

Article

Improving Streamflow Simulation with Coupled Conceptual and Data-Driven Hydrologic Modelling: A Case Study in the Woluo River Basin

Zhaohui He ¹, Tong Duan ^{1,*}, Penghao Shao ^{1,*}, Haoran Hao ^{2,3} and Ningpeng Dong ² 

¹ Yalong River Hydropower Development Company, Ltd., Chengdu 610051, China

² State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100048, China

³ State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin 300072, China

* Correspondence: duantong@ylhdc.com.cn (T.D.); shaopenghao@ylhdc.com.cn (P.S.)

Abstract

Accurate streamflow forecasting is critical for water resource management and disaster mitigation, yet conceptual models often suffer from systematic biases while machine learning lacks physical interpretability. This study proposes a coupled XAJ-LSTM model to integrate the mechanistic strengths of the Xin'anjiang (XAJ) model with the predictive capability of Long Short-Term Memory (LSTM) networks. Applied in the Woluo River basin using an 11-year daily record from 2012 to 2022, the framework was evaluated with leave-one-year-out validation. Results indicate that XAJ-LSTM achieved the highest full-period NSE of 0.885 and the lowest full-period RMSE of 19.48 m³/s, compared with NSE values of 0.867 for XAJ and 0.724 for LSTM. XAJ-LSTM also performed best during the flood period, with NSE of 0.820 and RMSE of 28.81 m³/s. The SHAP framework indicated that XAJ-simulated flow and precipitation remained the dominant predictors, suggesting that the LSTM relied strongly on the conceptual-model baseline and rainfall information. These findings indicate that XAJ-LSTM can improve selected full-period and flood-period diagnostics for daily streamflow simulation in the Woluo River Basin, while further tests with longer records and additional catchments are needed before broader application.

Keywords: hydrologic modelling; LSTM; Xin'anjiang; SHAP; leave-one-year-out validation



Academic Editors: Shimin Tian and Jia Jia

Received: 22 May 2026

Revised: 8 July 2026

Accepted: 14 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Precise prediction and modeling of river discharge are fundamental to optimizing water resource administration, mitigating flood risks, and preserving ecological systems. Such predictive capabilities are vital for sustainable water practices and societal resilience to severe hydrological extremes [1,2]. Conventional frameworks, including lumped approaches like the Xinanjiang (XAJ) and HBV models, as well as distributed systems such as CLHMS and VIC, are fundamentally driven by physical laws to simulate rainfall-runoff relationship and soil moisture behaviors [3–6]. The popularity of these models over the past decades stems largely from their mechanistic interpretability [7–9]. Nevertheless, the predictive accuracy of these physically based models is frequently constrained by uncertainties inherent in parameter calibration [10–12]. Simplified representations of complex hydrological processes further introduce systematic errors [13,14].

On the other hand, data-driven approaches like artificial neural networks (ANNs) [15,16], support vector machines (SVMs) [17], and long short-term memory (LSTM) networks [18] are highly effective at finding complex, nonlinear relationships in hydrological data without needing explicit physical rules. They also show good results when processing massive amounts of data [19–22]. For example, Li et al. [23] built an LSTM system combined with physical models that accurately predicted floods in the Jinhua River Basin, while Huang et al. [24] demonstrated that deep learning can merge various remote sensing datasets to generate detailed hydrological outputs [25]. Despite these advantages, such models cannot explain the physical processes behind their results, and they often perform poorly when predicting conditions that fall outside their original training data [10,26,27].

Coupled approaches provide an effective alternative by combining the physical reliability of traditional models with the predictive capabilities of machine learning [8,28–31]. For example, Wei [32] developed the PhyFill algorithm to restore global soil moisture records by integrating deep learning with physical processes. Similarly, Dong et al. [11] evaluated different methods for linking SWMM and LISFLOOD-FP models, highlighting the balance between simulation speed and accuracy. Recent differentiable modeling and physics-informed learning studies also show that physical constraints can be embedded more directly within learning architectures [33–35]. Nevertheless, many hydrological applications still adopt loose coupling or postprocessing structures because they are easier to implement when observations are limited and when the physical model is already operational.

Over recent decades, the development of LSTM networks has greatly improved hydrological modeling because these systems can effectively recognize long-term patterns in data [18,26,36]. For instance, Li et al. [23] showed that combining LSTM with physical models improves flood forecasting by utilizing mixed datasets. Despite these successes, LSTM networks cannot explain the physical causes of their outputs. This drawback restricts their application in studies that require a clear understanding of mechanisms [20,37]. To shed light on this, our study introduces a combined XAJ-LSTM framework, which is designed to improve prediction accuracy and stability while maintaining physical meaning [38].

This research has three main aims. First, it develops an XAJ-LSTM coupling framework that feeds XAJ-simulated streamflow, precipitation, and temperature into an LSTM network for daily streamflow simulation. Second, it evaluates whether this structure improves model performance relative to standalone XAJ and LSTM models using leave-one-year-out validation with flood-period and full-period diagnostics. Third, it applies SHAP analysis to diagnose the contribution of each input feature to shed light on the mechanisms on the improvements induced by model coupling.

2. Study Area and Data

2.1. Study Area

The Woluo River, also known as the Jiami River, has a total length of 166.1 km and is the largest tributary of the Litang River, which is itself a first-order tributary of the Yalong River. Originating on the western slope of Bailin Mountain in Yanyuan County, it flows into the Litang River in Muli County. The Jiami Hydrological Station is located in the middle reaches of the Woluo River, controlling a catchment area of 4017 km² at an elevation of 2100 m. The basin is mountainous, with steep slopes, valley channels, and seasonally concentrated rainfall–runoff response. Vegetation is mainly alpine shrubland, grassland, and forest patches, and soil storage conditions vary between hillslopes and valley areas. These characteristics support the use of a soil-moisture-based XAJ model while also requiring caution for flow recession and spatial heterogeneity. The geographical location and river network of the Woluo River Basin are illustrated in Figure 1.

