



Cause-driven Streamflow Forecasting Framework Based on Linear Correlation Reconstruction and Long Short-term Memory

Yani Lian¹ · Jungang Luo¹ · Wei Xue² · Ganggang Zuo¹ · Shangyao Zhang¹

Received: 24 December 2021 / Accepted: 22 February 2022 / Published online: 31 March 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

Reasonable runoff forecasting is the foundation of water resource management. However, the impact of environmental change on streamflow was not fully revealed due to the lack of enough streamflow features in many previous studies. In contrast, too many features also could lead cause undesired problems, including unstable model, interpretation difficulty, overfitting, high computational complexity, and high memory complexity. To address the above problems, this study proposes a cause-driven runoff forecasting framework based on linear-correlated reconstruction and machine learning model and refers to this framework as CSLM. We use variance inflation factor (VIF), pairwise linear correlation (PLC) reconstruction, and long short-term memory (LSTM) to realize this framework, referred to as VIF-PLC-LSTM. Four experiments were conducted to demonstrate the accuracy and efficiency of the proposed framework and its VIF-PLC-LSTM realization. Four experiments compare 1) different filter thresholds of driving factors, 2) different combination prediction features, 3) different reconstruction methods of linear-correlated features, and 4) different CSLM models. Experimental results on daily streamflow data from the Tangnaihai station at the Yellow River source and the Yangxian station at the Han River show that 1) data filtering has the risk of feature information loss, 2) when the streamflow, ERA5L, and meteorology data are used as inputs at the same time, the performance of the model is superior to the combination of other prediction features; the prediction effect of different prediction features, 3) the reconstruction of linear-correlated features is not only better than dimension reduction but also can improve the forecasting performance for streamflow prediction, and 4) among different CSLM models, LSTM is superior to other models.

Keywords Streamflow forecasting · Machine learning · Long short-term memory · Linear correlation reconstruction

✉ Jungang Luo
jgluo@xaut.edu.cn

¹ State Key Laboratory of Eco-Hydraulics in Northwest Arid Region, Xi'an University of Technology, Xi'an 710048, Shaanxi, China

² Hanjiang-To-Weihe River Water Diversion, Project Construction Co. Ltd, Shaanxi Province, Xi'an 710000, Shaanxi, China

1 Introduction

Reasonable runoff forecasting is the foundation of water resource management. It not only can support reservoir operation but also provide reference to water resources planning. However, streamflow is affected by regional environment factors under dynamic environments (Yang et al. 2021), leading to the formation mechanism. The evolution law of streamflow is constantly changing, which brings new challenges to streamflow prediction. Studies show that meteorology change, direct human activities, and underlying surface environment change are the main factors affecting runoff change (Jin et al. 2021; Su et al. 2021; Wang et al. 2021a; Li et al. 2009). Meteorology factors affect the streamflow, including precipitation, evaporation, and temperature (Zhang et al. 2021; López-Ballesteros et al. 2020). Underlying surface environment change factors also affect the streamflow, including terrain, vegetation cover, and land use (Chanapathi and Thatikonda 2020; Li et al. 2009). In addition, direct human activity, such as water use and drainage, can directly change the streamflow. Therefore, ideally, meteorology change, underlying surface change, and direct human activities should be considered in streamflow prediction.

The existing streamflow prediction methods are mainly physical and data-driven models (Yin et al. 2021). Physical mode required many boundary conditions and physical features to be determined in modeling hydrological processes with variabilities (Devia et al. 2015; Zuo et al. 2020). Further, most of the existing physical models were designed for the small watersheds, and thus it is laborious to verify the model in the large catchment due to the lack of a lot of relevant data (Nourani et al. 2011).

Recently, data-driven models have been widely used due to fewer requirements about the study area and low computational complexity. Data-driven models are mainly divided into machine learning (ML) models and time-series models. The linear assumption required for the time-series model limited its application for non-stationary and nonlinear streamflow prediction (Alizadeh et al. 2021). Long short-term memory (LSTM) (Alizadeh et al. 2021; Kim and Kim 2021), gradient boosting regression tree (GBRT) (He et al. 2020; Liao et al. 2020), deep neural networks (DNN) (Mahmoodzadeh et al. 2021), support vector regression (SVR) (Yang et al. 2017; Maity et al. 2010), and other ML models were explored for streamflow prediction.

Input plays an indispensable role in the modeling process. Reasonable input can decrease the complexity of the problem and improve the prediction performance of the algorithm. Previous studies on runoff prediction mainly focused on precipitation runoff as an input feature (Kratzert et al. 2018; Mao et al. 2021; Yokoo et al. 2022; Sedki et al. 2009). Because the selection of features can significantly affect the precision of the model (Vu et al. 2015), in recent years, the impact of meteorological features other than precipitation has attracted the attention of many researchers (Awotwi et al. 2021). (Xu et al. 2021) developed an ensemble streamflow forecast method under the considered rainfall, temperature, relative humidity, and land-use change. (Wang et al. 2021b) studied the contribution of soil, precipitation, land use, and temperature for streamflow increases in the high glacierized tributaries of Tarim River Basin, China. It is not hard to find that most studies only consider meteorology and underlying surface change because it is challenging to obtain data on the impact of direct human activities on runoff in real life. Therefore, this study considers the impact on streamflow from two aspects: meteorology and underlying surface change.

However, although many studies have considered meteorology and underlying surface change in runoff prediction, the selection of feature quantity or type is subjective, which

cannot guarantee the optimal feature selection without information loss. Fewer features cannot fully reflect the impact of meteorology and underlying surface change on streamflow. Although many features can provide more information, they can induce a higher computational complexity and computing resource requirements. Further, the model can be unstable and difficult to interpret due to multicollinearity. Thus, it is important to reflect the streamflow change with more detail while improving stability and computational efficiency.

