Climate-driven Model Based on Long Short-Term Memory and Bayesian Optimization for Multi-day-ahead Daily Streamflow Forecasting

Yani Lian
Xi'an University of Technology
Jungang Luo (jgluo@xaut.edu.cn)
Xi'an University of Technology
Jingmin Wang
Hanjiang-to-Weihe River Water Diversion Project Construction Co.Ltd.,Shaanxi Province
Ganggang Zuo
Xi'an University of Technology

Research Article

Keywords: Climate-driven model, streamflow forecasting, Bayesian optimization, Principal component analysis, Long short-term memory

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-216524/v1

Version of Record: A version of this preprint was published at Water Resources Management on November 4th, 2021. See the published version at https://doi.org/10.1007/s11269-021-03002-2.

Climate-driven Model Based on Long Short-Term Memory and Bayesian Optimization for Multi-day-ahead Daily Streamflow Forecasting

Yani Lian[1] · Jungang Luo[1]* · Jingmin Wang[2] · Ganggang Zuo[1]

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

[2]Hanjiang-to-Weihe River Water Diversion Project Construction Co.Ltd., Shaanxi Province, Xi'an, Shaanxi 710100, China

E-mail: yanilian@foxmail.com (Y.Lian); wangjingmin@hwrvwd.cn (J.Wang); zuoganggang@163.com (G.Zuo)

*Correspondence: jgluo@xaut.edu.cn

Abstract: Many previous studies have developed decomposition and ensemble models to improve runoff forecasting performance. However, these decomposition-based models usually introduce large decomposition errors into the modeling process. Since the variation in runoff time series is greatly driven by climate change, many previous studies considering climate change focused on only rainfall-runoff modeling, with few meteorological factors as input. Therefore, a climate-driven streamflow forecasting (CDSF) framework was proposed to improve the runoff forecasting accuracy. This framework is realized using principal component analysis (PCA), long short-term memory (LSTM) and Bayesian optimization (BO) referred to as PCA-LSTM-BO. To validate the effectiveness and superiority of the PCA-LSTM-BO method with which one autoregressive LSTM model and two other CDSF models based on PCA, BO, and

either support vector regression (SVR) or, gradient boosting regression trees (GBRT), namely, PCA-SVR-BO and PCA-GBRT-BO, respectively, were compared. A generalization performance index based on the Nash-Sutcliffe efficiency (NSE), called the GI(NSE) value, is proposed to evaluate the generalizability of the model. The results show that (1) the proposed model is significantly better than the other benchmark models in terms of the mean square error (MSE<=185.782), NSE>=0.819, and GI(NSE) <=0.223 for all the forecasting scenarios; (2) the PCA in the CDSF framework can improve the forecasting capacity and generalizability; (3) the CDSF framework is superior to the autoregressive LSTM models for all the forecasting scenarios; and (4) the GI(NSE) value is demonstrated to be effective in selecting the optimal model with a better generalizability.

Keywords: Climate-driven model; streamflow forecasting; Bayesian optimization; Principal component analysis; Long short-term memory

1. Introduction

Accurate runoff forecasting is vital for water resource planning and management. Therefore, the development of runoff prediction models has already attracted significant attention in recent decades. These models are mainly divided into physical models and data-driven models (Gong et al. 2016; Shirmohammadi et al. 2013; Wu et al. 2009). The Physical models have high requirements for the input meteorological data, information about physical properties as well as boundary conditions, and computational resources. Hence, it is rarely used in practical applications. Data-driven models are widely used because of their simplicity and low information requirements

(Fang et al. 2019; Wu and Huang 2009).

Previous data-driven streamflow forecasting has focused on time series models, such as Box-Jenkins and autoregressive moving average (ARMA) models (Ramaswamy and Saleh 2020; Sun et al. 2019; Zhang et al. 2011). However, these time series models fail to reasonably forecast the nonlinear runoff series due to the stationarity assumption (Dariane et al. 2018; He et al. 2019). Machine learning models such as support vector regression (SVR) (Cheng et al. 2015; Hadi and Tombul 2018; Lin et al. 2009; Maity et al. 2010; Vapnik et al. 1996), gradient boosting regression trees (GBRT) (He et al. 2020; Persson et al. 2017; Zhang and Haghani 2015) and artificial neural networks (ANNs) (Chua and Wong 2011; Gauch et al. 2020; Kisi et al. 2012; Kratzert et al. 2018) can address the nonstationary and nonlinear problems of runoff prediction (Friedman 2001; Kumar et al. 2019; Vapnik et al. 1996). However, the pure ML models still perform poorly for predicting complex nonlinear and nonstationary runoff series while the model is developed without input meteorological factors. In many study cases with available input meteorological data, only rainfall information was considered to predict runoff (Alizadeh et al. 2017; Chang et al. 2017; Kratzert et al. 2018; Sedki et al. 2009). Although rainfall-runoff modeling contributes greatly to improving runoff forecasting performance, more meteorological information is needed to further improve the streamflow forecasting performance. Additionally, signal decomposition algorithms such as ensemble empirical mode decomposition (EEMD), the discrete wavelet transform (DWT), singular spectrum analysis (SSA) and variational mode decomposition (VMD) have been introduced to handle the nonlinear

and nonstationary problems of raw runoff series. However, these algorithms introduce large decomposition errors when performing decomposition without using future information (Du et al. 2017; Quilty and Adamowski 2018; Tan et al. 2018; Zhang et al. 2015).

To address these problems, a climate-driven streamflow forecasting (CDSF) framework is proposed. In the CDSF framework, the meteorological data dimensionality is first reduced to save computational resources and modeling time and decrease the overfitting risk. A data-driven model is then used to implement the functionality of predicting runoff with these meteorological data as input. The hyperparameters of this data-driven model are finally tuned by an optimization algorithm. Variation analysis (Narayan and Ghosh 2021; Sheng et al. 2020), cluster analysis (Kourtit et al. 2021) and principal component analysis (PCA) (Cao et al. 2003; George and Vidyapeetham 2012), etc. could be used to reduce dimensionality. However, PCA is highly efficient and can transform the input variables into uncorrelated variables (Abdi and Williams 2010; Svante .Wold et al. 1987). Many data-driven models such as SVR, GBRT, backpropagation neural network (BPNN), and long short-term memory (LSTM) models, could implement the functionality of runoff forecasting. However, SVR and GBRT, which are shallow learning method, are very sensitive to hyperparameter selection and have a low capacity to represent distinct information (Li et al. 2020). BPNN models usually suffer from overfitting, local convergence, and low learning speed (Bisoyi et al. 2019). The LSTM model, a deep learning model, can address these drawbacks and learn long-term dependency from input meteorological

data (Bai et al. 2019; Bai et al. 2020; Yin et al. 2020). However, trial and error, grid search (GS), genetic algorithm (GA), random search (RS), Bayesian optimization (BO), etc. could be used for hyperparameter optimizations, BO is especially useful for expensive function evaluation (Bergstra and Bengio 2012; Dewancker et al. 2016; Rasmussen 2004; Snoek et al. 2012; Su et al. 2014), e.g., tuning the hyperparameters of LSTM in this study. Therefore, we realized this CDSF framework using PCA, LSTM, and BO, namely PCA-LSTM-BO.

To validate the effectiveness and superiority of the CDSF framework and the PCA-LSTM-BO model, one autoregressive LSTM model and two other CDSF models based on PCA, BO, SVR, and GBRT, namely PCA-SVR-BO and PCA-GBRT-BO, were compared. The first experiment compares the prediction performance of different the CDSF models to show the superiority of LSTM. The second experiment compares the proposed model with autoregressive models to prove the stability and high generalizability of the CDSF framework. The proposed model as well as the benchmark models are evaluated using daily runoff data and meteorological data collected from four stations located in the Huangshui River catchment, China.

2. Study Area and Data

The Huangshui River (see **Fig. 1**) is an important tributary of the upper reaches of the Yellow River, located eastern Qinghai Province, China. The Huangshui River has a total length of 374 km and a drainage area of 3.286×10^4 km^2. It originates from the Baohutu Mountains in Haiyan County, Qinghai Province, and flows through the longitudinal valley between the Datong-Daban and Laji Mountains in Qinghai Province.

The Huangshui River has a continental climate, and because of the great terrain difference in this area, the temporal and spatial variations in temperature are also large. The average annual runoff of the Huangshui River is 4.65×10^8 m^3, and that of the Minhe station on the mainstream is 1.79×10^8 m^3. Nearly 60% of Qinghai's population, 52% of cultivated land and more than 70% of industrial and mining enterprises are concentrated in the Hehuang Valley. It is also the main source of urban water in Lanzhou. Therefore, robust runoff prediction plays a vital role in production and life in this area.

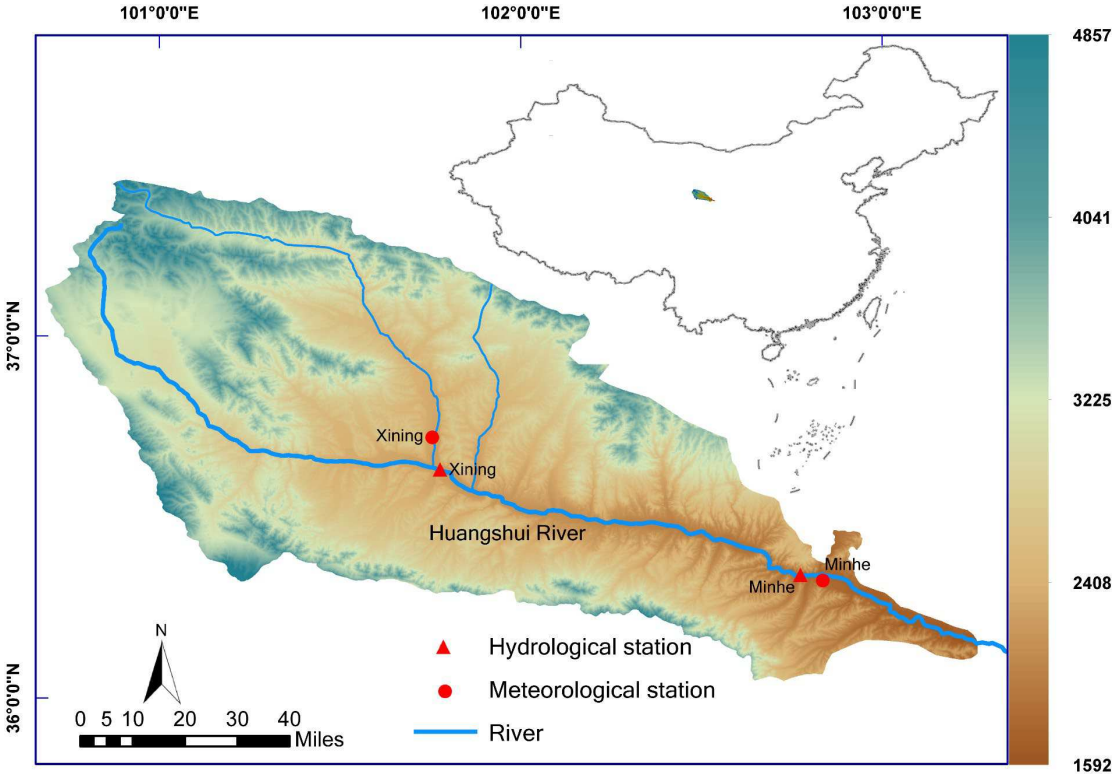

Fig. 1. A geographical overview of Huangshui River catchment in China.

In this study, the historical daily runoff data and daily climate data (see **Table 1**) of the Xining and Minhe stations from January 1, 2006, to December 31, 2013 (2922 records for each feature), are used to evaluate the proposed model and the benchmark models. The records were collected from the China Meteorological Data Network. The daily runoff data and daily climate data of each station are divided into three parts: a

training set, validation set and testing set. These sets account for approximately 60%, 20% and 20%, respectively, of the entire data.

Table 1 Modeling data at the Xining and Minhe stations.

Time series	Unit	Data range	Data partition
Average air temperature	℃	2006/01/01-2013/12/31	Training set (60%) validation set (20%) testing set (20%)
Maximum air temperature	℃		
Minimum air temperature	℃		
Average relative humidity,	%		
Minimum relative humidity	%		
Average wind speed	m/s		
Maximum wind speed	m/s		
Precipitation	mm		
Average pressure,	pa		
Maximum pressure,	pa		
Minimum pressure	pa		
Maximum surface temperature	℃		
Minimum surface temperature	℃		
Streamflow(Q)	m^3/s		

3. Methodology

3.1. Principal component analysis

PCA is a simple and useful tool to reduce the dimensionality of a set of correlated features while preserving as much original information as possible (Abdi and Williams 2010). PCA reduces the dimensionality by transforming a set of original variables into a smaller set of uncorrelated variables called principal components (PCs) (Helena et al. 2000). The process of PCA focuses on seeking a larger variance within the same variable but a small covariance among different variables. The PCA method has five main steps: (1) standardizing the original variables, (2) computing the covariance matrix of the standardized variables, (3) computing the eigenvectors and eigenvalues of the covariance matrix, (4) forming the feature vector from the eigenvectors and

eigenvalues, and (5) recasting the original variables to PCs. The PCs are new variables that are constructed as linear combinations of the original variables (Davis and Sampson 1986; Noori et al. 2011).

$$Z_i = \boldsymbol{a_{i1}} * X_1 + \boldsymbol{a_{i2}} * X_2 + \cdots + \boldsymbol{a_{iP}} * X_p \quad (1)$$

where Z_i represent the PCs; $\boldsymbol{a_{i1}}$ are the related eigenvectors; X_i are the input variables; and p represents the number of input variables. These parameters are calculated by the following formula:

$$|\boldsymbol{R} - \boldsymbol{I}\lambda| = 0 \quad (2)$$

where $\boldsymbol{R}$ is the variance-covariance matrix, $\boldsymbol{I}$ is the identity matrix and λ is the eigenvalue.

The number of PCs (or predictors) is the only parameter that should be predefined in PCA. The number of predictors for each station is different. To facilitate comparisons, the number of predictors is replaced by the number of excluded predictors. In this paper, the number of excluded predictors ranges from 0 to 18 or 19 (half of the total number of predictors at each station). We also estimated the optimal number of predictors by using maximum likelihood estimation (MLE) (Minka 2001).

3.2. Long short-term memory

LSTM is a special kind of neural network, that is generated to solve the problem of gradient explosion and disappearance of recurrent neural networks in long sequence training. Similar to recurrent neural networks, LSTM also has the structure of a neural network repeating module chain (Bengio et al. 1994). Its structure is shown in **Fig. 2**.

