j.jhydrol.2020.124631.

Kisi, O., Cimen, M., 2011. A wavelet-support vector machine conjunction model for monthly stream-flow forecasting. J. Hydrol. 399, 132–140. https://doi.org/10.1016/j.jhydrol.2010.12.041.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing nash-sutcliffe and kling-gupta efficiency scores. Hydrol. Earth Syst. Sci. 23, 4323–4331.

Kollat, J.B., Reed, P.M., Wagener, T., 2012. When are multiobjective calibration trade-offs in hydrologic models meaningful? Water Resour. Res. 48, W03520. https://doi.org/10.1029/2011WR011534.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23 (12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Ladson, A.R., Brown, R., Neal, B., Nathan, R., 2013. A standard approachto baseflow separation using the Lyne and Hollick filter. Australian J. Water Resour. 17 (1), 25–34. https://doi.org/10.7158/W12-028.2013.17.1.

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. Atmos. 99 (D7), 14415–14428. https://doi.org/10.1029/94JD00483.

Li, B., Sun, T., Tian, F.Q., Ni, G.H., 2023. Enhancing process-based hydrological models with embedded neural networks: a hybrid approach. J. Hydrol. 590, 125206. https://doi.org/10.1016/j.jhydrol.2023.130107.

Li, H., Zhang, C., Chu, W., Shen, D., Li, R., 2024. A process-driven deep learning hydrological model for daily rainfall-runoff simulation. J. Hydrol. 637, 131434. https://doi.org/10.1016/j.jhydrol.2024.131434.

Li, S., Xie, Q., Yang, J., 2022. Daily suspended sediment forecast by an integrated dynamic neural network. J. Hydrol. 604, 127258. https://doi.org/10.1016/j.jhydrol.2021.127258.

Liu, Q., Dai, H., Gui, D.W., Hu, B., Ye, M., Wei, G., Qin, J., Zhang, J., 2022. Evaluation and optimization of the water diversion system of ecohydrological restoration megaproject of Tarim River, China, through wavelet analysis and a neural network. J. Hydrol. 608, 127586. https://doi.org/10.1016/j.jhydrol.2022.127586.

Luo, C., Ding, W., Zhang, C., Yang, X., 2023. Exploiting multiple hydrologic forecasts to inform real-time reservoir operation for drought mitigation. J. Hydrol. 618, 129232. https://doi.org/10.1016/j.jhydrol.2023.129232.

Luo, L., Wood, E.F., Pan, M., 2007. Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. J. Geophys. Res. Atmos. 112 (D10). https://doi.org/10.1029/2006JD007655.

Luo, S., Tetzlaff, D., Smith, A., Soulsby, C., 2024. Long-term drought effects on landscape water storage and recovery under contrasting landuses. J. Hydrol. 636, 131339. https://doi.org/10.1016/j.jhydrol.2024.131339.

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., Shen, C., 2021. Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. Water Resour. Res. 57, e2020WR028600. https://doi.org/10.1029/2020WR028600.

Mao, D., Cherkauer, K.A., 2009. Impacts of land-use change on hydrologic responses in the Great Lakes region. J. Hydrol. 374 (1–2), 71–82. https://doi.org/10.1016/j.jhydrol.2009.06.016.

Martinez, G.F., Gupta, H.V., 2010. Toward improved identification of hydrological models: a diagnosticevaluation of the “abcd” monthly water balance model for the conterminous United States. Water Resour. Res. 46, W08507. https://doi.org/10.1029/2009WR008294.

Ma, X., Hu, H., Ren, Y., 2023. A hybrid deep learning model based on feature capture of water level influencing factors and prediction error correction for water level prediction of cascade hydropower stations under multiple time scales. J. Hydrol. 617. https://doi.org/10.1016/j.jhydrol.2022.129044.

McDonnell, J.J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M.L., Selker, J., Weiler, M., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43, W07301. https://doi.org/10.1029/2006WR005467.

Mohanty, A., Sahoo, B., Kale, R.V., 2024. A hybrid model enhancing streamflow forecasts in paddy land use-dominated catchments with numerical weather prediction model-based meteorological forcings. J. Hydrol. 635, 131225. https://doi.org/10.1016/j.jhydrol.2024.131225.

Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrol. Sci. J. 30 (2), 273–297. https://doi.org/10.1080/02626668509490989.

Nanda, T., Sahoo, B., Beria, H., Chatterjee, C., 2016. A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite-based rainfall products. J. Hydrol. 539, 57–73. https://doi.org/10.1016/j.jhydrol.2016.05.014.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Philos. Phenomenol. Res. 57 (3), e2020WR028091. https://doi.org/10.1029/2020WR028091.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19 (1), 209–223. https://doi.org/10.5194/hess-19-209-2015.

