ISSN: 2688-8645

Research on Daily Runoff Simulation Based on VMD-CNN-LSTM

Lulu Miao

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China

Abstract

Runoff simulation plays a crucial role in hydrological research and water resource management for scientific planning and flood control strategy formulation. To address runoff data non-stationarity, this study develops a VMD-CNN-LSTM ensemble model integrating Variational Mode Decomposition, Convolutional Neural Network, and Long Short-Term Memory network, aiming to enhance simulation accuracy and model generalization. Validated using 2008-2016 daily runoff data from the Wuding River Basin, the model demonstrates superior performance with training and testing period R^2 values of 0.955 and 0.946, and Nash-Sutcliffe efficiency coefficients of 0.945 and 0.938 respectively, outperforming both standalone LSTM and CNN-LSTM models. Notably, the integrated model shows enhanced capability in peak runoff simulation while maintaining stable accuracy, confirming its robust generalization capacity for hydrological applications.

Keywords

Variational Mode Decomposition; Convolutional Neural Network; Long Short-Term Memory Network; Daily Runoff Simulation.

1. Introduction

As an important component of the water resources system, accurate simulation of runoff is a key link in water resources management and flood control and disaster reduction. It is of great significance for flood control and disaster reduction, ensuring the rational use of water resources, and the sustainable development of ecosystems[1]. However, due to human activities and climate change, runoff changes exhibit complex nonlinear characteristics and high uncertainty[2]. Therefore, how to accurately simulate runoff processes and eliminate non-stationary data has always been a key issue in hydrological research[3].

At present, research methods for runoff simulation can be mainly divided into two categories: process driven models and data-driven models. Process driven models rely on the representation of physical processes, often requiring a large amount of data to calibrate model parameters, and have high requirements for data quality. The process is complex and time-consuming[4]. In recent years, with the development of computer technology, data-driven models have been widely used in runoff simulation due to their significant advantages in processing large amounts of hydro meteorological data[5]. For example, Deng et al. combined convolutional neural networks (CNN) and long short-term memory networks (LSTM) for daily runoff prediction, and the results showed that the CNN-LSTM model was superior to a single LSTM model in runoff prediction[6]; Yuan et al. applied Ensemble Empirical Mode Decomposition (EEMD) and Long Short Term Memory (LSTM) networks for daily runoff prediction, verifying that the EEMD-LSTM model can improve the accuracy of runoff prediction[7]. Yao et al. used a combination of CNN-LSTM and gated cyclic units to simulate runoff in the Bailong River Basin, demonstrating good simulation capabilities[8].

In summary, machine learning methods have gradually become a hot topic in runoff simulation research due to their excellent performance in dealing with complex nonlinear time series problems. Convolutional neural networks (CNN) have powerful feature extraction capabilities and can extract spatial features from complex hydro meteorological data, while long short-term memory networks (LSTM) excel at capturing long-term dependencies and dynamic changes in time series. However, directly applying these methods to simulate runoff may still lead to a decrease in prediction accuracy due to the non stationarity of input data. Therefore, how to effectively deal with the problem of non stationarity in data has become the key to improving model performance.

Variational Mode Decomposition (VMD), as an effective signal processing method, can decompose complex time series into several stationary sub modes, demonstrating significant advantages in eliminating non stationarity in data. The combination of VMD, CNN, and LSTM methods can fully utilize the signal decomposition ability of VMD, the spatial feature extraction ability of CNN, and the time-dependent modeling ability of LSTM, providing a new solution for runoff simulation. However, there is still limited research on runoff simulation based on VMD-CNN-LSTM, and further exploration is needed to determine its applicability and simulation capabilities in actual watersheds.

This article constructs a daily runoff simulation model based on the VMD-CNN-LSTM method, processes the non stationarity of runoff data through variational mode decomposition, extracts key features using convolutional neural networks, and uses long short-term memory networks for time series modeling to improve the accuracy and stability of daily runoff simulation and explore the effectiveness of this method in practical applications. Taking the Wuding River Basin as an example, evaluate the applicability and simulation accuracy of this method, and provide theoretical basis and methodological support for runoff simulation research in the Wuding River Basin.