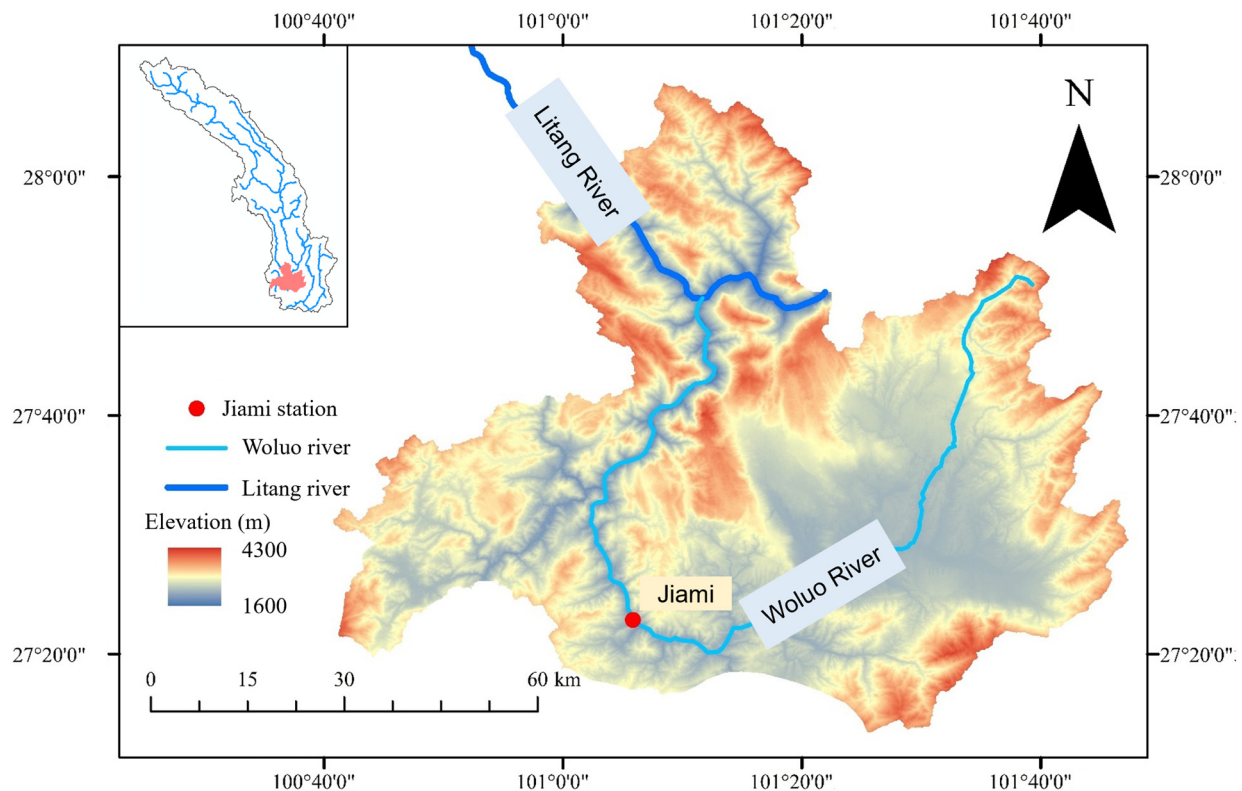


Figure 1. Location and river system map of the Woluo River Basin.

2.2. Data

The streamflow data used in this study were sourced from the Jiami Hydrological Station for an 11-year period from 2012 to 2022. The rainfall and temperature data employed for model calibration and training were daily observational data provided by the Chengdu Meteorological Bureau of Sichuan Province. Data before 2012 were not used because consistent daily meteorological forcing required by the coupled model was not available.

3. Methods

3.1. Deep Learning Model: Long Short-Term Memory Network (LSTM)

The Long Short-Term Memory (LSTM) network is an advanced type of Recurrent Neural Network (RNN). The primary distinction between LSTM and traditional RNN lies in the inability of RNN to effectively handle the long-term temporal dependency, which is prevalent in processing large datasets and is particularly common in hydrological science. In contrast, LSTM, due to its internal mechanisms, is better equipped to address this issue [26].

The LSTM architecture consists of a forget gate, an input gate, a cell state, and an output gate.

(1) The forget gate determines which information should be discarded. Specifically:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (1)$$

(2) The input gate determines which information should be recorded in the cell state. Specifically:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3)$$

(3) Update the cell state. Specifically:

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (4)$$

(4) The output gate determines which information is carried forward to the next neuron. Specifically:

$$Q_t = \sigma(W_Q[h_{t-1}, x_t] + b_Q) \quad (5)$$

$$h_t = Q_t \tanh(C_t) \quad (6)$$

In the equations, W_f, W_i, W_c, W_Q are weight matrices; b_f, b_i, b_c, b_Q are bias vectors; h_t, h_{t-1} are the input and output of the current module and the previous module, respectively; C_t, C_{t-1} are the cell state of the current module and the previous module, respectively; and $\sigma(\cdot)$ is the sigmoid function.

3.2. Process-Based Model: Xin'anjiang (XAJ) Model

The Xin'anjiang (XAJ) model is a widely recognized lumped hydrological model designed for streamflow simulation and forecasting in humid and semi-humid regions [39]. The XAJ model simulates the rainfall-runoff process by representing evapotranspiration, soil moisture dynamics, runoff generation, and routing in a conceptually simplified but physically interpretable manner. It was selected in this study because the Woluo River Basin has seasonally concentrated rainfall and soil-water storage controls that are consistent with the runoff generation assumptions of XAJ.

One strength of the XAJ model is its clear physical interpretability, which helps explain the processes driving runoff generation. However, its performance is sensitive to parameter calibration, and the lumped structure may oversimplify complex hydrological processes in basins with strong spatial heterogeneity. The calibrated XAJ parameters and their search ranges were as follows. WM1, WM2, and WM3 represent the upper-layer, lower-layer, and deep-layer tension water capacities, with ranges of 10 to 80, 10 to 120, and 10 to 120, respectively. K is the evapotranspiration coefficient, with a range of 0 to 1. Cd is the deep-layer evapotranspiration coefficient, with a range of 0.05 to 0.5. SM is the mean free water storage capacity used to partition surface runoff, with a range of 5 to 50. KSS and KG are the interflow and groundwater outflow coefficients, both with ranges of 0.2 to 0.85. The routing parameters uhn and uhk are Nash unit hydrograph shape parameters used to determine the catchment routing response, both with ranges of 0 to 10. KKSS and KKGF are the recession coefficients for interflow and groundwater flow in slope routing, both with ranges of 0.98 to 0.9999.

3.3. Coupled Model: XAJ-LSTM Coupled Model

The coupled XAJ-LSTM model was implemented as a loose coupling and postprocessing framework rather than a dynamically integrated hydrological model (Figure 2). Within each leave-one-year-out fold, the XAJ model was calibrated using all years except the held-out year and was then used to generate daily simulated streamflow. The LSTM network received three groups of predictors over a moving temporal window: daily precipitation, daily temperature, and the streamflow simulated by the XAJ model. Observed daily streamflow at Jiami Station was used as the prediction target.

The coupling procedure consisted of four steps. First, all meteorological and streamflow variables were quality controlled and normalized using parameters estimated from the training years of the current fold. Second, the XAJ model generated a physically based baseline streamflow series for each day. Third, the LSTM learned the nonlinear relationship between the predictor window and observed streamflow, with the XAJ output serving as

a physically informed baseline predictor. Fourth, the trained LSTM was applied to the held-out year, and the resulting XAJ-LSTM output was compared with standalone XAJ and LSTM simulations using identical evaluation metrics.

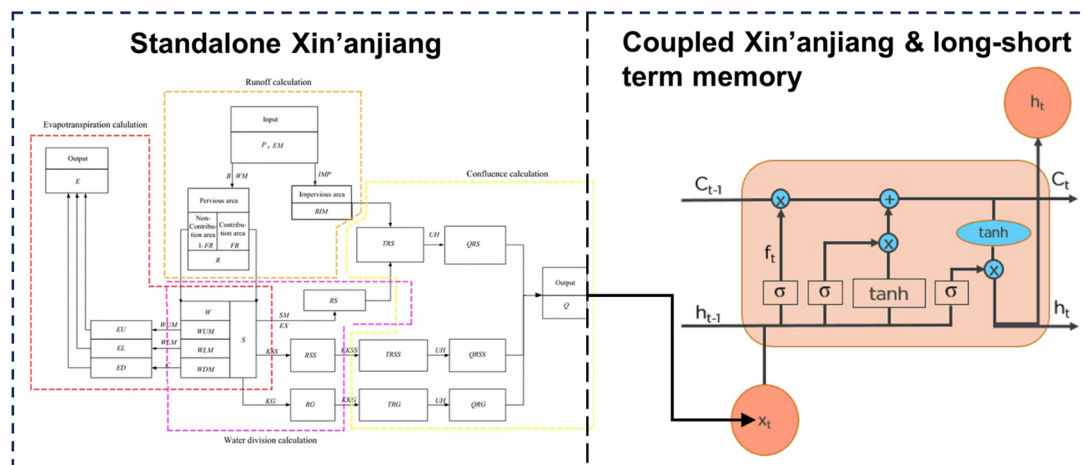


Figure 2. Structure of the XAJ-LSTM model.

