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

Supplementary Information

Contents of Supplementary Information

Supplementary Note 1: Study areas

The Wuding River, a key tributary of the Yellow River, is located in the northern part of Shaanxi Province, China, and is the largest river in the Yulin region of Shaanxi. It has a total length of 491.0 kilometers and a drainage area of 30,260 square kilometers. The Wuding River Basin experiences a semi-arid climate of the temperate continental type, with an average annual precipitation of approximately 400 millimeters that rises gradually from the northern to the southern parts. The mean annual flow of the Wuding River is approximately 1.53 billion cubic meters, accounting for 2.4% of the mean annual flow of the Yellow River Basin, which is 62.8 billion cubic meters. The basin area represents 4.2% of the Yellow River Basin area, making the river's streamflow relatively scarce.

Situated along the left bank of the lower section of the Yangtze River, the Chu River has its source near Liangyuan in Feidong County of Anhui Province. The main stream has a total length of approximately 269 kilometers, with a drainage area of about 8,057 square kilometers, including 6,250 square kilometers in Anhui Province and 1,750 square kilometers in Jiangsu Province. The Chu River has an average annual rainfall of 900-1000 millimeters, with the majority occurring in the months of June, July, and August, exhibiting significant interannual variability. The average annual evaporation ranges from 900 to 1000 millimeters, and there is generally a balance between rainfall and evaporation. The total annual runoff is 2.498 billion cubic meters, with 1.825 billion cubic meters occurring within Anhui Province, accounting for 73.1% of the total annual runoff of the Chu River Basin.

The Jianxi River, originating from the Xianxia Ridge in the Wuyi Mountains, constitutes the principal tributary of the upper Min River. Comprised of three major tributaries—Chongxi, Nanpuxi, and Songxi—it is the largest contributor to the upper Min River. It is located between 26°31' to 28°31' north latitude and 117°31' to 119°00' east longitude, with a total length of 635.6 kilometers. Covering 14,787 square kilometers, the basin makes up approximately 27% of the total Min River basin area. Positioned in a subtropical monsoon climate zone, the area encounters moist air and plentiful precipitation, with average annual rainfall varying between 1800 to 2200

millimeters. The annual runoff is approximately 15.8 billion cubic meters, accounting for about one-third of the total flow of the Min River.

Originating from the Shuwest Camp between the Balang Mountain and the Jiajin Mountain in the Qionglai Mountains, the Qingyi River serves as a tributary of the Min River, which, in turn, is a tributary of the Yangtze River. The main stream is 289 kilometers long with a drop of 2844 meters, covering a basin area of 12,897 square kilometers. The Qingyi River basin falls within a subtropical humid climate zone, characterized by an average annual temperature ranging from 15℃ to 18℃. Most of the basin is located in a region prone to heavy rain, with an average annual rainfall of around 2000 millimeters. However, there are significant variations across the area, generally increasing from northwest to southeast. Floods mostly occur from late June to mid-September, characterized by high peaks and large volumes, lasting for 3 to 5 days. The average annual flow at the Jiajiang Hydrological Station, a control station on the main stream, is approximately 482 m3/s, with an annual total runoff of about 15.2 billion cubic meters.

Fig. S1 shows the relationship and trend of daily runoff with rainfall and mean temperature for the river basins of the Wuding and Qingyi Rivers as examples.

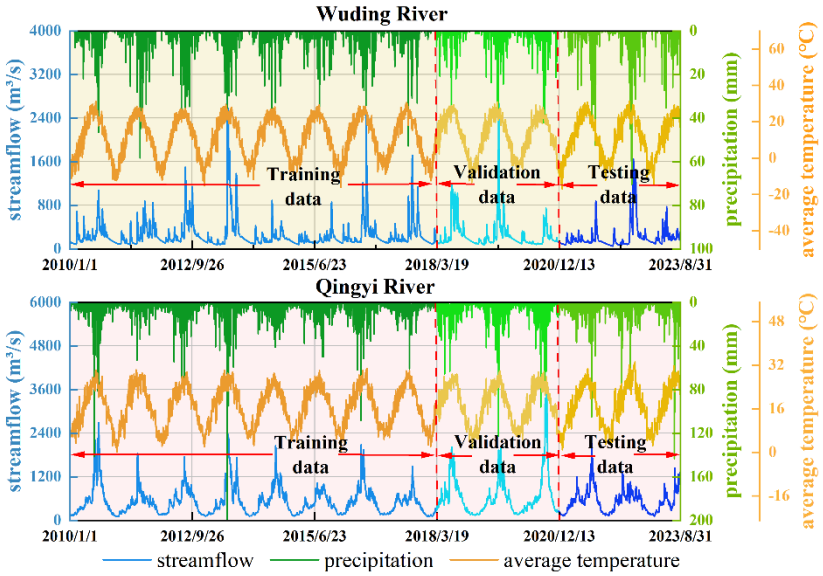

Fig. S1 Variations in streamflow, precipitation, and average temperature in Wuding River and Qingyi River basins.