To address the above problems, we propose a cause-driven streamflow forecasting framework based on linear-correlated reconstruction and machine learning model and refer to this framework as CSLM. The linear-correlated reconstruction includes variance inflation factor (VIF) and pairwise linear correlation (PLC) reconstruction. We use VIF, PLC reconstruction, and LSTM to realize this framework, referred to as VIF-PLC-LSTM. First, the prediction feature set is established based on the meteorology and ERA5-Land (referred to as ERA5L; see Sect. 2.2 for more information) data related to the cause of streamflow. Second, the redundant features are removed from the feature set in multicollinearity reconstruction by VIF, and then, the remaining features are reconstructed by PLC. Lastly, the reconstructed features are used for streamflow prediction. Four experiments were conducted to evaluate the accuracy and efficiency of the proposed CSLM framework and VIF-PLC-LSTM implementation. The first experiment showed the risk of feature information filtering by comparing driving factors screening based on different filtering thresholds MI. The second experiment analyzed the effects of different features: the streamflow, ERA5L, and meteorology data, on the prediction results. The results indicate that the best prediction performance can be derived with the existence of all three features. The third experiment explored the difference between the two linear correlation feature reconstruction methods based on principal component analysis (PCA) and VIF-PLC, showing the superiority of VIF-PLC. The fourth experiment compared the performance of SVR, GBRT, DNN, and LSTM based on the CSLM framework, showing that LSTM has the best prediction performance.

2 Study Area and Dataset

2.1 Study Area

In this study, we examine the streamflow forecasting performance of the two stations in the main study area, the source of the Yellow River, and the comparative study area, the upstream of the Han River in China, respectively. The source of the Yellow River (see Fig. S1), the second-largest river in China, is located in the Three Rivers national nature reserve of Qinghai Province. It is one of the most sensitive areas to meteorology change (Su et al. 2016; Sun et al. 2019) and is hardly affected by direct human activities. The previous research showed that meteorology change is the main factor leading to the streamflow change in the source of the three rivers, accounting for 90% (Jiang et al. 2017). The control basin above Tangnaihai station in the source area of the Yellow River covers an area of 12.19×10^4 km². Tangnaihai station is not only the control station on the mainstream of the Yellow River and the monitoring station of the Longyangxia Reservoir. Therefore, the predicted streamflow prediction of Tangnaihai station can be used to estimate the yield at the source of the Yellow River and the inflow of Longyangxia Reservoir, which is very important for the water resources management at the Yellow River.

The Han River (see Fig. S2), the largest tributary of the Yangtze River. The upstream of the Han River is the water source of the ‘Hanjiang-to-Weihe River Water Diversion Project.’ The Hanjiang-to-Weihe River Water Diversion Project is an inter-basin water diversion project in Shaanxi Province, transferring water from the Han River to the water-deficient area in Guanzhong of Shaanxi Province. Yangxian hydrological station has a catchment area of 14,484 km² and controls the mainstream of Hangjiang River with a length of 201.4 km. Therefore, the streamflow prediction of the Yangxian station upstream can effectively evaluate the adjustable water of the project. It has important strategic importance for alleviating the pressure of water resources in the Guanzhong area.

2.2 Dataset

The measured time-series data of Tangnaihaid and Yangxian stations are from the Yellow River Network and hydrological Yearbook. The measured streamflow of Tangnaihaid and Yangxian stations is obtained from 2006/05/02 to 2018/11/13 and 1981/01/02 to 2014/12/31, respectively.

The daily dataset of surface climatological data in China (V3.0) (hereinafter referred to as climatic dataset) is got from the China Meteorological Data Service Center. The climatic dataset includes 824 baseline and basic weather stations in China. The information of meteorological variables used in this study is shown in Table S1, where the small and large evaporations are converted to evaporation denoted by ‘EVP.’

ERA5L is a global land-surface dataset at 9 km resolution (Hereinafter referred to as the ERA5L dataset) obtained from ECMWF. The ERA5L is used to conduct lumped transformation on the same feature of different longitude and latitude in the control catchment area of Tangnaihaid and Yangxian stations, where the overall feature change in the catchment area is evaluated. The information of ERA5-Land variables is summarized in Table S2.

The total data for the Tangnaihaid station and Yangxian station is divided into the training, validation, and test sets with a ratio of 8:1:1. For Tangnaihaid station, the training, validation, test sets include data from 2006/05/02 to 2016/05/11, from 2016/05/12 to 2017/08/12, and from 2017/08/13 to 2018/11/13, respectively. For Yangxian station, the training, validation, and test sets include data from 1981/01/02 to 2008/03/14, from 2008/03/15 to 2011/08/07, and from 2011/08/08 to 2014/12/31, respectively.

3 Methods

3.1 Variance Inflation Factor

Variance inflation factor (VIF) (Vu et al. 2015) measures the severity of multicollinearity, defined as follows:

$$VIF = \frac{1}{1 - R^2} \quad (1)$$

where R^2 is the coefficient of determination. When the correlation between the target feature and other features is low, the closer the value of R^2 to 0. The closer the VIF value to 1, the lighter the multicollinearity. In general, a greater VIF value than 5 indicates a high linear correlation between the target feature and other features, which needs to be reconstructed (Vu et al. 2015).

3.2 Long Short-term Memory (LSTM)

LSTM was proposed to address the gradient disappearance problem that often occurs in recurrent neural networks when training long sequences (Hochreiter 1998). Similar to recurrent neural networks (Hochreiter and Schmidhuber 1997), LSTM is also structured with a neural network repeating module chain (Fig. S3) (Kratzert et al. 2018; Lian et al. 2022).

LSTM dominates the flow of information to the cell state through the forget gate (f_t), output gate (o_t), and input gate (i_t). The main process of LSTM is as follows:

1. In the forget gate, the sigmoid function σ decides which information to throw away. The f_t vector represents the preserved and abandoned information in the cell state C_{t-1} , represented as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

where W_f is the input weight matrix, and x_t is the current input vector. h_{t-1} is the last hidden cell state, and U_f is recurrent weight matrix. b_f represent the bias vector and $t \in (1, n)$ is each time step.

2. h_{t-1} and x_t are used to obtain new candidate cell information \tilde{C}_t through the \tanh layer, the process is shown in the following equation:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

where the \tanh represents activation function. The output of the \tanh function is 0, when the input is 0.

