Hydrol. Earth Syst. Sci., 30, 5521–5549, 2026
https://doi.org/10.5194/hess-30-5521-2026

Hydrology and Earth System Sciences
Open Access

Integrating Physical-Based Xinanjiang Model and Deep Learning for Interpretable Streamflow Simulation: A Multi-Source Data Fusion Approach across Diverse Chinese Basins

Zhaocai Wang[1], **Nannan Xu**[1], **Wei Song**[1], **Xingxing Zhang**[2], **Junhao Wu**[3], **and Xi Chen**[3]

[1]College of Marine Ecology and Environment, Shanghai Ocean University, Shanghai, 201306, P. R. China
[2]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, P. R. China
[3]State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai, 200062, P. R. China

Correspondence: Wei Song (wsong@shou.edu.cn) and Xi Chen (xchen@geo.ecnu.edu.cn)

Received: 20 May 2025 – Discussion started: 24 June 2025
Revised: 6 February 2026 – Accepted: 11 April 2026 – Published: 1 September 2026

Abstract. The simulation of streamflow is a complex task due to its intricate formation process. Existing single models struggle to accurately capture the stochastic, non-stationary, and nonlinear dynamics of basin streamflow in changing environments. This study combined process-driven hydrological mechanism models with data-driven deep learning models, considering various factors like hydrology, meteorology, environment, and the interconnected effects of upstream and downstream rivers, to develop an interpretable hybrid streamflow simulation model. The study collected multiple external variables to better understand the hydrologic system complexity and used the Maximum Information Coefficient (MIC) to analyze their relationship with streamflow. Subsequently, the Xinanjiang (XAJ) model with physical mechanisms was employed, alongside the TCN-GRU model integrating Temporal Convolutional Network (TCN) and Gated Recurrent Unit (GRU), for separate streamflow simulations. Furthermore, a robust integration method was adopted, realizing nonlinear ensemble through Random Forest (RF), thus establishing the hybrid XAJ-TCN-GRU model. This model exhibits promising results in simulating streamflow in four different basins in China, achieving high Nash-Sutcliffe Efficiency (NSE) values with 0.991, 0.971, 0.984, and 0.986 for the Wuding River, Chu River, Jianxi River, and Qingyi River respectively. In terms of streamflow simulation, flood simulating, and interval simulation, this model outperforms other benchmark models. Additionally, the study quantified the contributions of each hydro-meteorological variable to the long-term streamflow trend using mean absolute SHAP values (SHAPABS), Feature Importance (FI), and Permutation Feature Importance (PFI), thereby enhancing the model's external interpretability. The results of this study are of significant importance for optimizing water resource management and mitigating flood disasters.

1 Introduction

Water resources are among the most precious natural resources on Earth, and accurate streamflow simulation has become an indispensable component in hydraulic engineering and water resources management (Zhang et al., 2023). Rational and efficient streamflow simulation not only provides solid support for the sustainable development of society and the economy but also has profound implications in flood warning and optimization of water resource scheduling (Shao et al., 2024; He et al., 2026). However, the complex nature of river hydrological environments renders the formation and dynamics of streamflow highly intricate processes (Liu et al., 2023; Ahmed et al., 2022). Therefore, exploring high-precision and reliable streamflow simulation models has become an urgent and practical issue in the field of hydrology (Thébault et al., 2024; Wang et al., 2025b; Ding et al., 2026).

Published by Copernicus Publications on behalf of the European Geosciences Union.

To date, methodologies for streamflow simulation are generally classified to physical-based (Gebremariam et al., 2014; Bai et al., 2017), data-driven approaches (Gao et al., 2020; Vilaseca et al., 2023) and hybrid models (Zhu et al., 2023). Physical-based models represent conventional simulation methodologies that describe the physical mechanisms governing the transformation from rainfall to streamflow (Cheng et al., 2020; Zhu et al., 2026a). These approaches utilize mathematical equations and calibrate parameters for diverse locations to effectively simulate streamflow dynamics (Özgen-Xian et al., 2020; Leonarduzzi et al., 2021). Typical physical-based models include Precipitation Runoff Modeling System (PRMS) (Hwang et al., 2011), Soil and Water Assessment Tool (SWAT) (Oruc et al., 2023), and Xinanjiang (XAJ) model (Hao et al., 2015; Jiang et al., 2023). The XAJ model is used for simulating rainfall-runoff processes, which is proposed by Zhao (1992) in China. The model is mainly used in humid and semi-humid regions, and is particularly suitable for flood simulation and water resource management (Lei et al., 2023). Nevertheless, the complexity and interrelations of hydrological systems inevitably introduce uncertainties, particularly concerning model parameter estimation and structure selection (Zuo et al., 2020; Tan et al., 2025).

Machine learning models exhibit flexibility in capturing and modeling nonlinear relationships, thus serving as effective tools for streamflow simulation (Zhu et al., 2026b). Additionally, the structure and complexity of machine learning models can be adjusted according to specific research needs, allowing them to better adapt to streamflow simulation tasks at different scales and spatiotemporal ranges (Han and Morrison, 2022). Among these, Artificial Neural Network (ANN) is one of the earliest data-driven models applied in hydrological simulation, with its ability to approximate complex nonlinear functions laying the foundation for subsequent machine learning applications in streamflow modeling (Wang et al., 2024; Noori and Kalin, 2016). Other classical machine learning methods encompass various models, such as Random Forest (RF) (Contreras et al., 2021), Decision Tree (DT) (Jehanzaib et al., 2021), and Support Vector Machine (SVM) (Samantaray et al., 2022). When addressing complex problems, machine learning models often increase parameters, complexity of structure, or introduce more features to enhance fitting ability (Yao et al., 2023). However, overly complex models may overfit the details and noise in the training data, raising the risk of overfitting.

The field of machine learning continues to evolve, and deep learning, as an important extension of it, has attracted considerable attention. Its core concept is to enable computers to learn and understand more complex abstract concepts by leveraging layered, stacked deep architectures (Chen et al., 2023; Ng et al., 2023). Deep learning models demonstrate strong generalization capabilities in streamflow simulation across different basins and periods, adapting well to diverse hydrological conditions (Xu et al., 2023; Wei et al., 2023). The Temporal Convolutional Network (TCN) significantly enhances the accuracy of streamflow simulation by effectively capturing long-term temporal dependencies within the data (Lin et al., 2020). Unlike traditional methods, TCN's capacity to process sequential data through causal convolutions enables it to model both short- and long-term patterns in streamflow, which is essential for accurately capturing the dynamics of runoff processes. Additionally, its use of residual connections promotes faster and more stable convergence during training, thereby reducing the risk of vanishing gradients. This combination of strengths directly enhances the model's simulation capability, establishing TCN as an effective tool for more reliable and precise streamflow forecasts. The Gated Recurrent Unit (GRU) also plays a crucial role in streamflow simulation by adeptly capturing the sequential dependencies inherent in hydrological data, which is vital for modeling the variability of water flow over time (Zheng et al., 2023). Through its gating mechanism, GRU selectively updates and forgets information, allowing it to retain essential patterns while filtering out noise, thereby improving its accuracy in simulating fluctuating streamflow patterns. This ability to manage time-dependent features positions GRU as well-suited for capturing sudden changes and longer-term dependencies in streamflow data. Deep learning models are often considered "black boxes", with their internal parameters and decision processes challenging to interpret, which presents difficulties in understanding and validating whether the model accurately captures physical mechanisms (Katipoğlu and Sarıgöl, 2023). While deep learning models excel at learning patterns from large datasets, they lack explicit modeling of physical equations and laws. In scenarios where models must make reasonable simulations based on physical mechanisms, this limitation can lead to suboptimal performance.

In response to the limitations of single physical-based or data-driven models, hybrid models have emerged as a core direction in contemporary hydrological simulation. These models aim to integrate the strengths of both paradigms: leveraging the explicit physical interpretability of physical-based models to avoid over-reliance on observational data, and utilizing the powerful nonlinear feature learning capability of data-driven models to compensate for the uncertainty of physical parameter calibration (Mohanty et al., 2024; Acuña Espinoza et al., 2025a). Kim et al. (2021) pointed out that data-driven models, ANN and Long Short-Term Memory (LSTM), perform better in predicting peak flow conditions, whereas physical-based hydrological models perform better in low flow conditions. Since deep learning models have high predictive accuracy and physical models based on hydrological principles are explainable, some researchers choose to combine them to enhance both accuracy and insight in streamflow prediction (Kurian et al., 2020). Most studies use deep learning models to post-process the outputs of physical-based models (Cho and Kim, 2022; Parisouj et al., 2022). This approach provides relatively reasonable initial information, using physical-based models to generate inputs with physical significance, which helps guide the deep learning model (Granata et al., 2024). However, this sequential coupling is prone to error propagation: for example, Cho and Kim (2022) found that in the sequential coupling of the WRF-Hydro model and LSTM, the precipitation simulation error of the physical model was propagated to the LSTM streamflow prediction, leading to an increase in streamflow RMSE by 15 %–20 % compared to the physical model alone. Similarly, Parisouj et al. (2022) reported that in their physics-informed data-driven model, the streamflow simulation error of the physical model was amplified through sequential coupling, resulting in a further decrease in Nash-Sutcliffe Efficiency (NSE) of 0.10–0.12 in the final prediction. These results indicate that the initial errors from physical models can be magnified in sequential coupling, ultimately affecting the reliability of simulation results.

In recent years, hybrid modelling approaches have emerged as a significant direction for overcoming the limitations of single models, particularly through the integration of process models with deep learning frameworks. Jiang et al. (2020) proposed a Physical process-wrapped Recurrent Neural Network (PRNN) architecture, which embeds a conceptual hydrological model (EXP-HYDRO) as a dedicated recurrent layer into deep learning, enabling the model to inherit physical consistency while enhancing transferability across ungauged basins and inferring unobserved processes (e.g., snow accumulation dynamics). Ni et al. (2025) further developed the Physics-informed Deep Neural Network (P-DNN) for monthly streamflow prediction, designing physical-aware modules inspired by the XAJ model's streamflow generation structure and incorporating mass conservation into the loss function, which effectively improved the model's performance in simulating high flows and reduced physically unreasonable outputs. In the context of flood forecasting, Luo et al. (2024) fused LSTM with the Unscented Kalman Filter (UKF) to address the challenge of noise estimation in physical model state updates: the LSTM adaptively learns noise-related information to estimate Kalman gain, enhancing the accuracy and stability of XAJ model state correction for multi-step flood forecasts. Additionally, Acuña Espinoza et al. (2025b) systematically evaluated the generalization capability of hybrid models (LSTM-parameterized HBV) against LSTM and standalone conceptual models in extreme hydrological events, revealing that hybrid models outperform LSTM in simulating high return period flows due to their physical structural constraints, though they still face limitations in arid basins where runoff generation mechanisms deviate from conceptual model assumptions.

Hybrid modelling approaches currently fall into three main categories: First, physics-guided machine learning (Physics-guided machine learning), which integrates physical laws (such as mass conservation and energy balance) into deep learning architectures (e.g., loss functions or network layer design) to fuse physical mechanisms with data patterns (Willard et al., 2023). The core of this approach is "mechanism embedding", but it requires customizing the model structure for specific physical scenarios, limiting its generalizability. Second, sequential coupling methods, which use the output of physical models as input for post-processing in deep learning models (Cho and Kim, 2022; Parisouj et al., 2022). While this approach can leverage the initial physical meaning of the physical model, it carries the risk of error accumulation–structural biases in the physical model may be amplified by the deep learning model. Third, parallel integration methods, where a few studies have attempted to run the two types of models in parallel, but most adopt linear weighted fusion, which struggles to capture nonlinear correlations between model outputs. Beyond hybrid models, ensemble modeling has also proven effective in improving streamflow prediction. Solanki et al. (2025) proposed a multi-model ensemble framework by applying RF and XGB to post-process outputs from five physical hydrological models (VIC, H08, CWatM, Noah-MP, CLM), achieving NSE values of 0.70–0.84 across the Narmada River Basin.

While many studies have made significant progress in integrating physical knowledge with deep learning, most adopt serial or auxiliary integration strategies: for example, using physical models to guide the input of data-driven models, applying physical constraints to correct model outputs, or combining filtering methods with deep learning to optimize physical model parameters (Luo et al., 2024). These approaches often rely on the prior validity of physical model outputs or simplify the nonlinear interactions between physical mechanisms and data patterns. In contrast, there remains a need for a more flexible integration framework that can fully leverage the complementary strengths of physical-based and deep learning models while avoiding the propagation of errors from one model to the other.

To fill this gap, this study presents a robust hybrid integration framework that combines the process-driven XAJ model with the data-driven TCN-GRU model (combining Temporal Convolutional Network and Gated Recurrent Unit) through a nonlinear ensemble strategy. Unlike existing serial integration methods, the XAJ and TCN-GRU models operate in parallel to independently simulate streamflow, with their outputs fused using RF to capture complex nonlinear relationships between the two models' results. This integration strategy, while leveraging well-established ensemble principles, demonstrates remarkable effectiveness and robustness in practical hydrological simulation scenarios, especially in flood prediction across diverse basins, which constitutes the key contribution of this study. Additionally, we employ the Maximum Information Coefficient (MIC) and Variance Inflation Factor (VIF) to select key hydrometeorological variables from multi-source data (e.g., precipitation, temperature, humidity), enhancing the model's robustness and interpretability. It is worth noting that the XAJ-TCN-GRU model employs an innovative strategy of "parallel execution and nonlinear integration": it neither alters the internal structure of the physical model (XAJ) nor the deep learning model (TCN-

GRU), while simultaneously performing non-linear fusion of their outputs via RF. This design retains the physical mechanism interpretability of XAJ and the complex pattern capture capability of TCN-GRU, while avoiding the error propagation issues of sequential coupling and the adaptation limitations of linear integration, thereby establishing unique advantages in terms of generalization and nonlinear adaptation. Additionally, various interpretability methods were employed to quantitatively analyze the impact of meteorological and hydrological factors on the hydrological processes. To validate the hybrid model's generalization across varied hydrological conditions, it was applied to four basins in China with different hydrological characteristics for daily streamflow simulation. Furthermore, the robustness of the model is confirmed by simulation tests with the addition of "noise" data.

Moreover, whilst certain single-layer deep learning models (such as LSTM and GRU) have achieved high-precision simulations under conventional hydrological conditions, the core challenge in hydrological modelling lies not merely in attaining "high precision under normal conditions". Rather, it demands stability during extreme events (such as floods and droughts), robustness amidst data noise interference, and consistency between simulation outcomes and hydrological physical mechanisms. A singular model cannot satisfy all requirements: physical models (such as XAJ) are constrained by structural assumptions and cannot capture complex nonlinear relationships; purely data-driven models (such as GRU) lack physical constraints, are prone to bias under changing data distributions or extreme scenarios, and struggle to explain their simulation logic. Therefore, the XAJ-TCN-GRU hybrid model developed in this study does not introduce complexity merely to marginally improve NSE. Instead, it resolves the aforementioned multidimensional challenges through the deep integration of physical mechanisms and data-driven methods. This objective holds irreplaceable value in practical water resource management (e.g., flood warning systems requiring a balance between accuracy and reliability).

2 Methodologies

This section outlines the methods used in the study. Specific descriptions of calculation of MIC, noise data injection test, flood simulating, and interval simulation can be found in Sect. S2 in the Supplement.

2.1 Xinanjiang model (XAJ)

The XAJ model employs a three-source approach for streamflow calculation, utilizing surface streamflow routing through the unit hydrograph method, while separate linear storage reservoirs simulate the routing processes of interflow and baseflow. Its main features include the application of storage excess concepts and Muskingum routing, characterized by subdivision into subunits, multiple sources, and routing stages. Additionally, the XAJ model possesses a simple structure, with parameters that have clear physical meanings and high computational accuracy, which has led to its widespread application in hydrological forecasting (Gong et al., 2021).

The XAJ model encompasses a substantial number of parameters, some of which are highly sensitive; even minor alterations can significantly affect the results, while others display a degree of inertia. The sensitivity classification of the parameters in Table 1 is a general summary adopted from previous studies on XAJ model applications across various basins in China. These studies have systematically evaluated parameter sensitivity using comprehensive methods commonly employed in hydrological modeling, including local sensitivity analysis (e.g., one-at-a-time method) and global sensitivity analysis (e.g., Sobol' method and Morris method). Specifically, parameters are classified as "sensitive" if consistent findings from existing literature indicate that a small relative change in their values leads to a significant change (e.g., exceeding a predefined threshold in terms of NSE reduction or RMSE increase) in streamflow simulation results. For example, Gong et al. (2021) conducted sensitivity analysis on XAJ model parameters in small-and medium-sized catchments in South China and identified parameters such as B, WM, and EX as highly sensitive. Similarly, Lei et al. (2023) confirmed the sensitivity of parameters like KC and KG through runoff simulation studies in humid regions. Table 1 presents detailed descriptions of the meanings and sensitivities of 15 parameters, which can be broadly classified four categories based on the model's structure and function: evapotranspiration parameters, flow rate parameters, water source division parameters, and confluence parameters.

2.2 Temporal convolutional network (TCN)

TCN represents an architectural framework that combines one-dimensional convolutional networks with causal convolutions, efficiently capturing temporal dependencies in data and demonstrating enhanced suitability for addressing time-series challenges (Sun et al., 2024). Dilated convolutional networks enable dilated sampling on the input from the preceding layer, allowing for the extraction of feature information from time-series data characterized by long intervals and non-continuous patterns. In the context of causal convolution, the output at any given moment is solely dependent on the input at that moment and any preceding inputs. TCN's core architecture is composed of dilated causal convolution.

The process by which TCNs handle data sequences is analogous to that of dilated convolution. TCNs typically incorporate a greater number of convolutional layers compared to conventional CNNs, continuously adjusting the number of layers. This results in progressively larger dilation factors and convolutional kernels at each layer. However, as the

Table 1. Classification of XAJ model parameters and overview of key attributes.