Ng, K.W., Huang, Y.F., Koo, C.H., Chong, K.L., El-Shafie, A., Najah Ahmed, A., 2023. A review of hybrid deep learning applications for streamflow forecasting. J. Hydrol. 625 (2023), 130141. https://doi.org/10.1016/j.jhydrol.2023.130141.

Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., 2020. Streamflow and rainfall forecasting by two long short-term memory-based models. J. Hydrol. 583. https://doi.org/10.1016/j.jhydrol.2019.124296.

Nourani, V., Baghanam, A.H., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet-artificial intelligence models in hydrology: a review. J. Hydrol. 358–377. https://doi.org/10.1016/j.jhydrol.2014.03.057.

Piadeh, F., Behzadian, K., Alani, A.M., 2022. A critical review of real-time modelling of flood forecasting in urban drainage systems. J. Hydrol. 607, 127476. https://doi.org/10.1016/j.jhydrol.2022.127476.
Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. Phys 1 (5), 318–333. https://doi.org/10.1063/1.1745010.
Robertson, D.E., Wang, Q.J., 2012. A Bayesian approach to predictor selection for seasonal streamflow forecasting. J. Hydrometeor. 13, 155–171. https://doi.org/10.1175/JHM-D-10-05009.1.
Sankarasubramanian, A., Vogel, R.M., Limbrunner, J.F., 2001. Climate elasticity of streamflow in the United States. Water Resour. Res. 37 (6), 1771–1781. https://doi.org/10.1029/2000WR900330.
Schepen, A., Zhao, T., Wang, Q.J., Zhou, S., Feikema, P., 2016. Optimising seasonal streamflow forecast lead time for operational decision making in Australia. Hydrol. Earth Syst. Sci. 20, 4117–4128. https://doi.org/10.5194/hess-20-4117-2016.
SCS, 1972. Section 4: Hydrology. National engineering handbook, Soil Conservation Service, US Department of Agriculture. Retrieved from https://archive.org/details/CAT71334647003/page/n3/mode/2up.
Shamseldin, A.Y., O'Connor, K.M., 2001. A non-linear neural network technique for updating of river flow forecasts. Hydrol. Earth Syst. Sci. 5 (4), 577–597. https://doi.org/10.5194/hess-5-577-2001.
Shukla, S., Voisin, N., Lettenmaier, D.P., 2012. Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill. Hydrol. Earth Syst. Sci. 16 (8), 2825–2838. https://doi.org/10.5194/hess-16-2825-2012.
Sun, A.Y., Jiang, P., Mudunuru, M.K., Chen, X., 2021. Explore spatio-temporal learning of large sample hydrology using graph neural networks. Water Resour. Res. 57 (12). https://doi.org/10.1029/2021wr030394.
Sun, A.Y., Jiang, P., Yang, Z.-L., Xie, Y., Chen, X., 2022. A graph neural network (GNN) approach to basin-scale river network learning: the role of physics-based connectivity and data fusion. Hydrol. Earth Syst. Sci. 26 (19), 5163–5184. https://doi.org/10.5194/hess-26-5163-2022.
Takens, F., 1981. Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young (ed.). Dyn. Syst. Turbul., Lect. Notes Math. 898, Springer-Verlag. pp. 366–381.
Tripathy, K.P., Mishra, A.K., 2024. Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions. J. Hydrol. 628, 130458. https://doi.org/10.1016/j.jhydrol.2023.130458.
Troch, P.A., Carrillo, G., Sivapalan, M., Wagener, T., Sawicz, K., 2013. Climate–vegetation–soil interactions and long-term hydrologic partitioning: Signatures of catchment co-evolution. Hydrol. Earth Syst. Sci. 17 (6), 2209–2217. https://doi.org/10.5194/hess-17-2209-2013.
Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.L., 2021. Generating ensemble streamflow forecasts: a review of methods and approaches over the past 40 years. Water Resour. Res. 57 (7). https://doi.org/10.1029/2020wr028392.
Wang, D., 2018. A new probability density function for spatial distribution of soil water storage capacity leads to the SCS curve number method. Hydrol. Earth Syst. Sci. 22 (12), 6567–6578. https://doi.org/10.5194/hess-22-6567-2018.
Wang, D., Tang, Y., 2014. A one-parameter Budyko model for water balance captures emergent behavior in Darwinian hydrologic models. Geophys. Res. Lett. 41, 4569–4577. https://doi.org/10.1002/2014GL060509.
Wang, H., Qin, H., Liu, G., Liu, S., Qu, Y., Wang, K., Zhou, J., 2023. A novel feature attention mechanism for improving the accuracy and robustness of runoff forecasting. J. Hydrol. 618. https://doi.org/10.1016/j.jhydrol.2023.129200.
Wunsch, A., Liesch, T., Broda, S., 2021. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrol. Earth Syst. Sci. 25, 1671–1687. https://doi.org/10.5194/HESS-25-1671-2021.