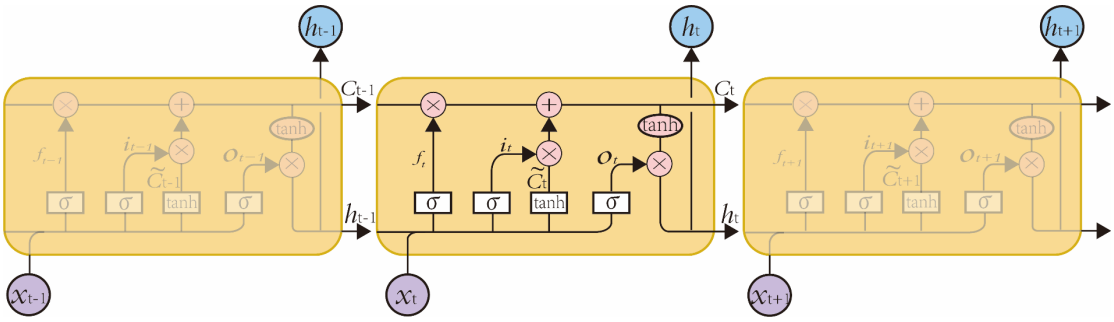

Fig. 2. Diagram of LSTM.

LSTM mainly controls the flow of information to the cell state by three "gates": a forget gate $(\boldsymbol{f}_t)$, an input gate $(\boldsymbol{i}_t)$ and an output gate $(\boldsymbol{o}_t)$. The first step is to use the sigmoid function σ of the forget gate to determine what information the cell state needs to discard. It outputs a vector $\boldsymbol{f}_t$ whose range is (0,1) through the information of $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_t$. The value of this vector indicates which information in the cell state $\boldsymbol{C}_{t-1}$ is retained and which is discarded (Kratzert et al. 2018).

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{3}$$

where t is each time step from $t = 1$ to $t = n$; $\boldsymbol{h}_{t-1}$ is the last hidden cell state; $\boldsymbol{x}_t$ is the current input vector. $\boldsymbol{W}_f$, $\boldsymbol{U}_f$ and $\boldsymbol{b}_f$ are the input weight matrix, recurrent weight matrix and bias vector, respectively.

The second step is to determine which new information is used to update the cell state by $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_t$ (Zuo et al. 2020a). Then, $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_t$ are used to obtain new candidate cell information $\widetilde{\boldsymbol{C}}_t$ through the tanh layer, which may be updated with cell information, as follows (Kratzert et al. 2018):

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{4}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_{\tilde{C}} \boldsymbol{x}_t + \boldsymbol{U}_{\tilde{C}} \boldsymbol{h}_{t-1} + \boldsymbol{b}_{\tilde{C}}) \tag{5}$$

where $\tanh(\cdot)$ is the tanh activation function, which maps a real number input to [-11]; when the input is 0, the output of the tanh function is also 0.

The third step is to update the old cell information $\boldsymbol{C}_{t-1}$ and obtain the new cell

information $\boldsymbol{C}_t$. The update rule is to choose to forget part of the old cell information through $\boldsymbol{f}_t$ and to add part of the candidate cell information $\widetilde{\boldsymbol{C}}_t$ through $\boldsymbol{i}_t$ to obtain new cell, as follows:

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

The fourth step, after updating the cell state, is to judge which state characteristics of the output cell are based on the input $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_t$; the judgment condition is obtained by inputting through the sigmoid function layer of $\boldsymbol{o}_t$. Then the cell state is passed through the tanh layer to obtain a vector. This vector is multiplied by the judgment condition obtained by $\boldsymbol{o}_t$ to obtain the final output of this unit as follows (Kratzert et al. 2018):

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \tag{8}$$

where $\boldsymbol{W}_o$, $\boldsymbol{U}_o$ and $\boldsymbol{b}_o$ are the input weight matrix, recurrent weight matrix and bias vector, respectively; The range of $\boldsymbol{h}_t$ is [-1,1].

3.2. Bayesian optimization

BO is a method commonly used to adjust hyperparameters, which can be used to optimize the black-box function. There are two main parts in the process of BO: a prior function (PF) and an acquisition function. If the distribution of the model is known, the optimal model can be selected according to experience; if not, the kernel function based on a Gaussian process (GP) can be used as the black-box function for self-learning. In this paper, it is assumed that the PF obeys a Gaussian distribution. Therefore, the function values $f(x_i)$ of the PF also obey a Gaussian distribution, as follows (Rasmussen 2004):

$$f(x_i) \sim GP(\lambda_i, \boldsymbol{K})(i = 1 \cdots n) \tag{9}$$

where $\{(x_i, y_i)\}_{i=1}^{n}$ is a known dataset, $y_i = f(x_i)$; $\{\lambda\}_{i=1}^{n}$ is the mean function set of GP and $\boldsymbol{K}$ is the covariance matrix described by the kernel function.

The acquisition function is also called the utility function. At present, there are three commonly used functions: the expected improvement (EI), upper confidence bound (UCB), and probability of improvement (PI). Here we choose the commonly used EI as the acquisition function as follows (Dewancker et al. 2015):

$$EI(x) = \begin{cases} \big(\mu(x) - f(x^+)\big)\Phi(z) + \sigma(x)\phi(z), & \sigma > 0 \\ 0, & \sigma = 0 \end{cases} \tag{10}$$

$$Z = \frac{\mu(x) - f(x^+)}{\sigma} \tag{11}$$

where $f(x^+)$ is the current maximum value, $\mu(x)$ is the posterior mean, and $\sigma(x)$ is the variance of $f(x)$. $\Phi(z)$ is the standard normal cumulative distribution function, and $\phi(z)$ is the standard normal probability density function. **Fig. 3** shows a flowchart of BO method based on a GP.

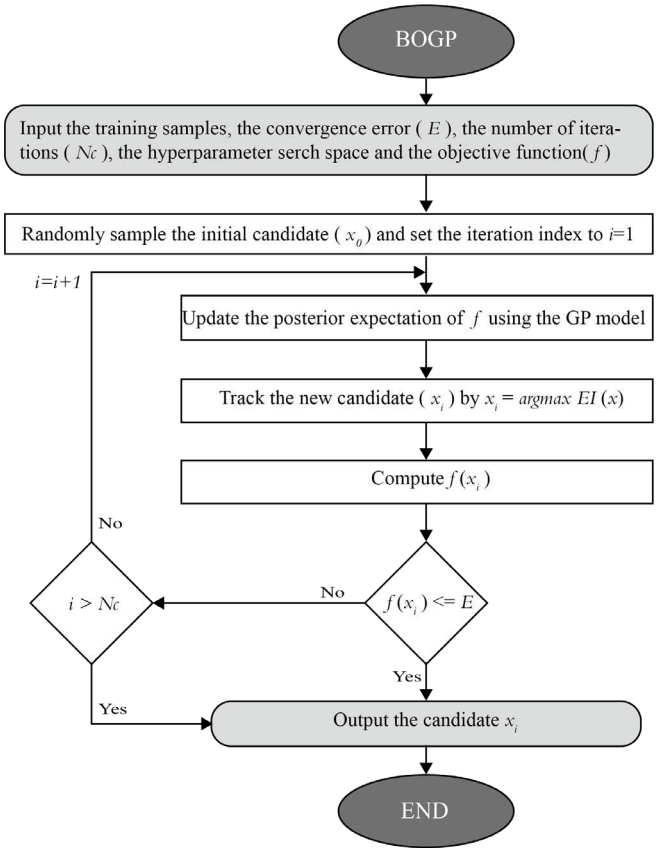

Fig. 3. A flowchart of the BOGP procedure.

3.4. The CDSF framework and the realization of PCA-LSTM-BO

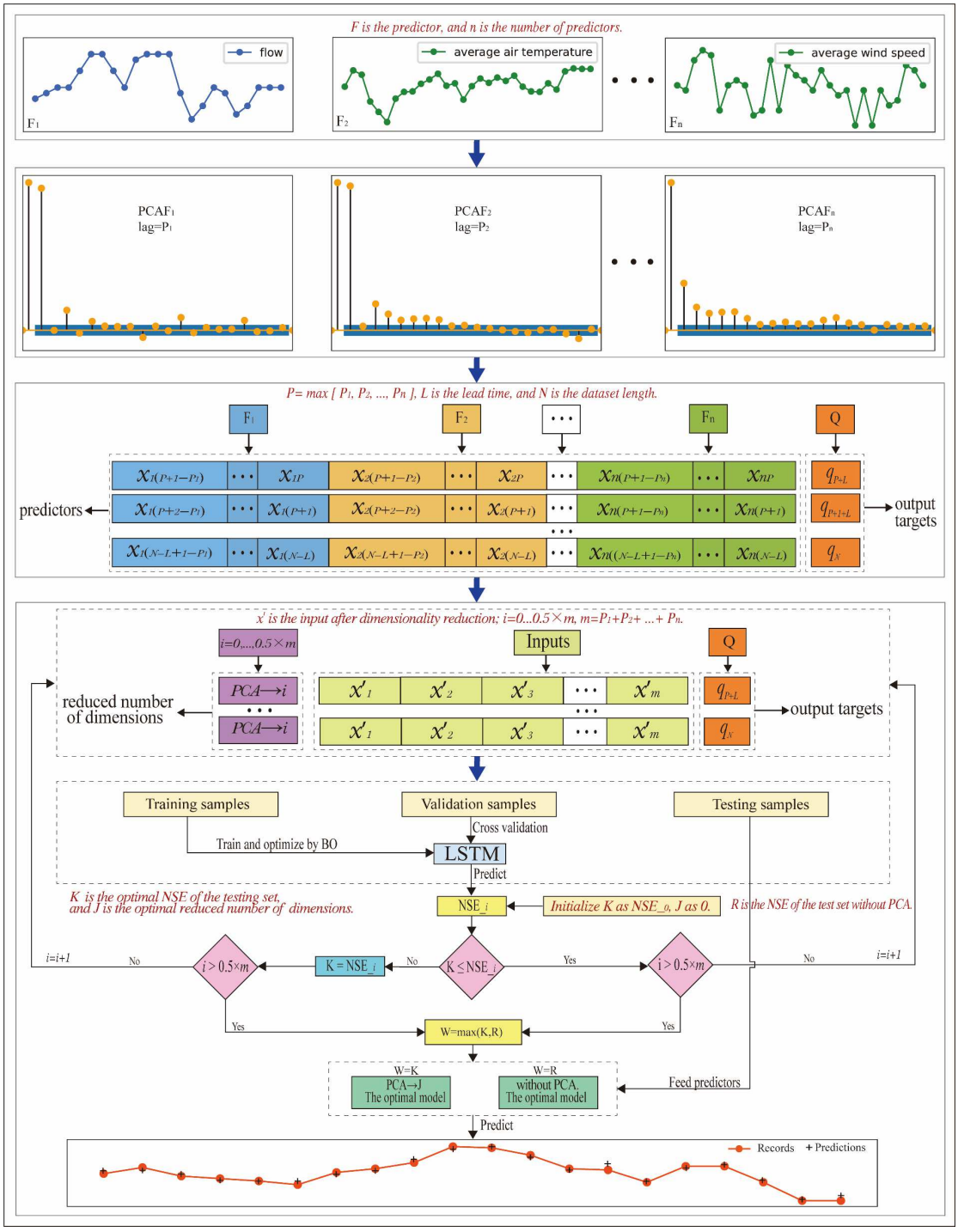

Fig. 4. The CDSF framework diagram with PCA-LSTM-BO realization.

Decomposition-based models and rainfall-runoff modeling have limited capacity to predict runoff series that have complex nonlinearity, high irregularity, and multiscale

variability. Therefore, this study develops a CDSF framework and realizes this framework with PCA-LSTM-BO to improve the runoff forecasting accuracy. The CDSF framework contains three main stages: (1) dimensionality reduction, (2) long-term dependency learning and (3) forecasting stage. During the first stage, PCA is used to reduce the input dimensionality to save time and computational resources and decrease the overfitting risk. In the second stage, the LSTM model and BO algorithm are used to learn the long-term dependency from meteorological data to runoff series. In the last stage, the optimized LSTM model is used to forecast the runoff series. The diagram of the CDSF framework and its PCA-LSTM-BO realization is illustrated in **Fig. 4** and is summarized as follows.

Step 1 Collect runoff time series and climate time series as inputs of the CDSF framework.

Step 2 Use the partial autocorrelation function (PACF) to select the optimal lag for each time series.

Step 3 Generate learning samples including input predictors and output target for different lead times based on the optimal lags obtained in Step 2.

Step 4 Use PCA to reduce the input dimensionality. The number of features achieved through dimensionality reduction is set from 0 to half of the number of input predictors.

Step 5 Divide the learning samples into training, validation and testing sets (accounting for 60%, 20% and 20%, respectively, of the total daily runoff samples in this study).

Step 6 Train the parameters of the PCA-LSTM-BO model with the training set and optimize the hyperparameters of the PCA-LSTM-BO model with the validation set and BO algorithm.

Step 7 Input the test sample into the optimized PCA-LSTM-BO model to predict the runoff time series.

4. Case Study

4.1. Predictors and predicted targets

The runoff series along with 13 climate time series (see **Table 1**) are selected to build the LSTM model. The input predictors and output targets are first determined to generate the learning samples. The PACF is widely used in determining the optimal input lags in ARMA models and machine learning models (He et al. 2020). One lag can be selected as the input of the LSTM model if it falls outside the 95% confidence interval. However, using some lags that pass the 95% confidence test but are insignificant leads to a high computational cost and modeling time. Therefore, we select all lags before the first insignificant lag as the optimal input.

The average air pressure at the Xining station is used as an example to reveal how to determine the optimal lags using the PACF plot from **Fig. 5.** The PACF value on the third day (lag 3) exceeds the boundary of the 95% confidence interval (light blue shaded area), and the lags after the third lag are all not insignificant. Therefore, $x_1(t)$, $x_1(t-1)$ and $x_1(t-2)$ are selected as the optimal input lags for the average air pressure. In this way, the optimal inputs of all the time series are selected. The optimal inputs of each series are merged as the final predictor of the PCA-LSTM-BO model. Additionally,

the predicted targets of the PCA-LSTM-BO model for 1-, 3-, 5-, and 7-day-ahead runoff prediction are the original daily runoff data $Q(t+1)$, $Q(t+3)$, $Q(t+5)$ and $Q(t+7)$.