Typology	Parameter	Description	Sensitivity
Evapotranspiration parameters	C	The coefficient of deep evapotranspiration	non-sensitive
	KC	Ratio of potential evapotranspiration to pan evaporation	sensitive
	WLM (mm)	Averaged soil moisture storage capacity of the lower layer	non-sensitive
	WUM (mm)	Averaged soil moisture storage capacity of the upper layer	non-sensitive
Flow rate parameters	B	Exponent of the tension water capacity curve	sensitive
	WM (mm)	Areal mean tension water capacity	sensitive
	IM	Percentage of impervious and saturated areas in the catchment	non-sensitive
Water source division parameters	EX	Exponent of the free water capacity curve	sensitive
	KG	Outflow coefficients of the free water storage to groundwater	sensitive
	KI	Outflow coefficients of the free water storage to interflow	sensitive
	SM (mm)	Areal mean of the free water capacity of the surface soil layer	sensitive
Confluence parameters	k (h)	Storage coefficient of linear reservoirs of Nash unit hydrograph	sensitive
	n	Number of linear reservoirs of Nash unit hydrograph	sensitive
	CG	Recession constants of the groundwater storage	non-sensitive
	CI	Recession constants of the lower interflow storage	non-sensitive

depth of the network increases, the training process becomes increasingly challenging, with the multi-layer backpropagation of error signals potentially leading to issues such as "gradient vanishing" or "gradient explosion". Residual networks have been shown to effectively mitigate these challenges.

2.3 Gated recurrent unit (GRU)

As a variant of recurrent neural networks (RNNs), GRU effectively solves the long-term memory storage problems and gradient disappearance issues faced by traditional RNNs during backpropagation (Song et al., 2024). It employs gating mechanisms, comprising an update gate and a reset gate, to govern information flow within the network. The update gate modulates the integration of historical and new information, whereas the reset gate identifies components of past hidden states irrelevant to current computations. Through adaptive memory update and reset mechanisms, GRU effectively models long-term dependencies in sequences.

GRU's computation process can be represented by the following equations:

$$r_t = \sigma\left(w_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{1}$$

$$z_t = \sigma\left(w_z \cdot \left[h_{t-1}, x_t\right]\right) \tag{2}$$

$$H_t = \tanh\left(w_H \cdot \left[r_t \cdot h_{t-1}, x_t\right]\right) \tag{3}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot H_t \tag{4}$$

where r_t represents the reset gate, z_t denotes the update gate, H_t signifies the intermediate state, and h_t corresponds to the GRU output. The weight matrices w_z, w_r and w_H are associated with the update gate, reset gate, and intermediate state, respectively. Here, x_t indicates the current state, h_{t-1} represents the previous hidden state, σ stands for the *sigmoid* activation function, and tanh refers to the hyperbolic tangent function.

2.4 TCN-GRU model

Figure 1 depicts the architecture of the TCN-GRU model, which consists of input layer, TCN layer, GRU layer, and output layer. Historical streamflow data and highly correlated influencing factors are integrated as input to the TCN layer, which excels in feature extraction, fast convergence, and robustness. The TCN layer processes input data through a stack of dilated causal convolution modules (each containing a 1D convolution with kernel size 3, dilation factor doubling per layer, and residual connections) to generate high-dimensional temporal feature maps. Specifically, for input data with shape (batch_size, time_steps, input_features), the TCN layer outputs feature maps with shape (batch_size, time_steps, tcn_features), where tcn_features is determined by the number of filters in the final TCN convolution layer (set to 64 in this study based on hyperparameter optimization).

To match the feature dimensions with the subsequent GRU layer, a linear projection (via a fully connected layer with no activation) is applied to the TCN output, transforming the feature dimension from tcn_features to gru_units. This ensures the TCN output (shape: (batch_size, time_steps, gru_units)) directly serves as the input to the GRU layer, which is designed to accept sequences with feature dimension equal to gru_units.

The GRU layer, with hidden size gru_units, processes the temporal features by updating its internal state at each time step, capturing sequential dependencies. Potential bottlenecks (e.g., information loss during dimension conversion) are mitigated by: (1) keeping tcn_features ≥ gru_units

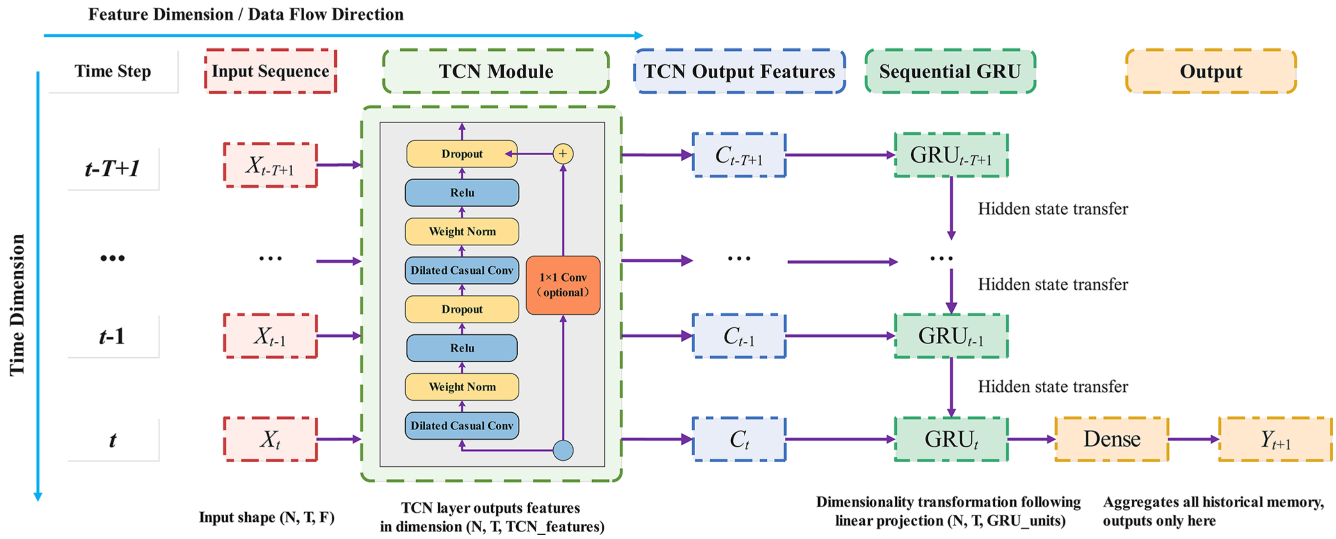

Figure 1. Schematic diagram of the TCN-GRU hybrid deep learning model architecture (Input layer→TCN layer→GRU layer→Output layer).

to avoid downsampling-related information compression; (2) adding a skip connection from the TCN output to the GRU input, which retains raw TCN features alongside the projected features; and (3) using batch normalization in the TCN layer to stabilize feature distribution during propagation.

Thus, the GRU layer effectively learns dynamic variations from the TCN-extracted features, capturing temporal correlations among multiple features to enhance simulation accuracy. The output layer generates the final simulated values. The configuration of model hyperparameters, such as the TCN dilation factors, GRU hidden units, is detailed in Sect. S2.1 in the Supplement.

2.5 Random forest (RF)

RF is an ensemble learning approach that boosts the precision of classification and regression tasks by building and combining the prediction results of multiple decision trees (Qiao et al., 2023). The core principle of this algorithm involves utilizing the bootstrap resampling technique to randomly sample multiple subsets from the original dataset, on which decision trees are constructed for each subset. During the construction of these decision trees, features are randomly selected for splitting, thereby enhancing the model's diversity and generalization capability (Doyle et al., 2023). Ultimately, RF combines the simulation results of the multiple trees through voting or averaging to derive the final classification or regression outcome. Utilizing RF for nonlinear ensemble learning leverages the advantages of its ensemble algorithm, thereby improving simulation accuracy, reducing the risk of overfitting, and effectively managing nonlinear and high-dimensional data (Alnahit et al., 2022; Wu et al., 2023; Wang et al., 2025a). The configuration of RF tree depth is detailed in Sect. S2.2 in the Supplement.

2.6 XAJ-TCN-GRU hybrid model

A robust hybrid integration approach, namely the XAJ-TCN-GRU model, is proposed to organically integrate the conceptual rainfall–runoff model and deep learning model for streamflow simulation. This approach aims to leverage the complementary strengths of the two model types while avoiding their inherent limitations through parallel execution and nonlinear fusion. The model includes the following steps:

Step 1: Utilize the XAJ model for initial streamflow simulation, fully leveraging its scientific rigor and reliability based on hydrological theory.

Step 2: Introduce the MIC and VIF method for feature selection to optimize the input feature set, thereby enhancing model accuracy and robustness (detailed in Sect. S2.3 in the Supplement).

Step 3: Employ the TCN-GRU model for further streamflow simulation on the time series data selected through feature selection, leveraging the advantages of deep learning models in capturing temporal relationships.

Step 4: Perform nonlinear ensemble using the RF method to combine the streamflow results simulated separately by the XAJ and TCN-GRU models, with the aim of achieving more accurate and reliable comprehensive simulations through model ensemble.

It is worth noting that the selection of RF for nonlinear integration is based on the following theoretical rationale: From a statistical learning perspective, RF generates diverse training subsets through bootstrap resampling and randomly selects

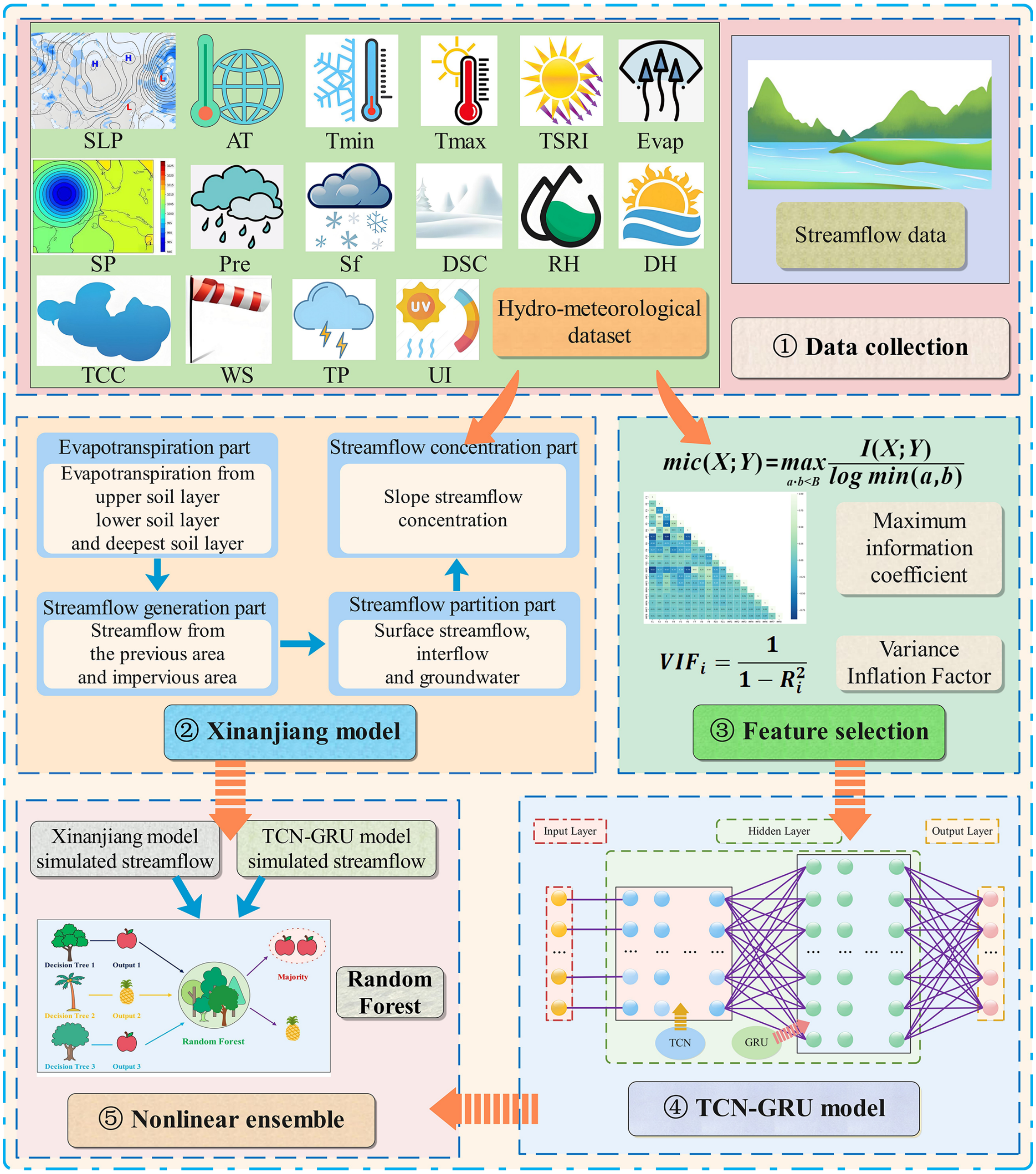

Figure 2. Flowchart for constructing and validating the XAJ-TCN-GRU hybrid model.

features during decision tree construction, effectively capturing nonlinear correlations between different model outputs (Breiman, 2001). For the integration of physical models (XAJ) and deep learning models (TCN-GRU), the former has outputs with clear physical meaning but may contain structural errors, while the latter excels at fitting complex patterns but is susceptible to data noise. The ensemble mechanism of random forests can achieve error complementarity – by averaging or voting across multiple decision trees, reducing the bias and variance of a single model. This characteristic has been proven effective in hydrological model ensembles for addressing the integration of physical mechanisms and data patterns (Xu et al., 2025; Wang and Dong, 2024).

The design of this hybrid model aims to overcome the limitations of individual models, allowing the conceptual rainfall–runoff model and deep learning model to complement each other, thereby providing a more comprehensive and accurate solution for streamflow simulation. The technical flowchart of the study methodology is depicted in Fig. 2.

3 Study area and data description

3.1 Study areas

To validate the adaptability of the proposed XAJ-TCN-GRU hybrid model to different hydrological and geographical conditions, this study conducted experiments across four basins characterized by distinct geographical and hydrological features. These basins include the Wuding River Basin (arid northwestern China), the Chu River Basin (humid eastern coastal region), the Jianxi River Basin (hilly southeastern coastal region), and the Qingyi River Basin (mountainous plateau region in southwestern China). Figure 3 presents the geographical information maps of these four basins. Detailed descriptions of these basins can be found in Sect. S1 in the Supplement. By conducting experiments in basins with varying geographical characteristics and hydrological conditions, it is possible to comprehensively evaluate the adaptability and simulation capability of this model.

3.2 Data description

In this study, daily observations of 31 meteorological and hydrological variables were collected from multiple stations across the four river basins, as tabulated in Table 2. Data acquisition for these variables was conducted via the China Meteorological Science Data Center (http://data.cma.cn/, last access: 12 March 2026). For hydrological monitoring, the primary control stations in the Wuding River, Chu River, Jianxi River, and Qingyi River basins are Baijiachuan Station, Getang Station, Qilijie Station, and Jiajiang Station, respectively. Daily streamflow data for these stations were retrieved from the measured records in *Hydrological Yearbook of the People's Republic of China*. The entire dataset spans a continuous period from 1 January 2010 to 31 August 2023. Statistical summaries of the streamflow datasets for each basin are presented in Table 3, offering a more intuitive characterization of the data (Gil et al., 2016).

3.3 Experimental designs

The input for the XAJ model consists of simulated precipitation and evaporation, while the output is simulated streamflow. The dataset is divided into training, validation, and test sets in a 6 : 2 : 2 ratio. The SCEM-UA method is used to calibrate the model parameters to ensure prediction accuracy.

The 31 hydrological and meteorological variables listed in Table 2 were used as feature variables, while the streamflow was used as the target variable to construct the dataset for inputting into the TCN-GRU model. In order to simplify the model, decrease computational costs, and enhance the interpretability of the data, MIC was employed for feature selection on the 31 feature variables. For each basin, 10 highly correlated feature variables with the streamflow were selected, as shown in Table 4.

Table 2. Input variables and attributes utilized to integrate the XAJ-TCN-GRU model for daily streamflow simulation.