This coupling strategy was selected because the available basin record is still limited for deep learning, which restricts the feasibility of calibrating a more complex process-level neural hydrological model. The two components do not exchange states during simulation, and the framework should therefore be interpreted as a correction-oriented coupled model rather than a fully integrated physical machine-learning system.

3.4. Sensitivity Analysis with the SHAP Model

This study leverages the Shapley Additive Explanations (SHAP) framework as a diagnostic attribution method for the coupled XAJ-LSTM model. SHAP quantifies how individual input features contribute to fitted model predictions for a given sample, but it does not by itself establish hydrological causality. For a given sample, let x_i denote the i -th feature of that sample, $f(x)$ represent the model prediction for the sample, and y_{base} signify the baseline prediction, typically the expected output across all samples. The SHAP value for a feature quantifies its marginal contribution to the prediction, and the relationship can be expressed as:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + f(x_{i3}) + \dots + f(x_{ij})$$

Here, $f(x_{ij})$ denotes the SHAP value of the j -th feature for the i -th sample, encapsulating its unique contribution to the prediction y_i . The magnitude of $f(x_{ij})$, taken as its absolute value, indicates the strength of the feature influence on the model output. A larger absolute SHAP value suggests that the corresponding feature has a more substantial impact on the prediction, while a smaller value implies a weaker effect.

The sign of the SHAP value provides further information about the direction of feature attribution in the fitted model. A positive SHAP value indicates that the feature increases the predicted value relative to the baseline, whereas a negative SHAP value indicates that the feature decreases the predicted value. In the context of streamflow simulation, these signs describe model attribution patterns rather than independent causal effects.

3.5. Data Processing

To ensure the reliability and consistency of the input data for the XAJ-LSTM coupled model, as well as the standalone XAJ and LSTM models, a comprehensive data standard-

ization and preprocessing procedure was implemented to mitigate issues related to varying scales, missing values, and noise in the hydrometeorological datasets. The following paragraphs detail the standardization and preprocessing techniques applied to the streamflow, precipitation, and temperature data from Jiami Station in the Woluo River basin.

Prior to standardization, a thorough quality assessment was conducted on the daily streamflow, precipitation, and temperature data collected from Jiami Station and the Chengdu Meteorological Bureau for the period from 2012 to 2022. The assessment involved checking for missing values, outliers, and inconsistencies. Missing data points, which accounted for less than 2% of the total dataset, were complemented using linear interpolation for streamflow and precipitation. For temperature, nearest-neighbor interpolation was applied. Outliers, defined as values exceeding three standard deviations from the mean, were identified and cross-verified with regional meteorological records. Approximately 0.5% of the data points were flagged as outliers and replaced with interpolated values to maintain dataset integrity.

To address the issue of differing scales among input variables (e.g., precipitation in millimeters, temperature in degrees Celsius, and streamflow in cubic meters per second), normalization was applied to standardize the data. The min-max normalization technique was used to rescale all input features to a common range of [0, 1], as defined by the following equation:

$$x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$$

where x is the original value, x_{min} and x_{max} are the minimum and maximum values of the variable in the training years, respectively, and x_{norm} is the normalized value. This normalization ensured that variables with larger numerical ranges, such as streamflow, did not disproportionately influence model training. The normalization parameters were computed using the training years only and then applied to the held-out period to prevent data leakage.

3.6. Model Training and Validation

The XAJ model was employed to simulate daily streamflow at Jiami Station using daily precipitation and temperature data from two sub-basins. Model validation followed a leave-one-year-out design (LOYO). For each test year from 2012 through 2022, that calendar year was held out, and the remaining ten years were used for model calibration or training. XAJ parameter optimization was performed using the Particle Swarm Optimization (PSO) algorithm, with the Nash-Sutcliffe Efficiency coefficient (NSE) serving as the objective function.

The standalone LSTM model used daily precipitation and temperature sequences as inputs, while the XAJ-LSTM model additionally used the XAJ-simulated streamflow sequence $q = q[1], q[2], \dots, q[N]$. For each timestep n , the input features are vectors spanning the past n_{seq} days and correspond to observed daily streamflow $o[n]$ at Jiami Station as the prediction target. The hyperparameter search considered learning rate from 0 to 1, integer epoch numbers, integer hidden-layer or hidden-unit settings, integer batch size, dropout from 0 to 1, optimizer type, activation function, and loss function. The selected standalone LSTM setting used learning rate 0.005, 50 epochs, hidden size 96, batch size 128, Adam optimizer, linear output activation, dropout 0.404810, NSE loss, and sequence length 45 days. The selected XAJ-LSTM setting used learning rate 0.005, 50 epochs, hidden size 12, batch size 128, Adam optimizer, linear output activation, dropout 0.24, NSE loss, and sequence length 75 days. In each LOYO fold, model fitting, normalization, and training were conducted without using observations from the held-out year.

3.7. Model Performance Evaluation Metrics

This study employed the Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency coefficient (NSE), Kling-Gupta Efficiency coefficient (KGE), and alpha-NSE to evaluate model prediction performance from different perspectives. The Sim-to-Obs ratio was also calculated to evaluate cumulative volume bias. Based on the seasonal rainfall–runoff regime of the Woluo River Basin, June to October was defined as the flood period. Model performance was reported for the flood period and the entire LOYO validation period. The calculation formulas for these evaluation metrics are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_i - q_i)^2} \quad (7)$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n (Q_i - q_i)^2}{\sum_{i=1}^n (Q_i - \bar{Q})^2} \quad (8)$$

$$\text{KGE} = 1 - \sqrt{(CC - 1)^2 + (BR - 1)^2 + (RV - 1)^2} \quad (9)$$

$$\text{AlphaNSE} = \frac{\sigma_{sim}}{\sigma_{obs}} \quad (10)$$

where Q_i represents the observed streamflow value, q_i denotes the predicted streamflow value, and the overbar indicates the mean of observed streamflow values. In the KGE formulation, CC represents the correlation coefficient, BR stands for the bias ratio, and RV indicates the relative variability. The Sim-to-Obs ratio was calculated as total simulated streamflow divided by total observed streamflow and multiplied by 100%. A value above 100% indicates overestimation, whereas a value below 100% indicates underestimation. Alpha-NSE was calculated as the ratio between simulated and observed standard deviation. Values close to 1 indicate that the simulated variability amplitude is similar to the observed variability, whereas values above or below 1 indicate overestimated or underestimated variability.

Because model validation was repeated over 11 held-out calendar years in one basin, model comparisons were interpreted using paired period-specific diagnostics, and annual LOYO behavior [40]. Formal statistical significance testing was not conducted for NSE and KGE because the experiment involved a single catchment and a limited number of annual folds.

To quantify uncertainty in the reported LOYO performance metrics, 95% confidence intervals were estimated using non-parametric bootstrap resampling at the held-out-year level. In each bootstrap replicate, 11 validation years were sampled with replacement from 2012 to 2022, the corresponding daily validation predictions were pooled, and RMSE, NSE, KGE, and alpha-NSE were recalculated. The 2.5th and 97.5th percentiles from 5000 bootstrap replicates were reported as the 95% confidence intervals.