3. C_{t-1} is updated to get the C_t as follows.

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (5)$$

4. Which state characteristics of the output cell are based on the input h_{t-1} and x_t is judged. Next, the cell state gets a vector through the \tanh level. The final output of this unit as following equation, where the range of h_t is $[-1, 1]$.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

3.3 The CSLM Framework and the VIF-PLC-LSTM Realization

Few features cannot fully express the basin information, and too many features can make the model inefficient and unstable due to multicollinearity. In order to solve these problems, a CSLM framework based on the cause law of streamflow is proposed in this paper, as shown in Fig. 1. The relevant parameters of the model are shown in Table S3. The specific process of the framework is as follows:

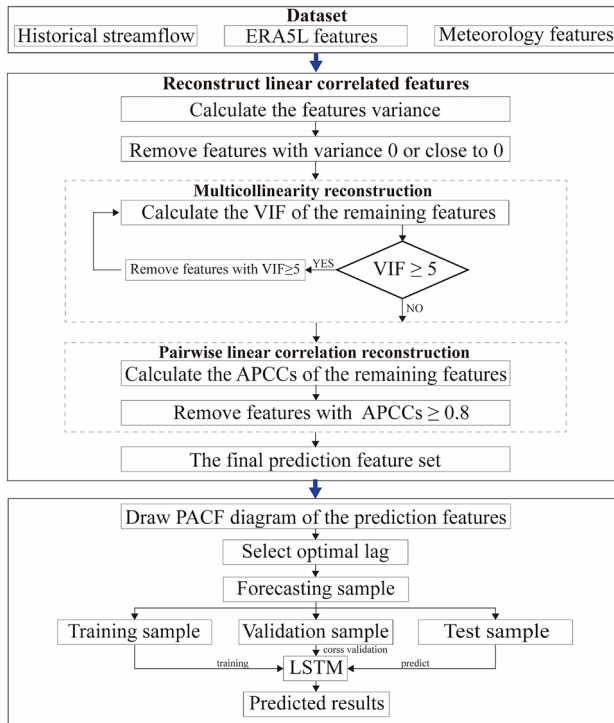


Fig. 1 The diagram of the CSLM framework, which is realized based on VIF-PLC-LSTM

Step 1 The meteorology, ERA5L, and streamflow data are collected and organized.

Step 2 The feature variances of meteorology and ERA5L are computed, respectively. Then, features whose lower variance (close to 0) is removed.

Redundant features with low variance are removed to avoid a complex model complexity and high computational time while ensuring sufficient information to prevent accuracy loss caused by the lack of features.

Step 3 VIF is used for multicollinearity reconstruction about meteorology and ERA5L features processed in Step2.

Despite the removal process of Step 2, there still remain redundant features that have multicollinearity. VIF further removes the multicollinearity features to avoid the instability and interpretation difficulty of the model.

Step 4 Judge whether the $VIF \geq 5$ of the feature in Step 3. If so, remove it, return to Step 3, calculate the VIF of the remaining features, and then judge again. Repeat Step 4 until the $VIF < 5$ of all the remaining features.

Step 5 The retained features in Step 4 are reconstructed by PLC.

The features with higher absolute Pearson correlation coefficients (APCCs) ($APCCs > 0.8$) are removed. The remaining features are fed into the prediction model as an input.

- Step 6 The partial autocorrelation function (PACF) diagram for each input feature data is drawn to select optimal lag. The forecasting sample is generated based on the optimal lag.
- Step 7 The forecasting samples are divided into training, validation, and test samples and then normalized with the same parameters.
- Step 8 The normalized training and validation samples are used to train and cross-validate the LSTM model. The prediction is conducted on the test samples.

3.4 Comparative Experimental Setups

The proposed CSLM framework and its VIF-PLC-LSTM realization are evaluated in the following four experiments. The first experiment compares the influence of different filter thresholds of driving factors on prediction results. The second experiment compares different prediction features: streamflow, ERA5L, and meteorology data on streamflow prediction results. The third experiment compares the effect of different reconstruction methods of linear-correlated features on streamflow prediction. The fourth experiment compares the different models based on the CSLM framework. The streamflow forecasting samples of the comparative experiment are generated based on the CSLM framework in Fig. 1. The sample sources of each experiment are summarized in Table S4. The details of the experiment are described in the following.

Experiment 1 Comparison of different filter thresholds of driving factors.

Ten filter thresholds (0.1 ~ 0.9) with the interval of 0.1 are compared on the normalized MI between predictors and streamflow in the S2 sample in Table S4.

Experiment 2 Comparison of different combinations of prediction features.

The sample S2 ~ S7 in Table S4 are generated based on the permutation and combination of streamflow, ERA5L, and meteorology data. Then, the prediction results of the streamflow, ERA5L, and meteorology data are compared.

Experiment 3 Comparison of different reconstruction methods of linear-correlated features.

S1 uses PCA and S2 uses VIF-PLC for the reconstruction of linear-correlated features. Thus, the comparison between S1 and S2 can analyze the PCA and VIF-PLC on reconstructing linear-correlated features.

Experiment 4 Comparison of different machine learning models.

Streamflow prediction by LSTM, DNN, GBRT, and SVR models for 1~7 days lead time are compared on S2 samples. Based on the comparison, the advantages and disadvantages of these models are analyzed.

4 Case Study

4.1 Calculation of Feature Variance

Figures S4 and S5 show the variance of meteorology and ERA5L features at Tangnaihahai station. They indicate no feature with a variance value of 0 but many features with small variance values. Also, there is a correlation between these selected features, which can easily lead to over-fitting. And, too many input variables often reduce the convergence speed. It means that these features contribute little to streamflow change and can be discarded. Therefore, in this paper, we remove the feature of variance less than 0.001.

4.2 Reconstruction of Linear-correlated Features

The remaining features after removing features whose difference is less than 0.001 are further tested and reconstructed. The process consists of the following two steps. Firstly, multicollinearity features are reconstructed by cycle removal. The feature with the highest VIF is removed repetitively until the remaining feature VIF is all less than 5. Table 1 depicts the process and results of the reconstruction of the multicollinearity meteorology features.

Then, the APCCs of the features, screened in the previous step, is computed; the results of meteorology features are shown in Fig. S6. As shown in Fig. S6, the APCCs between the MAX_T and MIN_ST exceeds 0.8, thus removing MAX_T. Therefore, the selected features of meteorology are MIN_ST and P2020 used to predict streamflow for the Tangnaihahai station. Since the APCCs between these features are all less than 0.8 after multicollinearity reconstruction on ERA5L features, no further reconstruction is unnecessary. smlt, evavt, sro, evabs, sd, sf, u10, and v10 are selected as the features of ERA5L in the prediction of the Tangnaihahai station.