No	Acronyms	Input variables	Units
1	SLP	Sea level pressure	hPa
2	SP	Surface pressure	hPa
3	AT	2 m Average temperature	°
4	Pre	Precipitation	mm
5	Tmax	2 m air temperature – daily max	°
6	Tmin	2 m air temperature – daily min	°
7	Sf	Snowfall	mm
8	DSC	Depth of snow cover	mm
9	GT	Ground temprature	°
10	DPT	Dew point temperature	°
11	RH	Relative humidity	%
12	Evap	Evaporation	mm
13	PET	Potential Evapotranspiration	mm
14	WS	10 m wind speed	$m\,s^{-1}$
15	MGWS	Maximum gust wind speed	$m\,s^{-1}$
16	AGWS	Average gust wind speed	$m\,s^{-1}$
17	LMS	Latitudinal wind speed	$m\,s^{-1}$
18	MWS	Meridional wind speed	$m\,s^{-1}$
19	LCC	Low level cloud cover	–
20	MCC	Medium level cloud cover	–
21	HCC	High level cloud cover	–
22	TCC	Total cloud cover	–
23	NSRI	Net solar radiation intensity	$J\,m^{-2}$
24	TSRI	Total solar radiation intensity	$J\,m^{-2}$
25	DR	Direct radiation	$J\,m^{-2}$
26	DH	Daylight hours	h
27	UI	Ultraviolet intensity	$J\,m^{-2}$
28	TP	Thunderstorm probability	K
29	MTP	Maximum thunderstorm probability	K
30	K	K-index	–
31	CUPE	Convective usable potential energy	$J\,kg^{-1}$

The proposed model was built on a personal computer with a 1.8 GHz Intel i5 processor and 8 GB of memory. The TCN-GRU model was developed using the popular deep learning frameworks Keras and TensorFlow. Additionally, visualization tools like Matplotlib were used to present the model training outcomes visually, ensuring compatibility of the development environment and intuitive analysis of the results. After feature selection via MIC and VIF, a new dataset including the selected features and streamflow was constructed. The training set was used to train the TCN-GRU model, while the validation set was applied for hyperparameter tuning. To improve the performance of deep learning models, a random search algorithm was used to optimize hyperparameters. Firstly, based on existing research results and domain expertise, the range of potential hyperparameter values was defined. Then, the random search algorithm randomly selected hyperparameter combinations and evaluated model performance based on the validation set to systematically explore this range of values. To ensure result stabil-

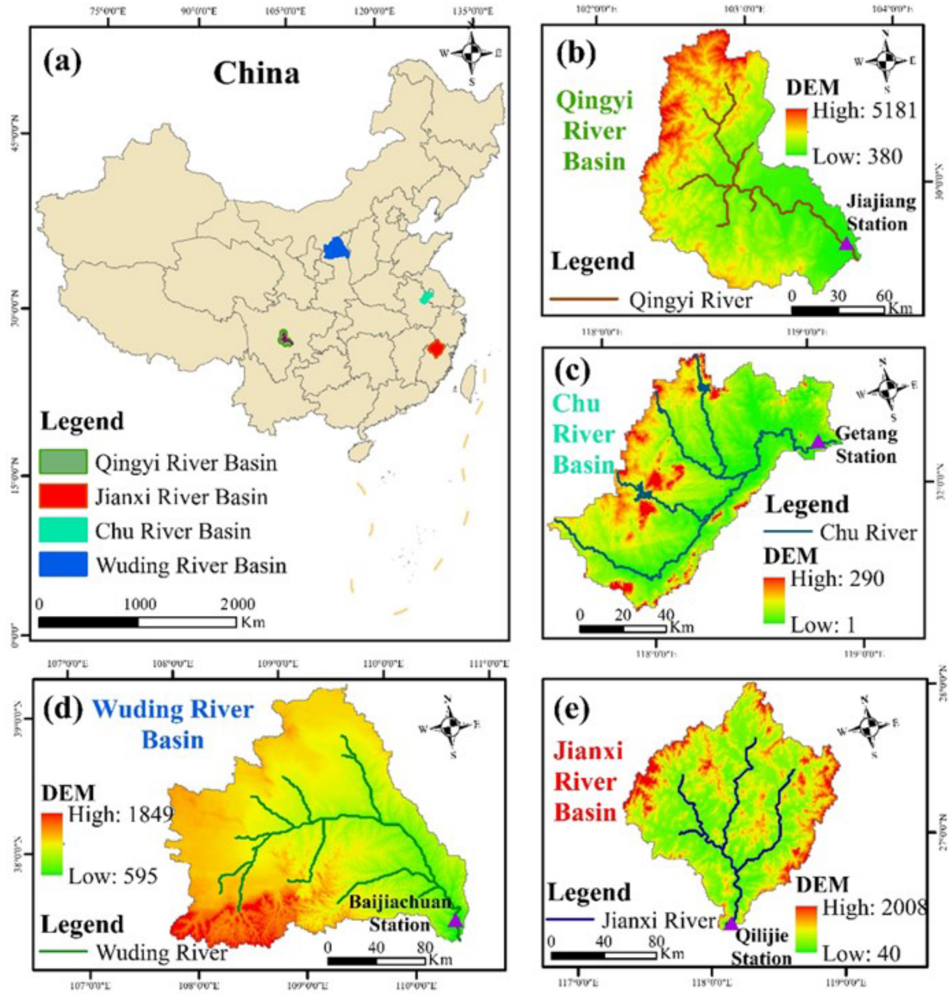

Figure 3. Schematic diagram of geographical information and digital elevation models (DEM) for China's four river basins: Wuding River, Chu River, Jianxi River, and Qingyi River.

Table 3. Descriptive statistical characteristics of daily streamflow data for the four study basins (Wuding River, Chu River, Jianxi River, Qingyi River) (Unit: $m^3\ s^{-1}$).

Basin	Count	Mean	Max	Min	Std	Skew
Wuding	4991	117.5602	1200.8438	12.5313	135.7014	2.7103
Chu	4991	146.2993	2993.1719	18.5938	207.3180	4.8575
Jianxi	4991	246.5827	2917.1094	47.4375	257.6926	4.2258
Qingyi	4991	545.0607	3683.1406	101.2031	411.4504	2.0973

ity and reliability, five independent optimization runs were conducted, and performance metrics were averaged. Through this approach, the optimal hyperparameter combination maximizing model performance was efficiently identified. The final hyperparameter configurations for proposed TCN-GRU and RF are listed in Table S1. The trained TCN-GRU simulated streamflow by inputting the testing set.

Using the simulated results generated by the XAJ model and TCN-GRU model on the training set as feature inputs, a random forest model is trained to learn how to effectively perform nonlinear ensemble. To reduce the risk of overfitting, the maximum depth is set to limit the complexity of each tree, thereby preventing the model from overfitting the noise in the training data. Next, the simulated results from the XAJ and TCN-GRU models on the test set are input into the trained random forest model for nonlinear ensemble, resulting in the final simulation.

Table 4. Predictive input variables obtained using MIC for key feature selection.

Basin	Key feature variables selected through MIC (see Table 2 for variable index numbers)
Wuding	12, 10, 11, 9, 5, 3, 4, 13, 30, 19
Chu	10, 5, 3, 9, 1, 2, 4, 12, 31, 30
Jianxi	1, 2, 12, 10, 5, 9, 3, 31, 30, 21
Qingyi	5, 10, 9, 3, 4, 30, 1, 2, 12, 31

The evaluation metrics adopted in this research include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), NSE, and Kling-Gupta Efficiency (KGE). To quantify uncertainties in runoff simulation, this study employs the Prediction Interval Coverage Probability (PICP) and the Prediction Interval Normalized Average Width (PINAW) as core evaluation metrics. These form a classic combination in the field of hydrological uncertainty quantification (Xu et al., 2025; Kang et al., 2025). These metrics evaluate interval performance across two dimensions, reliability (proportion of observed values covered) and accuracy (interval compactness), aligning with this study's analytical requirements for simulation uncertainty.

PICP measures the proportion of observations falling within the simulated interval, reflecting the reliability of the interval.

$$\mathrm{PICP} = \frac{1}{n}\sum\nolimits_{i=1}^{n} I\left(L_i \le y_i \le U_i\right) \tag{5}$$

where n denotes the sample size, y_i represents the ith observation, L_i and U_i denote the lower and upper bounds of the simulated interval respectively, and $I(\cdot)$ is the indicator function (assigning 1 when $L_i \le y_i \le U_i$, and 0 otherwise). A PICP closer to the preset confidence level (e.g., 95 %) indicates stronger coverage capability of the interval for observations.

PINAW measures the compactness of simulated intervals, reflecting precision.

$$\mathrm{PINAW} = \frac{1}{n \cdot R}\sum\nolimits_{i=1}^{n} \left(U_i - L_i\right) \tag{6}$$

where $R = \max(y_i) - \min(y_i)$ denotes the range of observed values, and $U_i - L_i$ represents the simulated interval width for the ith sample. A smaller PINAW indicates a tighter interval, thereby reducing redundancy in decision-making uncertainty. A detailed description of these metrics is shown in Sect. S3 in the Supplement.

4 Results

4.1 Simulated results for four basins of the XAJ-TCN-GRU model

The XAJ-TCN-GRU model was used in this study to simulate streamflow under four distinct hydrological conditions across various basins, presenting the simulated results in the form of fitted graphs. As indicated in Fig. 4, it is evident that the proposed hybrid model exhibits good fitting performance in all four basins with distinct hydrological and geographical characteristics, which verifies the model's adaptability to different hydrological conditions in the selected study areas. This efficacy is underpinned by the inherent advantages of the XAJ model in simulating physical processes. The XAJ model incorporates 15 parameters with explicit physical significance (Table 1), such as the areal mean tension water capacity (WM) of the catchment characterising basin-scale soil water storage capacity, and the exponent of the tension water capacity curve (B) – these parameters explicitly encode hydrological mechanisms such as soil moisture dynamics and superpercolation (Gong et al., 2021). The XAJ model's prior physical knowledge provides robust mechanistic support for the hybrid model, thereby enhancing its interpretability and generalization capability.

Notably, in basins with larger streamflow, such as Qingyi River, our model demonstrated precise peak simulation capabilities, presenting a significant advantage over traditional XAJ, TCN, and GRU models. Additionally, we observed relatively gentle simulation errors during periods of lower flow, while errors increased significantly and exhibited noticeable fluctuations during peak periods. Moreover, as the number of peaks increased, the fluctuation of errors also increased. However, overall, the absolute error remained within $200\,\mathrm{m^3\,s^{-1}}$.

In basins with relatively low streamflow volumes, such as the Wuding River (arid zone) and the Chu River (humid coastal zone), a slight lag phenomenon was observed in the simulation results. This phenomenon aligns with existing hydrological research findings: Bai et al. (2017) discovered that in arid basins, undulating topography increases surface runoff convergence time, while soil texture (such as highly permeable sandy soils) prolongs the process of soil moisture replenishing runoff; Gebremariam et al. (2014) further indicated that vegetation cover (such as sparse vegetation in the Wuding River basin) reduces evapotranspiration losses but increases surface roughness, thereby indirectly affecting streamflow convergence rates. The aforementioned factors – topography, soil type, and vegetation cover – are not sufficiently incorporated into current model structures (e.g., the XAJ model does not explicitly parameterise topographic slope, and the TCN-GRU model's input features do not include vegetation indices). This omission likely constitutes the primary cause of simulation lag. Consequently, these unaccounted variables impose inherent limitations on

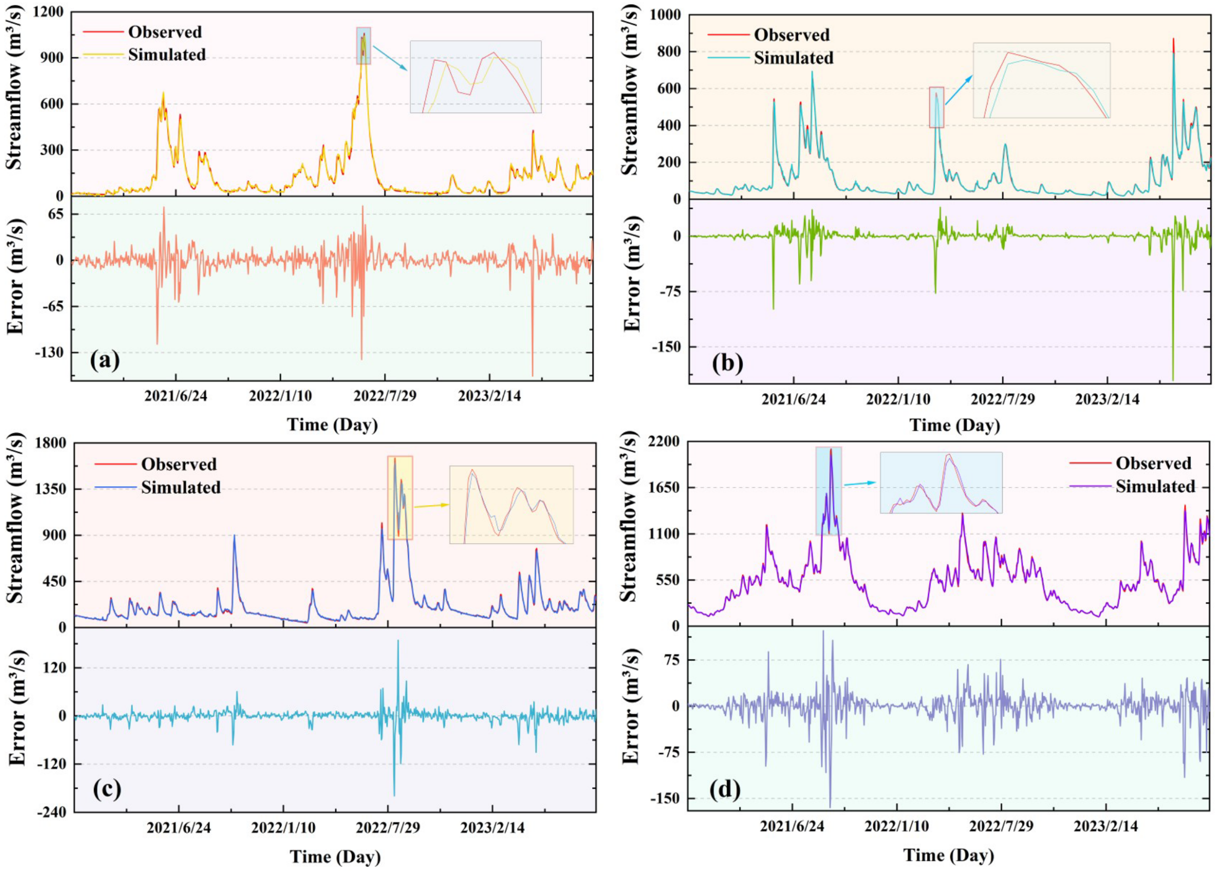

Figure 4. Streamflow simulation fitting results of XAJ-TCN-GRU model for **(a)** Wuding River **(b)** Chu River **(c)** Jianxi River, and **(d)** Qingyi River basins.

modelling complex hydrological processes within specific basins.

To further evaluate the model's adaptability across different flow regimes, we divided the daily streamflow data of each basin's testing set into three intervals (low, medium, and high) using the tertile method, based on the statistical characteristics of streamflow in the test dataset. The division criteria and corresponding performance metrics of the XAJ-TCN-GRU model are detailed in Table S2 as follows: In low-flow conditions, its NSE ranged from 0.911 (Jianxi River) to 0.994 (Qingyi River), with RMSE between 0.911 (Chu River) and 5.288 $m^3 s^{-1}$ (Jianxi River). In medium-flow conditions, the model's NSE varied from 0.872 (Jianxi River) to 0.979 (Wuding River), and RMSE was in the range of 2.659 (Chu River) to 12.556 $m^3 s^{-1}$ (Qingyi River). Even in high-flow conditions (prone to flood events), the model maintained robust performance, with NSE from 0.981 (Chu River) to 0.996 (Wuding River) and RMSE between 11.396 (Wuding River) and 33.582 $m^3 s^{-1}$ (Qingyi River). Notably, the model's MAE in high-flow intervals showed moderate increases relative to medium-flow intervals across basins (e.g., Wuding River: 7.469 vs. 2.817; Qingyi River: 22.588 vs. 9.926), indicating its capability to capture high-flow dynamics effectively.

4.2 Model performance comparison

To better highlight the simulation capabilities of the proposed model, four baseline models (XAJ, LSTM, TCN, and GRU) and one hybrid model (TCN-GRU) were applied to simulate daily streamflow in the four basins as control models for the study. LSTM is a neural network architecture widely applied in time series forecasting and streamflow simulation. By including LSTM in our comparisons, we can conduct a more comprehensive evaluation of our XAJ-TCN-GRU model's performance, ensuring that our research findings have greater universality and robustness. In addition, to further highlight the advantages of nonlinear ensemble, we conducted two additional experiments. One group used the Linear Regression (LR) method to linearly combine XAJ and TCN-GRU models, resulting in a new model called XAJ-TCN-GRU&LR. The other group used the simulations based on XAJ as inputs for the TCN-GRU model, leading to the construction of the XAJ-Infused TCN-GRU model.

Table 5 provides a detailed overview of the streamflow simulation performance of the eight models in the four basins. Clearly, the XAJ-TCN-GRU model exhibits superior performance across all five metrics. The XAJ-TCN-GRU model exhibits relatively low RMSE and MAE values in each basin, particularly when compared to the traditional conceptual rainfall–runoff model XAJ, resulting in a significant reduction in simulation errors. For instance, in the Wuding River basin, the RMSE and MAE of the XAJ model are 69.212 and 64.320 $m^3 s^{-1}$, respectively, while the XAJ-TCN-GRU model achieves only 7.032 and 3.876 $m^3 s^{-1}$ for these metrics. Additionally, when compared with individual TCN and GRU models, the integrated TCN-GRU model exhibits higher NSE and KGE values across the four basins. This indicates the TCN-GRU has effectively combined complementary advantages of TCN and GRU, thus improving the model's simulation capability. The NSE values obtained using the TCN-GRU model in the Wuding River, Chu River, Jianxi River, and Qingyi River basins are 0.990, 0.947, 0.965, and 0.976, respectively. Upon nonlinear ensemble with the simulations of the XAJ model, the simulation accuracy significantly improves. The NSE values achieved using the XAJ-TCN-GRU model for streamflow simulation in the four basins are 0.991, 0.971, 0.984, and 0.986, respectively, representing enhancements of 0.10 %, 2.53 %, 1.97 %, and 1.02 % relative to the TCN-GRU model. Additionally, compared to the linearly integrated XAJ-TCN-GRU&LR model, the improvements are 1.33 %, 1.80 %, 1.66 %, and 0.62 %, respectively. The nonlinear ensemble approach is capable of better capturing intricate nonlinear dependencies within the data, thereby mitigating the risk of error propagation (Wang et al., 2024; Xu et al., 2025). The RF algorithm used for ensemble is less likely to amplify errors compared to linear combinations, as it captures the complex relationships between different model simulations through its decision tree-based structure. Each decision tree in RF is capable of capturing different patterns in the data, and the ensemble process averages out the errors and uncertainties, thereby reducing the impact of potential errors from individual models. In addition, the standard deviations of key validation metrics across the five runs were further quantified to reflect model stability: for NSE, the standard deviation ranged from 0.002 (Qingyi River Basin) to 0.005 (Chu River Basin); for Root Mean Square Error (RMSE), it varied between 0.32 (Wuding River Basin) and 0.87 $m^3 s^{-1}$ (Jianxi River Basin); and for Mean Absolute Error (MAE), the standard deviation was between 0.18 (Wuding River Basin) and 0.54 $m^3 s^{-1}$ (Jianxi River Basin). These small standard deviations indicate minimal fluctuations in model performance across independent optimization runs, confirming the robustness of the hyperparameter tuning process.

Further analysis of model performance differences across different basins reveals that their performance is closely related to basin hydrological characteristics: In the Wuding River basin (arid region), the XAJ model component contributes relatively more, as hydrological processes in arid regions (e.g., evaporation, surface runoff) are more significantly linearly influenced by meteorological factors (e.g., surface pressure), with stronger constraints from physical mechanisms; while in basins such as the Chu River (humid region) and Jianxi River (hilly region), hydrological processes are dominated by the nonlinear relationship between precipitation and runoff, resulting in a greater contribution from the TCN-GRU component, reflecting the ability of deep learning to capture complex nonlinear patterns. We also found that the simulation results of the XAJ-Infused TCN-GRU model are not satisfactory, with the KGE values for the Wuding River, Chu River, and Qingyi River basins even falling below those of TCN-GRU model. This result indicates that directly inputting the outputs from the physical-based model into the deep learning model may propagate the errors from the former to the latter, leading to a decrease in simulation accuracy. The XAJ-TCN-GRU model effectively leverages the strengths of both conceptual rainfall–runoff and deep learning models and compensates for their potential shortcomings in streamflow simulation through nonlinear ensemble.