4. Results

4.1. Performance of Streamflow Simulation

The LOYO simulation results of the XAJ, LSTM, and XAJ-LSTM models at Jiami Station are presented in Table 1.

Using the combined LOYO validation results, the XAJ model produced a total simulated streamflow of 158,561.7 m³/s day for the entire period, equivalent to 100.4% of the observed volume. Its full-period RMSE, NSE, and KGE were 21.08 m³/s, 0.867, and 0.911, respectively. During the flood period, XAJ produced a total simulated streamflow of

135,982.6 m³/s day, equivalent to 99.1% of observations. Its flood-period RMSE, NSE, and KGE were 31.68 m³/s, 0.782, and 0.861.

Table 1. Metrics of LOYO streamflow simulation by the XAJ, LSTM, and XAJ-LSTM models.

Model	Period	Simulated Total Streamflow (m ³ /s day)	Sim-to-Obs Ratio (%)	RMSE (m ³ /s) [95% CI]	NSE [95% CI]	KGE [95% CI]	Alpha-NSE [95% CI]
XAJ	Flood period	135,982.6	99.1	31.68 [24.83, 38.23]	0.782 [0.721, 0.841]	0.861 [0.780, 0.890]	0.922 [0.810, 1.041]
	Entire period	158,561.7	100.4	21.08 [16.60, 25.38]	0.867 [0.836, 0.901]	0.911 [0.833, 0.931]	0.944 [0.852, 1.040]
LSTM	Flood period	137,048.0	99.9	45.81 [37.96, 53.40]	0.545 [0.453, 0.621]	0.757 [0.690, 0.799]	0.952 [0.852, 1.088]
	Entire period	158,638.9	100.3	30.31 [25.15, 34.92]	0.724 [0.673, 0.773]	0.854 [0.787, 0.879]	0.968 [0.875, 1.076]
XAJ-LSTM	Flood period	140,447.3	102.4	28.81 [23.44, 33.72]	0.820 [0.785, 0.853]	0.863 [0.791, 0.884]	0.898 [0.808, 0.969]
	Entire period	164,914.7	104.5	19.48 [15.90, 22.39]	0.885 [0.871, 0.908]	0.893 [0.837, 0.897]	0.923 [0.851, 0.996]

Note: Values in brackets for RMSE, NSE, KGE, and alpha-NSE are year-level bootstrap 95% confidence intervals.

The standalone LSTM model produced a total simulated streamflow of 158,638.9 m³/s day for the entire period, equivalent to 100.3% of observations. Its full-period RMSE, NSE, and KGE were 30.31 m³/s, 0.724, and 0.854. During the flood period, LSTM produced a total simulated streamflow of 137,048.0 m³/s day, equivalent to 99.9% of observations. The flood-period RMSE increased to 45.81 m³/s and the NSE decreased to 0.545, indicating weaker performance during high-flow conditions.

The XAJ-LSTM model produced a total simulated streamflow of 164,914.7 m³/s day, equivalent to 104.5% of observations. Its full-period RMSE, NSE, and KGE were 19.48 m³/s, 0.885, and 0.893. The coupled model achieved the best flood-period RMSE and NSE among the three models, with RMSE of 28.81 m³/s and NSE of 0.820. Its flood-period simulated streamflow was 140,447.3 m³/s day, equivalent to 102.4% of observations.

The LOYO results show that XAJ-LSTM improved full-period NSE and RMSE relative to XAJ and LSTM, and it showed the strongest flood-period NSE. However, XAJ retained the best full-period KGE and a smaller full-period volume bias. Therefore, the coupled model should be interpreted as improving selected diagnostics rather than uniformly outperforming XAJ across all metrics.

The bracketed values in Table 1 provide the corresponding year-level bootstrap 95% confidence intervals for RMSE, NSE, KGE, and alpha-NSE. These intervals support a cautious interpretation of the model comparison because several diagnostics show overlapping uncertainty ranges across models.

In terms of RMSE, XAJ-LSTM reduced full-period RMSE by 7.6% relative to XAJ and 35.7% relative to LSTM. During the flood period, XAJ-LSTM reduced RMSE by 9.1% relative to XAJ and 37.1% relative to LSTM. These reductions indicate that the coupled model improved squared-error behavior for the full-period and flood-period evaluations.

Regarding NSE, XAJ-LSTM recorded the highest full-period value of 0.885 and the highest flood-period value of 0.820. XAJ had the best full-period KGE of 0.911 and a flood-period KGE of 0.861, which was close to the XAJ-LSTM value of 0.863. The LSTM model had lower NSE than XAJ and XAJ-LSTM in both reported periods.

Regarding alpha-NSE, values for all three models were close to 1 over the entire period, with XAJ at 0.944, LSTM at 0.968, and XAJ-LSTM at 0.923. During the flood period, alpha-NSE values were 0.922 for XAJ, 0.952 for LSTM, and 0.898 for XAJ-LSTM, indicating that all three models slightly underestimated flood-period variability amplitude.

Figure 3 depicts observed streamflow and simulated streamflow across the LOYO validation years. The XAJ and XAJ-LSTM models both reproduced the main rising and falling streamflow patterns better than the standalone LSTM model. The XAJ-LSTM hydrograph showed improved agreement during several high-flow periods, although some peak timing and magnitude errors remained.