4.3 Determination of Input Predictors

In ML model, PACF is widely used to select the optimal input lags (He et al. 2020). However, some lags that pass the 95% confidence test but are insignificant may lead to a high modeling time and computational cost (Zuo et al. 2020). Based on this, we select all lags before the first insignificant lag as the optimal input. Meteorology features MIN_ST, P2020, and ERA5L features smlt, evavt, sro, evabs, sd, sf, u10, v10 are used to predict the streamflow of the Tangnaihahai station in Sect. 4.1. After obtaining the feature dataset, the forecasting samples are generated. According to the CSLM framework in Fig. 1, PACF is used to select the optimal lag of features. The meteorology feature, smlt, of Tangnaihahai station is taken as an example to illustrate how to get the optimal lag. According to Fig. S7, The PACF of the lags after the seventh day is all around the blue line (95% confidence interval) and insignificant. Therefore, $x_{1(t)}$, $x_{1(t-1)}$, $x_{1(t-2)}$, $x_{1(t-3)}$, $x_{1(t-4)}$, $x_{1(t-5)}$, $x_{1(t-6)}$, and $x_{1(t-7)}$ are selected as the optimal input predictors of the smlt. The optimal input predictors of all features are selected and combined to obtain the final predictors of CSLM framework

Table 1 The process and results of the reconstruction of the multicollinearity meteorology feature at the Tangnathai station

fe(VIF) i=0	fe(VIF) i=1	fe(VIF) i=2	fe(VIF) i=3	fe(VIF) i=4	fe(VIF) i=5	fe(VIF) i=6	fe(VIF) i=7	fe(VIF) i=8	fe(VIF) i=9	fe(VIF) i=10	fe(VIF) i=11	fe(VIF) i=12	fe(VIF) i=13
AVG_PS (1.4E+06)	MAX_PS (1.3E+05)	MAX_W (4.5E+02)	MIN_PS (3.3E+02)	AVG_ST (1.8E+02)	MIN_T (9.3E+01)	AVG_T (8.5E+01)	MAX_ST (5.0E+01)	EXT_W (4.5E+01)	EVP (3.2E+01)	SSA (1.6E+01)	RHU (1.2E+01)	AVG_W (7.6E+00)	P2020 (1.7E+00)
MIN_PS (1.1E+06)	MIN_PS (1.3E+05)	EXT_W (4.1E+02)	RHU (2.3E+02)	MAX_ST (1.5E+02)	AVG_T (8.6E+01)	RHU (7.3E+01)	EXT_W (4.5E+01)	AVG_W (3.6E+01)	SSA (2.3E+01)	MAX_T (1.4E+01)	AVG_W (1.0E+01)	MAX_T (5.0E+00)	MAX_T (1.6E+00)
MAX_PS (1.6E+05)	MAX_W (4.5E+02)	MIN_PS (3.3E+02)	AVG_ST (1.8E+02)	MIN_ST (1.4E+02)	RHU (8.5E+01)	MAX_ST (5.0E+01)	EVP (3.7E+01)	EVP (3.3E+01)	AVG_W (1.9E+01)	RHU (1.3E+01)	MAX_T (6.4E+00)	MIN_ST (3.3E+00)	MIN_ST (1.0E+00)
MAX_W (4.5E+02)	EXT_W (4.1E+02)	RHU (2.3E+02)	MAX_ST (1.6E+02)	MIN_T (9.3E+01)	MIN_ST (7.1E+01)	EXT_W (4.6E+01)	MAX_T (3.7E+01)	SSA (2.3E+01)	MAX_T (1.9E+01)	AVG_W (1.1E+01)	MIN_ST (3.6E+00)	P2020 (2.1E+00)	
EXT_W (4.2E+02)	RHU (2.3E+02)	AVG_ST (1.8E+02)	MIN_ST (1.4E+02)	AVG_T (8.6E+01)	MAX_ST (5.9E+01)	MAX_T (4.4E+01)	AVG_W (3.6E+01)	MAX_T (1.9E+01)	RHU (1.5E+01)	MIN_ST (6.7E+00)	P2020 (2.5E+00)		
RHU (2.3E+02)	AVG_ST (1.8E+02)	MAX_ST (1.6E+02)	MIN_T (9.6E+01)	RHU (8.6E+01)	MAX_T (5.0E+01)	EVP (3.7E+01)	SSA (2.3E+01)	RHU (1.6E+01)	MIN_ST (6.7E+00)	P2020 (3.1E+00)			
AVG_ST (1.8E+02)	MAX_ST (1.6E+02)	MIN_ST (1.4E+02)	AVG_T (9.2E+01)	MAX_T (5.3E+01)	EXT_W (4.6E+01)	AVG_W (3.6E+01)	RHU (1.7E+01)	MIN_ST (8.1E+00)	P2020 (3.1E+00)				
MAX_ST (1.6E+02)	MIN_T (1.4E+02)	MIN_T (9.6E+01)	MAX_T (5.9E+01)	EXT_W (5.0E+01)	EVP (3.8E+01)	SSA (2.3E+01)	MIN_ST (9.0E+00)	P2020 (3.1E+00)					
MIN_ST (1.4E+02)	MIN_T (1.4E+02)	AVG_T (9.6E+01)	EXT_W (5.0E+01)	EVP (4.0E+01)	AVG_W (3.7E+01)	MIN_ST (1.2E+01)	P2020 (3.1E+00)						
MIN_T (1.4E+02)	AVG_T (9.6E+01)	MAX_T (5.9E+01)	AVG_W (4.1E+01)	AVG_W (3.7E+01)	SSA (3.1E+01)	P2020 (3.1E+00)							
AVG_T (9.3E+01)	MAX_T (5.9E+01)	AVG_W (4.4E+01)	EVP (4.1E+01)	SSA (3.3E+01)	P2020 (3.1E+00)								
MAX_T (5.9E+01)	AVG_W (4.4E+01)	EVP (4.1E+01)	SSA (3.3E+01)	P2020 (3.1E+00)									
AVG_W (4.4E+01)	EVP (4.1E+01)	SSA (3.3E+01)	P2020 (3.1E+00)										

Table 1 (continued)

fe(VIF) i=0	fe(VIF) i=1	fe(VIF) i=2	fe(VIF) i=3	fe(VIF) i=4	fe(VIF) i=5	fe(VIF) i=6	fe(VIF) i=7	fe(VIF) i=8	fe(VIF) i=9	fe(VIF) i=10	fe(VIF) i=11	fe(VIF) i=12	fe(VIF) i=13
EVP	SSA	P2020											
(4.1E+01)	(3.4E+01)	(3.1E+00)											
SSA	P2020												
(3.4E+01)	(3.1E+00)												
P2020													
(3.1E+00)													

fe (VIF) represents the feature (VIF)

for the Tangnaihahi station. Then, the 1-day lead time as an example, the predicted target is $Q_{(t+1)}$ of the measured daily streamflow.