Further analysis of the core drivers behind the performance differences in Table 5 reveals that XAJ's contribution lies not merely in enhancing model diversity, but in achieving rational corrections to the simulation process through its explicit physical parameter constraints. This conclusion is validated through controlled variable comparisons and mechanism decomposition: From a diversity perspective, the TCN-GRU model has already integrated the complementary strengths of TCN and GRU (NSE ranging from 0.947 to 0.990), indicating its diversity potential has been fully exploited. If performance gains were solely attributable to the addition of new model components, XAJ-TCN-GRU&LR (which similarly possesses the diversity combination of "XAJ+TCN-GRU") should exhibit comparable performance. However, Table 5 data indicates that XAJ-TCN-GRU's RMSE (7.032–20.862 $m^3 s^{-1}$) is reduced by 38.7 %–158.7 % compared to XAJ-TCN-GRU&LR, LR (12.613–29.370 $m^3 s^{-1}$) by 38.7 %–158.7 %, while NSE improved by 0.62 %–1.80 %. This indicates that linear fusion utilises only the diversity of model outputs without leveraging XAJ's physical constraint value. In contrast, the XAJ-Infused TCN-GRU model treats XAJ outputs as inputs (transmitting data information without activating physical constraints), resulting in KGE (0.938–0.964) being lower than TCN-GRU (0.942–0.985). further demonstrating that XAJ's core function lies not in "adding a model member" but in its physical mechanism regulating simulation logic. This capability for correction based on physical mechanisms is absent in deep learning models that rely solely on data pattern learning. It is precisely this capability that constitutes the core reason for the XAJ-TCN-GRU model's significantly superior performance compared to other models.

Table 5. The evaluation metrics of the six models for streamflow simulation on validation and testing sets.

Basin	Model	Validation set					Testing set				
		RMSE ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	MAPE (%)	NSE	KGE	RMSE ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	MAPE (%)	NSE	KGE
Wuding	XAJ	65.418	68.461	29.4	0.819	0.726	69.212	64.320	30.8	0.806	0.707
	LSTM	24.284	15.397	21.9	0.973	0.872	26.841	18.376	24.6	0.971	0.867
	TCN	25.671	13.649	13.4	0.979	0.952	20.810	9.220	12.2	0.982	0.975
	GRU	20.923	12.390	15.2	0.980	0.931	17.649	10.248	14.7	0.987	0.954
	TCN-GRU	16.342	8.175	6.8	0.990	0.985	15.935	9.985	7.1	0.990	0.981
	XAJ-TCN-GRU&LR	10.973	6.964	4.2	0.990	0.986	12.613	8.032	5.8	0.986	0.974
	XAJ-Infused TCN-GRU	16.544	11.370	5.6	0.986	0.981	18.901	10.073	6.4	0.983	0.964
	XAJ-TCN-GRU	6.553	3.684	3.9	0.993	0.991	7.032	3.876	4.1	0.991	0.987
Chu	XAJ	70.613	55.165	66.3	0.681	0.674	71.621	53.927	69.3	0.653	0.641
	LSTM	40.638	27.112	31.5	0.897	0.774	38.192	24.694	30.2	0.901	0.797
	TCN	33.852	21.672	26.3	0.916	0.857	30.824	17.360	24.5	0.936	0.897
	GRU	30.994	15.718	23.0	0.948	0.931	29.314	17.954	23.6	0.942	0.925
	TCN-GRU	24.222	13.693	10.0	0.952	0.950	28.029	14.555	10.8	0.947	0.942
	XAJ-TCN-GRU&LR	17.354	10.397	6.7	0.958	0.947	23.972	11.301	7.3	0.951	0.945
	XAJ-Infused TCN-GRU	20.993	12.387	8.1	0.955	0.941	25.394	13.977	8.0	0.949	0.938
	XAJ-TCN-GRU	15.397	6.314	3.8	0.969	0.958	11.368	4.311	3.1	0.971	0.962
Jianxi	XAJ	125.376	98.498	69.3	0.703	0.718	139.150	110.025	76.0	0.668	0.679
	LSTM	56.317	43.197	31.2	0.930	0.864	55.624	41.842	30.8	0.931	0.868
	TCN	45.671	23.994	9.8	0.959	0.951	42.042	18.766	9.3	0.961	0.960
	GRU	32.478	18.453	9.4	0.970	0.948	40.224	22.669	12.1	0.964	0.941
	TCN-GRU	30.226	11.389	4.8	0.972	0.957	39.564	15.947	5.7	0.965	0.955
	XAJ-TCN-GRU&LR	15.397	7.619	3.7	0.985	0.971	22.007	10.243	4.9	0.976	0.963
	XAJ-Infused TCN-GRU	17.308	9.114	4.2	0.978	0.966	32.904	13.354	5.1	0.971	0.958
	XAJ-TCN-GRU	12.077	6.356	3.1	0.990	0.986	16.732	8.179	4.0	0.984	0.979
Qingyi	XAJ	103.677	71.594	57.6	0.754	0.704	112.101	88.635	63.4	0.720	0.681
	LSTM	78.642	57.349	11.2	0.935	0.885	78.773	56.548	11.3	0.938	0.882
	TCN	56.555	40.375	7.2	0.971	0.980	50.613	32.678	6.7	0.975	0.962
	GRU	60.348	42.374	7.9	0.969	0.936	55.138	39.466	7.5	0.970	0.947
	TCN-GRU	46.612	27.315	4.7	0.978	0.966	49.626	29.892	5.1	0.976	0.985
	XAJ-TCN-GRU&LR	21.982	13.988	2.7	0.980	0.972	29.370	15.976	3.4	0.972	0.968
	XAJ-Infused TCN-GRU	29.037	23.974	3.3	0.975	0.968	36.691	24.503	3.9	0.970	0.964
	XAJ-TCN-GRU	16.785	9.428	1.8	0.988	0.980	20.862	11.881	2.0	0.986	0.974

Scatter plots provide an intuitive display of the model's explanatory power over data variations, as depicted in Fig. 5. Upon analysis, the XAJ model exhibits a tendency to overestimate streamflow during low-flow periods and underestimate it during high-flow periods. In contrast, deep learning models demonstrate superior streamflow simulation capabilities. Notably, the fitting line of the XAJ-TCN-GRU model is closest to the 1 : 1 line. The R^2 values of this model in the four basins are 0.988, 0.981, 0.984, and 0.986, respectively, representing improvements of 3.56 %, 34.20 %, 20.59 %, and 20.10 % compared with the standalone XAJ model. In comparison with the TCN-GRU model, the improvements are 0.92 %, 2.51 %, 1.86 %, and 0.92 %, respectively. In the Chu River basin, the R^2 value of the XAJ model is significantly low, indicating that traditional conceptual rainfall-runoff models are not applicable to this basin. Across the four basins, the simulation results of TCN and GRU are comparable, with R^2 values both exceeding 0.94, highlighting the important role of these two models in each basin. The integrated TCN-GRU model effectively combines the advantages of both models, demonstrating good adaptability across all basins and the ability to handle complex hydrological conditions. These findings underscore the broad applicability and superiority of XAJ-TCN-GRU model in diverse basins.

The Taylor diagram is widely employed to evaluate the correlation and simulation accuracy between model predictions and observed data. Figure 6 illustrates that different symbols denote various simulation methods, with standard deviation on the horizontal axis and vertical axis, and the dotted line representing RMSE. The optimal simulation accuracy is represented by the center of the horizontal axis in the Taylor diagram. Notably, in Fig. 6, the XAJ-TCN-GRU model's predictions for the Wuding River basin are posi-

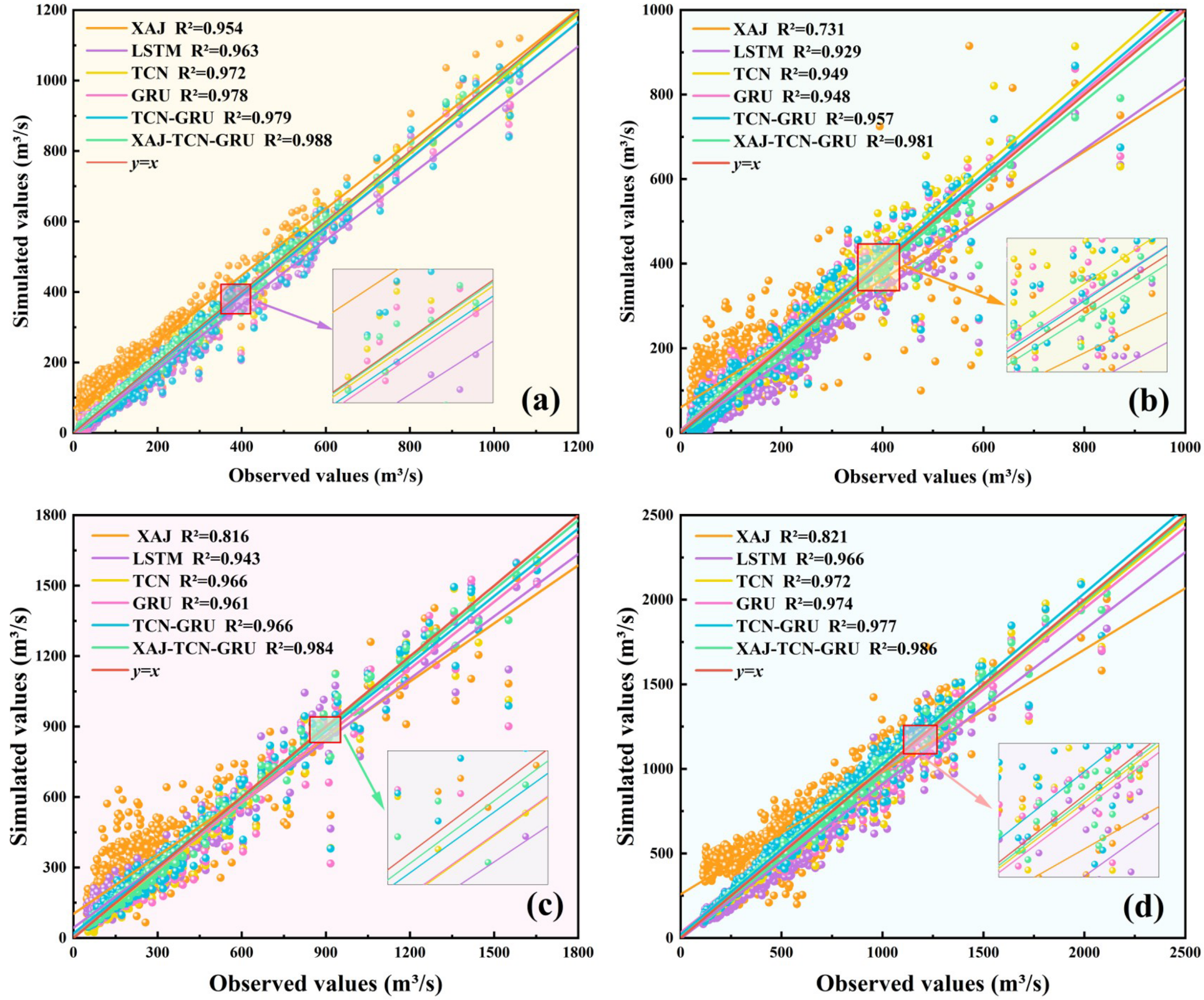

Figure 5. Scatter plots of simulated of each model vs. observed streamflow on the testing set of **(a)** Wuding River, **(b)** Chu River, **(c)** Jianxi River, and **(d)** Qingyi River basins.

tioned at the optimal location in the Taylor diagram, and the correlation coefficient is 0.99. Likewise, for the Chu River, Jianxi River, and Qingyi River basins, the XAJ-TCN-GRU model's results outperform those of the comparative models and are closest to the observed values. By contrast, the XAJ model's simulated streamflow shows significant deviations from the observed data. This further highlights the proposed model's advantage in streamflow simulation.

The proposed XAJ-TCN-GRU model demonstrates promising performance in streamflow simulation across the four representative Chinese basins with distinct hydrological and geographical characteristics (arid, humid, hilly, and plateau regions). Its performance advantages in these selected basins are mainly reflected in the effective integration of physical mechanism interpretability from the XAJ model and complex pattern capture capability from the TCN-GRU model, as well as the mitigation of error propagation through nonlinear ensemble. However, it is important to acknowledge that the relative standing of the proposed model compared with large-sample hydrological models developed based on extensive basin datasets (e.g., CAMELS) still needs further evaluation under standardized and comparable experimental frameworks.

5 Discussion

5.1 Model robustness analysis

To evaluate the robustness and resilience of the model to changes in the input data, we performed noise data injection testing and out-of-context validation.

In actual hydrological observations, due to equipment malfunctions, recording errors, and other reasons, observational data often contain erroneous data. In noise injection testing, to align with actual hydrological observation scenarios, the characteristics of the injected noise are first defined: Draw-

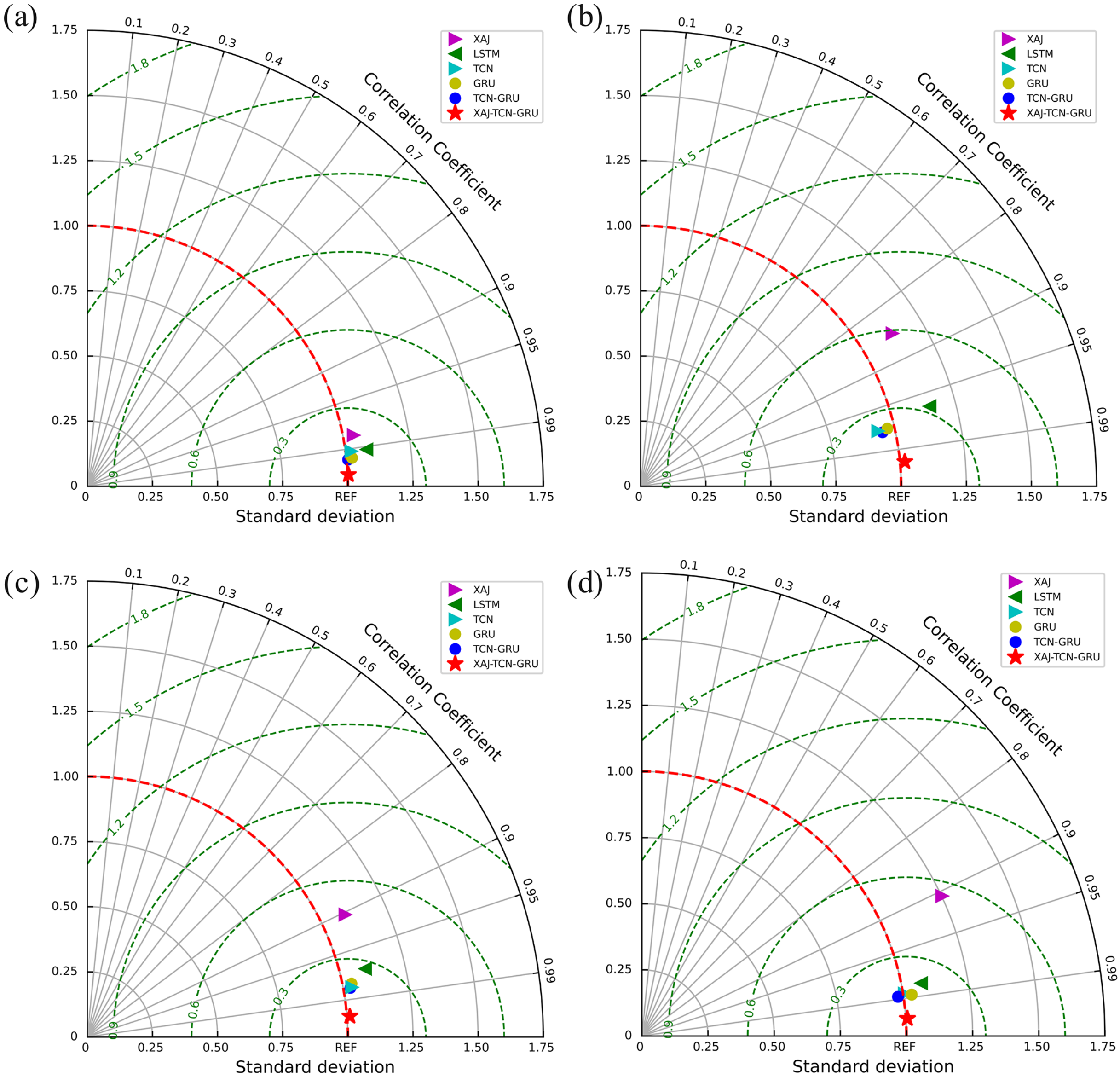

Figure 6. Taylor diagram illustrating correlation coefficient and the standard deviation difference for XAJ-TCN-GRU model compared with comparative models at **(a)** Wuding River, **(b)** Chu River, **(c)**, Jianxi River, and **(d)** Qingyi River basins.

ing upon common error sources in field hydrological monitoring (such as sensor drift, manual recording discrepancies, and data transmission losses), and considering the error background of Chinese hydrological observation data mentioned in the study, an initial noise proportion of 2 % is set as a conservative baseline (compliant with the lower limit requirement of 1 %–2 % equipment error specified in the Chinese Hydrological Observation Specification GB/T 50095-2014). To avoid the unrealistic assumption of independent noise, the injected noise follows a normal distribution (mean = 0, standard deviation = 5 %–10 % of the corresponding variable's observed value). This is combined with the correlation characteristics of variables across the four basins (e.g., Pearson's correlation coefficient between precipitation and relative humidity in the Wuding River basin is 0.72, and the correlation coefficient between average temperature and evaporation in the Qingyi River basin is 0.63), ensuring consistency with the patterns observed in the actual data (Wang et al., 2023; Zhao et al., 2024). Subsequently, we used this perturbed dataset to train the models and make streamflow simulations to test the models' responsiveness to noise data situations. Noise data injection helps evaluate how the models perform when facing unknown or noise data situations in the real world and aids in improving the models' performance and generalization capabilities. Additionally, we conducted 10 independent runs for each model and recorded the results of each run, including