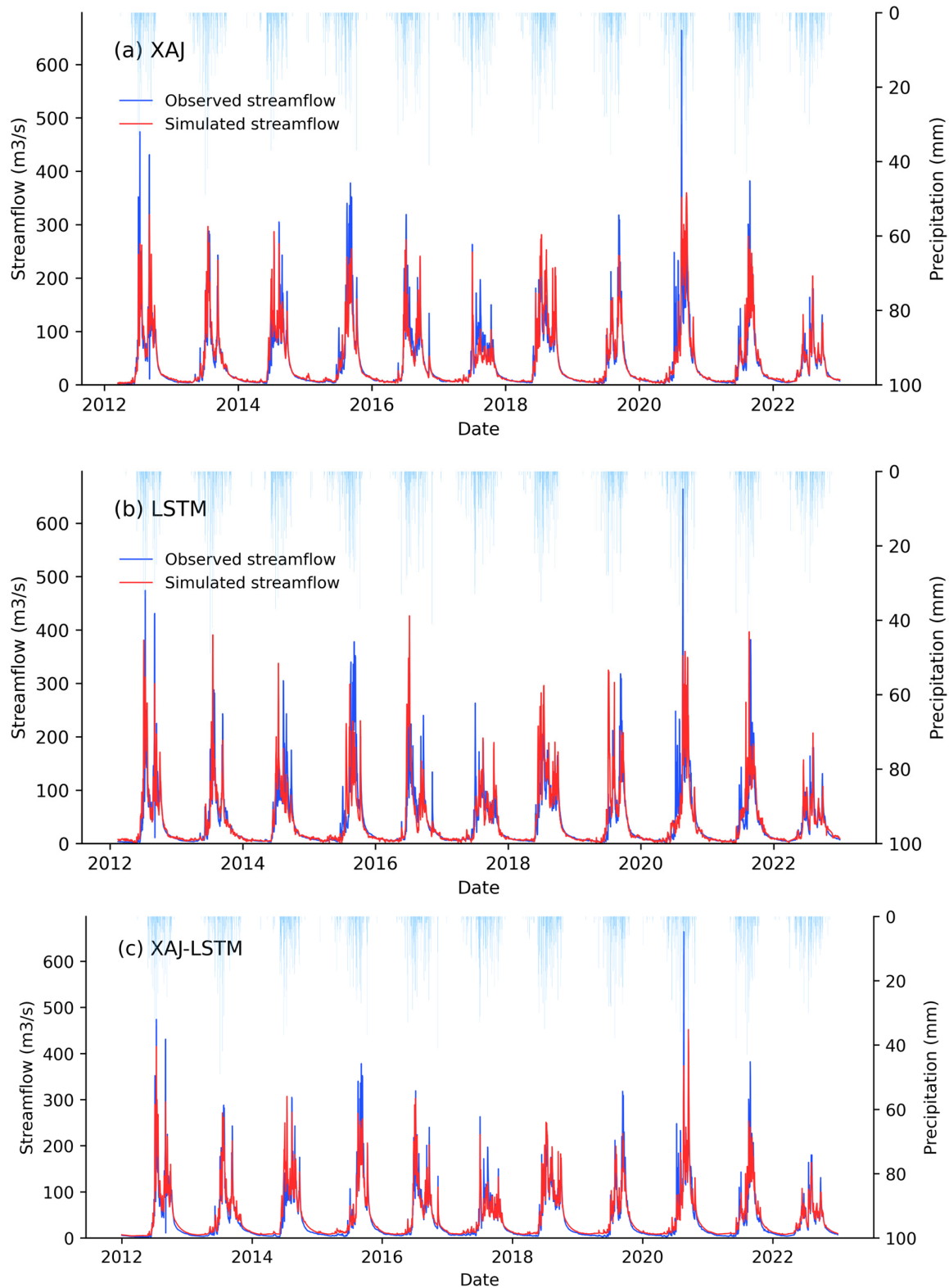


Figure 3. Observed and simulated streamflow under LOYO validation at Jiami by three models, (a) XAJ; (b) LSTM; (c) XAJ-LSTM.

4.2. Error Analysis of Streamflow Simulation

Figure 4 illustrates the sorted distribution of streamflow simulation errors for the three models at Jiami Station across the combined LOYO validation points. Over the entire period, the LSTM model had the widest error range, from -329.38 to 282.28 m^3/s , with a median absolute error of 3.15 m^3/s . The XAJ model ranged from -312.91 to 252.52 m^3/s , with a median absolute error of 2.21 m^3/s . The XAJ-LSTM model ranged from -290.63 to 207.45 m^3/s , with a median absolute error of 6.52 m^3/s . The full-period error distribution indicates that XAJ-LSTM showed relatively smaller positive and negative errors compared with LSTM model.

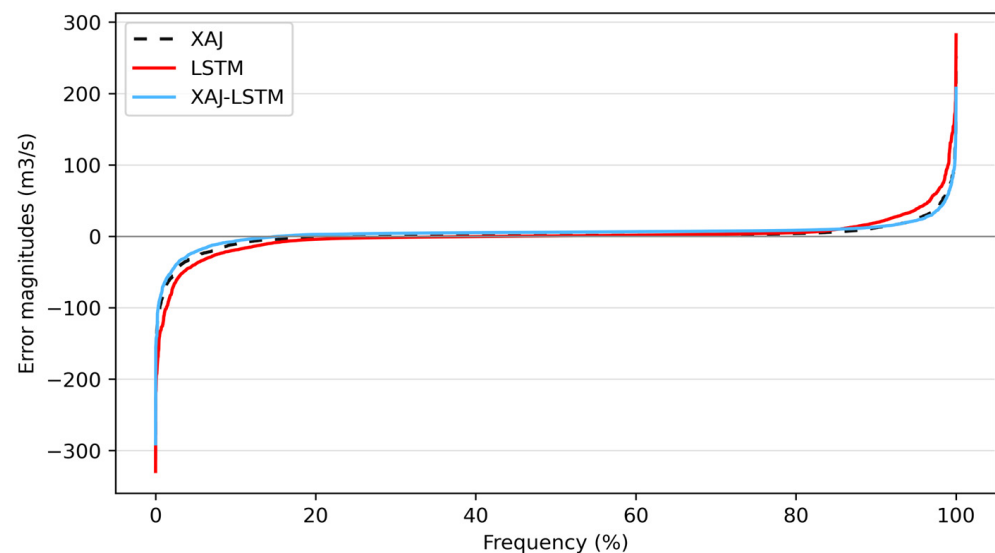


Figure 4. Distribution of streamflow simulation error magnitudes over the LOYO validation period at Jiami Station.

According to the analysis of Figure 4, the error profiles at Jiami Station reveal distinct performance patterns across the three models. Over the entire LOYO period, XAJ-LSTM had the narrowest overall error range, while LSTM had the broadest error range. XAJ had the smallest median absolute error over the entire period, indicating that the physical model remained highly competitive for typical daily errors.

During the flood period, XAJ-LSTM had a smaller error range than both standalone models, ranging from -290.63 to 207.45 m^3/s . Its median absolute error was 9.76 m^3/s , close to the XAJ value of 9.33 m^3/s and lower than the LSTM value of 16.89 m^3/s .

These results explain why the coupled model improved flood-period RMSE and NSE while XAJ remained comparable for typical flood-season errors. Overall, Figure 4 supports the conclusion that XAJ-LSTM reduced some large errors but did not dominate every error-distribution diagnostic.

The predictive accuracy and error patterns of the XAJ, LSTM, and XAJ-LSTM models are illustrated through the scatter plots and regression analyses in Figure 5. The conceptual XAJ model follows the regression equation $y = 0.8792x + 5.15$. The standalone LSTM model follows $y = 0.8305x + 7.11$ and shows greater dispersion, consistent with its lower NSE values in Table 1.

The XAJ-LSTM framework follows the regression equation $y = 0.8617x + 9.90$. Although the coupled model had the highest full-period NSE and lowest full-period RMSE, its intercept was larger than that of XAJ, which is consistent with its higher full-period Sim-to-Obs ratio. Therefore, the coupled model should not be described as superior under every diagnostic criterion.

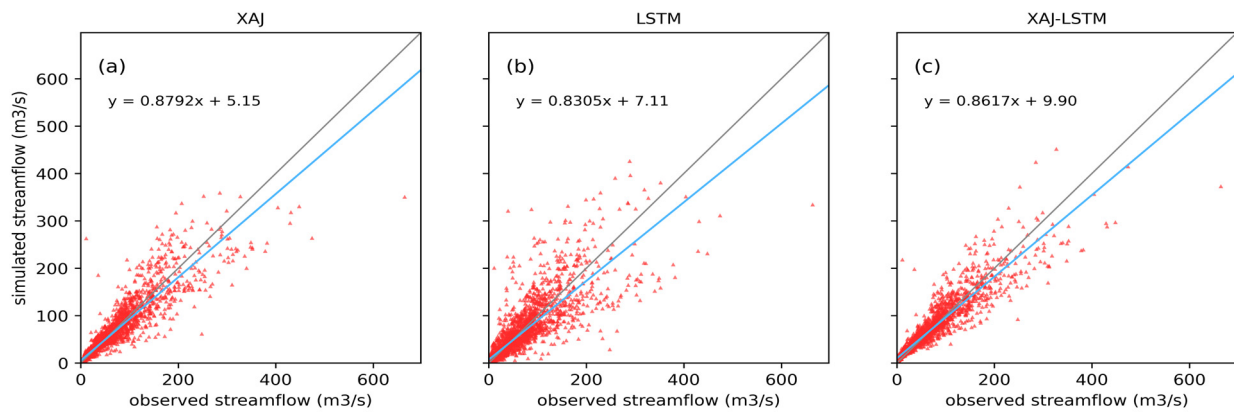


Figure 5. Scatter plots of observed and simulated streamflow under LOYO validation by three models, (a) XAJ; (b) LSTM; (c) XAJ-LSTM.