Supplementary Note 2: Methodologies

2.1 Calculation of MIC

The fundamental idea behind MIC is that if there is a certain relationship between two variables, a grid can be placed over a scatterplot of these variables to segment and capture this relationship. For two-dimensional variables *x* and *y*, MIC is calculated as follows.

(1) Partition *x* and *y* into *m* and *n* different intervals, discretizing the sample space into an *m* × *n* grid Q. Estimate the joint probability density and marginal probability densities through the sample count within the grid and the sample count within the intervals as a proportion of the sample capacity. Subsequently, calculate the mutual information for this grid.

$$I(x;y|Q) = \sum_x \sum_y p(x,y)\, lg \frac{p(x,y)}{p(x)p(y)} \tag{S1}$$

where $I(x;y|Q)$ represents the mutual information value of variables *x* and *y* when discretized in the grid Q, $p(x,y)$ represents the joint probability density function, and $p(x)$ and $p(y)$ represent the marginal density functions.

(2) Due to the possibility of different discretization methods within the same scale of *m* × *n*, select the maximum value of mutual information under different partitioning methods.

$$I_{m\times n}(x,y) = \max_{Q\in m,n} I(x,y|Q) \tag{S2}$$

(1) Build grids of different scales, *m* × *n*, and calculate the maximum mutual information values for different scale grids sequentially according to Eq. (S1) and Eq. (S2). Normalize these values and obtain the maximum normalized mutual information value as the final M_{IC}. The calculation method is as follows:

$$M_{IC}(x,y) = \max_{m\times n<B} \frac{I_{m\times n}(x,y)}{\lg \min(m,n)} \tag{S3}$$

where *m* × *n* represents the constraint condition for the total number of grid partitions. The MIC delivers strong performance in real-world scenarios when the B is set to the 0.6th power of the total sample size. Hence, this specific value was employed in all experiments conducted within our research.

2.2 Outlier injection test

In this study, to evaluate the robustness of the model on data containing outliers, we injected random outliers into the training set. The detailed process is as follows:

(1) Based on the common error range in field data (1% - 2%), the outlier injection ratio was set to 2% to ensure the model's robustness and stability under higher error conditions.

(2) Assuming there are N samples in the training set, a random sampling function was used to draw $0.02N$ positions from the training set index list without replacement, ensuring that each sample point is selected only once.

(3) For each selected sample point, outliers were generated based on the overall data distribution. Specifically, the standard deviation σ and mean of the original training set μ were calculated, and a random deviation value was generated for each selected observation. Outliers were set within a $\pm 3\sigma$ range of the original observation to simulate extreme conditions. The specific formula for generation is:

$$\text{Outlier} = \mu + r \times \sigma \tag{S4}$$

where r is a random number sampled uniformly from a specific range (e.g., -3 to +3) to ensure that the injected outliers have significant deviations.

(4) The generated outliers were used to replace the original observation values. After replacement, the training set contained 2% outlier observations. The model was trained on the training set with random outliers to evaluate its learning capability and robustness in an anomalous data environment. By training under these conditions, the model can learn to adapt to noise and anomalies in the data, thereby improving its generalization ability.

2.3 Flood simulation

This study conducted an in-depth analysis of river flow simulations during flood and non-flood seasons across four different basins, with the following specific steps:

(1) Based on historical precipitation and flow data, the flow variation patterns of each basin were analyzed to determine the time range of the flood season in the test set for each basin.

(2) The test sets for each basin were divided into flood and non-flood periods,

resulting in two independent test sets for each basin.

(3) Using the trained models, simulations were made for both the flood and non-flood periods, and the results were recorded separately.

In addition, this study pays special attention to accurately simulating extreme hydrological events. The specific methods are as follows:

(1) In each basin, the standard deviation of the test set was calculated, and the time periods during which river flow exceeds four times the standard deviation were defined as flood events. This criterion was used to identify potential flood risks.

(2) The feature variables corresponding to these flood event periods were input into the trained model to obtain the simulated flow values during the flood events, which were then compared and analyzed against the actual observed values.