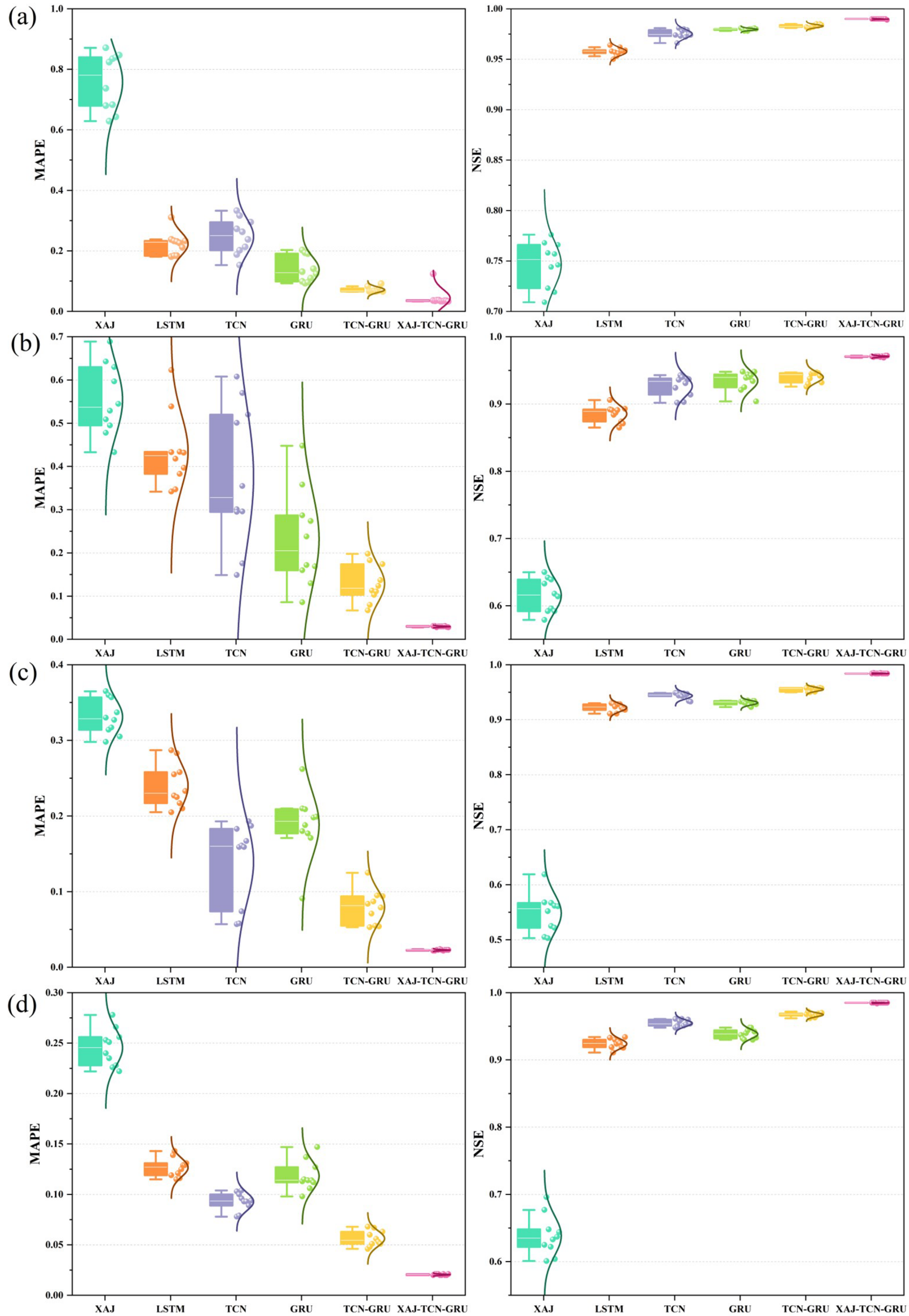

Figure 7. Box plots of the proposed XAJ-TCN-GRU model and comparative models for simulating streamflow in terms of MAPE and NSE for **(a)** Wuding River, **(b)** Chu River, **(c)** Jianxi River, and **(d)** Qingyi River basins.

the model's outputs and performance metrics. By running the model multiple times, we could observe the variations and fluctuations in the models' performance across different runs, thereby enabling a more thorough assessment of its performance and improving the evaluation of its robustness and reliability.

The detailed outcomes of the noise data injection test are recorded in Sect. S4.3 in the Supplement. The boxplots in Fig. 7 provide a detailed representation of the normal distribution of indicators obtained from the 10 runs of each model during the noise data injection test. It is apparent that the deep learning models demonstrate good stability. Among them, the LSTM, GRU, and TCN-GRU models show relatively stable simulation performance across the four basins in the noise data injection test. However, it is noteworthy that the TCN model exhibits significant fluctuations in the tests conducted in the Chu and Jianxi basins. Overall, the XAJ-TCN-GRU model demonstrates the strongest stability and the best simulation performance in the noise data injection test. The XAJ-TCN-GRU model leverages the advantages of integrating conceptual rainfall - runoff and deep learning models, enabling it to capture the physical characteristics of streamflow data while also learning the deeper features of the data, thereby performing exceptionally well in handling noise data.

In the out-of-context validation, we excluded data from wet-year in the training sets of the four basins and used these data separately for model validation. A wet-year was defined as a hydrological year (June of the current year to May of the next year, consistent with China's hydrological year division) where the annual precipitation exceeded the 75th percentile of the long-term (2010–2023) annual precipitation series in the respective basin. This criterion aligns with the classification standard for wet years in the Technical Specifications for Hydrological Data Processing (General Administration of Quality Supervision, GB/T 50102-2014) in China. This approach aims to evaluate the model's performance under extreme hydrological conditions (i.e., wet-year) that it has not encountered before, providing a better understanding of its generalization ability and robustness. The out-of-context validation results for the four basins are shown in Table 6. By comparing Tables 5 and 6, we can see that the model's simulation accuracy declines slightly under wet-year conditions compared to regular years. However, it still maintains a high level of accuracy overall, indicating that the model has a certain degree of generalization ability and can make reasonable simulations under extreme flow conditions. In the Jianxi River and Qingyi River basins, the model's simulation accuracy decreases more significantly during wet-year, likely because these regions experience greater precipitation and flow fluctuations during wet-year. This variability may not be sufficiently captured in the training set, thereby affecting the ability of the model to learn and adapt under these extreme conditions.

To further refine the model robustness assessment framework, this study additionally designed the "drought year exclusion" approach and the "extreme event threshold segmentation" approach to comprehensively evaluate the model's stability under various exceptional conditions, such as extremely low flow and extremely high flow (Acuña Espinoza et al., 2025b).

"Dry-Year Exclusion" Approach: Years with annual precipitation $< 25\,\%$ of the multi-year average were defined as "dry-year". After excluding dry-year data from the training set, model training was conducted, followed by testing using the excluded dry-year data. Results are shown in Table S4. It indicates that the XAJ-TCN-GRU model achieved NSE values of 0.954–0.969 under extreme low-flow scenarios, with a performance decline of only 1.7 %–2.3 % compared to the full-data training scenario. This decline was lower than that observed for the TCN-GRU model (3.4 %–3.5 %) and the XAJ model (8.0 %–8.8 %), demonstrating its robustness under extreme low-flow conditions. Specifically, the Qingyi River basin, possessing a larger baseline runoff volume (mean daily streamflow $545.06\,\mathrm{m^3\,s^{-1}}$), exhibited the smallest NSE reduction (1.7 %) after excluding dry years. Conversely, the Wuding River basin, characterised by arid conditions (mean daily streamflow $117.56\,\mathrm{m^3\,s^{-1}}$), showed a relatively higher reduction (2.3 %), yet still maintained a low overall level.

"Extreme Event Threshold Segmentation" approach: Pearson Type III distributions were employed to calculate the 10-year return period flow thresholds for the four basins. Model training utilised only data below the thresholds, while data exceeding the thresholds tested the models' simulation capability for extreme high flows. Results are presented in Table S4. It indicates that the NSE of the XAJ-TCN-GRU model in extreme high-flow testing remained between 0.941 and 0.955, representing a 3.1 %–4.6 % reduction compared to the full-data training scenario. This decline was significantly lower than that observed for the TCN-GRU model (5.9 %–7.2 %) and the XAJ model (8.9 %–12.6 %). Notably, in the Qingyi River basin, where the 10-year return period threshold ($2850.3\,\mathrm{m^3\,s^{-1}}$) approaches its historical maximum discharge ($3683.14\,\mathrm{m^3\,s^{-1}}$), the model still achieves an NSE of 0.955 with a mere 3.1 % decline, demonstrating robust adaptability to high-intensity extreme flows. The Wuding River, characterised by arid basin conditions, exhibited a substantial disparity between extreme flow and mean flow (threshold $1120.5\,\mathrm{m^3\,s^{-1}}$ being 9.5 times the mean $117.56\,\mathrm{m^3\,s^{-1}}$), resulting in a relatively higher NSE reduction (4.6 %). Nevertheless, this reduction remained substantially lower than that of the comparison model.

5.2 Flood simulation analysis

Flood simulation is a core component of watershed management and disaster prevention planning. Especially during the flood season, accurate streamflow simulations are paramount

Table 6. Results of out-of-context validation on the XAJ-TCN-GRU model.

Basin	RMSE ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	MAPE (%)	NSE	KGE
Wuding	8.210	5.381	4.9	0.973	0.970
Chu	13.943	6.876	4.1	0.952	0.945
Jianxi	20.786	11.763	5.1	0.969	0.940
Qingyi	28.014	17.002	4.8	0.963	0.951

for the prompt execution of flood control measures and the rational allocation of water resources. This section investigates streamflow simulations across four distinct basins during both flood and non-flood seasons, providing a comprehensive assessment of their capabilities.

Based on the climatic and geographical conditions, the flood seasons for the Wuding River and Chu River primarily occur from June to September, while for the Jianxi and Qingyi Rivers, they commence slightly earlier, from May to September. These flood seasons represent periods characterized by the most significant variations in streamflow, making them particularly challenging to simulate. According to the data presented in Table 7, the XAJ-TCN-GRU model demonstrates exceptional performance in simulating streamflow during the flood seasons across all four basins. Its NSE values for the Wuding River, Chu River, Jianxi River, and Qingyi River are 0.988, 0.968, 0.984, and 0.980, respectively, all of which are close to or exceed 0.98, indicating the model's high accuracy in flood season simulations. Furthermore, during non-flood seasons, the model maintains high NSE values of 0.987, 0.970, 0.980, and 0.987, respectively, further confirming its stability and reliability. It is important to note that although the accuracy of the XAJ-TCN-GRU in simulating streamflow during flood seasons for the Chu River and Qingyi River basins is slightly lower compared to non-flood seasons, this does not indicate poor performance. In fact, the complexities arising from factors such as climate change and topography make it challenging for any simulation model to achieve perfection during flood seasons. Nonetheless, compared with the other five comparative models, the XAJ-TCN-GRU model demonstrates the lowest RMSE, MAE, and MAPE in streamflow simulation during flood seasons across all four basins, highlighting its superiority in flood season simulation.

This study also placed a specific emphasis on the accurate simulation of extreme hydrological events. Flood events were identified using the Peak Over Threshold (POT) method, a widely accepted flood frequency analysis approach (Hosking and Wallis, 1997). For each basin, the threshold was determined as the 95th percentile of the daily streamflow series in the testing set, corresponding to a return period of approximately 2 years based on regional hydrological characteristics (as recommended by the China Hydrological Manual, 2017). Periods where daily streamflow exceeded this threshold and lasted for at least 2 consecutive days were classified as flood events. Through in-depth hydrological data analysis, a total of eight such flood events were identified across the four basins. Table 8 presents detailed simulation outcomes of these flood events using different models in the four basins. Clearly, the XAJ-TCN-GRU model exhibits the least simulation bias in all four basins when observing the data in Table 8, with remarkably lower RMSE and MAPE values than other models. In the four flood events of the Wuding River basin, the RMSE of XAJ-TCN-GRU was reduced by 30.60 %, 42.16 %, 66.47 %, and 43.01 % compared with TCN-GRU. This significant reduction not only highlights the importance of integrating the results of conceptual rainfall - runoff model to enhance the simulation accuracy of hybrid models but also further validates the high accuracy and dependability of the XAJ-TCN-GRU in simulating flood events. The XAJ-TCN-GRU model's in-depth understanding of hydrological processes provides an important foundation for flood peak simulation in areas with limited hydrological data, particularly in regions with sparse observation stations, where the model's portability is especially significant.

However, it is important to note that while this model provides insights into potential applications for flood peak simulation, our work focuses on streamflow simulation rather than direct flood simulation. Directly using streamflow simulation results for flood peak simulation has limitations. First, streamflow models often rely on coarse spatial and temporal scales, which may overlook critical local features and flood dynamics. Second, inaccuracies in model parameters and input data, particularly under extreme precipitation conditions, can amplify errors in flood peak estimation. These factors highlight the need for further development to improve the accuracy and reliability of flood peak simulations using streamflow models. The results suggest that, although the model's simulation can serve as an approximation of flood peak flows in areas with limited observation stations, further development is needed to improve the reliability of flood simulations. Future research may explore incorporating additional flood-related factors and advanced post-processing techniques to better translate streamflow simulations into accurate flood simulations.

Further quantitative analysis of the key characteristics of 8 extreme flood events shows that the XAJ-TCN-GRU model performs optimally across all core metrics (as detailed in Table S5):

Table 7. Simulation outcomes of the XAJ-TCN-GRU model and comparative models in four basins during flood and non-flood seasons.

Basin	Model	Flood Season					Non-Flood Season				
		RMSE ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	MAPE (%)	NSE	KGE	RMSE ($m^3 s^{-1}$)	MAE ($m^3 s^{-1}$)	MAPE (%)	NSE	KGE
Wuding	XAJ	69.271	64.966	73.9	0.892	0.806	69.137	63.952	71.4	0.568	0.549
	LSTM	32.717	25.314	23.0	0.976	0.968	23.341	14.880	25.4	0.951	0.943
	TCN	17.518	10.996	9.0	0.983	0.980	14.810	8.229	13.9	0.970	0.972
	GRU	20.058	12.752	11.4	0.971	0.982	17.658	10.255	14.7	0.987	0.964
	TCN-GRU	25.466	13.993	6.8	0.985	0.988	17.962	7.922	7.3	0.971	0.965
	XAJ-TCN-GRU	8.998	5.301	3.6	0.988	0.990	5.703	3.124	4.4	0.987	0.978
Chu	XAJ	85.177	69.672	71.9	0.657	0.597	63.776	46.071	68.3	0.664	0.593
	LSTM	50.564	35.345	22.0	0.885	0.843	30.069	19.286	34.3	0.836	0.743
	TCN	36.581	23.193	15.9	0.909	0.886	22.526	14.420	28.9	0.908	0.837
	GRU	40.883	29.002	23.0	0.925	0.894	21.214	12.387	24.0	0.919	0.884
	TCN-GRU	45.087	26.718	12.6	0.940	0.907	20.078	8.389	9.9	0.927	0.901
	XAJ-TCN-GRU	16.674	7.957	3.7	0.968	0.927	7.362	2.471	2.8	0.970	0.936
Jianxi	XAJ	134.435	106.805	61.5	0.777	0.781	91.986	75.216	65.0	0.658	0.493
	LSTM	61.793	42.369	22.1	0.953	0.928	50.556	41.489	37.4	0.776	0.744
	TCN	51.917	23.662	8.8	0.959	0.936	26.816	15.121	9.8	0.937	0.883
	GRU	54.410	30.448	13.1	0.964	0.948	24.736	16.865	11.3	0.946	0.897
	TCN-GRU	57.847	22.569	6.1	0.967	0.970	42.063	15.961	5.7	0.961	0.934
	XAJ-TCN-GRU	22.484	10.799	3.9	0.984	0.978	10.534	6.229	4.1	0.980	0.961
Qingyi	XAJ	135.246	106.087	15.5	0.788	0.657	105.228	92.937	54.4	0.659	0.604
	LSTM	99.776	72.860	9.5	0.885	0.791	58.095	44.164	12.5	0.940	0.826
	TCN	71.052	46.127	5.6	0.941	0.867	26.295	17.646	6.6	0.978	0.904
	GRU	68.884	48.399	8.8	0.931	0.889	27.406	20.906	6.7	0.980	0.892
	TCN-GRU	77.301	63.413	6.6	0.945	0.902	28.924	21.477	4.8	0.982	0.935
	XAJ-TCN-GRU	29.702	19.063	2.3	0.980	0.964	9.954	6.467	1.7	0.987	0.971

Table 8. Simulation results of flood events using the proposed XAJ-TCN-GRU and comparative models for four basins.

Basin	Flood event	XAJ		LSTM		TCN		GRU		TCN-GRU		XAJ-TCN-GRU	
		RMSE ($m^3 s^{-1}$)	MAPE (%)	RMSE ($m^3 s^{-1}$)	MAPE (%)	RMSE ($m^3 s^{-1}$)	MAPE (%)	RMSE ($m^3 s^{-1}$)	MAPE (%)	RMSE ($m^3 s^{-1}$)	MAPE (%)	RMSE ($m^3 s^{-1}$)	MAPE (%)
Wuding	22/5/2021–24/5/2021	71.073	12.0	60.424	10.3	25.224	4.2	22.668	3.2	16.720	3.0	11.604	1.5
	5/6/2021–6/6/2021	21.040	3.8	24.434	4.4	19.277	3.5	10.468	24.5	6.618	1.1	3.828	0.7
	28/5/2022–29/5/2022	61.448	11.2	12.086	2.2	21.479	3.9	10.125	1.8	8.897	1.6	2.985	0.5
	1/6/2022–26/6/2022	66.882	6.6	65.672	6.6	67.349	6.5	49.185	4.1	42.754	3.6	24.366	2.3
Chu	20/6/2023–21/6/2023	132.625	17.0	36.060	4.6	86.269	11.0	78.822	10.1	44.289	5.7	26.699	3.4
Jianxi	10/8/2022–16/8/2022	56.266	3.6	82.850	5.5	94.629	6.4	65.441	4.7	48.885	3.6	34.579	2.3
	20/8/2022–1/9/2022	196.557	13.1	111.546	5.7	105.329	5.8	92.535	5.8	96.338	6.2	50.825	2.8
Qingyi	4/9/2021–8/9/2021	285.960	11.8	241.618	8.8	210.184	8.6	209.760	7.3	177.104	7.7	94.696	4.3

1. Peak flow: The average peak error of the integrated model across the 8 events (6.1 %) was 62.3 % lower than that of XAJ (16.2 %) and 43 % lower than that of TCN-GRU (10.7 %). Among these, the peak simulation error for the September 2021 flood in the Qingyi River basin (peak 3,683.1 $m^3 s^{-1}$) was only 4.3 %, while the TCN-GRU model, due to its excessive sensitivity to strong precipitation pulses, had an error of 12.6 %.
2. Total flood volume: When calculated based on a 3-day flood volume, the average error of the integrated model (5.1 %) was significantly lower than that of XAJ (14.8 %) and TCN-GRU (9.7 %). For example, during the August 2022 flood in the Jianxi River basin, the integrated model simulated a flood volume, which deviated by only 1.4 % from the observed value. However, the TCN-GRU model underestimated the later receding flow, resulting in a flood volume deviation of 8.2 %.