4.3. Sensitivity of Predictors of the XAJ-LSTM Model

The sensitivity of the coupled XAJ-LSTM model to various input features was assessed using the SHAP framework as a diagnostic analysis. Figure 6a presents the average absolute SHAP values, which summarize the attribution of different predictors in the fitted LSTM model. XAJ-simulated streamflow showed the largest average SHAP value of approximately 7.5. This high attribution is partly expected because XAJ-simulated streamflow is included as a baseline hydrological predictor in the model design. Precipitation followed with an average SHAP value of about 5.0, while temperature had a lower average SHAP value of around 2.0.

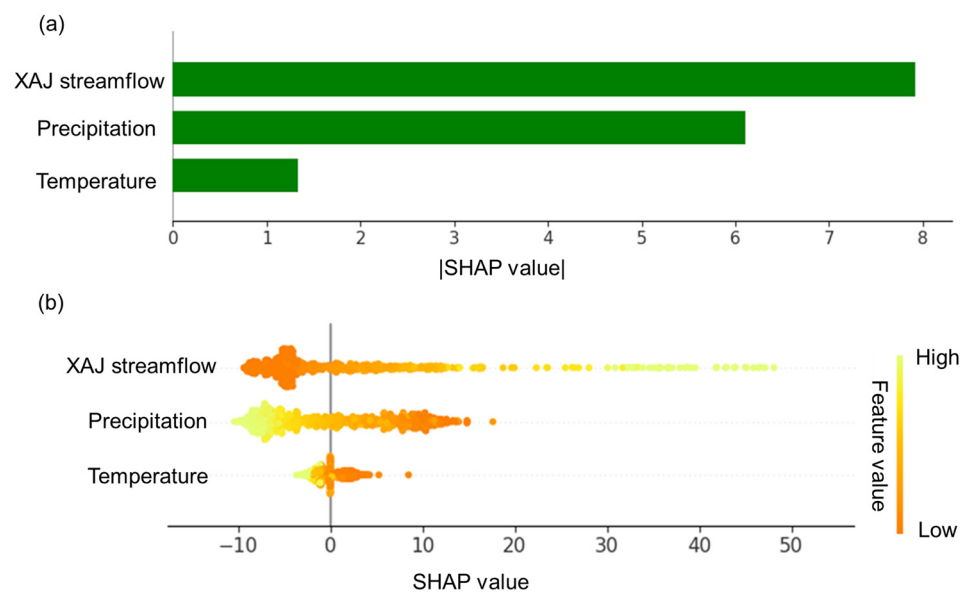


Figure 6. (a) Distribution of influential factors on streamflow prediction (SHAP value reflecting the importance of different elements); (b) SHAP value distribution for streamflow sensitivity.

The beeswarm plot in Figure 6b provides a detailed distribution of SHAP values, with a color gradient ranging from green for low feature values to yellow for high feature values. For XAJ-simulated streamflow, SHAP values range from −10 to 50, with many positive values in the high-value range, indicating that higher XAJ-simulated streamflow generally increases the final predicted streamflow. Precipitation shows a broader range of SHAP values from −20 to 40, and higher precipitation values are frequently associated with positive contributions. In contrast, temperature SHAP values span from −15 to 20,

with higher temperatures often associated with negative contributions, which is consistent with increased evaporative demand or related seasonal effects.

Overall, the SHAP analysis provides a diagnostic interpretation of how the trained LSTM used its inputs in the Woluo River basin. The results should not be interpreted as independent proof of new hydrological mechanisms. Instead, they show that the coupled model relied most strongly on the XAJ baseline simulation and precipitation, while temperature made a smaller and context-dependent contribution. Because SHAP was originally developed for additive feature attribution and the LSTM uses temporally dependent input sequences, the results should be interpreted with caution.

5. Discussion

5.1. Understanding the Advantages and Limitations of the Coupled Model

Instead of indicating a uniform replacement of the conceptual model, the LOYO validation results define where the coupled framework is most useful. Across the combined validation points, XAJ-LSTM achieved the highest full-period NSE and the lowest full-period RMSE and improved flood-period RMSE and NSE, indicating that the LSTM component helped reduce large nonlinear residual errors left by XAJ. However, XAJ retained the highest full-period KGE and the smaller full-period volume bias, showing that the physically based model still preserved water-balance characteristics more robustly. The coupled model should therefore be interpreted as providing metric-dependent gains, especially for dynamic and high-flow errors, rather than as a universally superior alternative under every diagnostic criterion.

The main value of the XAJ-LSTM structure is that it uses XAJ-simulated streamflow as hydrological prior information while allowing the LSTM to learn nonlinear corrections from precipitation, temperature, and the XAJ baseline [8,28,29,41]. This interpretation is consistent with hybrid model studies in which process-model outputs help constrain data-driven models and reduce the data burden, while machine learning compensates for structural and parameter errors in the hydrological model. In the present case, the improvement in RMSE and NSE is likely related to correction of timing and magnitude errors during variable-flow and flood periods, whereas the larger intercept and larger water balance bias indicate that residual bias can remain when the LSTM correction is not explicitly constrained by mass balance. Such bias can arise from lumped soil-water storage, groundwater recession, evapotranspiration losses, routing parameters, and limited representation of spatial heterogeneity.

This finding also refines the broader literature on machine learning and hybrid hydrological modelling. Previous studies have shown that neural networks and LSTM-based models can improve rainfall–runoff simulation and flood prediction, particularly when nonlinear memory effects and extreme events are important [7,37,42,43]. Hybrid frameworks have further demonstrated that combining hydrological model outputs with machine learning can improve forecast skill under limited observations or complex watershed responses [8,23,28,29,41]. At the same time, studies on hydrological model evaluation and process-guided learning emphasize that better error statistics do not automatically imply better water-balance realism, transferability, or process representation [9,14,27,33]. The Woluo River results therefore position loose XAJ-LSTM coupling as a useful residual-learning and flood-error reduction strategy, while confirming that the conceptual XAJ model remains important for hydrological interpretability and bias control.

Several limitations should be recognized. First, LOYO validation provides a stronger test than a single holdout year, but it still uses data from only one basin and 11 annual folds. It therefore cannot fully test transferability across different catchments or climate regimes. Second, the method relies on a loose coupling strategy in which the XAJ output is used as

an input variable for LSTM, and the two components do not exchange states. Third, the lumped structure limits representation of spatial heterogeneity. Fourth, the current input set is limited to precipitation and temperature together with XAJ-simulated flow, which may omit relevant controls such as soil moisture, evapotranspiration, snow condition, land cover, reservoir regulation, and other human influences [44]. Fifth, full predictive uncertainty, parameter sensitivity, and daily prediction intervals were not quantified in this deterministic experiment, although year-level bootstrap confidence intervals were added for the reported performance metrics.