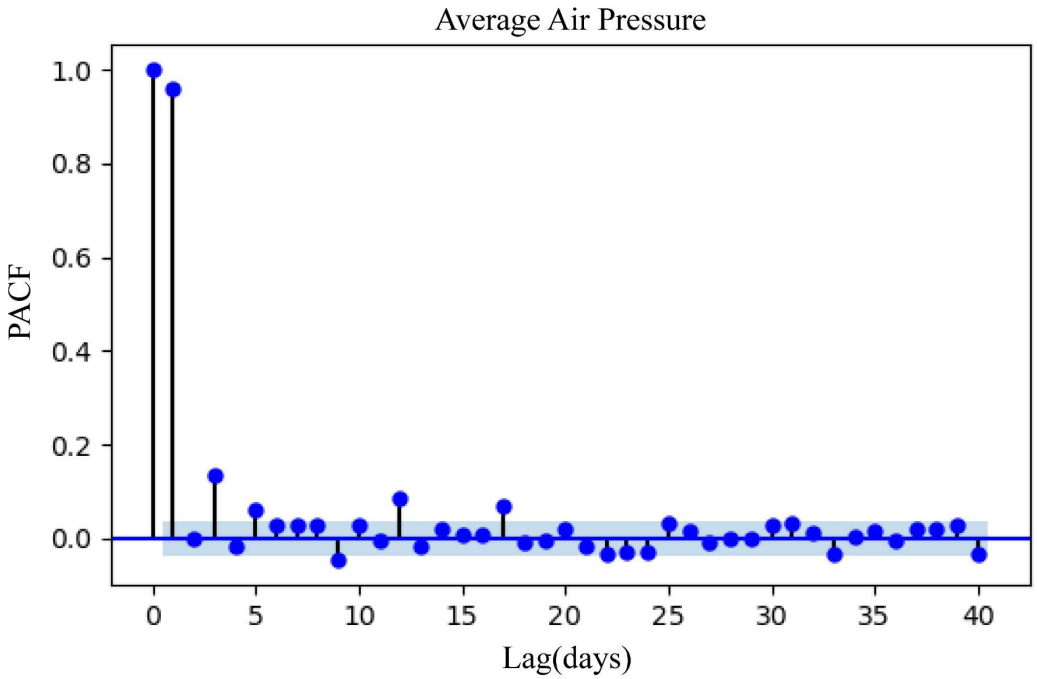

Fig. 5. PACF of the average air pressure sample series for the Xining station.

4.2. Sample normalization

To improve the convergence speed of the model in BO, all the samples are normalized to [-1,1] using max-min normalization; the formula is as follows (Zuo et al. 2020b):

$$y = 2 * \frac{x - x_{min}}{x_{max} - x_{min}} - 1 \tag{12}$$

where x is the original value, y is the normalized value, and x_{max} and x_{min} are the maximum and minimum values, respectively, in the original samples. The parameters x_{max} and x_{min} are obtained based on the training samples, and to avoid using the information of the validation and test samples, these parameters are used to normalize the validation and test samples.

4.3. Criteria for performance evaluation

To evaluate the performance of the model, the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970) and the mean squared error (MSE) (H. Marmolin 1986) were applied in this study. In addition, we propose a generalization index based on the NSE (GI(NSE)). The mathematical formulas of these three criteria are as follows:

$$NSE = \frac{\sum_{t=1}^{T}(x(t)-\hat{x}(t))^2}{\sum_{t=1}^{T}(x(t)-\bar{x}(t))^2} \tag{13}$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(x(t)-\hat{x}(t))^2 \tag{14}$$

$$GI(NSE) = a * (3 - (NSE_{train} + NSE_{val} + NSE_{test})) + b * (|NSE_{train} - NSE_{val}| + |NSE_{train} - NSE_{test}| + |NSE_{val} - NSE_{test}|) \tag{15}$$

where $x(t)$, $\hat{x}(t)$ and $\bar{x}(t)$ are the recorded, predicted and average of the measured samples, respectively, and T is the number of samples. For the GI(NSE), NSE_{train}, NSE_{val}, and NSE_{val} are the NSE values of the training set, validation set and testing set, respectively. a is the weight of the sum of the distances between NSE_{train}, NSE_{val} and NSE_{val} and 1. b is the weight of the sum of the distances among NSE_{train}, NSE_{val}, and NSE_{val}. The closer the value of $GI(NSE)$ is to 0, the better the generalizability of the model.

4.4. Parameter optimization

In this study, we test the performance of LSTM by comparing the prediction results of PCA-LSTM-BO with those of PCA-SVM-BO and PCA-GBRT-BO. For each climate-driven model, the relevant hypermeter settings and search ranges are shown in **Table 2**. Each hypermeter has been adjusted to minimize the MSE, and finally the climate-driven model with the smallest MSE is selected.

Table 2 The hyperparameters, tuning strategies, and search ranges for the compared climate-driven models.

Data-driven model	Tuning strategy	Hyperparameter	Search space
LSTM	BOGP	Batch size	512
		Optimizer	Adam
		Learning rate	$[1e-4, 1e-1]$
		Activation function	ReLU
		Number of hidden layers	$[1, 3]$
		Number of hidden units	$[8, 64]$
		Dropout rate	$[0.0, 0.5]$
SVR	BOGP	Weight penalty (C)	$[0.1, 200]$
		Error tolerance (ε)	$[1e-6, 1]$
		Width control coefficient (σ)	$[1e-6, 1]$
GBRT	BOGP	Learning rate	$[1e-5, 1]$
		Maximum depth,	$[1, 25]$
		Maximum feature,	$[1, n\_features]$
		Minimum sample split	$[2, 100]$
		Minimum sample leaf.	$[1, 100]$

Note: *n_features* IS the number of features.

5. Results and Discussion

5.1. Forecasting results with PCA-LSTM-BO

Fig. 6 shows the comparison between the NSE of PCA-LSTM-BO with and without dimensionality reduction for 1-day ahead streamflow forecasting. As seen from **Fig. 6**(a) and (b), the overall gap between the training (or validation) NSE values and the testing NSE values of the PCA-LSTM-BO with dimensionality reduction is smaller than that of the PCA-LSTM-BO without dimensionality reduction; and with the increases in the number of excluded predictors, this gap also shows a decreasing trend. Additionally, the testing NSE values are larger than 0.93, indicating that the PCA-LSTM-BO methods forecasts the unseen data reasonably well. These results indicate that the dimensionality reduction can improve the generalizability of PCA-LSTM-BO,

because PCA can transform the correlated predictors into uncorrelated predictors and reduce the overfitting risk.

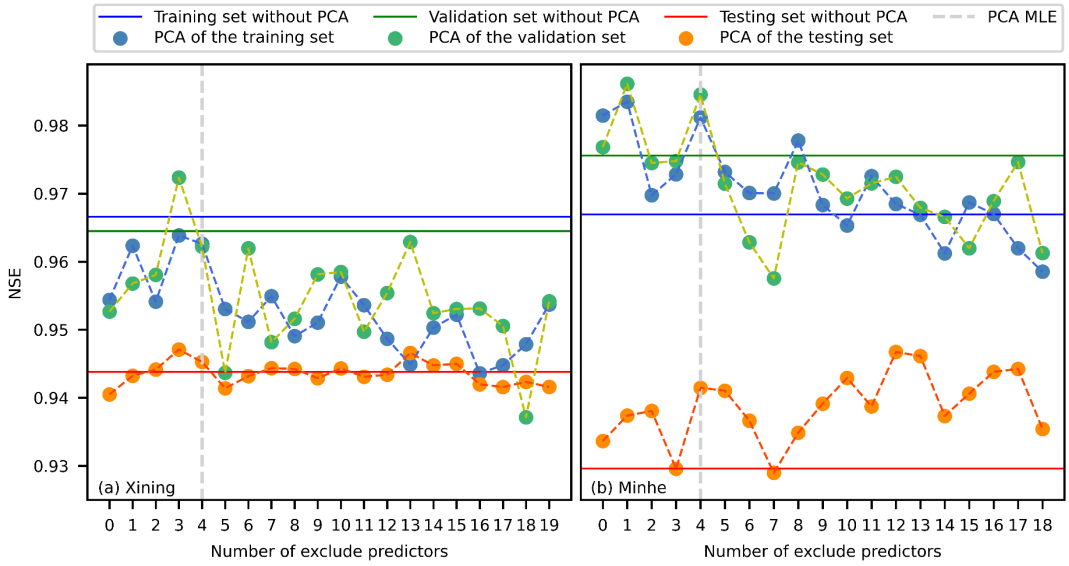

Fig. 6. The NSE of 1-day-ahead streamflow forecasting for the PCA-LSTM-BO model.

The GI(NSE) values of different dimensionality-reduction scenarios of the prediction results of the PCA-LSTM-BO at the Xining and Minhe stations 1-, 3-, 5-, and 7-day-ahead are shown in **Table 3** and **Table 4**, where "pca-0" means that the predictors are not excluded but transformed into uncorrelated predictors.

Table 3 The GI(NSE) values of the prediction results of the PCA-LSTM-BO at the Xining station.

Numbers of excluded predictors	1-day-ahead	3-day-ahead	5-day-ahead	7-day-ahead
pca-0	0.07766	0.204668	0.280841	0.39391
pca-1	0.077982	0.214576	0.350733	0.433557
pca-2	0.074173	0.212652	0.287483	0.468371
pca-3	0.076974	0.200728	0.292248	0.461177
pca-4	0.072778	0.196327	1.064446	0.538341
pca-5	0.078727	0.186064	0.284074	0.469206
pca-6	0.08003	0.208546	0.295245	0.460214
pca-7	0.07369	0.223795	0.277747	0.471209
pca-8	0.070871	0.197849	0.341172	0.448091
pca-9	0.077484	0.199261	0.363836	0.450177
pca-10	0.072772	0.200805	**0.266474**	0.510767
pca-11	0.07406	0.184521	0.301654	**0.377366**
pca-12	0.075447	0.191017	0.277423	0.387595

pca-13	0.079846	**0.182093**	0.299965	0.446094
pca-14	0.070213	0.202631	0.279562	0.464691
pca-15	**0.069566**	0.209953	0.31866	0.506516
pca-16	0.07789	0.201491	0.350117	0.417641
pca-17	0.076011	0.191812	0.327705	0.504364
pca-18	0.081968	0.198577	0.303907	0.485478
pca-19	0.075293	0.198323	0.31268	0.463934
Without PCA	0.077391	0.207283	0.419459	0.508536

Table 4 The GI(NSE) values of the prediction results of the PCA-LSTM-BO at the Minhe station.

Numbers of excluded predictors	**1-day-ahead**	**3-day-ahead**	**5-day-ahead**	**7-day-ahead**
pca-0	0.100621	0.157603	0.221081	0.262768
pca-1	0.095694	0.168789	0.225561	0.256162
pca-2	0.090793	0.188879	**0.204795**	0.250288
pca-3	0.103379	0.1696	0.215747	0.28299
pca-4	0.088828	0.152133	0.254784	0.255699
pca-5	0.084285	0.166015	0.207295	0.288337
pca-6	0.092314	0.155831	0.229265	0.30292
pca-7	0.106631	0.161925	0.21422	0.275764
pca-8	0.096595	0.162602	0.220754	0.274793
pca-9	0.088302	0.160891	0.241878	0.268716
pca-10	0.080634	**0.151385**	0.25799	0.270077
pca-11	0.087482	0.15439	0.241352	0.284734
pca-12	0.075809	0.152626	0.28856	0.275357
pca-13	**0.073768**	0.170935	0.224054	0.300767
pca-14	0.089056	0.182431	0.266341	0.318614
pca-15	0.085207	0.17895	0.263824	0.329677
pca-16	0.078214	0.16658	0.264852	0.290609
pca-17	0.08414	0.165721	0.249843	0.309032
pca-18	0.088918	0.196062	0.269823	0.274055
Without PCA	0.106342	0.180127	0.28081	**0.223024**

The black bold NSE values in **Table 3** and **Table 4** represent the optimal PCA settings of PCA-LSTM-BO models for 1-, 3-, 5- and 7-day-ahead streamflow forecasting at the Xining and Minhe stations. **Fig. 7** and **Fig. 8** show the final predicted results and scatter plots of the optimal PCA settings for the two stations. As seen from **Fig. 7,** the forecasted values can vary with the testing set and are consistent with the observed values, but they underestimate the observed values at the peak runoff and

valley runoff. The scatter plot shows that the observed and forecasted value clusters concentrated near the ideal fit of 1-day-ahead streamflow forecasting and the angle between the linear fit and ideal fit is small, which indicates that the forecasted values have a greater consistency with the observations. However, the PCA-LSTM-BO correlation values are dispersed around the ideal fit with a large angle between the ideal and linear fits for forecasting runoff 3-, 5- and 7-day-ahead. However, the angles of linear fit and ideal 1-, 3-, 5- and 7-day-ahead fist are relatively small, which further indicates that this model has better accuracy in daily runoff prediction to a certain extent. Similar results can be observed from **Fig. 8**.

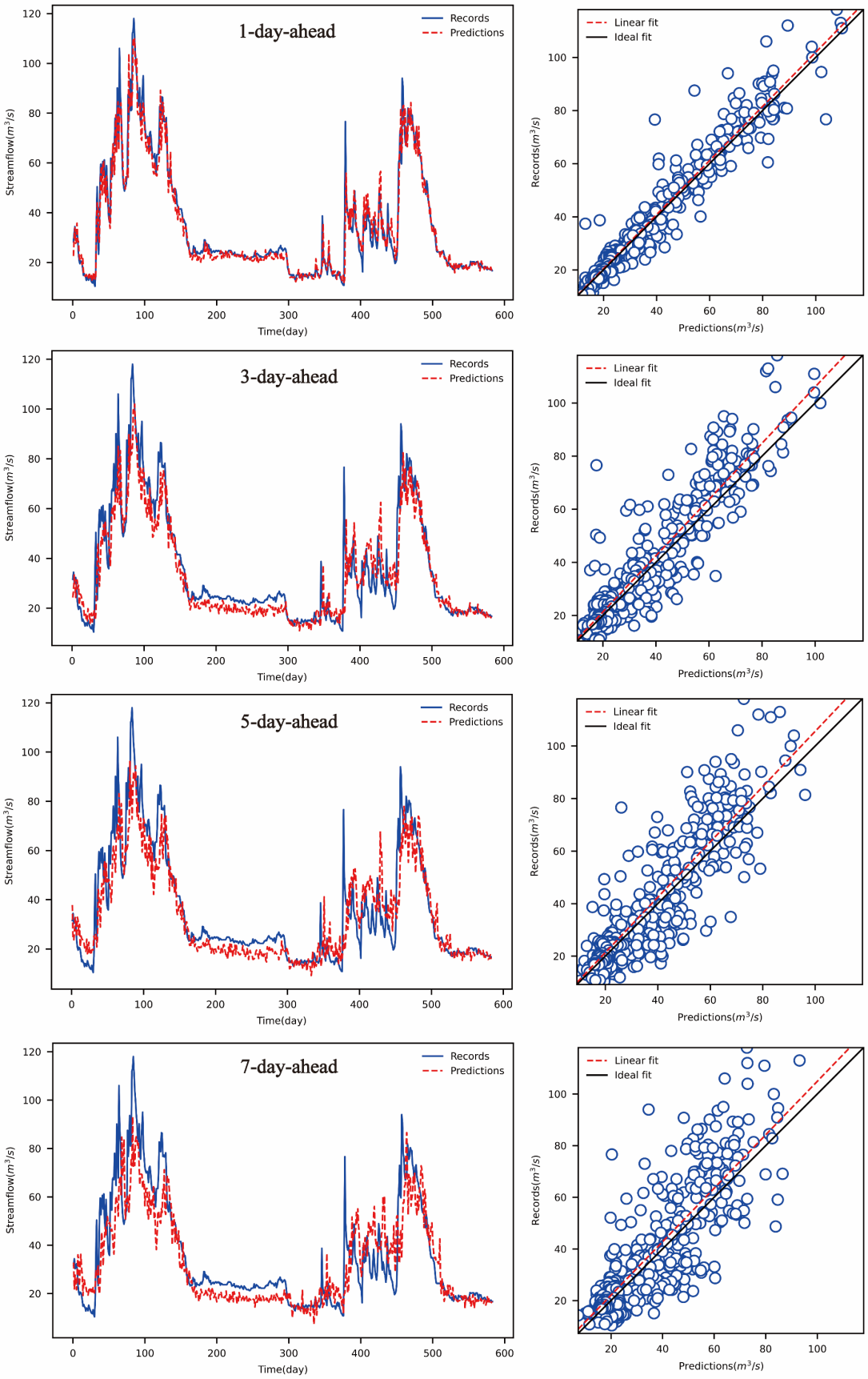

Fig. 7. Forecasting results and scatter plots of the optimal PCA settings for the testing set for the Xining station.