2.4 Interval simulation

The specific steps for interval simulation in this study are as follows:

(1) Extract the flow simulation results based on point forecasts. These results will serve as the foundation for interval simulation.

(2) Calculate the errors between the point forecast values and the actual observed values to form a simulation error sequence. This sequence is used to assess the accuracy of the model's simulations and provides essential information for interval simulation.

(3) Fit the simulation error sequence using the logistic distribution function. Through this fitting process, determine the distribution characteristics of the errors in order to generate the corresponding confidence intervals.

(4) At the specified significance level (e.g., 95%), calculate the simulation intervals using the fitted logistic distribution function. Specifically, construct the upper and lower bounds based on the forecast values and the corresponding error distribution to form the final simulation intervals.

(5) Use two evaluation metrics, Prediction Interval Coverage Probability (PICP) and Prediction Interval Normalized Average Width (PINAW), to assess the accuracy of the interval simulations.

Supplementary Note 3: Model performance assessment

The evaluation of regression models is typically based on the function between simulated values and observed values. In this study, two error metrics were chosen to evaluate the performance of the model: Mean Absolute Error (MAE, unit: m^3/s) and Root Mean Square Error (RMSE, unit: m^3/s). Additionally, two efficiency metrics were selected to assess the goodness of fit of the model: Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE). The corresponding mathematical expressions are shown in **Eqs. (S5) ~ (S8)**.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{S5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{S6}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y}_i)^2} \tag{S7}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\bar{\hat{y}}}{\bar{y}} - 1\right)^2 + \left(\frac{\sigma_{\hat{y}}}{\sigma_y} - 1\right)^2} \tag{S8}$$

Additionally, the Mean Absolute Percentage Error (MAPE) is a highly suitable evaluation metric in simulation applications. MAPE is independent, meaning it is not influenced by the magnitude or units of the data. This feature makes it widely applicable in flow simulation across different flow levels, particularly demonstrating significant value in comparing and evaluating simulation performance across different basins. The mathematical expression of MAPE is shown in **Eq. (S9)**.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{S9}$$

In the above expression, n represents the number of samples, r represents the correlation coefficient between observed and simulated values, y_i and $\hat{y}_i$ represent the observed and simulated values of the i-th element, while $\bar{y}$ and $\bar{\hat{y}}$ represent their respective means, and σ_y and $\sigma_{\hat{y}}$ represent their respective standard deviations.

In assessing interval simulation performance, two metrics, PICP and PINAW, are

utilized. PICP represents the proportion of true values falling within the upper and lower bounds of the simulation interval, while PINAW measures the narrowness of the simulation interval. **Eqs. (S10) ~ (S11)** respectively define these two metrics.

$$\mathrm{PICP} = \frac{1}{n}\sum_{i=1}^{n} c_i \,, c_i = \begin{cases} 1, y_i \in \{L_i, U_i\} \\ 0, otherwise \end{cases} \tag{S10}$$

$$\mathrm{PINAW} = \frac{1}{nR}\sum_{i=1}^{n} (U_i - L_i) \tag{S11}$$

Where R is the range of the target values (used for normalization), U_i is the upper limit of the simulation interval, and L_i is the lower limit.

Supplementary Note 4: Results and discussion

4.1 Setting of model parameters in this study

Table S1 Parameterization of the models.

Basin	Model	Parameter setting
Wuding	TCN-GRU	timesteps = 1, nb_filters = 64, kernel_size = 3, units = 128, dropout = 0.2, optimizer = 'adam', epochs = 200, batch_size = 64
	RF	n_estimators = 100, criterion = "mse"
Chu	TCN-GRU	timesteps = 1, nb_filters = 64, kernel_size = 4, units = 256, dropout = 0.2, optimizer = 'adam', epochs = 200, batch_size = 64
	RF	n_estimators = 100, criterion = "mse"
Jianxi	TCN-GRU	timesteps = 1, nb_filters = 64, kernel_size = 3, units = 128, dropout = 0.3, optimizer = 'adam', epochs = 180, batch_size = 32
	RF	n_estimators = 100, criterion = "mse"
Qingyi	TCN-GRU	timesteps = 1, nb_filters = 128, kernel_size = 3, units = 128, dropout = 0.2, optimizer = 'adam', epochs = 180, batch_size = 32
	RF	n_estimators = 100, criterion = "mse"