4.4 Normalization of Learning Samples

All samples are normalized through (8) to find the optimal hyperparameter for the model.

$$y = 2 * \frac{x - x_{min}}{x_{max} - x_{min}} - 1 \quad (8)$$

where x_{min} , x_{max} , y , and x are the minimum, maximum, normalized, and original values, respectively. Please note that these parameters of the training set are also used to normalize the test and validation samples to avoid using the information of the test and validation samples.

4.5 Evaluation Indicators

The accuracy of the model is evaluated with three criteria: the peak percentage of threshold statistics (PPTS) (Lohani et al. 2014), the normalized root mean squared error (NRMSE) (He et al. 2020), and Nash–Sutcliffe efficiency (NSE) (He et al. 2019). Let denote $x(t)$, $\hat{x}(t)$ and $\bar{x}(t)$ be the measured, predicted, and average of the measured samples, respectively, and T be the number of samples. In PPTS, the samples are arranged in descending order, γ is the threshold level, which denotes the percentage of the arranged samples from the largest value. m is the number of values above γ . For instance, PPTS(5) represents that the top 5% peak samples are selected from the descending order of the measured sequence to calculate the index value.

$$NRMSE = \frac{\sqrt{\sum_{t=1}^T (x(t) - \hat{x}(t))^2 / T}}{\sum_{t=1}^T x(t) / T} \quad (9)$$

$$PPTS(\gamma) = \frac{100}{\gamma} \frac{1}{T} \sum_{t=1}^m \left| \frac{x(t) - \hat{x}(t)}{x(t)} \right| \quad (10)$$

$$NSE = 1 - \frac{\sum_{t=1}^T (x(t) - \hat{x}(t))^2}{\sum_{t=1}^T (x(t) - \bar{x}(t))^2} \quad (11)$$

5 Results and Discussion

5.1 Comparison of Different Filter Thresholds of Driving Factors

Figures S8 and S9 show the results of Experiment 1 (Sect. 3.4), where PPTS(5), NRMSE, and NSE of LSTM on 1-day ahead for the Tangnaihahi and Yangxian stations are depicted, respectively. As shown in Fig. S8, 1) the NSE fluctuates within a narrow range with the change of mutual information, 2) when the normalized mutual information is 0.1 or 0.2, the

NRMSE value of LSTM is lower, and 3) the NRMSE value increases with the normalized mutual information increase. PPTS also has a similar change law. As shown in Fig. S9, the LSTM has lower NRMSE, higher NSE, and PPTS (5) when the normalized mutual information is 0. It indicates a risk of information loss when the key predictors-driven are screened based on the normalized mutual information. Comparing with Figs. S8 and S9, it can be found that the daily streamflow prediction for the Tangnaihahi station has significantly lower PPTS(5) and NRMSE and higher NSE than the Yangxian station. It indicates the prediction results of the Tangnaihahi station are better than the Yangxian station because the streamflow of the Yangxian station maybe more affected by human activities than that of the Tangnaihahi station.

5.2 Comparison of Different Combinations of Prediction Features

The results of Experiment 2 (Sect. 3.4) are shown in Fig. S10. The details of S2~S7 samples used in this experiment are shown in Table S4. Figure S10 shows NSE, NRMSE, and PPTS(5) of LSTM on the 1-day ahead based on training, validation, and test samples for the Tangnaihahi and Yangxian stations, respectively.

From Fig. S10, the following results can be found. 1) The mean NSE of S2, S3, and S4 are all significantly higher than that of S5, S6, and S7, while the mean NRMSE and PPTS(5) of S2, S3, and S4 are significantly lower than that of S5, S6, and S7. It indicates that the streamflow series can substantially improve the prediction accuracy of LSTM. The NSE mean of S2 is larger than that of S3 and S4, and the NSE mean of S5 is larger than that of S6 and S7. 2) The mean NRMSE and PPTS(5) of S2 are generally lower than that of S3 and S4, and the mean NRMSE and PPTS(5) of S5 are also generally lower than that of S6 and S7. The results show that utilizing meteorology and ERA5L data simultaneously is beneficial to improve the runoff prediction effect. 3) The mean NSE of S4 and S7 are higher than S3 and S6, respectively. The mean NRMSE and PPTS(5) of S4 and S7 are lower than S3 and S7, respectively. It indicates that the streamflow forecasting accuracy of ERA5L is slightly better than meteorology data. 4) The mean NSE of the Tangnaihahi station is higher than the Yangxian station, indicating that the influence of direct human activities on streamflow change is significant.

When using only the meteorology or ERA5L data, the streamflow prediction is worse than using the streamflow data due to the following reasons: 1) the quality of data is affected by interpolation and observation; 2) the model needs to fit the errors from different sources; 3) the sample of streamflow prediction based on the statistical law cannot reflect the law of runoff generation and concentration; 4) the streamflow is affected by direct human activities. The forecasting accuracy of ERA5L is slightly better than that of meteorology data due to the reanalysis based on measure data, higher spatial and temporal resolution, and more comprehensive response to relevant characteristic information in runoff generation and concentration.

5.3 Comparison of Different Reconstruction Methods of Linear-Correlated Features

Experiment 3 (Sect. 3.4) results are shown in Fig. S11, where PPTS(5), NRMSE, and NSE of LSTM on 1-day ahead based on training, validation, and test samples for the Tangnaihahi and Yangxian stations are depicted. In this experiment, S1 and S2 samples (Table S4) are generated by PCA and VIF-PLC, respectively. The results show that the reconstruction of linear-correlated features (S2) has a higher mean, lower interquartile

range of NSE, and lower average and interquartile range of NRMSE than feature dimension reduction (S1). The Tangnaihahi station obtains a lower mean and higher interquartile range of PPTS(5). In contrast, the Yangxian station obtains a higher mean and lower interquartile range of PPTS(5).