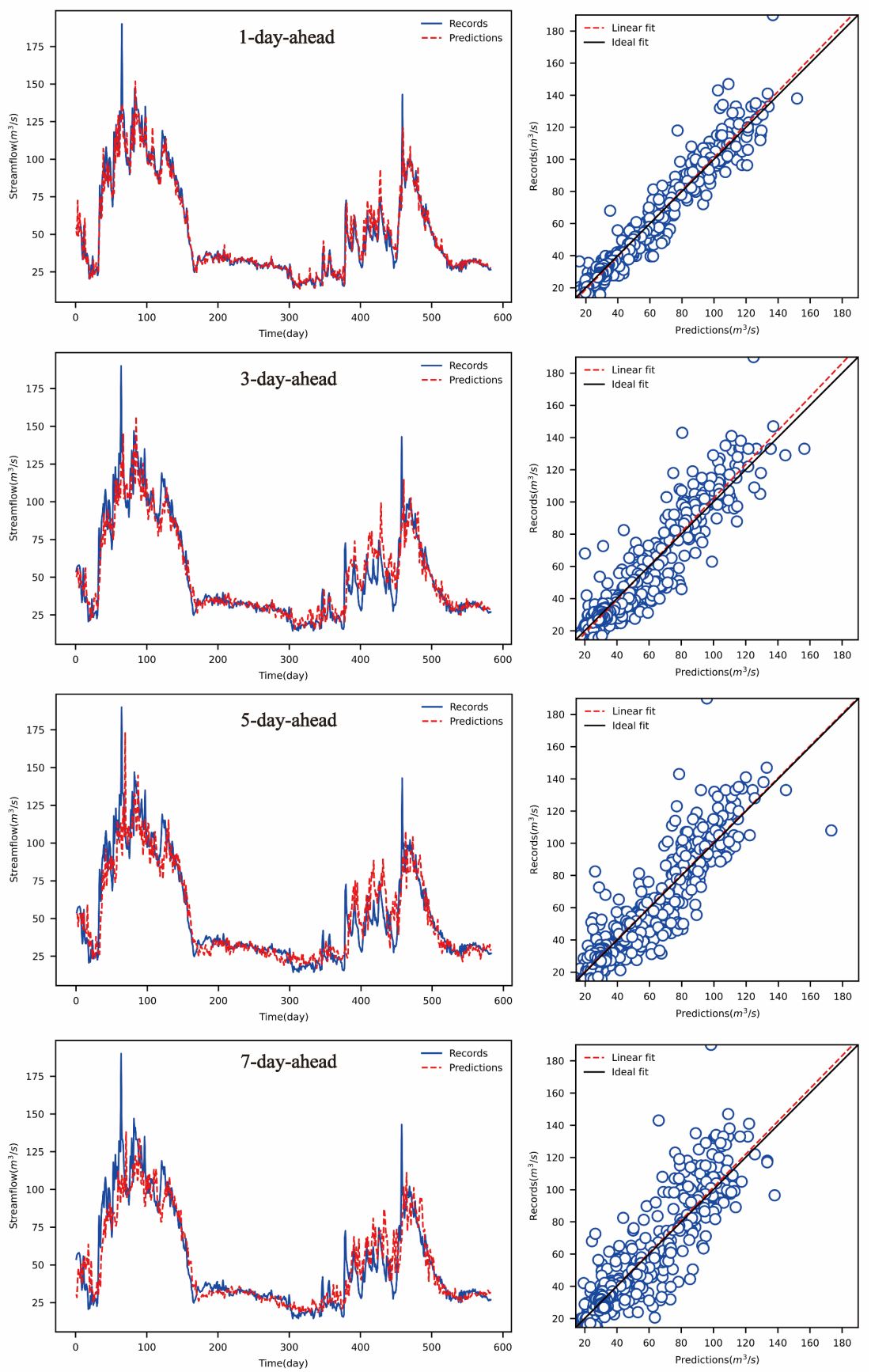

Fig. 8. Forecasting results and scatter plots of the optimal PCA settings for the testing set for the Minhe station.

5.3. Comparative analysis

To evaluate the superiority of the proposed PCA-LSTM-BO method, two CDSF realizations based on SVR and GBRT, namely, PCA-SVR-BO and PCA-GBRT-BO are compared using the same dataset. In addition, an autoregressive LSTM model is also compared with these CDSF realizations. **Table 5** shows the quantitative evaluation results of these optimized models. As seen that the GI(NSE) values of the PCA-LSTM-BO method for 1-, 3-, 5- and 7-day-ahead streamflow forecasts are lower than those of the PCA-SVR-BO, PCA-GBRT-BO and LSTM methods, illustrating that PCA-LSTM-BO is superior to the other models in terms of the generalizability.

Table 5 Comparison of the generalizability performances using different models by the GI(NSE).

Hydrological stations	Model	1-day-ahead	3-day-ahead	5-day-ahead	7-day-ahead
Xining	**PCA-LSTM-BO**	**0.069566**	**0.182093**	**0.266474**	**0.377366**
	PCA-SVR-BO	0.078173	0.215962	0.347667	0.400676
	PCA-GBRT-BO	0.111974	0.262021	0.391713	0.431187
	LSTM	0.100691	0.257702	0.36481	0.458753
Minhe	PCA-LSTM-BO	**0.073768**	**0.151385**	**0.204795**	**0.223024**
	PCA-SVR-BO	0.095294	0.172747	0.249221	0.314126
	PCA-GBRT-BO	0.087293	0.170901	0.222945	0.295155
	LSTM	0.091877	0.185808	0.233487	0.329197

To further validate the performance of these models, the performance gap in terms of the NSE and MSE for these models are compared and the results are presented in **Fig. 9**. As shown in **Fig. 9**(a) and (b), all the CDSF realizations show similar trends for forecasting daily streamflow 1-, 3- and 5-day-ahead at the Xining station, but the proposed PCA-LSTM-BO has lower NSE and higher MSE values compared with the PCA-SVR-BO and PCA-GBRT-BO methods and has much higher NSE and lower MSE values compared with the LSTM method. As can be observed from **Fig. 9**(c) and (d),

similar results are obtained for the Minhe station, illustrating the superiority of the PCA-LSTM-BO model. Overall, the forecasting performances can be ranked as PCA-LSTM-BO > PCA-SVR-BO ≈ PCA-GBRT-BO > LSTM. Moreover, the CDSF realizations are all superior to the single LSTM method, indicating the advantages of the CDSF framework on daily runoff forecasting. In addition, with increasing lead time, the NSE(MSE) gap between the LSTM and the PCA-LSTM-BO gradually increases from 0.0179 (8.913) to 0.0622 (30.807) at the Xining station, indicating that the CDSF framework is more stable than the autoregressive LSTM model for longer lead times. Overall, the above results sufficiently illustrate that the proposed PCA-LSTM-BO model has the best forecasting performance among these models.

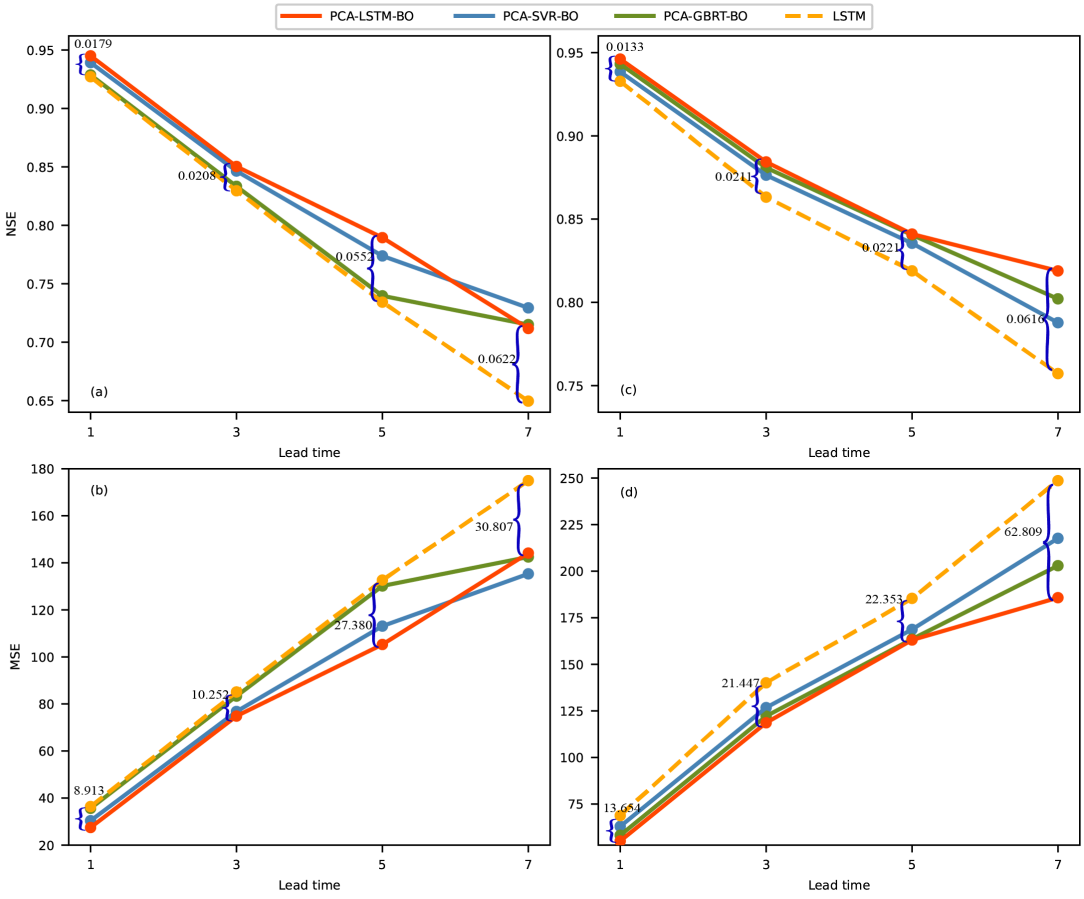

Fig. 9. Evaluation results of the forecasting performance of the different models for the Xining (a, b) and Minhe (c, d) stations.

Fig. 10 and **Fig. 11** display the forecasting results generated by the PCA-LSTM-BO, PCA-SVR-BO, PCA-GBRT-BO and LSTM models. As shown in **Fig. 10**(b), (c) and (d), PCA-SVR-BO and PCA-GBRT-BO have a certain ability to fit the trend and periodicity but have a poor tracking ability for predicting peak runoff and capturing random variations. As seen from **Fig. 10**, the LSTM model is better at forecasting the periodicity but performs poorly for forecasting the peak and valley runoff. Furthermore, the PCA-LSTM-BO method generally follows the runoff trend and has a better tracking ability for forecasting the observed peak runoff. In general, the PCA-LSTM-BO method has a better forecasting accuracy and generalizability performance than the other three models for forecasting daily runoff at the Xining and Minhe stations.

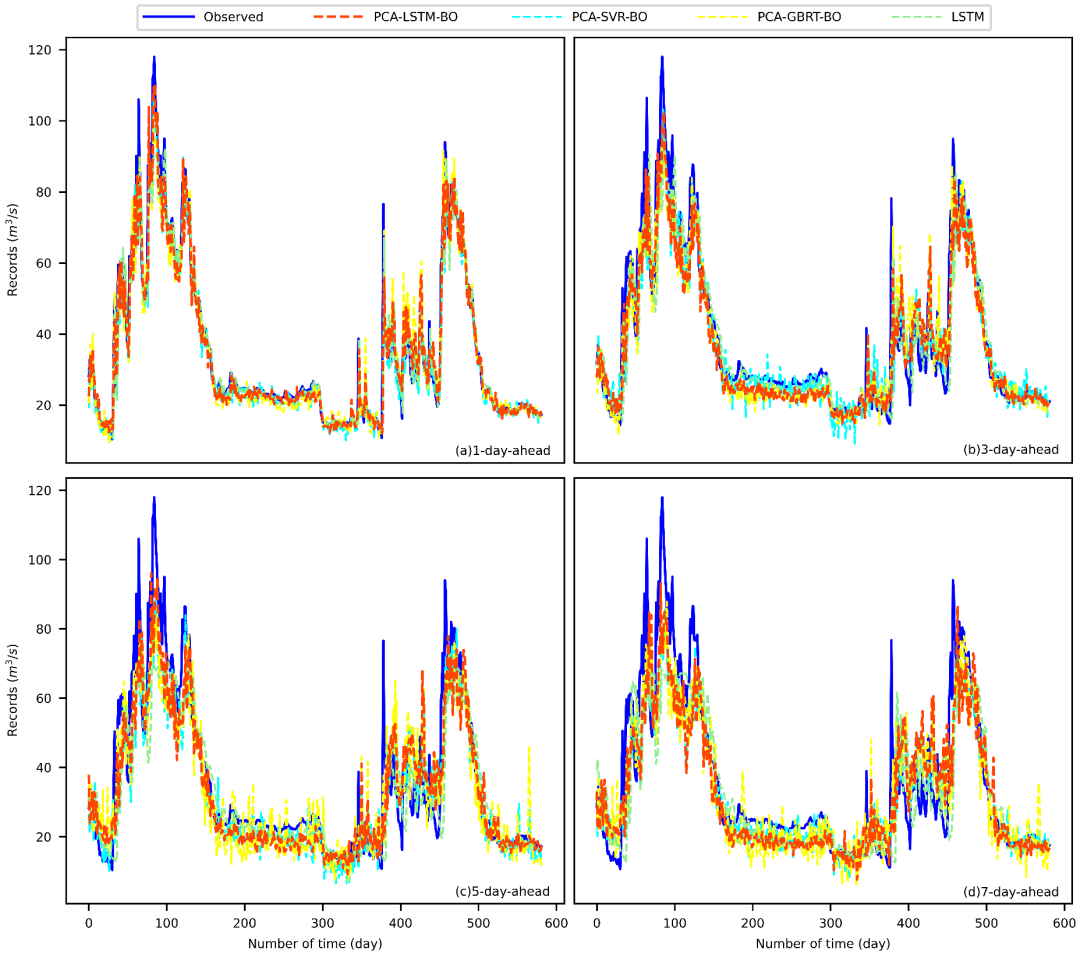

Fig. 10. Forecasted and observed results for the testing set at the Xining station.