4.2 Model performance comparison

Fig. S2 provides a more intuitive comparison of the streamflow simulation performance of the six models in the four basins. The results indicate that deep learning models exhibit higher simulation accuracy compared to the traditional physical model XAJ, a trend confirmed across all basins. Specifically, TCN and GRU models outperform LSTM in streamflow simulation, demonstrating better performance. However, there is no clear superiority between TCN and GRU. In the comparison between TCN and GRU, their simulation accuracy is similar across all basins, showing comparable performance levels. Additionally, it is noteworthy that integrating the simulations of the physical model with those of the deep learning model leads to a significant improvement in simulation performance. Taking the KGE metric as an example, the hybrid XAJ-TCN-GRU model exhibits a relative improvement of 39.60%, 50.08%, 44.18%, and 43.02% compared to using only the

XAJ model, and a relative improvement of 0.61%, 2.12%, 2.51%, and 1.25% compared to using the TCN-GRU model, across the four basins.

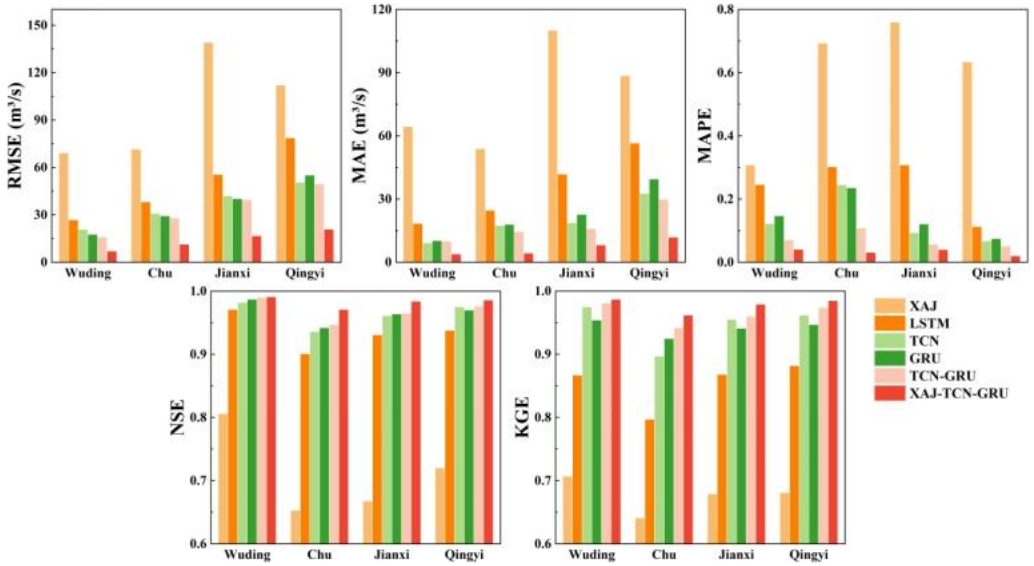

Fig. S2 Model performance for forecasting daily streamflow (m³/s) in terms of RMSE, MAE, MAPE, NSE and KGE

To visualize the streamflow trend simulation, we plotted line graphs showing the simulated values and actual observations for the six models, as depicted in **Fig. S3**. As shown in **Fig. S3**, it's evident that the proposed XAJ-TCN-GRU model demonstrates superior goodness of fit.