These results indicate that the VIF-PLC outperforms the PCA in terms of streamflow prediction. The possible reasons led this result are as follows: 1) some principal components (PC) with smaller variance are excluded by PCA, resulting in a certain degree of information loss; 2) in the feature dimensionality reduction, more PCs of meteorology data and fewer PCs of ERA5L data are selected. The reconstruction of linear-correlated features include fewer meteorology features and more ERA5L features. As can be seen from Sect. 5.2, since the accuracy of ERA5L is better than meteorology data on streamflow prediction, a higher proportion of ERA5L could result in better accuracy.

5.4 Comparison of Different Machine Learning Models

The performance of different streamflow prediction models is compared in Experiment 4 (see Sect. 3.4), where S2 (Table S4) samples are used. The LSTM, DNN, GBRT, and SVR are used as the compared models in predicting the daily streamflow on the Tangnaihahi and Yangxian stations with 1~7 days lead time. Figures S12 and S13 depict the NSE box figure of different models. Figures 2 and 3 compare the measured results and predicted results of different models with 1, 3, 5, and 7 days ahead.

The results found from Fig. S12 are as follows. 1) the NSE interquartile range of the LSTM and DNN is significantly smaller than that of the GBRT and SVR. It indicates the better generalization performance of the LSTM and DNN than the GBRT and SVR. 2) the NSE of LSTM is higher than DNN, indicating that the streamflow prediction accuracy of LSTM is superior to the DNN. The results found from Fig. S13 are as follows. 1) the NSE interquartile range of the LSTM and DNN is slightly smaller than that of the GBRT and SVR, indicating that the streamflow forecasting results of the LSTM and DNN is slightly better than the GBRT and SVR. 2) the GBRT and SVR have a higher mean NSE but a larger interquartile range, indicating the poor generalization performance of GBRT and SVR. The same conclusion can be reached for PPTS(5) and NRMSE.

As shown in Fig. 2, the four models, the LSTM, SVR, GBRT, and DNN, provide good trend and periodicity fitting abilities. Also, Fig. 3a shows that these four models can capture random changes to predict daily runoff with one day ahead at the Yangxian station. However, as seen from Fig. 3b–d, the tracking ability of these models to the peak runoff gradually decreased with increasing lead time.

From Figs. S12, S13, 2 and 3, it can be seen that the forecasting accuracy of the streamflow forecasting model mainly depends on the sample quality. The streamflow series of the Tangnaihahi station is almost not affected by human activities, thus inducing better prediction. However, the streamflow series of the Yangxian station is greatly affected by human activities, consequently resulting in poor prediction accuracy. The GBRT and SVR have a larger NSE interquartile range and a mean higher NSE due to the larger NSE values of training and development samples but smaller NSE values of the test samples. In summary, LSTM has great potential for streamflow prediction based on causes because the LSTM model can simulate the change of hydrometeorological time series and the long-term dependence between predictors and predicted targets.

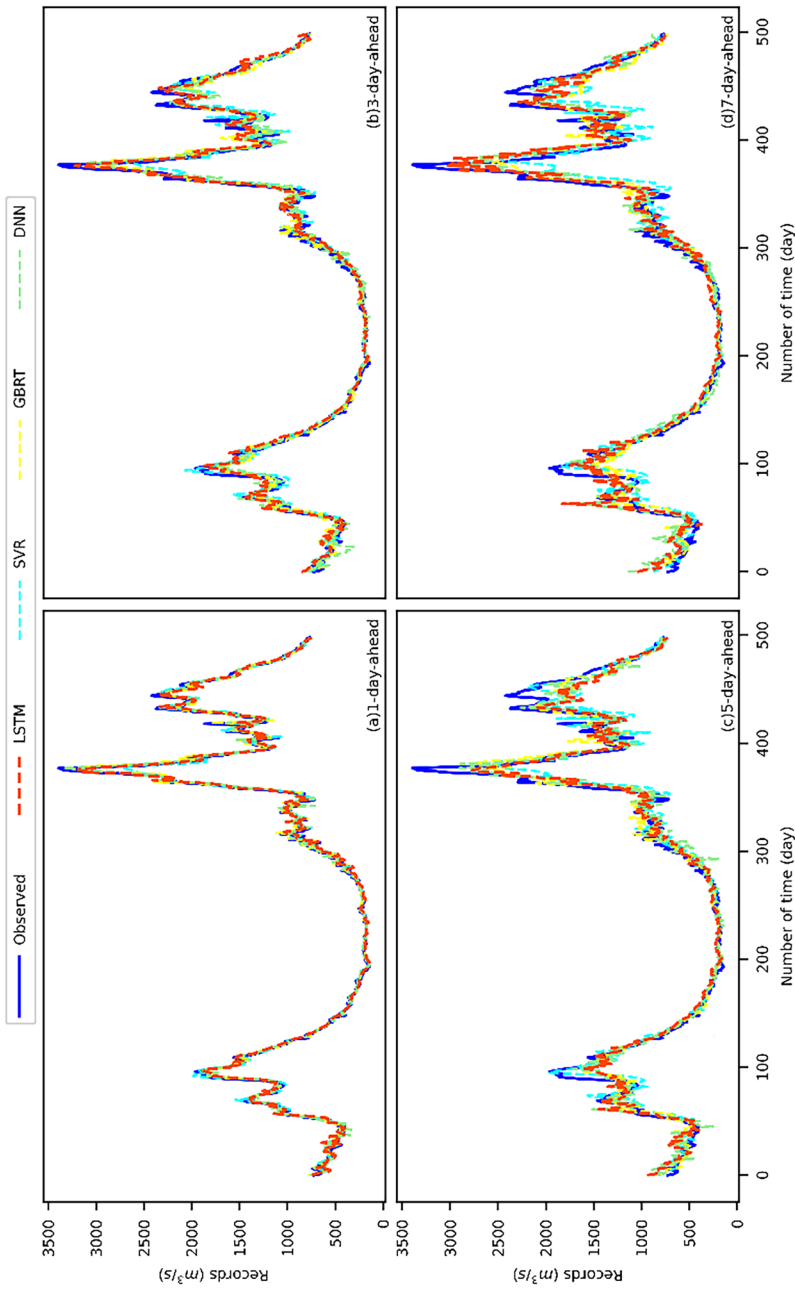


Fig. 2 Predicted and measured results for the testing sample of S2 samples at the Tangnaihai station