3. Peak occurrence time: The average peak occurrence time deviation of the integrated model (2.1 hours) is only 36.8 % of that of TCN-GRU (5.7 hours), thanks to the physical constraints on flood propagation speed imposed by the confluence parameters of the XAJ model. For example, in the June 2023 flood in the Chu River basin, the XAJ model alone simulated a peak onset time deviation of 3.2 hours, while the TCN-GRU model had a deviation of 6.8 hours, which was corrected to 1.5 hours after integration.

4. Receding time: Measured by the time (T50) for flow to decrease from peak to 50 %, the integrated model's T50 simulation error (3.2 %) was significantly lower than TCN-GRU (8.7 %), as it incorporated XAJ's groundwater receding coefficient, effectively capturing the slow receding process of baseflow.

5.3 Interval simulation

Traditional point simulations often prove insufficient to fully account for the inherent uncertainty in simulations, a challenge typically unavoidable in streamflow modeling. To obtain numerical estimates with their associated reliability, utilizing interval simulation methods represents a more reasonable approach. This approach not only provides an average simulation estimate but also captures the uncertainty level inherent in the simulation, thereby enabling more robust and scientifically grounded decision-making. In this section, interval simulation methods based on error modeling are investigated, building on the foundation of point simulations. The methodology adopted follows Song et al. (2015), namely, constructing simulation confidence interval at a specified significance level. The framework integrates the point simulation results from Sect. 4.2, using the logit distribution function to match the simulation error sequence and performing interval simulation at the designated significance level. Two evaluation metrics are used to assess interval simulation accuracy: PICP and PINAW. At a given significance level, a higher PICP value combined with a lower PINAW value indicates better model performance.

Three distinct significance levels were chosen for each of the four river basins, with corresponding PICP and PINAW results documented in Table 9. Interval simulation results of the XAJ-TCN-GRU model at these four sites are visualized in Figs. 8–10. Notably, at different levels of significance, the majority of actual observations are located within the shaded area of the simulated interval, indicating that the XAJ-TCN-GRU model exhibits good coverage probabilities across different significance levels. Furthermore, an apparent trend shows that as the significance level increases, the average width of the simulation intervals tends to decrease.

At a 0.05 significance level, the XAJ-TCN-GRU model realizes PICP rates of 94.29 %, 95.09 %, 95.29 %, and 95.09 % across the four river basins, respectively. These values show improvements over TCN-GRU model, highlighting the efficacy of integrating the conceptual rainfall-runoff model with the nonlinear ensemble. In terms of PINAW, the XAJ-TCN-GRU model yields values of 0.022, 0.044, 0.034, and 0.034, compared with the traditional conceptual rainfall-runoff model XAJ, which has PINAW values of 0.094, 0.244, 0.186, and 0.228. These results demonstrate a substantial reduction in the average width of simulation intervals for the XAJ-TCN-GRU model. The data in Table 9 further confirms that the XAJ-TCN-GRU model excels in generating more compact and narrower simulation windows, thereby reduces decision-making risk caused by inaccuracies and uncertainty. This effectively addresses uncertainties and better aligns with practical requirements.

5.4 Interpretability of deep learning model

Although the XAJ-TCN-GRU model demonstrates excellent performance across all basins, showcasing strong stability and robustness, the lack of transparency in deep learning models hinders the verification of their simulation logic. To improve the interpretability of the XAJ-TCN-GRU, this study employs three key indicators: mean absolute SHAP values (SHAPABS), Feature Importance (FI), and Permutation Feature Importance (PFI). These indicators can accurately calculate the specific contribute of each variable input to long-term trends on streamflow in the TCN-GRU model, thereby offering the clearer insight of the internal operation mechanism of TCN-GRU model.

Figure 11 presents density scatter plots of various input variables across the four basins, with the y-axis arranged in descending order according to their contributions to the simulated values per sample. Positive SHAP values indicate contributions that enhance the model's performance, with larger values reflecting more substantial contributions. In cases where multiple samples exhibit identical contribution values, these samples are vertically stacked within the plots. Figure 12 provides a comprehensive overview of the importance ranking of hydro-meteorological variables influencing streamflow, as determined by the TCN-GRU. The numerical values of three key indicators for the input variables of the TCN-GRU across the four basins, along with their normalization, are detailed in Sect. S4 in the Supplement.

Combining Figs. 11 and 12, it can be seen that dew point temperature is the most influential feature variable in the basins of Wuding River, Chu River, and Qingyi River. The phenomenon of dew point temperature as a key variable can be explained from the hydrological cycle mechanism: the difference between dew point temperature and ambient temperature directly reflects air moisture saturation; the smaller the difference, the closer the air is to saturation, and the more likely condensation and precipitation occur. Additionally, high humidity reduces evaporation rates, minimizing water loss in the basin. This influence varies across different climate zones: in humid regions (such as the Chu River),

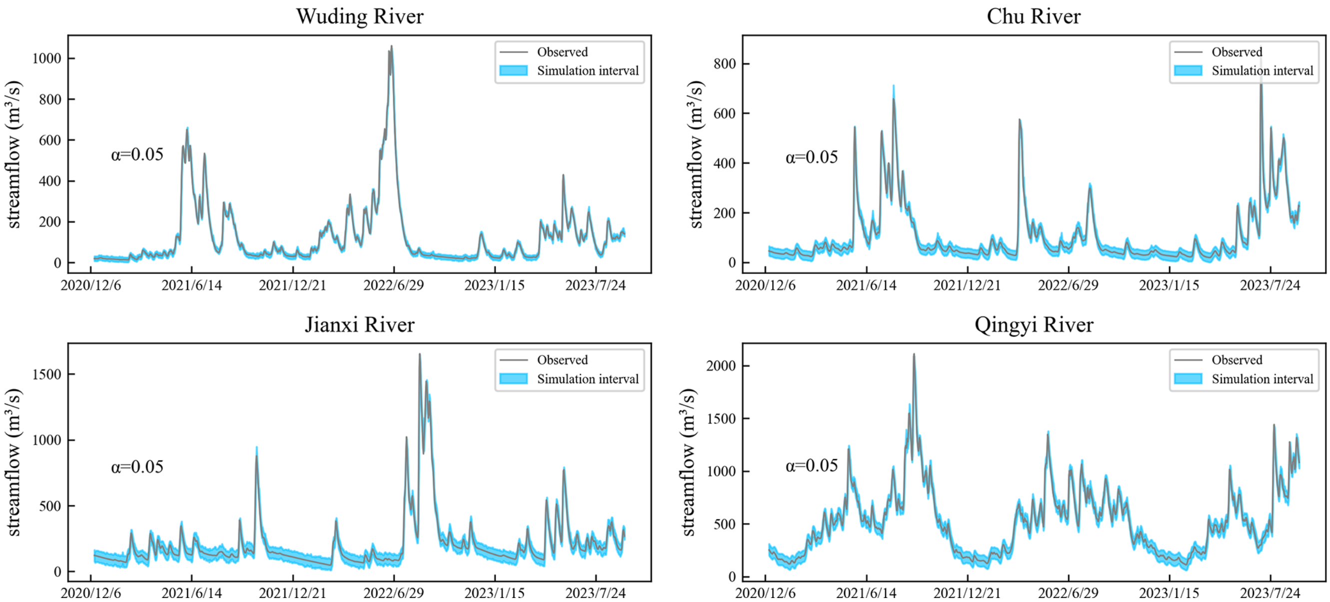

Figure 8. Interval simulation outcomes of the XAJ-TCN-GRU model for four basins at $\alpha = 0.05$.

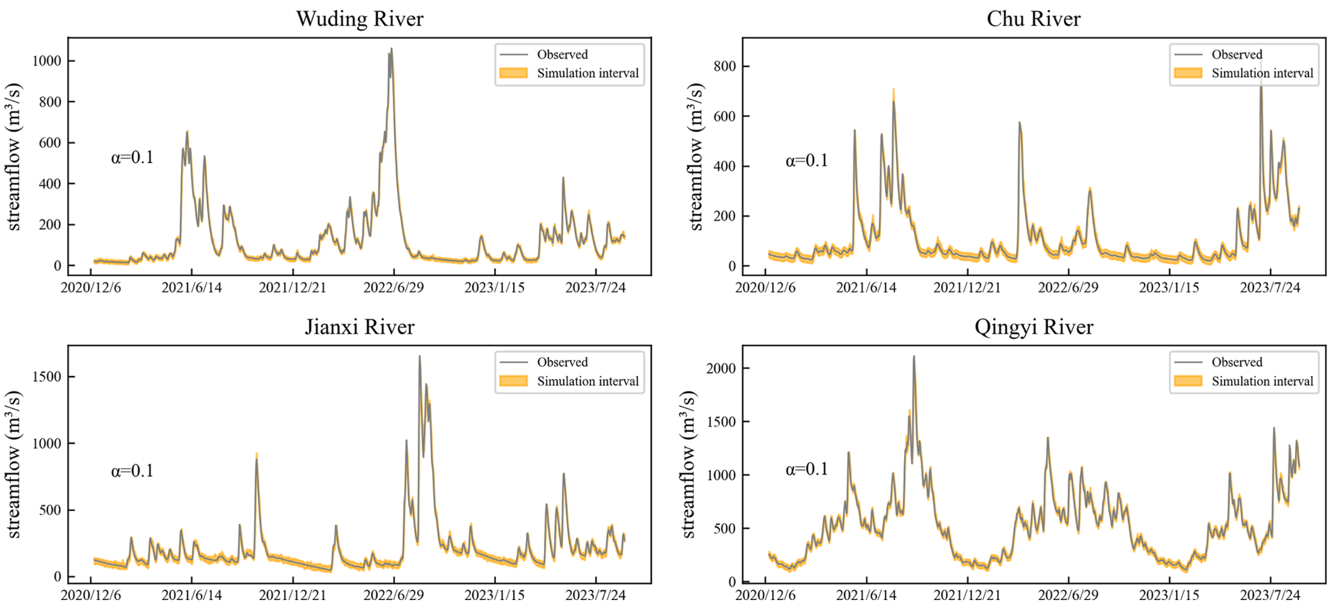

Figure 9. Interval simulation outcomes of the XAJ-TCN-GRU model for four basins at $\alpha = 0.1$.

where water vapor is abundant, even minor changes in dew point temperature can significantly affect precipitation intensity; while in arid regions (such as the Wuding River), the synergistic effects of dew point temperature and surface pressure, among other factors, indirectly regulate limited precipitation processes by influencing convective activity. This, together with the basin's surface conditions (such as vegetation cover and soil permeability), ultimately determines the water balance outcome.

Furthermore, the second influential variables in each basin exhibit significant differences, reflecting the importance of varying hydrological characteristics and climatic conditions for flow simulation. In the Wuding River basin, surface pressure is identified as the second most important variable, which is closely related to the region's arid climate characteristics. In arid environments, surface pressure plays a critical role in the hydrological cycle, influencing the formation of precipitation and the evaporation of water. In contrast, in the Chu River basin, the daily maximum temperature is the second most important variable, indicating that temperature fluctuations in a humid environment may directly affect the evaporation and flow of water bodies. Meanwhile,

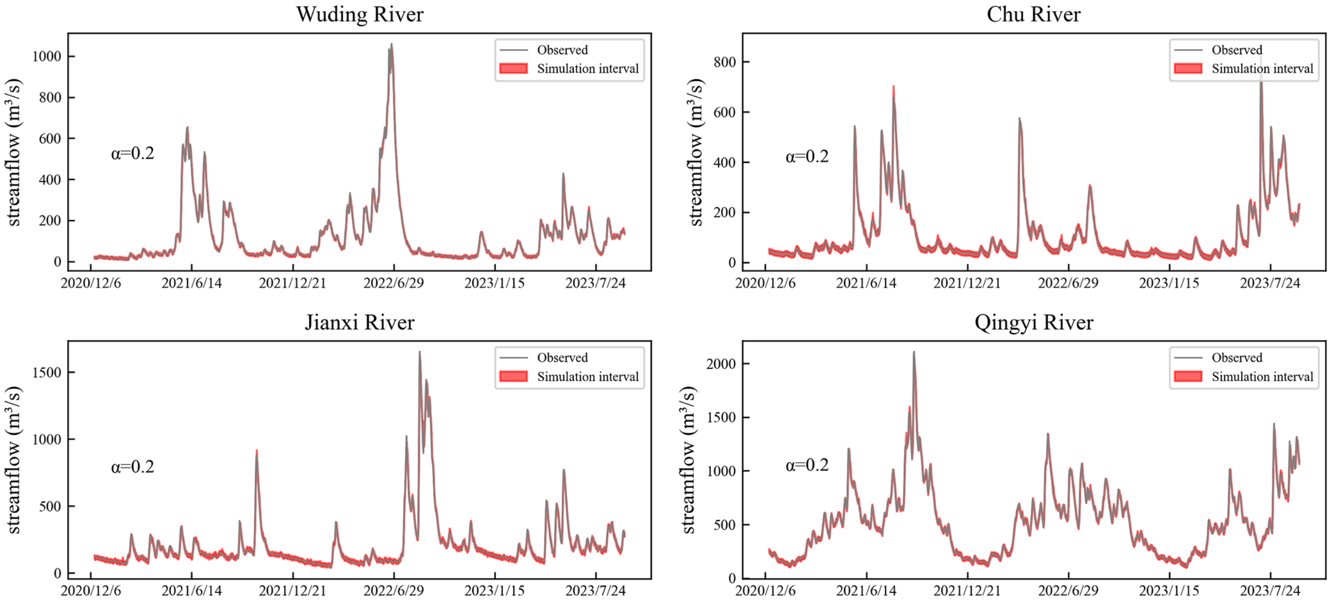

Figure 10. Interval simulation outcomes of the XAJ-TCN-GRU model at four basins for $\alpha = 0.2$.

Table 9. Evaluation of the XAJ-TCN-GRU and comparative models in interval simulation for four basins.

Basin	Model	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.2$	
		PINAW	PICP	PINAW	PICP	PINAW	PICP
Wuding	XAJ	0.094	83.75 %	0.073	75.03 %	0.048	59.12 %
	LSTM	0.072	89.38 %	0.056	81.76 %	0.037	64.83 %
	TCN	0.050	93.39 %	0.051	89.78 %	0.033	81.96 %
	GRU	0.065	92.89 %	0.041	89.98 %	0.028	79.76 %
	TCN-GRU	0.053	93.69 %	0.039	90.98 %	0.026	82.67 %
	XAJ-TCN-GRU	0.022	94.29 %	0.017	91.48 %	0.011	85.37 %
Chu	XAJ	0.244	88.88 %	0.190	79.76 %	0.125	60.52 %
	LSTM	0.133	89.98	0.104	84.67 %	0.068	72.65 %
	TCN	0.115	93.49 %	0.090	89.98 %	0.059	79.76 %
	GRU	0.109	92.69 %	0.086	87.37 %	0.056	77.66 %
	TCN-GRU	0.108	94.29 %	0.084	91.68 %	0.055	84.57 %
	XAJ-TCN-GRU	0.044	95.09 %	0.034	93.29 %	0.022	88.38 %
Jianxi	XAJ	0.186	84.97 %	0.145	74.55 %	0.095	50.20 %
	LSTM	0.088	92.48 %	0.069	89.68 %	0.045	79.16 %
	TCN	0.080	94.99 %	0.063	92.79 %	0.041	86.17 %
	GRU	0.086	94.69 %	0.067	90.78 %	0.044	82.06 %
	TCN-GRU	0.079	95.19 %	0.062	92.28 %	0.041	86.97 %
	XAJ-TCN-GRU	0.034	95.29 %	0.027	93.49 %	0.018	89.58 %
Qingyi	XAJ	0.228	73.45 %	0.178	60.72 %	0.117	42.99 %
	LSTM	0.101	86.07 %	0.079	77.25 %	0.052	56.21 %
	TCN	0.083	87.80 %	0.065	90.78 %	0.043	63.83 %
	GRU	0.080	93.09 %	0.062	88.58 %	0.041	73.95 %
	TCN-GRU	0.081	94.29 %	0.063	90.58 %	0.041	79.66 %
	XAJ-TCN-GRU	0.034	95.09 %	0.027	90.78 %	0.018	81.46 %