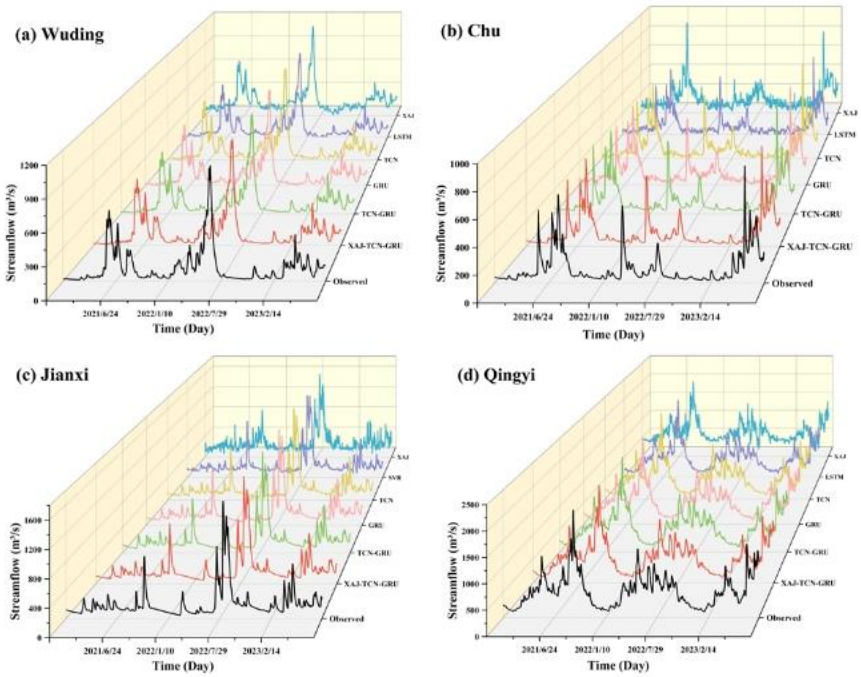

Fig. S3 Comparison between forecasted and observed streamflow during model testing using the proposed model (i.e., XAJ-TCN-GRU) and other comparative models for four basins.

4.3 Model robustness analysis

Taking the Wuding River basin as an example, the RMSE values of the XAJ-TCN-GRU model range from 7.978 m^3/s to 16.280 m^3/s, with an average of 9.445 m^3/s. This value represents a significant reduction compared to the XAJ model and the TCN-GRU model, decreasing by 69.638 m^3/s and 11.069 m^3/s, respectively, validating its superior simulation accuracy. Furthermore, the XAJ-TCN-GRU model performs exceptionally well on the MAE and MAPE metrics, achieving the lowest levels among all models. Additionally, its highest NSE value further highlights the outstanding performance of this model for streamflow simulation, as shown in **Table S2**.

Table S2 Evaluation metrics for streamflow simulation by six models in noise data injection testing.

Basin	Model	RMSE			MAE			MAPE			NSE			KGE		
		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Wuding	XAJ	79.083	74.439	84.804	51.597	47.056	57.679	0.759	0.629	0.871	0.747	0.709	0.776	0.541	0.506	0.566
	LSTM	32.394	29.871	35.109	17.895	16.591	20.886	0.223	0.181	0.311	0.958	0.950	0.964	0.846	0.831	0.868
	TCN	24.635	21.791	28.908	16.510	12.381	20.910	0.248	0.153	0.333	0.975	0.966	0.981	0.886	0.837	0.923
	GRU	22.342	21.518	23.141	10.303	10.030	11.329	0.138	0.093	0.203	0.980	0.978	0.981	0.953	0.932	0.986
	TCN-GRU	20.514	19.425	21.665	11.951	10.033	15.180	0.072	0.065	0.092	0.983	0.981	0.985	0.957	0.922	0.992
	XAJ-TCN-GRU	9.445	7.978	16.280	4.813	4.027	10.339	0.044	0.033	0.124	0.990	0.989	0.990	0.984	0.975	0.986
Chu	XAJ	77.610	71.921	83.749	56.870	47.227	67.797	0.555	0.433	0.689	0.616	0.579	0.650	0.508	0.435	0.672
	LSTM	41.096	37.332	44.619	28.261	23.259	34.607	0.435	0.342	0.623	0.885	0.865	0.906	0.797	0.719	0.832
	TCN	32.717	28.984	38.056	22.607	15.600	30.806	0.378	0.149	0.608	0.927	0.902	0.943	0.840	0.743	0.931
	GRU	30.964	27.739	37.682	17.179	10.017	27.509	0.232	0.086	0.448	0.935	0.904	0.948	0.874	0.756	0.939
	TCN-GRU	29.776	28.045	33.163	13.670	9.845	19.146	0.129	0.067	0.198	0.940	0.926	0.947	0.916	0.857	0.959
	XAJ-TCN-GRU	11.632	10.914	12.503	4.420	4.128	4.641	0.030	0.028	0.031	0.965	0.963	0.972	0.955	0.952	0.970
Jianxi	XAJ	142.217	130.779	149.315	75.667	70.988	80.321	0.331	0.298	0.365	0.548	0.503	0.619	0.402	0.361	0.479
	LSTM	59.095	55.943	63.274	39.167	35.527	41.221	0.240	0.205	0.287	0.922	0.911	0.930	0.867	0.808	0.913
	TCN	50.124	47.603	54.740	27.377	16.398	36.668	0.140	0.057	0.193	0.945	0.933	0.949	0.888	0.840	0.922
	GRU	55.630	53.875	58.664	34.696	22.976	38.718	0.189	0.091	0.262	0.931	0.923	0.935	0.874	0.823	0.916
	TCN-GRU	44.704	43.185	47.226	19.656	15.617	26.825	0.080	0.053	0.125	0.955	0.950	0.958	0.943	0.891	0.970
	XAJ-TCN-GRU	16.1339	15.363	16.707	6.445	6.164	6.820	0.023	0.022	0.024	0.976	0.970	0.985	0.968	0.965	0.981
Qingyi	XAJ	190.664	175.033	200.468	135.814	128.521	144.947	0.246	0.222	0.278	0.639	0.601	0.696	0.590	0.567	0.623
	LSTM	87.439	81.796	94.541	63.482	57.966	71.828	0.127	0.115	0.143	0.924	0.911	0.934	0.858	0.832	0.893
	TCN	67.533	62.533	72.242	45.105	40.700	49.512	0.093	0.078	0.104	0.955	0.948	0.961	0.908	0.882	0.940
	GRU	78.600	72.048	84.158	56.155	48.268	63.389	0.118	0.098	0.147	0.938	0.930	0.948	0.876	0.825	0.934
	TCN-GRU	57.378	53.536	62.106	32.803	27.954	38.445	0.056	0.046	0.068	0.967	0.962	0.972	0.936	0.911	0.956