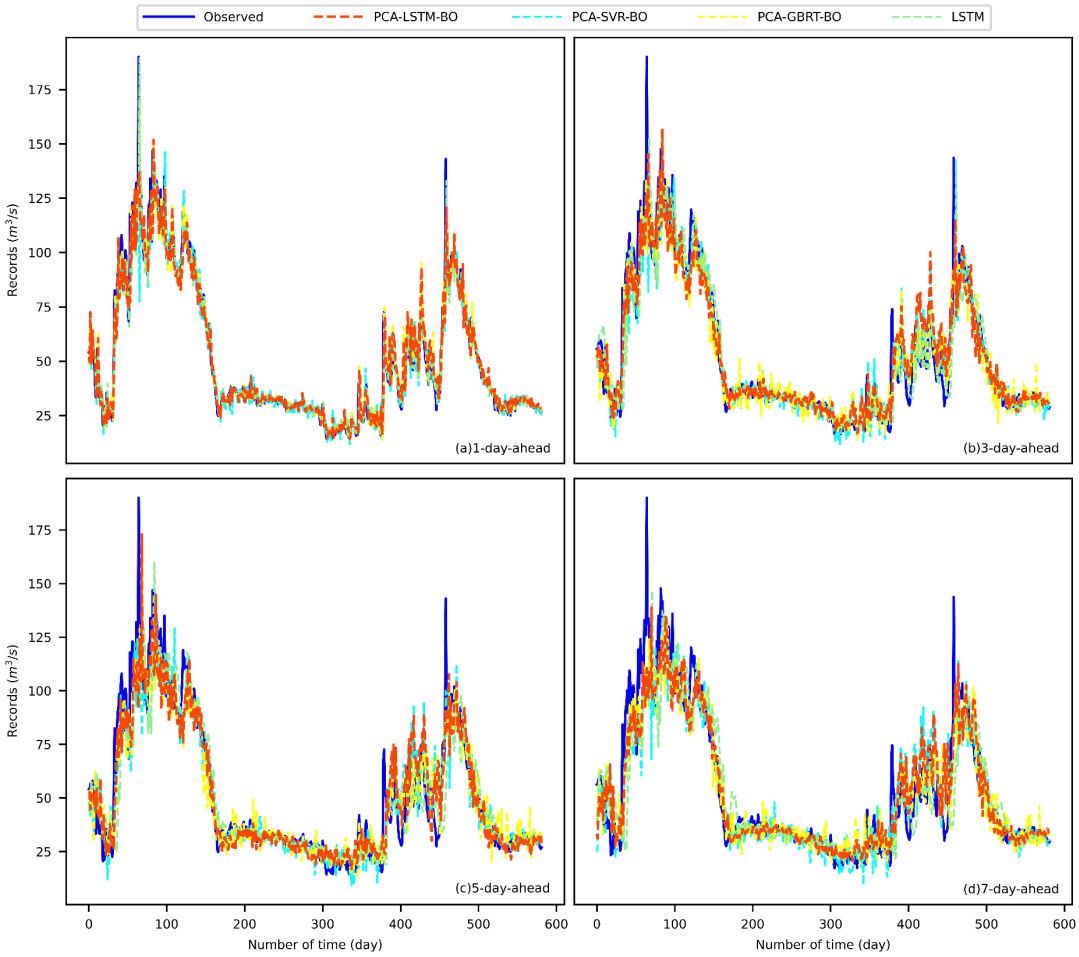

Fig. 11. Forecasted and observed results for the testing set at the Minhe station.

6. Conclusion

To avoid the introduced decomposition errors of decomposition-based models and the drawbacks of rainfall-runoff modeling in runoff forecasting, this study proposed a novel CDSF framework realized with PCA-LSTM-BO. There are four stages in implementing the CDSF framework: (1) Determine the optimal lag for each variable by PACF to form predictors; (2) reduce the input dimension by PCA to decrease the overfitting risk; (3) tune the hyperparameters of an LSTM model using BO; and (4) forecast the future streamflow using the optimized LSTM model. CDSF realizations, including PCA-LSTM-BO, PCA-SVR-BO and PCA-GBRT-BO, and an autoregressive

LSTM were compared for forecasting daily streamflow 1-, 3-, 5-, and 7-day-ahead at the Xining and Minhe stations of Huangshui River, China. The main conclusions are summarized as follows.

(1) PCA in the CDSF framework can improve the forecasting capacity and generalizability.

(2) The CDSF framework is superior to the autoregressive LSTM models for all the forecasting scenarios.

(3) The PCA-LSTM-BO has the best performance in terms of the generalizability performance among all the CDSF realizations.

(4) The GI(NSE) value is demonstrated to be effective in selecting the generalizability of the model.

Overall, the CDSF framework and the PCA-BO-LSTM method are useful for daily runoff prediction for nonlinear and nonstationary runoff time series. However, the streamflow is also greatly affected by human activity. Further research will consider the impact of climate and human activity on runoff to build a streamflow forecasting model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 51679186, 51979221 and 51709222), the Natural Science Basic Research Program of Shaanxi (Program No. 2019JLZ-15), and the Research Fund of the State Key Laboratory of Eco-hydraulics in Northwest Arid Region, Xi'an University of Technology (Grant No. 2019KJCXTD-5).

Funding

National Natural Science Foundation of China (Grant Nos. 51679186, 51979221 and 51709222), the Natural Science Basic Research Program of Shaanxi (Program No. 2019JLZ-15), and the Research Fund of the State Key Laboratory of Eco-hydraulics in Northwest Arid Region, Xi'an University of Technology (Grant No. 2019KJCXTD-5).

Competing Interests

None.

Data and Code Availability

Data used in the research work have been acknowledged, and data and code are available on request.

Author information

Affiliations

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

Yani Lian, Jungang Luo, Ganggang Zuo

Hanjiang-to-Weihe River Water Diversion Project Construction Co.Ltd., Shaanxi Province, Xi'an, Shaanxi 710100, China

Jingmin Wang

Contributions

Conceptualization: Lian YN; Methodology: Lian YN and Zuo GG; Writing - original draft preparation: Lian YN and Wang JM; Writing - review and editing: Luo JG and Zuo GG; Funding acquisition: Luo JG.

Ethics declarations

Ethical Approval and Consent to Participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to Publish

All authors have consented to publish this manuscript.

References

Abdi H, Williams LJ (2010) Principal component analysis. WIREs Comp Stat 2:433–459. https://doi.org/10.1002/wics.101

Alizadeh MJ, Kavianpour MR, Kisi O, Nourani V (2017) A new approach for simulating and forecasting the rainfall-runoff process within the next two months. Journal of Hydrology 548:588–597. https://doi.org/10.1016/j.jhydrol.2017.03.032

Bai Y, Bezak N, Sapač K, Klun M, Zhang J (2019) Short-Term Streamflow Forecasting Using the Feature-Enhanced Regression Model. Water Resour Manage 33:4783–4797. https://doi.org/10.1007/s11269-019-02399-1

Bai P, Liu X, Xie J (2020) Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models. Journal of Hydrology 592:125779. https://doi.org/10.1016/j.jhydrol.2020.125779

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5:157–166. https://doi.org/10.1109/72.279181.

Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research 13:281–305

Bisoyi N, Gupta H, Padhy NP, Chakrapani GJ (2019) Prediction of daily sediment discharge using a back propagation neural network training algorithm: A case study of the Narmada River, India. International Journal of Sediment Research 34:125–135. https://doi.org/10.1016/j.ijsrc.2018.10.010

Cao LJ, Chua KS, Chong WK, Lee HP, Gu QM (2003) A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing 55:321–336. https://doi.org/10.1016/S0925-2312(03)00433-8

Chang TK, Talei A, Alaghmand S, Ooi MP-L (2017) Choice of rainfall inputs for event-based rainfall-runoff modeling in a catchment with multiple rainfall stations using data-driven techniques. Journal of Hydrology 545:100–108. https://doi.org/10.1016/j.jhydrol.2016.12.024

Cheng C, Feng Z, Niu W, Liao S (2015) Heuristic Methods for Reservoir Monthly Inflow

Forecasting: A Case Study of Xinfengjiang Reservoir in Pearl River, China. Water 7:4477–4495. https://doi.org/10.3390/w7084477

Chua LH, Wong TS (2011) Runoff forecasting for an asphalt plane by Artificial Neural Networks and comparisons with kinematic wave and autoregressive moving average models. Journal of Hydrology 397:191–201. https://doi.org/10.1016/j.jhydrol.2010.11.030

Dariane AB, Farhani M, Azimi S (2018) Long Term Streamflow Forecasting Using a Hybrid Entropy Model. Water Resour Manage 32:1439–1451. https://doi.org/10.1007/s11269-017-1878-0

Davis JC, Sampson RJ (1986) Statistics and data analysis in geology. John Wiley & Sons, New York

Dewancker I, McCourt M, Clark S (2015) Bayesian optimization primer. URL https://app. sigopt. com/static/pdf/SigOpt_ Bayesian_Optimization_Primer. pdf

Dewancker I, McCourt M, Clark S (2016) Bayesian optimization for machine learning: A practical guidebook. arXiv preprint arXiv:1612.04858

Du K, Zhao Y, Lei J (2017) The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. Journal of Hydrology 552:44–51. https://doi.org/10.1016/j.jhydrol.2017.06.019

Fang H-T, Jhong B-C, Tan Y-C, Ke K-Y, Chuang M-H (2019) A Two-Stage Approach Integrating SOM- and MOGA-SVM-Based Algorithms to Forecast Spatial-temporal Groundwater Level with Meteorological Factors. Water Resour Manage 33:797–818. https://doi.org/10.1007/s11269-018-2143-x

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics 29:1189–1232. https://doi.org/10.2307/2699986

Gauch M, Mai J, Lin J (2020) The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. Environmental Modelling & Software 135:104926. https://doi.org/10.1016/j.envsoft.2020.104926

George A (2012) Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM. International Journal of Computer Applications 47

George A, Vidyapeetham AV (2012) Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM. International Journal of Computer Applications 47:5–8. https://doi.org/10.5120/7470-0475

Gong Y, Zhang Y, Lan S, Wang H (2016) A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida. Water Resour Manage 30:375–391. https://doi.org/10.1007/s11269-015-1167-8

H. Marmolin (1986) Subjective MSE Measures. IEEE Transactions on Systems, Man, and Cybernetics 16:486–489. https://doi.org/10.1109/TSMC.1986.4308985

Hadi SJ, Tombul M (2018) Forecasting Daily Streamflow for Basins with Different Physical

Characteristics through Data-Driven Methods. Water Resour Manage 32:3405–3422. https://doi.org/10.1007/s11269-018-1998-1

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 Manage 33:1571–1590. https://doi.org/10.1007/s11269-019-2183-x

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 Manage 34:865–884. https://doi.org/10.1007/s11269-020-02483-x

Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Research 34:807–816. https://doi.org/10.1016/S0043-1354(99)00225-0

Huang S, Chang J, Huang Q, Chen Y (2014) Monthly streamflow prediction using modified EMD-based support vector machine. Journal of Hydrology 511:764–775. https://doi.org/10.1016/j.jhydrol.2014.01.062

Kisi O, Nia AM, Gosheh MG, Tajabadi MRJ, Ahmadi A (2012) Intermittent Streamflow Forecasting by Using Several Data Driven Techniques. Water Resour Manage 26:457–474. https://doi.org/10.1007/s11269-011-9926-7

Kourtit K, Marinescu Pele MM, Nijkamp P, Traian Pele D (2021) Safe cities in the new urban world: A comparative cluster dynamics analysis through machine learning. Sustainable Cities and Society 66:102665. https://doi.org/10.1016/j.scs.2020.102665

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. 22:6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kumar A, Kumar P, Singh VK (2019) Evaluating Different Machine Learning Models for Runoff and Suspended Sediment Simulation. Water Resour Manage 33:1217–1231. https://doi.org/10.1007/s11269-018-2178-z

Li Y, Sun H, Yan W, Zhang X (2020) Multi-output parameter-insensitive kernel twin SVR model. Neural Netw 121:276–293. https://doi.org/10.1016/j.neunet.2019.09.022

Lin G-F, Chen G-R, Huang P-Y, Chou Y-C (2009) Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. Journal of Hydrology 372:17–29. https://doi.org/10.1016/j.jhydrol.2009.03.032

Luo X, Yuan X, Zhu S, Xu Z, Meng L, Peng J (2019) A hybrid support vector regression framework for streamflow forecast. Journal of Hydrology 568:184–193. https://doi.org/10.1016/j.jhydrol.2018.10.064

Maity R, Bhagwat PP, Bhatnagar A (2010) Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrol. Process. 24:917–923. https://doi.org/10.1002/hyp.7535

Minka TP (2001) Automatic Choice of Dimensionality for PCA: Advances in NIPS. Advances in Neural Information Processing Systems:598–604

Narayan RK, Ghosh SK (2021) Analysis of variations in morphological characteristics of orbito-meningeal foramen: An anatomical study with clinical implications. Translational Research in Anatomy 24:100108. https://doi.org/10.1016/j.tria.2020.100108

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Noori R, Karbassi AR, Moghaddamnia A, Han D, Zokaei-Ashtiani MH, Farokhnia A, Gousheh MG (2011) Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology 401:177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021

Persson C, Bacher P, Shiga T, Madsen H (2017) Multi-site solar power forecasting using gradient boosted regression trees. Solar Energy 150:423–436. https://doi.org/10.1016/j.solener.2017.04.066