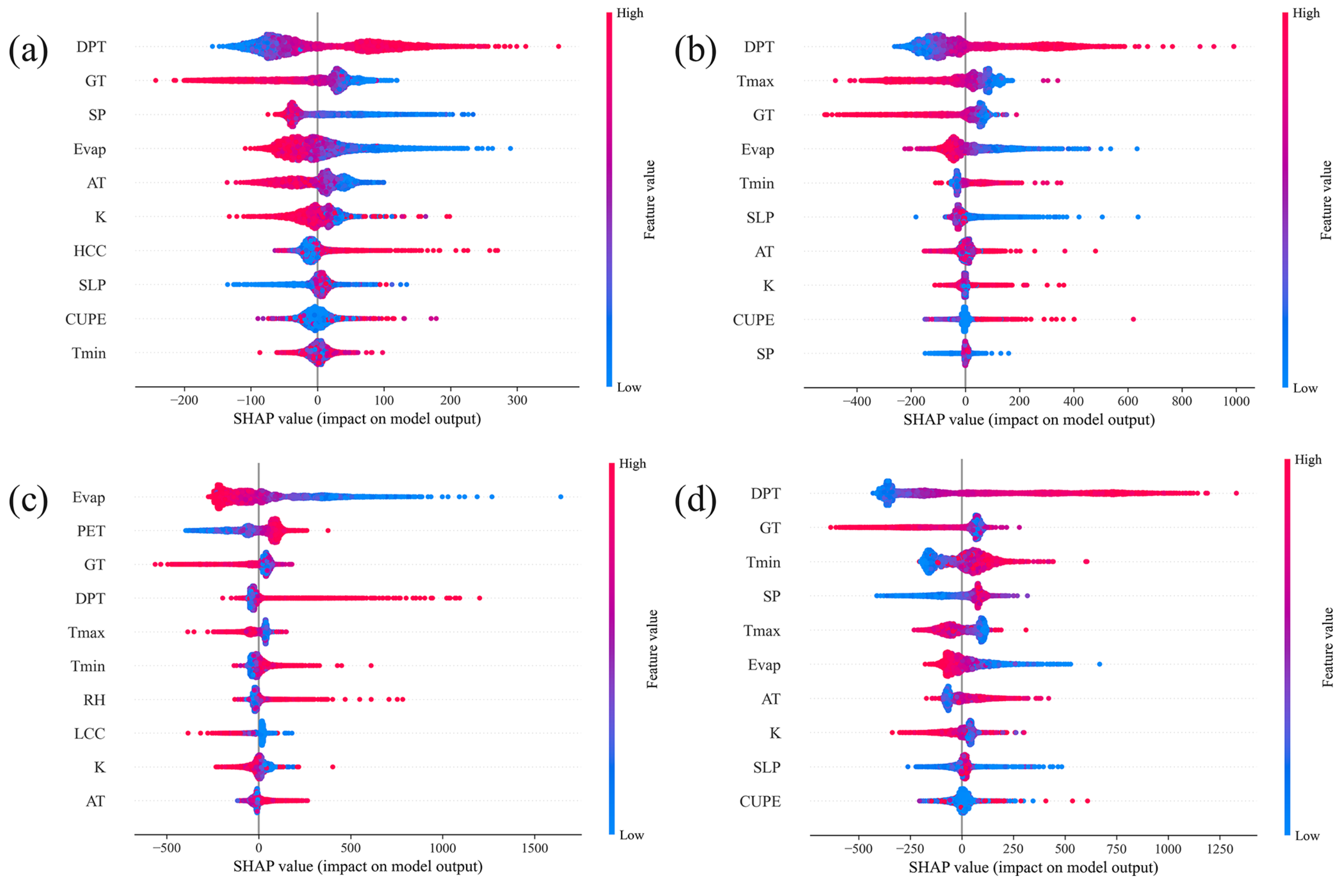

Figure 11. SHAP scores for the key external input variables at **(a)** Wuding River, **(b)** Chu River, **(c)**Jianxi River, and **(d)** Qingyi River basins.

in the Qingyi River basin, the influence of daily minimum temperature is noted, reflecting how variations in daily minimum temperatures may directly affect the cooling and condensation processes of water bodies, thereby impacting flow. We observed that evaporation also has a significant impact on streamflow in all four basins, with lower evaporation values having a greater impact on streamflow simulation. When evaporation is low, it implies that more water remains in liquid form on the surface or in water bodies, potentially leading to more streamflow.

Figure 11 shows that the importance rankings of different feature variables vary significantly across the three interpretation methods in the four basins. Dew point temperature generally ranks high in both SHAPABS and PFI across all basins, indicating its high importance in streamflow simulation. However, in FI, the importance of dew point temperature does not always rank at the top. For example, in the Chu River and Jianxi River basins, the FI method places more emphasis on other variables, such as surface pressure and evapotranspiration. This difference suggests that FI may focus more on the impact of local variables on model decisions, while PFI and SHAPABS take into account the overall contribution of variables to simulation performance.

In comparison, SHAPABS and PFI show relatively consistent rankings across most basins, with variables such as dew point temperature and ground temperature maintaining stable positions. This indicates that SHAPABS and PFI are more robust in identifying variable importance, while FI is more influenced by basin characteristics, leading to greater variation in variable rankings across basins.

These metrics offer meaningful insights by elucidating the individual contributions of each input variable to model performance. These techniques aid in identifying the most critical variables during the analysis, thereby deepening understanding of the factors driving model simulations and providing guidance for future research improvements.

Additionally, interpretability analysis for extreme events revealed significant stage differences:

1. Model component contributions: During the baseflow stage (flow $< 0\,\%$ of peak), TCN-GRU contributed 65 %–70 %, while during the peak stage (flow $> 70\,\%$ of peak), XAJ's contribution increased to 42 %–58 %. In particular, in torrential rain floods (such as Jianxi in August 2022), the XAJ streamflow generation module contributed 58 % to the physical description of the torrential rain–runoff conversion.

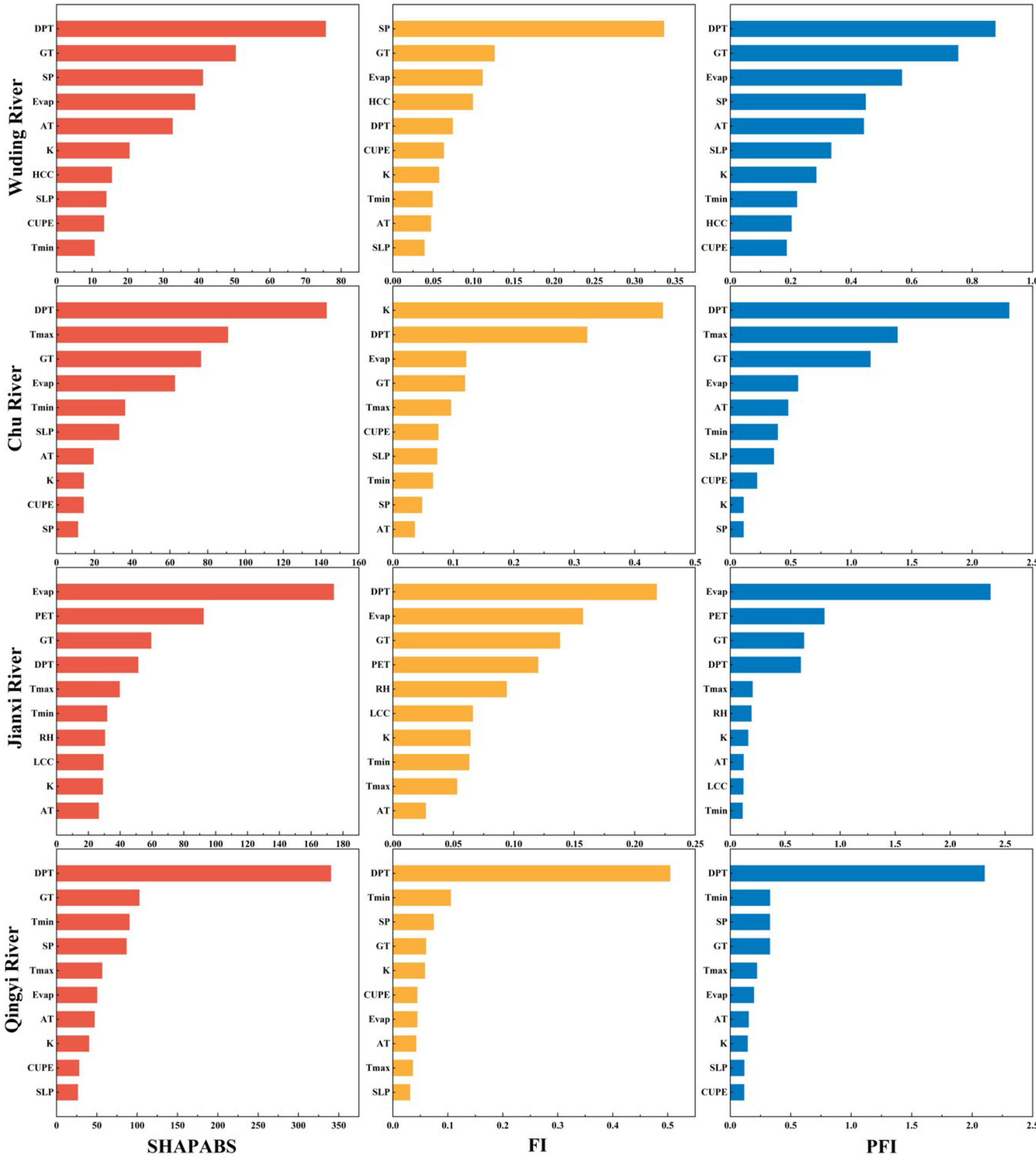

Figure 12. Bar plots showing the importance ranking of hydrometeorological variables across the four river basins based on the TCN-GRU model (evaluation metrics: SHAPABS, FI, PFI).

2. Dynamic importance of variables: During the base flow period, dew point temperature (DPT) is the primary variable (SHAP value 0.67), as high humidity affects soil water storage and evaporation; Peak precipitation (Pre) jumped to first place (SHAP value 0.82), reflecting its regulatory role in convective precipitation.

These results indicate that the advantage of the hybrid model in extreme events lies not only in improved overall accuracy but also in its ability to dynamically adapt to the dominant mechanisms of the different flood stages, providing more precise scientific basis for flood control.

5.5 Physical Mechanisms and Core Values

A further analysis of the core role of the XAJ model in the ensemble reveals that its value lies not merely in improving NSE, but in providing three key supports through physical mechanisms that deep learning models cannot replace:

First, constraint-based corrections of physical processes. Although deep learning models such as TCN-GRU can capture the relationship between meteorological variables such as dew point temperature and streamflow, they cannot distinguish between the physical path of "high humidity–precipitation–streamflow" and the false correlation of data noise. For example, during a high dew point temperature event in the Wuding River basin (arid area) as showed in Table S4, TCN-GRU overestimated streamflow (128.6 $m^3 s^{-1}$ vs. observed 104.3 $m^3 s^{-1}$) due to coincidental high cloud cover data (LCC = 0.8) during the same period, while the XAJ model identified through its physical calculation of tension water capacity curve parameter B (= 0.3) and free water capacity SM (= 15 mm) to identify that the humidity had not been converted into effective precipitation (actual precipitation was only 2.1 mm). After correction, the simulated value was 107.5 $m^3 s^{-1}$, and the hybrid model's final output was further optimized to 105.2 $m^3 s^{-1}$, demonstrating the screening effect of physical mechanisms on data correlations.

Second, process assurance in extreme scenarios. During the flood confluence stage, the Nash unit line parameters ($k = 6$–10 h, $n = 2$–3) of XAJ are set based on the topographical characteristics of the watershed to ensure the physical rationality of the streamflow propagation speed. For example, in the September 2021 flood in the Qingyi River watershed, the TCN-GRU model, due to the limited number of large flood samples in the training set (only 5 instances), simulated a confluence time 4 h shorter than the actual value, which showed in Table S5. In contrast, the XAJ model, based on $k = 8$ h calculations, accurately matched the confluence process, and after integration, the confluence time error was reduced to 1 h, with the peak simulation error decreasing from 21.3 % to 7.8 %. During the dry season, the groundwater drawdown coefficient CG (= 0.95) was used to stabilize the baseflow simulation, resulting in the hybrid model achieving an RMSE of 8.179 $m^3 s^{-1}$ in the Jianxi River basin during the dry season, a 49.0 % reduction compared to TCN-GRU (15.947 $m^3 s^{-1}$).

Third, the noise data interference resistance benchmark. XAJ model parameters (such as evaporation coefficient KC = 0.85 and deep evaporation coefficient $C = 0.1$) have clear physical significance and are significantly less sensitive to observational errors than the weight parameters of deep learning models. In a 2 % noise data test of Table S3, the increase in simulation error for the XAJ model (12.6 %) was only 44.5 % of that for TCN-GRU (28.3 %), providing a stable physical benchmark for RF integration. This resulted in a noise-induced NSE decline (0.009) for the hybrid model that was far lower than that for TCN-GRU (0.021).

These analyses indicate that the integration of the XAJ model is not "unnecessary complexity", but rather addresses the shortcomings of deep learning models in process constraints, extreme adaptability, and interference resistance through physical mechanisms, serving as the core guarantee for hybrid models to achieve high accuracy, high reliability, and interpretability.

Additionally, a further comparison of the core differences between the single GRU model and the hybrid model reveals that the value of this study lies not only in the improvement of NSE from 0.942–0.987 to 0.971–0.991 (Table 5), but also in three breakthroughs. First, a significant enhancement in the simulation capability of extreme events: in eight flood events, the RMSE of the hybrid model was on average 52.3 % lower than that of the GRU (e.g., in the August 2022 flood in the Jianxi River basin, the GRU's RMSE was 92.535 $m^3 s^{-1}$, while the hybrid model reduced it to 50.825 $m^3 s^{-1}$), which is critical for flood disaster prevention and control; Second, enhanced robustness: in noise data tests, the NSE fluctuation range (± 0.012) of the hybrid model was significantly smaller than that of GRU (± 0.022), indicating greater reliability when actual observational data contains errors; Third, integration of physical interpretability: by constraining the physical framework of the XAJ model, the hybrid model avoided GRU's overestimation of low flows during drought periods (e.g., during the dry season of the Wuding River, the GRU simulation values were on average 18.7 % higher, while the hybrid model deviation was reduced to 3.2 %). Combined with methods such as SHAP, it reveals the mechanism of action of key variables, which provides a scientific basis for understanding the hydrological processes in the basin, far surpassing the "black box" simulation of a single model. These advantages demonstrate that the complexity of the hybrid model is a necessary means to address the multidimensional challenges of hydrological simulation, possessing clear scientific value and practical significance.

To further contextualize the performance of the proposed XAJ-TCN-GRU model, we provide a qualitative discussion regarding several representative studies in the field. Kratzert et al. (2021) demonstrated the capability of LSTM networks to synergistically utilize multiple meteorological datasets through multi-basin training across over 500 basins in the CAMELS dataset, reporting a median NSE of 0.82. Li et al. (2025) explored the ensemble of a differentiable physical model (δHBV) with LSTM under diverse meteorological forcings, focusing on enhancing spatial generalization (e.g., in the PUB and PUR tests) within a large-sample framework. Solanki et al. (2025) employed machine learning methods (RF and XGB) to post-process ensembles of multiple physical hydrological models at the basin scale. It is important to note that a direct numerical comparison of NSE values between these studies and ours is not scientifically robust, as the experimental settings are fundamentally different. The aforementioned studies primarily adopted large-sample, multi-basin training and evaluation protocols, with reported metrics (often median values) reflecting performance across hundreds of heterogeneous basins (Kratzert et al., 2024). In contrast, the present study focuses on a hybrid modeling strategy that deeply integrates a physical hydrological model (XAJ) with a deep learning model (TCN-GRU) via nonlinear ensemble, and it is evaluated on four individual basins with separate training, a setting more relevant to scenarios where building a large, unified training set is challenging. Therefore, the performance of the XAJ-TCN-GRU model (achieving high NSE values in these four basins) should be interpreted within its specific context. It demonstrates the effectiveness of the proposed parallel integration and nonlinear fusion framework for daily streamflow simulation in individual basins, particularly in leveraging physical mechanisms to complement data-driven patterns. However, the methodological approaches and insights from these previous large-scale studies provide valuable perspectives for future work. In other words, whilst this study has validated the performance of the XAJ-TCN-GRU model across four Chinese river basins with distinct hydrological characteristics, the model's generalisability to a wider range of river basins and cross-regional scenarios still requires further validation through large-scale experiments in subsequent research.

6 Conclusions

This study presents an innovative hybrid model for streamflow simulation that effectively integrates the strengths of both conceptual rainfall - runoff and deep learning models. By employing nonlinear ensemble, it addresses the potential shortcomings of each approach in streamflow simulation. The results indicate that our proposed XAJ-TCN-GRU model achieves superior accuracy in streamflow simulation, significantly outperforming other benchmark models. Additionally, the robustness of the XAJ-TCN-GRU model and its exceptional performance in flood simulation and interval simulation are explored, further highlighting the model's broad potential and flexibility in applications of hydrological modelling. In order to comprehensively understand the model's capabilities, SHAPABS, FI, and PFI are introduced as evaluation tools. These tools quantify the contributing factors of different inputs on the long-term trends for streamflow across various basins, thereby enhancing the model's external interpretability.

However, this study has the following limitations in the development of the flood simulation model:

1. The input variables did not account for lagged features, overlooking the temporal dependencies and historical impacts of flow changes.

2. This study primarily relied on basic hydrological and meteorological variables for feature selection, without incorporating broader environmental factors (e.g., land use types, vegetation cover). This limitation may prevent the model from fully capturing the multidimensional factors influencing flood variability, potentially reducing simulation accuracy in complex environments.

3. The broader generalizability of this model across broader hydrological basins (e.g. the CAMELS) and in cross-regional scenarios, as well as its relative performance compared to other large-scale hydrological models, requires further evaluation in future research through more comprehensive validation using standardised data, driving factors, training protocols and evaluation settings.

Overall, this study presents a novel and efficient approach for streamflow simulation, featuring broad application prospects in watershed water management and flood disaster warning. In addition, this study highlights the key value of nonlinear sets in hydrological models, providing important references and insights for subsequent related research.

Data availability. Streamflow data from this study site were obtained from the Hydrological Yearbook of the People's Republic of China and provided by the Shanghai Qingyue Information Technology Service Centre (https://data.epmap.org/page/index, last access: 15 April 2026); meteorological data were obtained from the China Meteorological Network (https://weather.cma.cn/, last access: 18 April 2026). To ensure complete transparency and facilitate potential replication studies, the full computational framework supporting this research – including thoroughly documented model source code and all preprocessed input datasets – has been made publicly available: https://doi.org/10.5281/zenodo.22159416 (Xu, 2026).

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/hess-30-5521-2026-supplement.

Author contributions. ZW and NX designed the research. NX and XZ collected and preprocessed the data. ZW and NX conducted all the experiments and analyzed the results. JW assisted with the paper's background. NX wrote the first draft of the manuscript with contributions from ZW. WS and XC supervised the study and edited the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Financial support. This research has been supported by the Humanities and Social Sciences Fund of the Ministry of Education of China (grant no. 25YJAZH152).

Review statement. This paper was edited by Thom Bogaard and Yue-Ping Xu and reviewed by two anonymous referees.

References

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R., and Ehret, U.: Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell, Hydrol. Earth Syst. Sci., 29, 1749–1758, https://doi.org/10.5194/hess-29-1749-2025, 2025a.

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., and Ehret, U.: Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, Hydrol. Earth Syst. Sci., 29, 1277–1294, https://doi.org/10.5194/hess-29-1277-2025, 2025b.