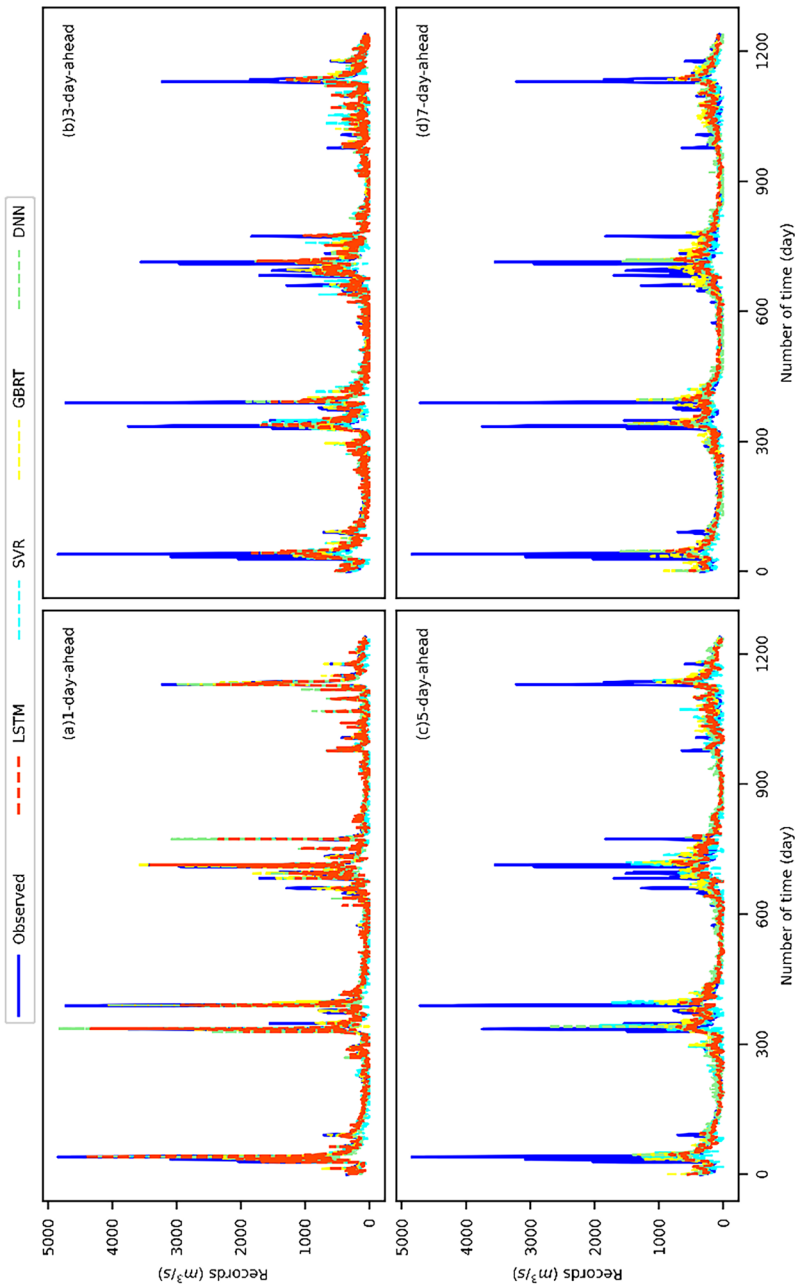


Fig. 3 Predicted and measured results for the testing sample of S2 samples at the Yangxian station

6 Conclusion

This paper proposes the CSLM framework and its VIF-PLC-LSTM realization for streamflow prediction based on linear correlation reconstruction. The CSLM input feature is based on the cause-driven features of streamflow, which can better reflect catchment changes. The CSLM reconstructs the linear correlation features to avoid serious collinearity, making the model stable. The CSLM prediction stage reduces the risk of overfitting while improving the model efficiency and prediction accuracy. Four experiments demonstrate that the proposed framework and realization can simulate daily streamflow with competitive performance compared to benchmark models. With the first set experiment, we evaluate the risk of filtering features in streamflow prediction. With the second set experiment, we evaluate the necessity of meteorology, ERA5L, and streamflow features for streamflow prediction. With the third set experiment, we evaluate the difference between PCA and VIF-PLC for linear correlation reconstruction. With the last set experiment, we evaluate the forecasting performance of the different CSLM models. The main conclusions summarized as follows.

1. There is a risk of feature information loss in driving factors filtering based on different filtering thresholds MI .
2. Historical streamflow is an essential predictor in streamflow prediction. The prediction effect is poor to use features excluding streamflow to predict streamflow. Both meteorological feature and ERA5L feature can significantly improve the performance of runoff forecasting. The predicted effect of the ERA5L feature is better than the meteorological feature.
3. The VIF-PLC is not only better than PCA but also can improve the prediction performance for runoff forecasting.
4. The prediction performance of LSTM is better than SVR and GBRT based on the same CSLM framework.

The contribution of this study is to propose a prediction framework based on linear correlation reconstruction, considering the impact on streamflow from meteorology and underlying surface change. In addition, the variance inflation factor is used to remove the redundant features, and pairwise linear correlation reconstruction is used for feature reconstruction. The results of this paper demonstrate the superiority of the CSLM framework and the ability of the VIF-PLC-LSTM in runoff prediction.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11269-022-03097-1>.

Acknowledgements I sincerely appreciate the language editing by EDITIDEAS and the editors and reviewers for their comments.

Authors Contributions Conceptualization: Lian YN and Zuo GG; Supervision and Resources: Luo JG; Data curation: Xue W; Writing-original draft preparation: Zuo GG and Lian YN; Software: Zhang SY.

Funding This work was supported by the Research Fund of the State Key Laboratory of Eco-hydraulics in Northwest Arid Region, Xi'an University of Technology (Grant No. 2019KJCXTD-5), the Natural Science Basic Research Program of Shaanxi (Grant No. 2019JLZ-15), and the National Natural Science Foundation of China (Grant Nos. 51979221).

Data and Code Availability All data are available on the given website and Hydrological Yearbook.

Declarations

Ethical Approval None.

Consent to Participate All participants agreed.

Consent to Publish The publication of this manuscript was approved by all authors.

Competing Interests None.