Quilty J, Adamowski J (2018) Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. Journal of Hydrology 563:336–353. https://doi.org/10.1016/j.jhydrol.2018.05.003

Ramaswamy V, Saleh F (2020) Ensemble Based Forecasting and Optimization Framework to Optimize Releases from Water Supply Reservoirs for Flood Control. Water Resour Manage 34:989–1004. https://doi.org/10.1007/s11269-019-02481-8

Rasmussen CE (ed) (2004) Gaussian Processes in Machine Learning. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg

Sedki A, Ouazar D, El Mazoudi E (2009) Evolving neural network using real coded genetic algorithm for daily rainfall–runoff forecasting. Expert Systems with Applications 36:4523–4527. https://doi.org/10.1016/j.eswa.2008.05.024

Sheng Y, Yao K, Qin Z (2020) Continuity and variation analysis of fractional uncertain processes. Chaos, Solitons & Fractals 140:110250. https://doi.org/10.1016/j.chaos.2020.110250

Shirmohammadi B, Vafakhah M, Moosavi V, Moghaddamnia A (2013) Application of Several Data-Driven Techniques for Predicting Groundwater Level. Water Resour Manage 27:419–432. https://doi.org/10.1007/s11269-012-0194-y

Su J, Wang X, Liang Y, Chen B (2014) GA-Based Support Vector Machine Model for the Prediction of Monthly Reservoir Storage. J. Hydrol. Eng. 19:1430–1437. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000915

Sun Y, Niu J, Sivakumar B (2019) A comparative study of models for short-term streamflow forecasting with emphasis on wavelet-based approach. Stoch Environ Res Risk Assess 33:1875–1891. https://doi.org/10.1007/s00477-019-01734-7

Svante .Wold, Kim .Esbensen, Paul .Geladi (1987) Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2:37–52.

https://doi.org/10.1016/0169-7439(87)80084-9

Tan Q-F, Lei X-H, Wang X, Wang H, Wen X, Ji Y, Kang A-Q (2018) An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. Journal of Hydrology 567:767–780. https://doi.org/10.1016/j.jhydrol.2018.01.015

Vapnik V, Golowich S, Smola A (1996) Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. Advances in Neural Information Processing Systems 9:281–287

Wu Z, Huang NE (2009) ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD. Advances in Adaptive Data Analysis:1–41

Wu CL, Chau KW, Li YS (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour. Res. 45. https://doi.org/10.1029/2007WR006737

Yin J, Deng Z, Ines AV, Wu J, Rasu E (2020) Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM). Agricultural Water Management 242:106386. https://doi.org/10.1016/j.agwat.2020.106386

Zhang Q, Wang B-D, He B, Peng Y, Ren M-L (2011) Singular Spectrum Analysis and ARIMA Hybrid Model for Annual Runoff Forecasting. Water Resour Manage 25:2683–2703. https://doi.org/10.1007/s11269-011-9833-y

Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58:308–324. https://doi.org/10.1016/j.trc.2015.02.019

Zhang X, Peng Y, Zhang C, Wang B (2015) Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. Journal of Hydrology 530:137–152. https://doi.org/10.1016/j.jhydrol.2015.09.047

Zhao X, Chen X (2015) Auto Regressive and Ensemble Empirical Mode Decomposition Hybrid Model for Annual Runoff Forecasting. Water Resour Manage 29:2913–2926. https://doi.org/10.1007/s11269-015-0977-z

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

Zuo G, Luo J, Wang N, Lian Y, He X (2020b) Two-stage variational mode decomposition and support vector regression for streamflow forecasting. Hydrol. Earth Syst. Sci. 24:5491–5518. https://doi.org/10.5194/hess-24-5491-2020

Figures

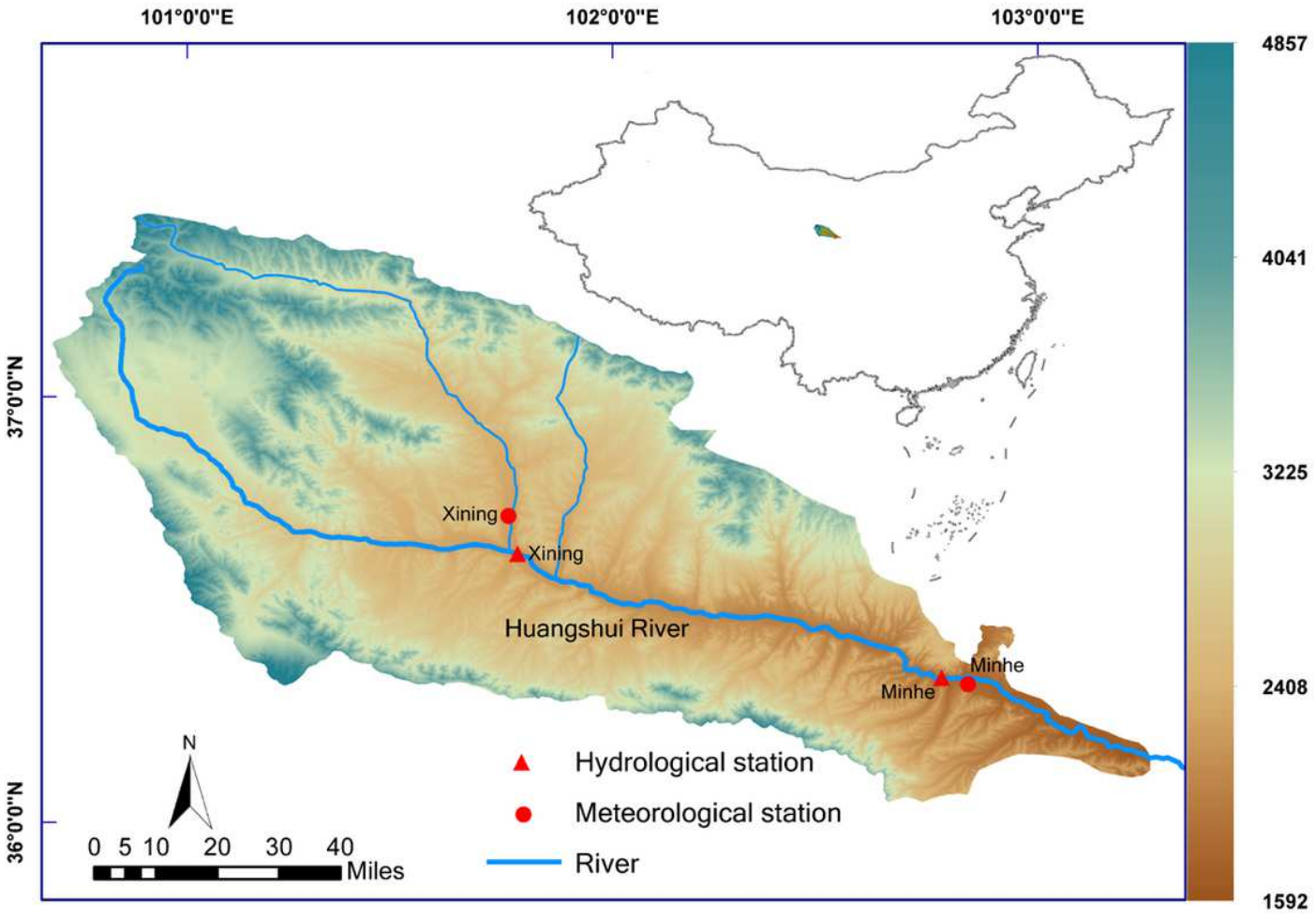

Figure 1

A geographical overview of Huangshui River catchment in China. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

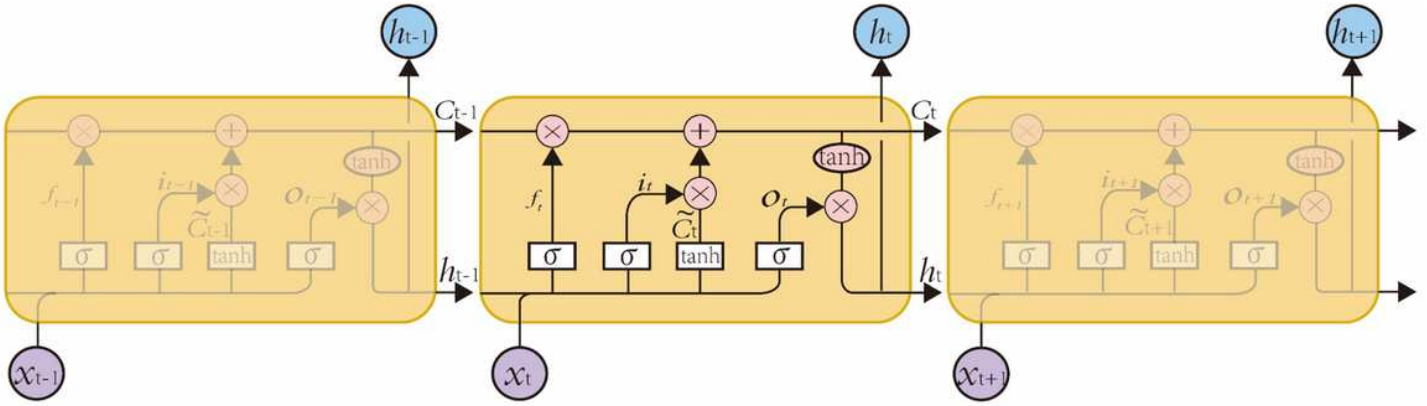

Figure 2

A geographical overview of Huangshui River catchment in China.

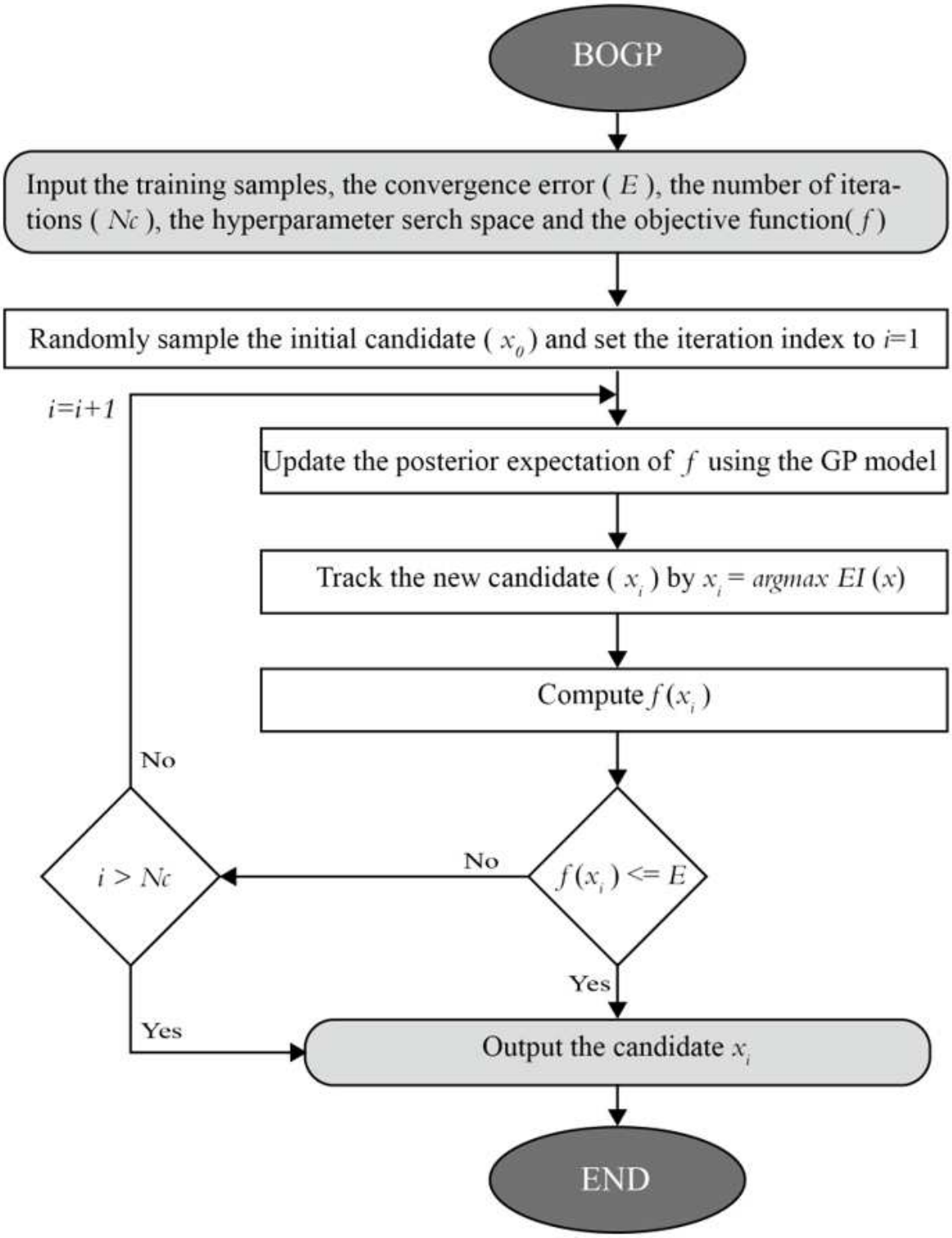

Figure 3

A flowchart of the BOGP procedure.