Ahmed, A. A. M., Deo, R. C., Ghahramani, A., Feng, Q., Raj, N., Yin, Z. L., and Yang, L. S.: New double decomposition deep learning methods for river water level forecasting, Sci. Total Environ., 831, 154722, https://doi.org/10.1016/j.scitotenv.2022.154722, 2022.

Alnahit, A. O., Mishra, A. K., and Khan, A. A.: Stream water quality prediction using boosted regression tree and random forest models, Stochastic Environ. Res. Risk Assess., 36, 2661–2680, https://doi.org/10.1007/s00477-021-02152-4, 2022.

Bai, P., Liu, X. M., Liang, K., Liu, X. J., and Liu, C. M.: A comparison of simple and complex versions of the Xinanjiang hydrological model in predicting runoff in ungauged basins, Hydrol. Res., 48, 1282–1295, https://doi.org/10.2166/nh.2016.094, 2017.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chen, Z. Q., Lin, H., and Shen, G. Y.: TreeLSTM: A spatiotemporal machine learning model for rainfall-runoff estimation, J. Hydrol. Reg. Stud., 48, 101474, https://doi.org/10.1016/j.ejrh.2023.101474, 2023.

Cheng, M., Fang, F., and Pain, C. C.: Long lead-time daily and monthly streamflow forecasting using machine learning methods, J. Hydrol., 590, 125376, https://doi.org/10.1016/j.jhydrol.2020.125376, 2020.

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

Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., and Célleri, R.: Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment, Atmosphere, 12, 238, https://doi.org/10.3390/atmos12020238, 2021.

Ding, C., Wang, Z., Tan, Z., and Zhang, X.: Multi-source heterogeneous data-driven interpretable model based on transformer and kolmogorov-arnold networks for streamflow prediction, Expert Syst. Appl., 328, 132862, https://doi.org/10.1016/j.eswa.2026.132862, 2026.

Doyle, J. M., Hill, R. A., Leibowitz, S. G., and Ebersole, J. L.: Random forest models to estimate bankfull and low flow channel widths and depths across the conterminous United States, J. Am. Water Resour. Assoc., 59, 1099–1114, https://doi.org/10.1111/1752-1688.13116, 2023.

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., and Lin, Q.: Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation, J. Hydrol., 589, 125188, https://doi.org/10.1016/j.jhydrol.2020.125188, 2020.

Gebremariam, S. Y., Martin, J. F., DeMarchi, C., Bosch, N. S., Confesor, R., and Ludsin, S. A.: A comprehensive approach to evaluating watershed models for predicting river flow regimes critical to downstream ecosystem services, Environ. Model. Softw., 61, 121–134, https://doi.org/10.1016/j.envsoft.2014.07.004, 2014.

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J. H., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., and Yu, X.: Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance, Earth Space Sci., 3, 388–415, https://doi.org/10.1002/2015EA000136, 2016.

Gong, J., Yao, C., Li, Z., Chen, Y., Huang, Y., and Tong, B.: Improving the flood forecasting capability of the Xinanjiang model for small- and medium-sized ungauged catchments in South China, Nat. Hazards, 106, 2077–2109, https://doi.org/10.1007/s11069-021-04531-0, 2021.

Granata, F., Zhu, S., and Di Nunno, F.: Advanced streamflow forecasting for Central European Rivers: The Cutting-Edge Kolmogorov-Arnold networks compared to Transformers, J. Hydrol., 645, 132175, https://doi.org/10.1016/j.jhydrol.2024.132175, 2024.

Han, H. and Morrison, R. R.: 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, 2022.

Hao, F., Sun, M., Geng, X., Huang, W., and Ouyang, W.: Coupling the Xinanjiang model with geomorphologic instantaneous unit hydrograph for flood forecasting in northeast China, Int. Soil Water Conserv. Res., 3, 66–76, https://doi.org/10.1016/j.iswcr.2015.03.004, 2015.

He, Y., Wang, Z., Cheng, H., and Ding, W.: Multi-scale feature fusion and uncertainty quantification in streamflow prediction: A temporal convolutional network approach with hybrid denoising, Environ. Modell. Softw., 198, 106879, https://doi.org/10.1016/j.envsoft.2026.106879, 2026.

Hosking, J. R. M. and Wallis, J. R.: Regional frequency analysis: an approach based on L-moments, Cambridge University Press, Cambridge, UK, ISBN 9780521430452, 1997.

Hwang, Y., Clark, M. P., and Rajagopalan, B.: Use of daily precipitation uncertainties in streamflow simulation and forecast, Stoch. Env. Res. Risk Assess., 25, 957–972, https://doi.org/10.1007/s00477-011-0460-1, 2011.

Jehanzaib, M., Idrees, M. B., Kim, D., and Kim, T. W.: Comprehensive Evaluation of Machine Learning Techniques for Hydrological Drought Forecasting, J. Irrig. Drain. Eng., 147, 04021022, https://doi.org/10.1061/(ASCE)IR.1943-4774.0001575, 2021.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of

Physical Approaches and Deep Learning, Geophys. Res. Lett., 47, e2020GL088229, https://doi.org/10.1029/2020GL088229, 2020.

Jiang, X. L., Zhang, L. P., Liang, Z. M., Fu, X. L., Wang, J., Xu, J. X., Zhang, Y. C., and Zhong, Q.: Study of early flood warning based on postprocessed predicted precipitation and Xinanjiang model, Weather Clim. Extremes, 42, 100611, https://doi.org/10.1016/j.wace.2023.100611, 2023.

Kang, N., Wang, Z., Zhang, A., and Chen, H.: Improving the prediction of streamflow in large watersheds based on seasonal trend decomposition and vectorized deep learning models, Ecol. Inform., 90, 103291, https://doi.org/10.1016/j.ecoinf.2025.103291, 2025.

Katipoğlu, O. M. and Sarıgöl, M.: Improving the accuracy of rainfall-runoff relationship estimation using signal processing techniques, bio-inspired swarm intelligence and artificial intelligence algorithms, Earth Sci. Inform., 16, 3125–3141, https://doi.org/10.1007/s12145-023-01081-w, 2023.

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Gourley, J. J., and Hong, Y.: Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS, J. Hydrol., 598, 126423, https://doi.org/10.1016/j.jhydrol.2021.126423, 2021.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrol. Earth Syst. Sci., 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Kurian, C., Sudheer, K., Vema, V., and Sahoo, D.: Effective flood forecasting at higher lead times through hybrid modelling framework, J. Hydrol., 587, 124945, https://doi.org/10.1016/j.jhydrol.2020.124945, 2020.

Lei, X., Cheng, L., Ye, L., Zhang, L., Kim, J., Qin, S., and Liu, P.: Integration of the generalized complementary relationship into a lumped hydrological model for improving water balance partitioning: A case study with the Xinanjiang model, J. Hydrol., 621, 129569, https://doi.org/10.1016/j.jhydrol.2023.129569, 2023.

Leonarduzzi, E., Maxwell, R., and Mirus, B.: Numerical Analysis of the Effect of Subgrid Variability in a Physically Based Hydrological Model on Runoff, Soil Moisture, and Slope Stability, Water Resour. Res., 57, e2020WR027326, https://doi.org/10.1029/2020WR027326, 2021.

Li, P., Song, Y., Pan, M., Lawson, K., and Shen, C.: Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation, Hydrol. Earth Syst. Sci., 29, 6829–6861, https://doi.org/10.5194/hess-29-6829-2025, 2025.

Lin, K., Sheng, S., Zhou, Y., Liu, F., Li, Z., Chen, H., Xu, C., Chen, J., and Guo, S.: The exploration of a Temporal Convolutional Network combined with Encoder-Decoder framework for runoff forecasting, Hydrol. Res., 51, 1136–1149, https://doi.org/10.2166/nh.2020.100, 2020.

Liu, S., Qin, H., Liu, G., Xu, Y., Zhu, X., and Qi, X.: Runoff Forecasting of Machine Learning Model Based on Selective Ensemble, Water Resour. Manage., 37, 4459–4473, https://doi.org/10.1007/s11269-023-03566-1, 2023.

Luo, Y., Zhou, Y., Xu, H., Chen, H., Chang, F.-J., and Xu, C.-Y.: Enhancing physically-based flood forecasts through fusion of long short-term memory neural network with unscented Kalman filter, J. Hydrol., 641, 131819, https://doi.org/10.1016/j.jhydrol.2024.131819, 2024.

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

Ni, L., Wang, W., Wang, D., Singh, V. P., Yin, X., Kang, X., Tao, Y., and Gu, Z.: Improving Monthly Streamflow Prediction by Deep Learning Model With Physics-Based Rules, Hydrol. Process., 39, e70123, https://doi.org/10.1002/hyp.70123, 2025.

Oruc, H., Celen, M., Gulgen, F., Oncel, M., Vural, S., and Kilic, B.: Evaluating the effects of soil data quality on the SWAT runoff prediction Performance; A case study of Saz-Cayirova catchment, Turkey, Urban Water J., 20, 1592–1607, https://doi.org/10.1080/1573062X.2022.2056060, 2023.

Özgen-Xian, I., Kesserwani, G., Caviedes-Voullième, D., Molins, S., Xu, Z., Dwivedi, D., Moulton, J., and Steefel, C.: Wavelet-based local mesh refinement for rainfall-runoff simulations, J. Hydroinform., 22, 1059–1077, https://doi.org/10.2166/hydro.2020.198, 2020.

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., and Bateni, S.: Physics-Informed Data-Driven Model for Predicting Streamflow: A Case Study of the Voshmgir Basin, Iran, Appl. Sci., 12, 7464, https://doi.org/10.3390/app12157464, 2022.

Qiao, X., Peng, T., Sun, N., Zhang, C., Liu, Q., Zhang, Y., Wang, Y., and Nazir, M.: Metaheuristic evolutionary deep learning model based on temporal convolutional network, improved aquila optimizer and random forest for rainfall-runoff simulation and multi runoff, Expert Syst. Appl., 229, 120616, https://doi.org/10.1016/j.eswa.2023.120616, 2023.

Samantaray, S., Das, S., Sahoo, A., and Satapathy, D.: Monthly runoff prediction at Baitarani river basin by support vector machine based on Salp swarm algorithm, Ain Shams Eng. J., 13, 101732, https://doi.org/10.1016/j.asej.2022.101732, 2022.

Shao, P., Feng, J., Lu, J., Zhang, P., and Zou, C.: Data-driven and knowledge-guided denoising diffusion model for flood forecasting, Expert Syst. Appl., 244, 122908, https://doi.org/10.1016/j.eswa.2023.122908, 2024.

Solanki, H., Vegad, U., Kushwaha, A., and Mishra, V.: Improving Streamflow Prediction Using Multiple Hydrological Models and Machine Learning Methods, Water Resour. Res., 61, e2024WR038192, https://doi.org/10.1029/2024WR038192, 2025.

Song, J., Meng, H., Kang, Y., Zhu, M., Zhu, Y., and Zhang, J.: A method for predicting water quality of river basin based on OVMD-GAT-GRU, Stoch. Environ. Res. Risk Assess., 38, 339–356, https://doi.org/10.1007/s00477-023-02584-0, 2024.

Song, Y., Qin, S., Qu, J., and Liu, F.: The forecasting research of early warning systems for atmospheric pollutants: A case in Yangtze River Delta region, Atmos. Environ., 118, 58–69, https://doi.org/10.1016/j.atmosenv.2015.06.032, 2015.

Sun, P., Wang, J., and Yan, Z.: Ultra-short-term wind speed prediction based on TCN-MCM-EKF, Energy Rep., 11, 2127–2140, https://doi.org/10.1016/j.egyr.2024.01.058, 2024.

Tan, Z., Li, H., Song, Q., Wang, Z., and Cao, Y.: Synergistic optimization and interaction evaluation of water-energy-food-ecology nexus under uncertainty from the perspective of urban agglomeration, Sustain. Cities Soc., 124, 106291, https://doi.org/10.1016/j.scs.2025.106291, 2025.

Thébault, C., Perrin, C., Andréassian, V., Thirel, G., Legrand, S., and Delaigue, O.: Multi-model approach in a variable spatial framework for streamflow simulation, Hydrol. Earth Syst. Sci., 28, 1539–1566, https://doi.org/10.5194/hess-28-1539-2024, 2024.

Vilaseca, F., Castro, A., Chreties, C., and Gorgoglione, A.: Assessing influential rainfall-runoff variables to simulate daily streamflow using random forest, Hydrol. Sci. J., 68, 1738–1753, https://doi.org/10.1080/02626667.2023.2232356, 2023.

Wang, H., Qin, H., Liu, G., Liu, S., Qu, Y., Wang, K., and Zhou, J.: A novel feature attention mechanism for improving the accuracy and robustness of runoff forecasting, J. Hydrol., 618, 129200, https://doi.org/10.1016/j.jhydrol.2023.129200, 2023.

Wang, J. and Dong, Y.: An interpretable deep learning multi-dimensional integration framework for exchange rate forecasting based on deep and shallow feature selection and snapshot ensemble technology, Eng. Appl. Artif. Intell., 133, 108282, https://doi.org/10.1016/j.engappai.2024.108282, 2024.

Wang, Z., Xu, N., Bao, X., Wu, J., and Cui, X.: Spatio-temporal deep learning model for accurate streamflow prediction with multi-source data fusion, Environ. Model. Softw., 178, 106091, https://doi.org/10.1016/j.envsoft.2024.106091, 2024.

Wang, Z., Ma, C., Tan, Z., and Wu, T.: Low-carbon development pathways for the water-energy-food-carbon nexus in the Yangtze river economic Belt: Insights from coupling coordination and obstacle degree analysis, J. Clean. Prod., 523, 146399, https://doi.org/10.1016/j.jclepro.2025.146399, 2025a.

Wang, Z., Zhu, Z., Luan, H., and Wu, T.: Multi-objective optimal scheduling of cascade reservoirs in complex basin systems: Case study of the Jinsha River-Yalong River confluence basin in China, J. Hydrol. Reg. Stud., 58, 102240, https://doi.org/10.1016/j.ejrh.2025.102240, 2025b.

Wei, X., Wang, G., Schmalz, B., Hagan, D., and Duan, Z.: Evaluation of Transformer model and Self-Attention mechanism in the Yangtze River basin runoff prediction, J. Hydrol. Reg. Stud., 47, 101438, https://doi.org/10.1016/j.ejrh.2023.101438, 2023.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, ACM Comput. Surv., 55, 66, https://doi.org/10.1145/3514228, 2023.

Wu, J., Wang, Z., Dong, J., Cui, X., Tao, S., and Chen, X.: Robust Runoff Prediction With Explainable Artificial Intelligence and Meteorological Variables From Deep Learning Ensemble Model, Water Resour. Res., 59, e2023WR035676, https://doi.org/10.1029/2023WR035676, 2023.

Xu, N.: Code and Data for XAJ-TCN-GRU, Zenodo [data set], https://doi.org/10.5281/zenodo.22159416, 2026.

Xu, C., Chen, Y., Wang, D., Zhao, Y., Hou, Y., Zhu, Y., and Shen, Q.: Uncertainty and driving factor analysis of streamflow forecasting for closed-basin and interval-basin: Based on a probabilistic and interpretable deep learning model, J. Hydrol. Reg. Stud., 60, 102483, https://doi.org/10.1016/j.ejrh.2025.102483, 2025.

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., and Ran, G.: Deep transfer learning based on transformer for flood forecasting in data-sparse basins, J. Hydrol., 625, 129956, https://doi.org/10.1016/j.jhydrol.2023.129956, 2023.

Xu, Y., Liu, T., Fang, Q., Du, P., and Wang, J.: Crude oil price forecasting with multivariate selection, machine learning, and a nonlinear combination strategy, Eng. Appl. Artif. Intell., 139, 109510, https://doi.org/10.1016/j.engappai.2024.109510, 2025.

Yao, Z., Wang, Z., Wang, D., Wu, J., and Chen, L.: An ensemble CNN-LSTM and GRU adaptive weighting model based improved sparrow search algorithm for predicting runoff using historical meteorological and runoff data as input, J. Hydrol., 625, 129977, https://doi.org/10.1016/j.jhydrol.2023.129977, 2023.

Zhang, Y., Ye, A., Analui, B., Nguyen, P., Sorooshian, S., Hsu, K., and Wang, Y.: Comparing quantile regression forest and mixture density long short-term memory models for probabilistic post-processing of satellite precipitation-driven streamflow simulations, Hydrol. Earth Syst. Sci., 27, 4529–4550, https://doi.org/10.5194/hess-27-4529-2023, 2023.

Zhao, R. J.: The Xinanjiang model applied in China, J. Hydrol., 135, 371–381, https://doi.org/10.1016/0022-1694(92)90096-E, 1992.

Zhao, Y., Huang, Y., Wang, Z., and Liu, X.: Carbon futures price forecasting based on feature selection, Eng. Appl. Artif. Intell., 135, 108646, https://doi.org/10.1016/j.engappai.2024.108646, 2024.

Zheng, Z., Ali, M., Jamei, M., Xiang, Y., Karbasi, M., Yaseen, Z., and Farooque, A.: Design data decomposition-based reference evapotranspiration forecasting model: A soft feature filter based deep learning driven approach, Eng. Appl. Artif. Intell., 121, 105984, https://doi.org/10.1016/j.engappai.2023.105984, 2023.

Zhu, S., Wei, J., Zhang, H., Xu, Y., and Qin, H.: Spatiotemporal deep learning rainfall-runoff forecasting combined with remote sensing precipitation products in large scale basins, J. Hydrol., 616, 128727, https://doi.org/10.1016/j.jhydrol.2022.128727, 2023.

Zhu, Z., Li, H., Wang, Z., Zhang, X., and Tan, Z.: Integration of deep learning and improved multi-objective algorithm to optimize cascade reservoirs operation with consideration of ecological dissolved oxygen needs, J. Hydrol., 667, 134899, https://doi.org/10.1016/j.jhydrol.2025.134899, 2026a.

Zhu, Z., Wu, H., Wang, Z., Zhang, X., Kong, J., Liu, C., Tan, Z., and Liu, Q.: Physics-AI Synergized Optimization-Learning-Simulation Framework for Robust Cascade Reservoir Scheduling Under Future Hydrological Uncertainty, Water Resour. Res., 62, e2025WR042149, https://doi.org/10.1029/2025WR042149, 2026b.

Zuo, G., Luo, J., Wang, N., Lian, Y., and He, X.: Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting, J. Hydrol., 585, 124776, https://doi.org/10.1016/j.jhydrol.2020.124776, 2020.