Wunsch, A., Liesch, T., Cinkus, G., Ravbar, N., Chen, Z., Mazzilli, N., Jourde, H., Goldscheider, N., 2022. Karst spring discharge modeling based on deep learning using spatially distributed input data. Hydrol. Earth Syst. Sci. 26 (9), 2405–2430. https://doi.org/10.5194/hess-26-2405-2022.
Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J. Hydrol. 603, 127043. https://doi.org/10.1016/j.jhydrol.2021.127043.
Xu, C., Zhong, P., Zhu, F., Xu, B., Wang, Y., Yang, L., Wang, S., Xu, S., 2024. A hybrid model coupling process-driven and data-driven models for improved real-time flood forecasting. J. Hydrol. 638. https://doi.org/10.1016/j. jhydrol.2024.131494.
Xu, W., Chen, J., Corzo, G., 2025. Combining data augmentation and hybrid modeling approaches for deep learning-based monthly streamflow forecasting. J. Hydrol. 659, 133318. https://doi.org/10.1016/j.jhydrol.2025.133318.
Xu, W., Chen, J., Zhang, X.J., Xiong, L., Chen, H., 2022. A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model. J. Hydrol. 614. https://doi.org/10.1016/j.jhydrol.2022.128599.
Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., Zhao, B., 2020. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. J. Hydrol. 590 (March), 125206. https://doi.org/10.1016/j.jhydrol.2020.125206.
Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., Liu, C., 2019. Evaluation and machine learning improvement of global hydrological model-based flood simulations. Environ. Res. Lett. 14 (11), ab4d5e. https://doi.org/10.1088/1748-9326/.
Yao, L., Libera, D.A., Kheimi, M., Sankarasubramanian, A., Wang, D., 2020. The roles of climate forcing and its variability on streamflow at daily, monthly, annual, and long-term scales. Water Resour. Res. 56, e2020WR027111. https://doi.org/10.1029/2020WR027111.
Yao, L., Wang, D., 2022. Hydrological basis of different Budyko equations: the spatial variability of available water for evaporation. Water Resour. Res. 58, e2021WR030921. https://doi.org/10.1029/2021WR030921.
Ye, S., Yaeger, M., Coopersmith, E., Cheng, L., Sivapalan, M., 2012. Exploring the physical controls of regional patterns of flow duration curves-part 2: role of seasonality, the regime curve, and associated process controls. Hydrol. Earth Syst. Sci. 16 (11), 4447–4465. https://doi:hess-16-4447-2012.
Yu, Q., Jiang, L., Schneider, R., Zheng, Y., Liu, J., 2024. Deciphering the mechanism of better predictions of regional LSTM models in ungauged basins. Water Resour. Res. 60 (7), e2023WR035876. https://doi.org/10.1029/2023WR035876.
Yu, Q., Jiang, L., Wang, Y., Liu, J., 2023. Enhancing streamflow simulation using hybridized machine learning models in a semi-arid basin of the Chinese loess plateau. J. Hydrol. 617, 129115. https://doi.org/10.1016/j.jhydrol.2023.129115.
Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short-term streamflow forecasting using artificial neural networks. J. Hydrol. 214, 32–48. https://doi.org/10.1016/S0022-1694(98)00242-X.
Zhang, W., Liu, P., Wang, H., et al., 2017. Reservoir adaptive operating rules based on both of historical streamflow and future projections. J. Hydrol. 553, 691–707. https://doi.org/10.1016/j.jhydrol.2017.08.031.
Zhao, R.-J., 1992. The Xinanjiang model applied in China. J. Hydrol. 135 (1–4), 371–381. https://doi.org/10.1016/0022-1694(92)90096-E.
Zhao, T., Cai, X., Yang, D., 2011. Effect of streamflow forecast uncertainty on real-time reservoir operation. Adv. Water Resour. 34 (4), 495–504. https://doi.org/10.1016/j.advwatres.2011.01.004.
Zhao, X., Lv, H., Lv, S., Sang, Y., Wei, Y., Zhu, X., 2021. Enhancing robustness of monthly streamflow forecasting model using gated recurrent unit based on improved grey wolf optimizer. J. Hydrol. 01, 126607. https://doi.org/10.1016/j.jhydrol.2021.126607.
Zhao, Y., Chadha, M., Barthlow, D., Yeates, E., McKnight, C.J., Memarsadeghi, N.P., 815, Gugaratshan, G., Todd, M.D., Hu, Z., 2024. Physics-enhanced machine 816 learning models for streamflow discharge forecasting. J. Hydroinform. 26 (10), 2506–2537. https://doi.org/10.2166/hydro.2024.061.
Zuo, G., Luo, J., Wang, N., Lian, Y., He, X., 2020. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. J. Hydrol. 585. https://doi.org/10.1016/j.jhydrol.2020.124776.