To address these limitations, future research should evaluate the method using longer records, more basins, and more diverse hydroclimatic conditions. Tighter coupling strategies could be explored by embedding learning modules within specific XAJ subprocesses, such as soil moisture accounting or flow routing, where sufficient data are available [34,45]. Additional benchmark models, including GRU, CNN-LSTM, Random Forest, XGBoost, linear correction, and quantile mapping [46–48], should be included to isolate the benefit of nonlinear LSTM learning. Future applications should also evaluate high-flow events, drought diagnostics, peak timing, hourly and monthly scales, climate-change scenarios, data-scarce basins, and operational computational cost.

5.2. Interpretation of Model Attribution

The SHAP results indicate how the trained XAJ-LSTM model used its input features in the Woluo River basin. XAJ-simulated streamflow had the largest average SHAP value, approximately 7.5, and exerted a positive contribution when its values were high. This pattern is consistent with the model architecture because the XAJ output was included as a baseline hydrological predictor. The result therefore confirms that the LSTM relied strongly on the conceptual model output, but it should not be interpreted as independent evidence that the physical model controls streamflow.

Precipitation had the second largest attribution, with an average SHAP value of about 5.0. Higher precipitation values were generally associated with positive SHAP values, which is consistent with rainfall–runoff processes represented by the XAJ model and by the LSTM input sequence. Temperature had a lower average SHAP value of around 2.0, and higher temperatures were often associated with negative contributions, which may reflect evaporative demand and seasonal water balance effects. These interpretations are plausible for the study basin, but they require additional testing before being generalized to other regions.

The attribution results help explain why including a physical baseline can improve selected diagnostics in the LOYO experiment. They also show the limits of interpretation. SHAP identifies how the fitted model allocated importance among inputs, but it does not fully reveal the internal hydrological states learned by the LSTM. It does not replace independent hydrological mechanism analysis. For this reason, the SHAP analysis is presented as diagnostic support for model interpretation rather than as definitive mechanistic evidence [49,50].

6. Conclusions

This study developed an XAJ-LSTM coupling framework for daily streamflow simulation in the Woluo River basin using data from 2012 to 2022. The framework used XAJ-simulated streamflow, daily precipitation, and temperature as inputs to an LSTM network, with observed daily streamflow at Jiami Station as the target. Model performance was evaluated using leave-one-year-out validation, RMSE, NSE, KGE, alpha-NSE, Sim-to-Obs ratio, and graphical diagnostics, while SHAP analysis was used to diagnose feature attribution in the trained coupled model.

Across the combined LOYO validation points, XAJ-LSTM achieved the highest full-period NSE of 0.885 and the lowest full-period RMSE of 19.48 m³/s. It also achieved the best flood-period RMSE and NSE among the three models. However, XAJ retained the highest full-period KGE and the smaller full-period volume bias. These results indicate that XAJ-LSTM improves selected full-period and flood-period diagnostics, but the standalone XAJ model remains competitive and should not be replaced under all conditions.

Future research should extend the validation to more basins and longer hydrometeorological records and should test the approach in arid, semi-arid, humid, and highly heterogeneous basins. Further development should also examine tighter coupling strategies, distributed hydrological models, uncertainty quantification, benchmark models, event-level high-flow diagnostics, drought diagnostics, peak timing, and additional predictors such as soil moisture, evapotranspiration, snow variables, solar radiation, and human regulation. These steps are necessary to evaluate the robustness and transferability of XAJ-LSTM coupling beyond the present case study.

Author Contributions: Conceptualization, Z.H., T.D. and P.S.; Data curation, H.H. and Z.H.; Funding acquisition, N.D.; Investigation, P.S. and T.D.; Methodology, Z.H. and T.D.; Software, Z.H.; Supervision, T.D.; Writing, original draft, Z.H. and T.D.; Writing, review and editing, P.S., N.D. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: Ningpeng Dong received funding for this study through the National Key Research and Development Program of China (2023YFC3081000), the National Natural Science Foundation of China (42401053), and the Young Elite Scientist Sponsorship Program by CSHE (CSHE-YESS-2026006). This study is also funded by the Yunnan Key Laboratory of Water Security (20254916CE340051), Research Project of the State Key Laboratory of Water Cycle and Water Security (SKL2024YJZD03; SKL2025TDGG04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Zhaohui He, Tong Duan and Penghao Shao were employed by the Yalong River Basin Hydropower Development Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Cloke, H.L.; Pappenberger, F. Ensemble flood forecasting: A review. *J. Hydrol.* **2009**, *375*, 613–626. [\[CrossRef\]](#)
2. Balint, G.; Csik, A.; Barthá, P.; Gauzer, B.; Bonta, I. Application of meteorological ensembles for Danube flood forecasting and warning. In *Transboundary Floods: Reducing Risks Through Flood Management*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 57–68. [\[CrossRef\]](#)
3. Gassman, P.W.; Reyes, M.R.; Green, C.H.; Arnold, J.G. The Soil and Water Assessment Tool: Historical development, applications, and future research directions. *Trans. ASABE* **2014**, *57*, 1211–1250. [\[CrossRef\]](#)
4. Dong, N.; Wei, J.; Yang, M.; Yan, D.; Yang, C.; Gao, H.; Arnault, J.; Laux, P.; Zhang, X.; Liu, Y.; et al. Model estimates of China's terrestrial water storage variation due to reservoir operation. *Water Resour. Res.* **2022**, *58*, e2021WR031787. [\[CrossRef\]](#)
5. Dong, N.; Yang, M.; Wei, J.; Arnault, J.; Laux, P.; Xu, S.; Wang, H.; Yu, Z.; Kunstmann, H. Toward improved parameterizations of reservoir operation in ungauged basins: A synergistic framework coupling satellite remote sensing, hydrologic modeling, and conceptual operation schemes. *Water Resour. Res.* **2023**, *59*, e2022WR033026. [\[CrossRef\]](#)
6. Dong, N.; Yang, M.; Wei, J.; Xu, S.; Zhao, Y.; Zhang, X.; Liu, B.; Wu, M.; Wang, H.; Kunstmann, H. Reconstructing China's natural streamflow at 1-km resolution. *Water Resour. Res.* **2026**, *62*, e2025WR042606. [\[CrossRef\]](#)
7. Abrahart, R.J.; Anctil, F.; Coulibaly, P.; Dawson, C.W.; Mount, N.J.; See, L.M.; Shamseldin, A.Y.; Solomatine, D.P.; Toth, E.; Wilby, R.L. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Prog. Phys. Geogr.* **2012**, *36*, 480–513. [\[CrossRef\]](#)