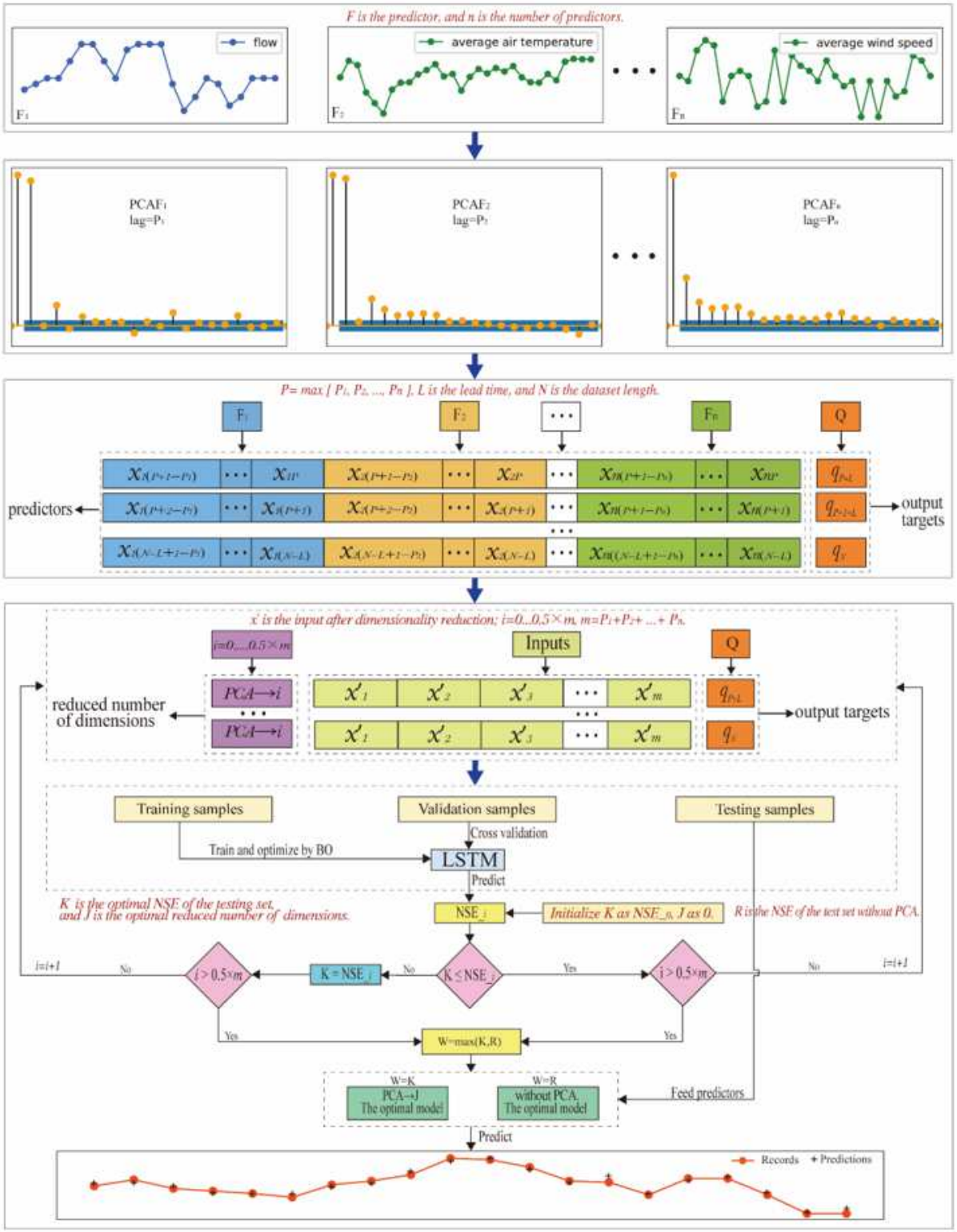

Figure 4

The CDSF framework diagram with PCA-LSTM-BO realization.

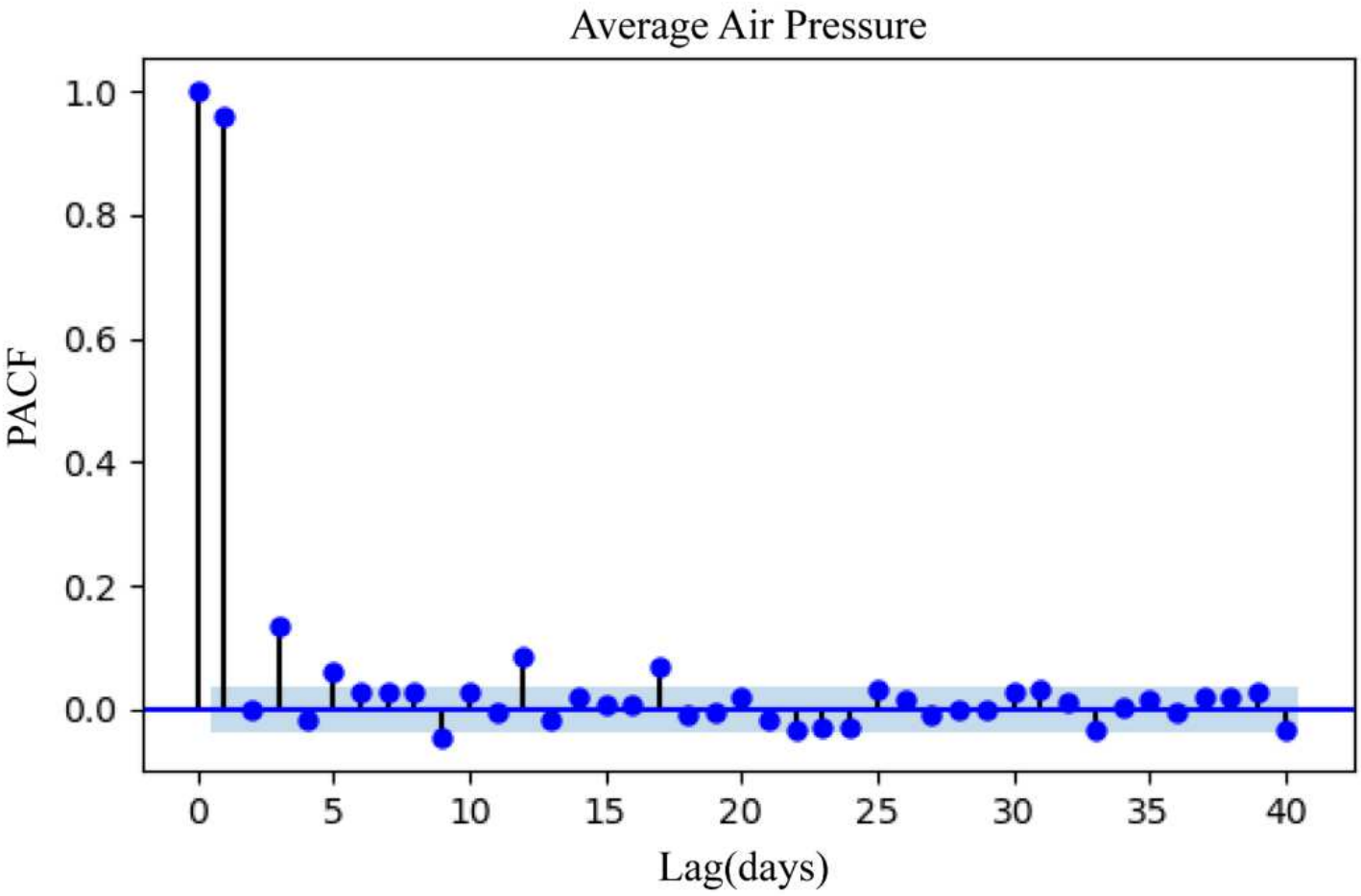

Figure 5

PACF of the average air pressure sample series for the Xining station.

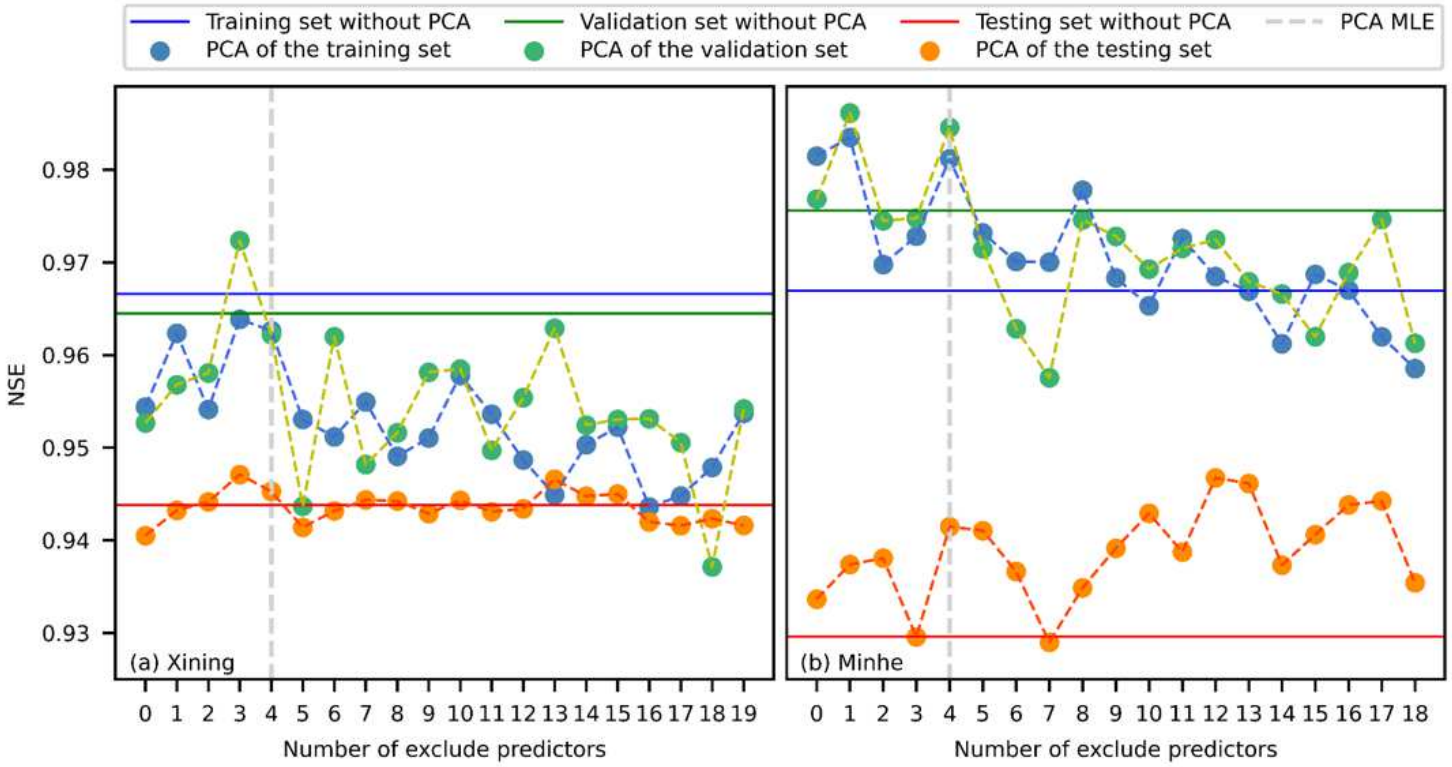

Figure 6

The NSE of 1-day-ahead streamflow forecasting for the PCA-LSTM-BO model.

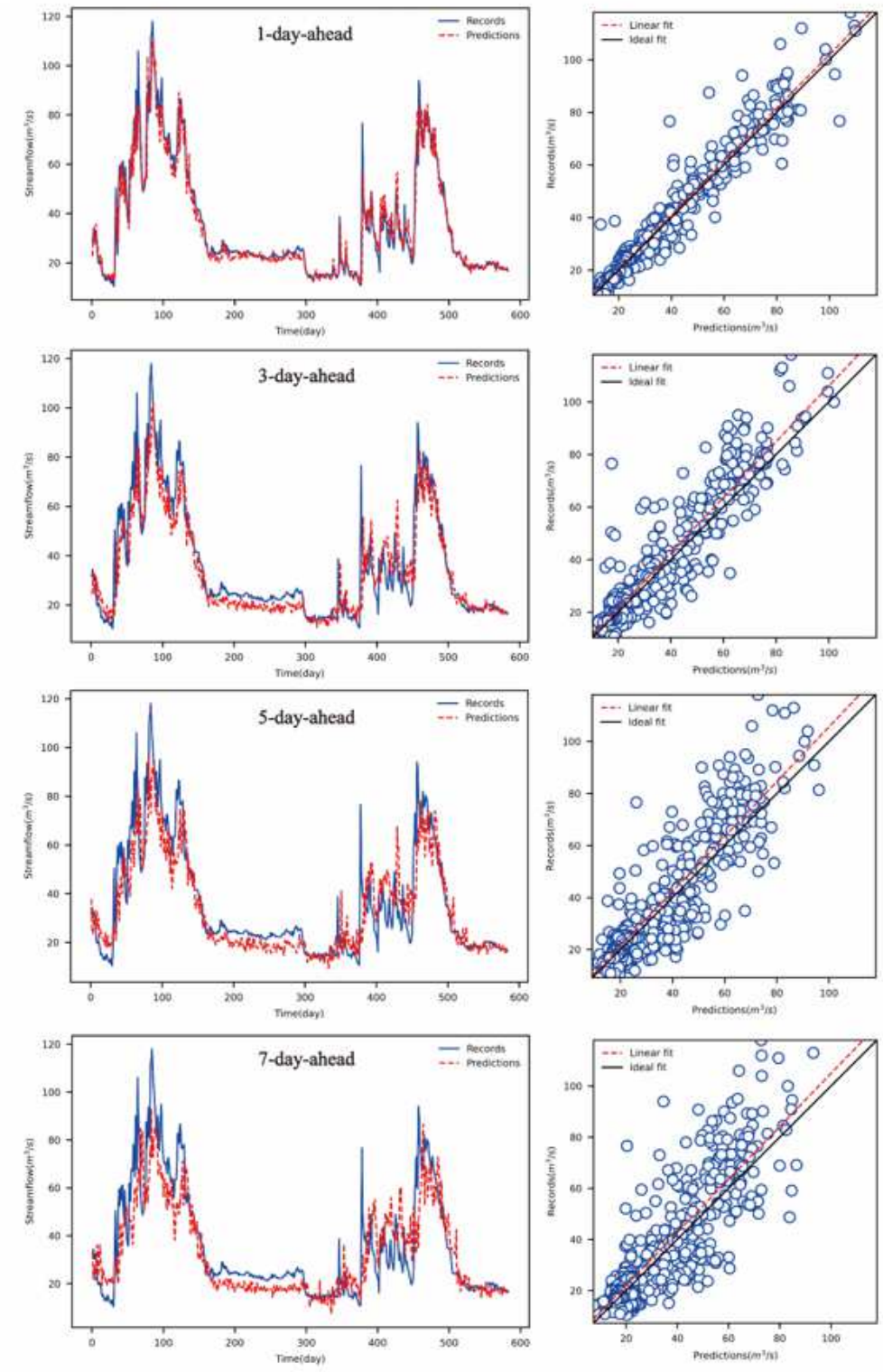

Figure 7

Forecasting results and scatter plots of the optimal PCA settings for the testing set for the Xining station.