References

- Alizadeh B, Bafti AG, Kamangir H, Zhang Y, Wright DB, Franz KJ (2021) A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. *J Hydrol* 601:126526
- Awotwi A, Annor T, Anornu GK, Quaye-Ballard JA, Agyekum J, Ampadu B, Nti IK, Gyampo MA, Boakyee E (2021) Climate change impact on streamflow in a tropical basin of Ghana, West Africa. *J Hydrol Reg Stud* 34:100805
- Chanapathi T, Thatikonda S (2020) Investigating the impact of climate and land-use land cover changes on hydrological predictions over the Krishna river basin under present and future scenarios. *Sci Total Environ* 721:137736
- Devia GK, Ganasri BP, Dwarakish GS (2015) A review on hydrological models. *Aquat Procedia* 4:1001–1007
- He X, Luo J, Li P, Zuo G, Xie J (2020) A hybrid model based on variational mode decomposition and gradient boosting regression tree for monthly runoff forecasting. *Water Resour Manag* 34(2):865–884
- He X, Luo J, Zuo G, Xie J (2019) Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks. *Water Resour Manag* 33(4):1571–1590
- Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions
- Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Jiang C, Li D, Gao Y, Liu W, Zhang L (2017) Impact of climate variability and anthropogenic activity on streamflow in the Three Rivers Headwater Region, Tibetan Plateau, China. *Theor Appl Climatol* 129(1–2):667–681
- Jin S, Zheng Z, Ning L (2021) Separating variance in the runoff in Beijing's river system under climate change and human activities. *Phys Chem Earth Parts A/B/C* 123:103044
- Kim C, Kim C-S (2021) Comparison of the performance of a hydrologic model and a deep learning technique for rainfall-runoff analysis. *Trop Cyclone Res Rev* 10(4):215–222
- Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol Earth Syst Sci* 22(11):6005–6022
- Li Z, Liu W, Zhang X, Zheng F (2009) Impacts of land use change and climate variability on hydrology in an agricultural catchment on the Loess Plateau of China. *J Hydrol* 377(1–2):35–42
- Lian Y, Luo J, Wang J, Zuo G, Wei N (2022) Climate-driven model based on long short-term memory and bayesian optimization for multi-day-ahead daily streamflow forecasting. *Water Resour Manag* 36(1):21–37
- Liao S, Liu Z, Liu B, Cheng C, Jin X, Zhao Z (2020) Multistep-ahead daily inflow forecasting using the ERA-Interim reanalysis data set based on gradient-boosting regression trees. *Hydrol Earth Syst Sci* 24(5):2343–2363
- Lohani AK, Goel NK, Bhatia K (2014) Improving real time flood forecasting using fuzzy inference system. *J Hydrol* 509:25–41
- López-Ballesteros A, Senent-Aparicio J, Martínez C, Pérez-Sánchez J (2020) Assessment of future hydrologic alteration due to climate change in the Arachthos River basin (NW Greece). *Sci Total Environ* 733:139299
- Mahmoodzadeh A, Mohammadi M, Noori KMG, Khishe M, Ibrahim HH, Ali HFH, Abdulhamid SN (2021) Presenting the best prediction model of water inflow into drill and blast tunnels among several machine learning techniques. *Autom Construct* 127:103719

- Maity R, Bhagwat PP, Bhatnagar A (2010) Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrol Process* 24(7):917–923
- Mao G, Wang M, Liu J, Wang Z, Wang K, Meng Y, Zhong R, Wang H, Li Y (2021) Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. *Phys Chem Earth Parts A/B/C* 123:103026
- Nourani V, Kisi Ö, Komasi M (2011) Two hybrid Artificial Intelligence approaches for modeling rainfall-runoff process. *J Hydrol* 402(1–2):41–59
- Sedki A, Ouazar D, El Mazoudi E (2009) Evolving neural network using real coded genetic algorithm for daily rainfall-runoff forecasting. *Expert Syst Appl* 36(3):4523–4527
- Su F, Zhang L, Ou T, Chen D, Yao T, Tong K, Qi Y (2016) Hydrological response to future climate changes for the major upstream river basins in the Tibetan Plateau. *Global Planet Change* 136:82–95
- Su X, Li X, Niu Z, Wang N, Liang X (2021) A new complexity-based three-stage method to comprehensively quantify positive/negative contribution rates of climate change and human activities to changes in runoff in the upper Yellow River. *J Clean Prod* 287:125017
- Sun L, Wang Y-Y, Zhang J-Y, Yang Q-L, Bao Z-X, Guan X-X, Guan T-S, Chen X, Wang G-Q (2019) Impact of environmental change on runoff in a transitional basin: Tao River Basin from the Tibetan Plateau to the Loess Plateau, China. *Adv Clim Chang Res* 10(4):214–224
- Vu DH, Muttaqi KM, Agalgaonkar AP (2015) A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. *Appl Energy* 140:385–394
- Wang Q, Guan Q, Lin J, Luo H, Tan Z, Ma Y (2021a) Simulating land use/land cover change in an arid region with the coupling models. *Ecol Indic* 122:107231
- Wang X, Luo Y, Sun L, Shafeeqe M (2021b) Different climate factors contributing for runoff increases in the high glacierized tributaries of Tarim River Basin, China. *J Hydrol Reg Stud* 36:100845
- Xu ZP, Li YP, Huang GH, Wang SG, Liu YR (2021) A multi-scenario ensemble streamflow forecast method for Amu Darya River Basin under considering climate and land-use changes. *J Hydrol* 598:126276
- Yang T, Asanjan AA, Welles E, Gao X, Sorooshian S, Liu X (2017) Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resour Res* 53(4):2786–2812
- Yang W, Jin F, Si Y, Li Z (2021) Runoff change controlled by combined effects of multiple environmental factors in a headwater catchment with cold and arid climate in northwest China. *Sci Total Environ* 756:143995
- Yin H, Zhang X, Wang F, Zhang Y, Xia R, Jin J (2021) Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. *J Hydrol* 598:126378. <https://doi.org/10.1016/j.jhydrol.2021.126378>
- Yokoo K, Ishida K, Ercan A, Tu T, Nagasato T, Kiyama M, Amagasaki M, (2022) Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM. *Sci Total Environ* 802:149876
- Zhang L, Yuan F, Wang B, Ren L, Zhao C, Shi J, Liu Y, Jiang S, Yang X, Chen T, Liu S (2021) Quantifying uncertainty sources in extreme flow projections for three watersheds with different climate features in China. *Atmosph Res* 249:105331
- Zuo G, Luo J, Wang N, Lian Y, He X (2020) Two-stage variational mode decomposition and support vector regression for streamflow forecasting. *Hydrol Earth Syst Sci* 24(11):5491–5518

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.