XAJ-TCN-GRU	22.622	21.818	23.579	12.468	12.163	12.910	0.021	0.020	0.021	0.980	0.978	0.985	0.969	0.965	0.984

Moreover, in the tests conducted in the Chu River, Jianxi River, and Qingyi River basins, the XAJ-TCN-GRU model also demonstrates robust simulation capabilities. Its average MAPE values are 0.030, 0.023, and 0.021, respectively, lower than the other five models. Similarly, its average KGE values are 0.965, 0.978, and 0.973, respectively, higher than the other models, indicating the model's universal applicability across different basins.

It is evident that the NSE values of the XAJ model decreased in all four basins after the outlier injection test, with reductions of 0.059, 0.037, 0.120, and 0.081, respectively. This outcome suggests that the physical model's ability to handle outliers is relatively weak and susceptible to interference. In contrast, the XAJ-TCN-GRU model shows minimal changes in all four indicators after the outlier injection test, demonstrating its robustness against disturbances. This characteristic enables the XAJ-TCN-GRU model to maintain stable simulation performance when facing sudden events and complex variations in the real world.

4.4 Interpretability of deep learning model

Table S3 Values of the three key indices of the input variables of the TCN-GRU and their normalized sums.

Basin	Input variable	Index			
		SHAPABS	FI	PFI	Normalized sum
Wuding	DPT	75.634	0.074	0.876	2.118
	SP	41.089	0.336	0.447	1.847
	GT	50.307	0.126	0.753	1.725
	Evap	38.894	0.111	0.567	1.229
	AT	32.570	0.047	0.441	0.734
	K	20.457	0.057	0.284	0.354
	HCC	15.503	0.099	0.202	0.300
	SLP	13.961	0.039	0.333	0.264
	CUPE	13.306	0.063	0.186	0.122
	Tmin	10.623	0.049	0.220	0.083
Chu	DPT	142.834	0.321	2.304	2.695
	Tmax	90.627	0.096	0.096	1.329
	GT	76.342	0.119	1.157	1.175
	K	14.314	0.446	0.108	1.023
	Evap	62.577	0.121	0.559	0.803

	Tmin	36.187	0.066	0.391	0.392
	SLP	33.076	0.073	0.359	0.371
	AT	19.452	0.036	0.477	0.231
	CUPE	14.254	0.075	0.220	0.169
	SP	11.225	0.048	0.108	0.029
Jianxi	Evap	174.124	0.157	2.364	2.681
	DPT	51.205	0.218	0.639	1.402
	PET	92.281	0.120	0.854	1.263
	GT	59.418	0.138	0.668	1.052
	RH	30.358	0.094	0.191	0.413
	Tmax	39.533	0.053	0.201	0.265
	K	28.980	0.064	0.161	0.233
	LCC	29.359	0.066	0.117	0.227
	Tmin	31.693	0.063	0.110	0.224
	AT	26.448	0.027	0.119	0.004
Qingyi	DPT	340.134	0.504	2.101	3.000
	Tmin	90.211	0.105	0.327	0.468
	GT	102.513	0.060	0.326	0.411
	SP	86.617	0.074	0.326	0.390
	Tmax	56.397	0.036	0.219	0.160
	Evap	50.075	0.044	0.194	0.144
	K	40.070	0.058	0.142	0.116
	AT	47.000	0.042	0.151	0.109
	CUPE	27.811	0.044	0.113	0.032
	SLP	26.242	0.031	0.114	0.001