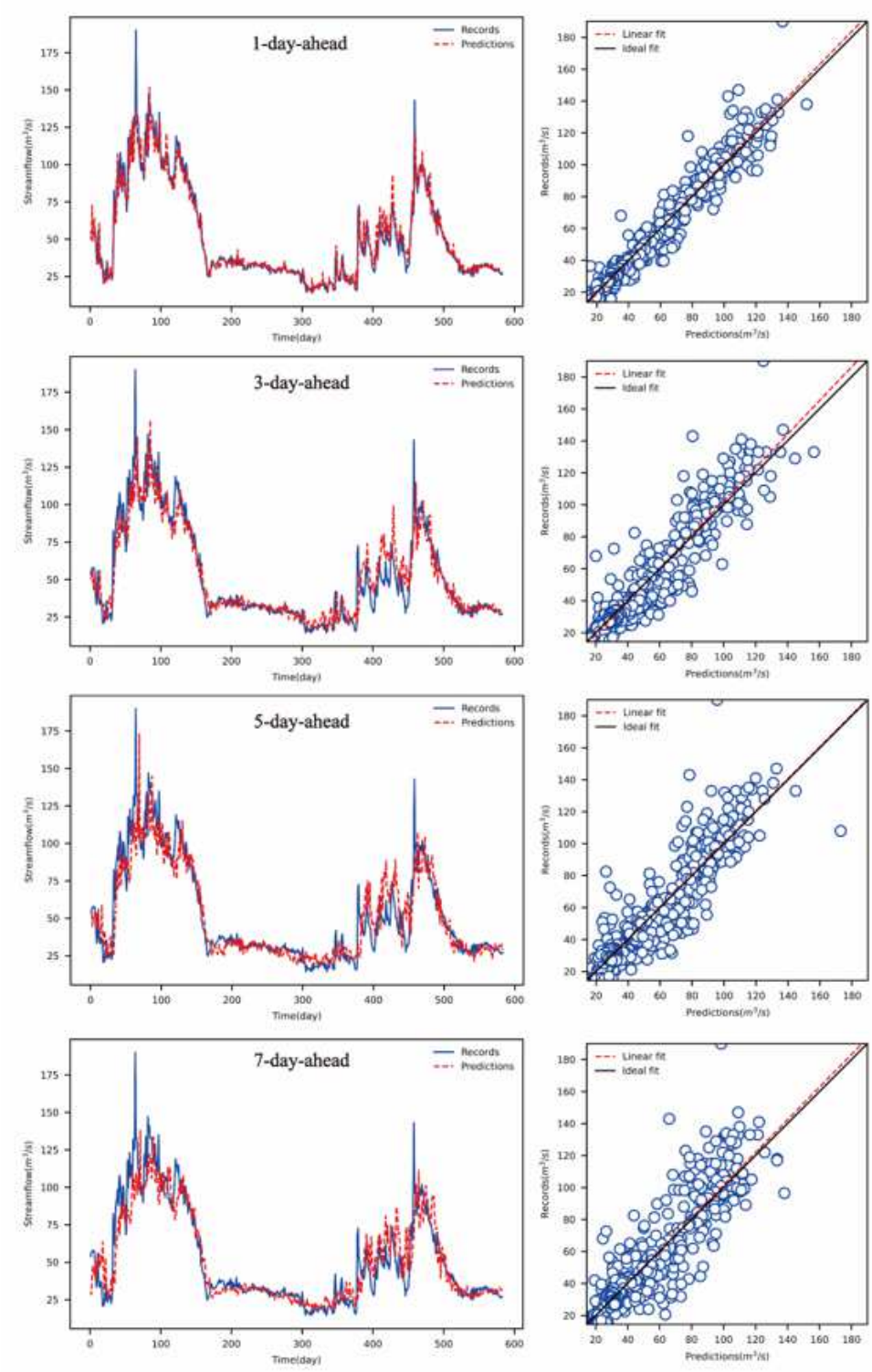

Figure 8

Forecasting results and scatter plots of the optimal PCA settings for the testing set for the Minhe station.

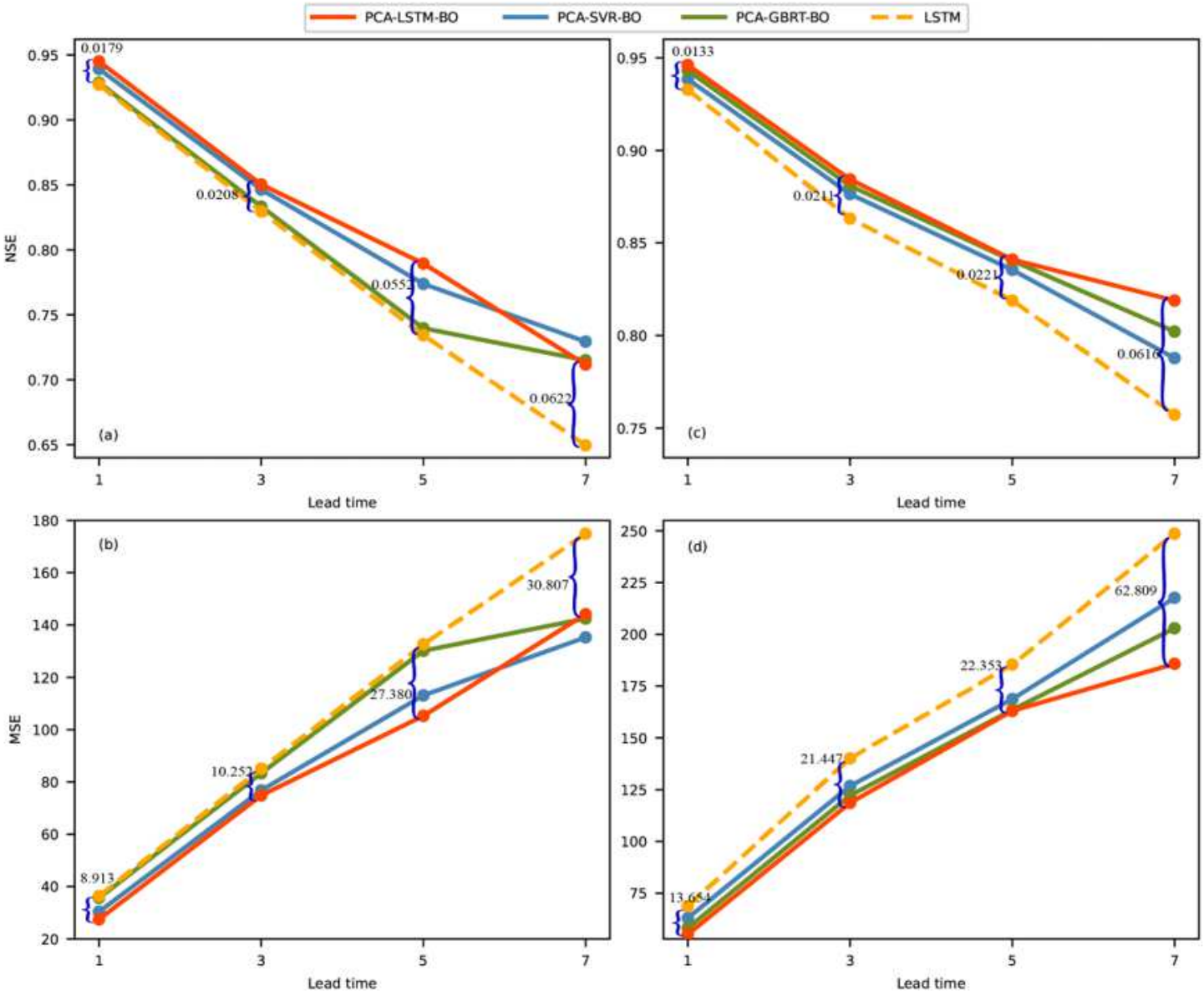

Figure 9

Evaluation results of the forecasting performance of the different models for the Xining (a, b) and Minhe (c, d) stations.

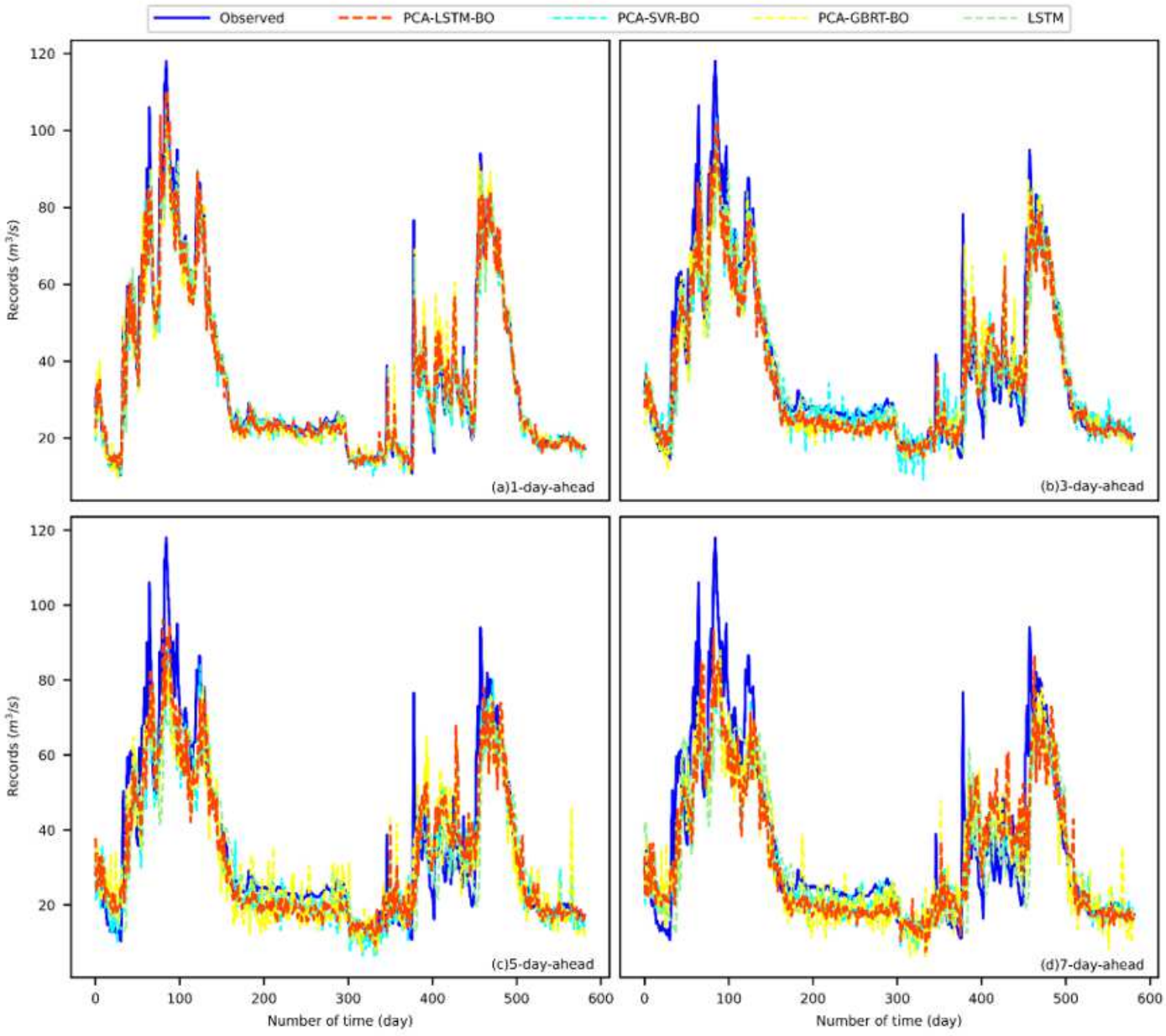

Figure 10

Forecasted and observed results for the testing set at the Xining station.

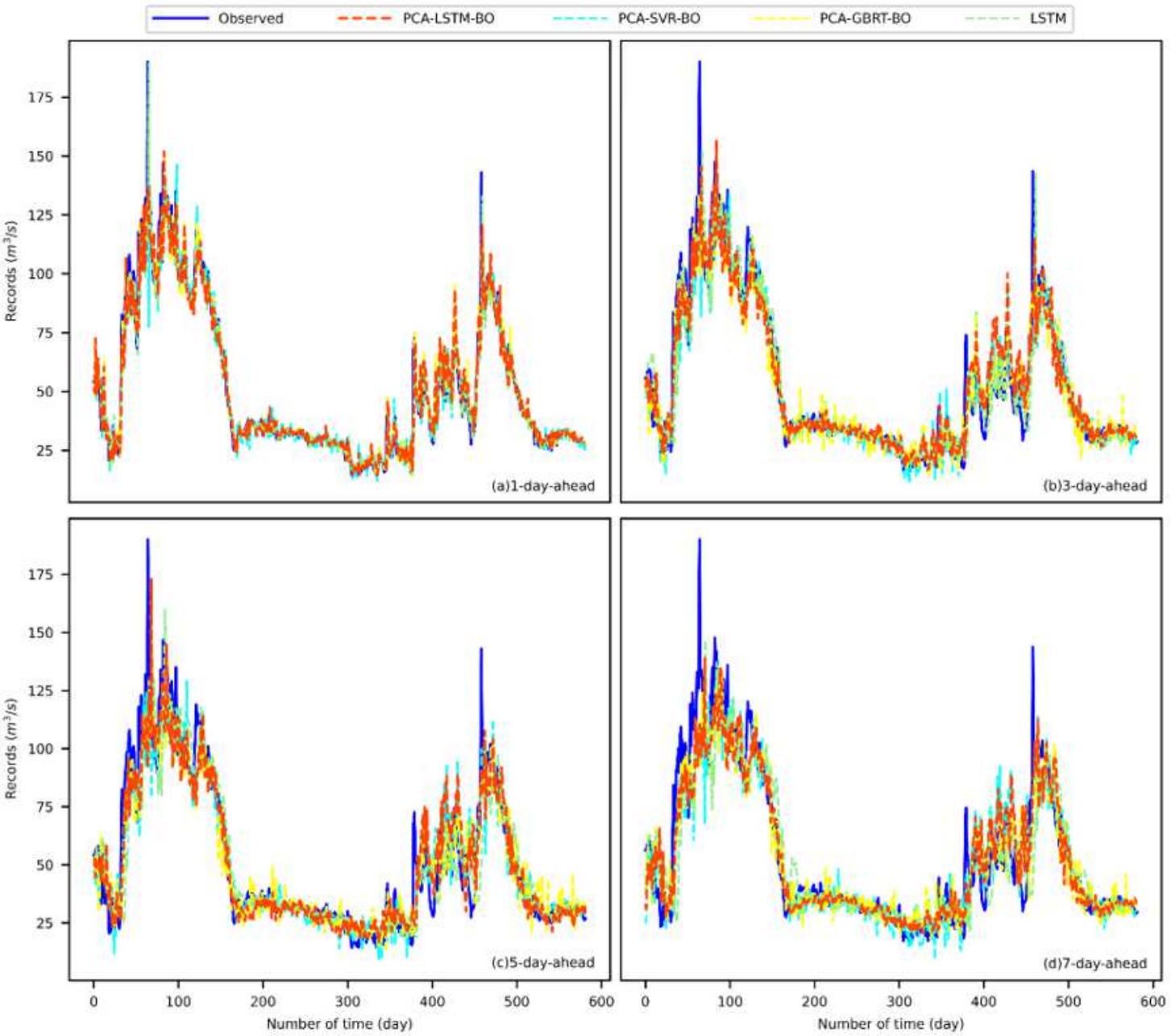

Figure 11

Forecasted and observed results for the testing set at the Minhe station.