8. Humphrey, G.B.; Gibbs, M.S.; Dandy, G.C.; Maier, H.R. A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *J. Hydrol.* **2016**, *540*, 623–640. [[CrossRef](#)]
9. Refsgaard, J.C.; Knudsen, J. Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res.* **1996**, *32*, 2189–2202. [[CrossRef](#)]
10. Addor, N.; Do, H.X.; Alvarez-Garretón, C.; Coxon, G.; Fowler, K.; Mendoza, P.A. Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrol. Sci. J.* **2020**, *65*, 712–725. [[CrossRef](#)]
11. Dong, W.; Zhang, D.; Wang, W.; Li, M.; Feng, X.; Sun, D. Analysis of computational accuracy and efficiency of urban flood models under different coupling modes. *J. China Inst. Water Resour. Hydropower Res.* **2025**, *23*, 307–317. [[CrossRef](#)]
12. Gupta, H.V.; Sorooshian, S.; Yapo, P.O. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrol. Eng.* **1999**, *4*, 135–143. [[CrossRef](#)]
13. Castellano-Méndez, M.A.; González-Manteiga, W.; Febrero-Bande, M.; Prada-Sánchez, J.M.; Lozano-Calderón, R. Modelling of the monthly and daily behaviour of the runoff of the Xallas river using Box–Jenkins and neural networks methods. *J. Hydrol.* **2004**, *296*, 38–58. [[CrossRef](#)]
14. Beven, K. *Rainfall-Runoff Modelling: The Primer*, 2nd ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2021.
15. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **1989**, *2*, 303–314. [[CrossRef](#)]
16. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1991**, *4*, 251–257. [[CrossRef](#)]
17. Kisi, O. Streamflow forecasting using different artificial neural network algorithms. *J. Hydrol. Eng.* **2007**, *12*, 532–539. [[CrossRef](#)]
18. Hinton, G.E. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
19. Adnan, R.M.; Liang, Z.; Trajkovic, S.; Zounemat-Kermani, M.; Li, B.; Kisi, O. Daily streamflow prediction using optimally pruned extreme learning machine. *J. Hydrol.* **2019**, *577*, 123981. [[CrossRef](#)]
20. Tan, W. *Runoff Forecasting and Interpretability Study in the Yangtze River Basin Based on Deep Learning Methods*; Nanjing University of Information Science & Technology: Nanjing, China, 2024. (In Chinese)
21. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Model. Softw.* **2000**, *15*, 101–124. [[CrossRef](#)]
22. Aoulmi, Y.; Marouf, N.; Rasouli, K.; Panahi, M. Runoff predictions in a semiarid watershed by convolutional neural networks improved with metaheuristic algorithms and forced with reanalysis and climate data. *J. Hydrol. Eng.* **2023**, *28*. [[CrossRef](#)]
23. Li, X.; Zhang, Y.; Wang, J.; Chen, Z. Optimizing flood predictions by integrating LSTM and physical-based models with mixed historical and simulated data. *Heliyon* **2024**, *10*, e33669. [[CrossRef](#)] [[PubMed](#)]
24. Huang, J.; Zhao, K.; Li, W.; Liu, X. A novel fusion method for generating surface soil moisture data with high accuracy, high spatial resolution, and high spatio-temporal continuity. *Water Resour. Res.* **2022**, *58*, e2021WR030827. [[CrossRef](#)]
25. Brocca, L.; Filippucci, P.; Hahn, S.; Ciabatta, L.; Massari, C.; Camici, S.; Wagner, W. SM2RAIN-ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. *Earth Syst. Sci. Data* **2019**, *11*, 1583–1601. [[CrossRef](#)]
26. Xu, Q. *Research on Mid- and Long-Term Runoff Forecasting Model Based on Machine Learning and System Integration*; Huazhong University of Science and Technology: Wuhan, China, 2023. (In Chinese)
27. Nearing, G.S.; Kratzert, F.; Sampson, A.K.; Pelissier, C.S.; Klotz, D.; Frame, J.M.; Gupta, H.V. What role does hydrological science play in the age of machine learning? *Water Resour. Res.* **2021**, *57*, e2020WR028091. [[CrossRef](#)]
28. Liu, J.; Yuan, X.; Zeng, J.; Jiao, Y.; Li, Y.; Zhong, L.; Yao, L. Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 265–278. [[CrossRef](#)]
29. Yang, S.; Yang, D.; Chen, J.; Santisirisomboon, J.; Zhao, B. A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *J. Hydrol.* **2020**, *590*, 125206. [[CrossRef](#)]
30. Raftery, A.E.; Gneiting, T.; Balabdaoui, F.; Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **2005**, *133*, 1155–1174. [[CrossRef](#)]
31. Zhang, D.; Lin, J.; Peng, Q.; Wang, D.; Yang, T.; Sorooshian, S.; Liu, X. Modeling hydrological processes using deep convolutional neural networks. *J. Hydrometeorol.* **2018**, *19*, 1919–1938. [[CrossRef](#)]
32. Wei, Z. Bridging spatio-temporal discontinuities in global soil moisture mapping by coupling physics in deep learning. *Remote Sens. Environ.* **2024**, *302*, 113845.
33. Shen, C.; Appling, A.P.; Gentile, P.; Band, L.E.; Gupta, H.V. Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **2023**, *4*, 552–567. [[CrossRef](#)]
34. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [[CrossRef](#)]
35. Klingaman, N.P.; Peatman, S.C.; Vidale, P.L. Process-based machine learning for weather and climate prediction: Challenges and opportunities. *Philos. Trans. R. Soc. A* **2023**, *381*, 20220151. [[CrossRef](#)] [[PubMed](#)]
36. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
37. Sit, M.; Demiray, B.Z.; Xiang, Z.; Ewing, G.J.; Sermet, Y.; Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* **2020**, *82*, 2635–2670. [[CrossRef](#)] [[PubMed](#)]

38. Razavi, S.; Tolson, B.A. A new formulation for feedforward neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 81–92. [[CrossRef](#)]
39. Zhao, R.J. The Xin'anjiang model applied in China. *J. Hydrol.* **1992**, *135*, 371–381. [[CrossRef](#)]
40. Ferranti, L.; Corti, S.; Janousek, M. Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Q. J. R. Meteorol. Soc.* **2018**, *144*, 317–326. [[CrossRef](#)]
41. Hwang, S.-H.; Ham, Y.-G.; Lee, M.-I. Improving hydrological predictions with hybrid machine learning and physical models: A case study in the Han River Basin. *Water Resour. Res.* **2022**, *58*, e2022WR032456. [[CrossRef](#)]
42. Kratzert, F.; Klotz, D.; Herrnegger, M.; Sampson, A.K.; Hochreiter, S.; Nearing, G.S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **2019**, *55*, 11344–11354. [[CrossRef](#)]
43. Frame, J.M.; Kratzert, F.; Klotz, D.; Gauch, M.; Nearing, G.S. Deep learning rainfall-runoff predictions of extreme events. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 3377–3392. [[CrossRef](#)]
44. Hao, H.; Dong, N.; Yang, M.; Wei, J.; Zhang, X.; Xu, S.; Yan, D.; Ren, L.; Leng, G.; Chen, L.; et al. The changing hydrology of an irrigated and dammed Yangtze River: Streamflow, extremes, and lake hydrodynamics. *Water Resour. Res.* **2024**, *60*, e2024WR037841. [[CrossRef](#)]
45. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [[CrossRef](#)] [[PubMed](#)]
46. Bremnes, J.B. Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Weather Rev.* **2020**, *148*, 403–414. [[CrossRef](#)]
47. Scheuerer, M.; Hamill, T.M. Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **2015**, *143*, 4578–4596. [[CrossRef](#)]
48. Taillardat, M.; Mestre, O.; Zamo, M.; Naveau, P. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Weather Rev.* **2016**, *144*, 2375–2393. [[CrossRef](#)]
49. Bai, T.; Wang, X.-S.; Han, P.-F. Controls of Groundwater-Dependent Vegetation Coverage in the Yellow River Basin, China: Insights from Interpretable Machine Learning. *J. Hydrol.* **2024**, *631*, 130747. [[CrossRef](#)]
50. Xu, Y.; Lin, K.; Hu, C.; Wang, S.; Wu, Q.; Zhang, J.; Xiao, M.; Luo, Y. Interpretable Machine Learning on Large Samples for Supporting Runoff Estimation in Ungauged Basins. *J. Hydrol.* **2024**, *639*, 131598. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.