2. Research Area and Data

2.1. Overview of the Research Area

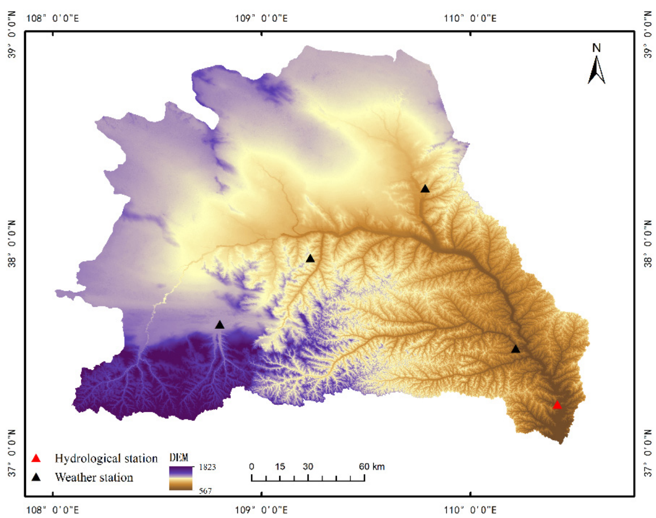


Figure 1. Overview of Wuding River Basin

Wuding River, a first level tributary of the Yellow River, is located in the northern part of Shaanxi Province, China. It is the largest river in the Yulin area of Shaanxi Province. It originates from the northern foot of Baiyushan in Dingbian County, and is called Hongliu River upstream. After passing through Jingbian New Bridge, it is called Wuding River. The total length is 491 kilometers, flowing through Dingbian, Jingbian, Wushen Banner, Hengshan District, Mizhi, Suide and Qingjian County, and flowing into the Yellow River from northwest to southeast. The upstream Hongliu River originates from the eastern foothills of Changchun Liang in the southeast of Dingbian, flows southeast, and joins tributaries such as Yuxi River, Luhe River, Dali River, and Huaining River along the way. It flows into the Yellow River at the mouth of Qingjian County. The location of the Wuding River Basin is shown in Figure 1. The entire basin is located between 107 ° 49 ′ E-110 ° 57 ′ E, 37 ° 06 ′ N-39 ° 27 ′ N, with a total length of approximately 491.2 km. The river flows from northwest to southeast through 15 counties and cities in Inner Mongolia and Shaanxi provinces, with a total drainage area of 30261 square kilometers, of which 21859 square kilometers are within Shaanxi province, accounting for approximately 72.24% of the entire drainage area.

2.2. Data Sources

The runoff and sediment data used in this article is the daily average flow rate of Baijiachuan Hydrological Station in the Wuding River Basin from 2008 to 2016, sourced from the "Hydrological Yearbook of the People's Republic of China Yellow River Basin Hydrological Data"; The meteorological data used in this article comes from the daily meteorological (precipitation, temperature, relative humidity) data of four meteorological stations in the basin provided by China Meteorological Data Network (Figure 1) from 2008 to 2016. The average meteorological data of the basin surface is calculated using the Thiessen polygon method. Use 7 years of data from 2008-2014 for model training and 2015-2016 for model testing.

3. Research Methods

3.1. Variational Mode Decomposition

Variational Mode Decomposition (VMD) is an adaptive signal decomposition method proposed by Dragomiretskiy and Zosso in 2014[9]. The goal of VMD is to decompose a complex signal into a set of Intrinsic Mode Functions (IMFs) with finite bandwidth, enabling multi-scale feature extraction of the signal. Compared with traditional Empirical Mode Decomposition (EMD)[10], VMD overcomes the problem of mode aliasing and has higher robustness and decomposition accuracy.

The core idea of VMD is based on the variational framework, which optimizes the center frequency and bandwidth of each mode, minimizes the first derivative of the complex envelope Fourier transform of each mode, and finally obtains the decomposition result. The objective function is as follows:

$$\min_{u_k,\omega_k}\left\{\sum_{k=1}^{K}\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\cdot e^{-j\omega_k t}\right]\right\|_2^2\right\} \tag{1}$$

In the formula, $u_k(t)$ is the kth mode function, ω_k is its corresponding center frequency, and K is the number of decomposed modes.

Due to the use of frequency domain decomposition method and the introduction of adaptive update mechanism, VMD exhibits superior noise robustness and decomposition result consistency, and is widely used in signal processing, fault diagnosis, time series analysis and other fields.

3.2. CNN Model

Convolutional Neural Network (CNN) is an important model in the field of deep learning, first proposed by LeCun et al., for solving complex image processing and feature extraction tasks[11]. The core structure of CNN includes convolutional layers, pooling layers, and fully connected layers. The convolutional layer extracts local features of data through the local perception mechanism of convolutional kernels, while the pooling layer is used to reduce the spatial size of feature maps, thereby reducing computational complexity and preventing model overfitting. The fully connected layer maps the extracted features to the final prediction results. CNN improves the feature learning and generalization performance of the model by introducing a

ISSN: 2688-8645

layer by layer stacking structure of convolutional layers, pooling layers, and fully connected layers to extract local features from high-dimensional data in a layer by layer manner.

3.3. LSTM Model

Long Short Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) proposed by Hochreiter and Schmidhuber in 1997[12]. The LSTM network introduces unique memory units and their control mechanisms, including forget gates, input gates, and output gates, which can dynamically adjust the storage and updating of information, thereby capturing long-term dependencies in time series and solving the problems of gradient vanishing and exploding that traditional RNNs often encounter when processing long sequence data. Due to its superior capabilities, LSTM is widely used in tasks such as time series prediction, natural language processing, and speech recognition.

3.4. Construction of VMD-CNN-LSTM Model

The formation of runoff is often influenced by multiple driving factors. In typical runoff simulation studies, early runoff, precipitation, average temperature, average wind speed, relative humidity, evapotranspiration, radiation, and sunshine hours are used as input variables for the model [13], in order to comprehensively capture the impact of various driving factors on runoff formation. Due to limitations in data acquisition and quality, this article only uses daily runoff data, daily precipitation, daily average temperature, daily average relative humidity, daily average wind speed, and sunshine hours as input variables for the model. In addition, there is currently no unified method or clear standard for selecting the lag time of input variables in runoff simulation. This article refers to the research of Deng[14] and Li Wenjia[15], and sets the lag time of input variables to 6 days through comprehensive consideration of data characteristics and research needs.

By combining the advantages of VMD, CNN, and LSTM, a VMD-CNN-LSTM model is constructed to achieve efficient decomposition, feature extraction, and time-dependent modeling of complex signals. The basic process is as follows:

Step 1: Decompose the original daily runoff sequence using VMD algorithm to generate several IMFs components with different frequency characteristics. Provide stationary subsequences for subsequent modeling.

Step 2: Combine the IMF components obtained from VMD decomposition with meteorological data as input variables, input them into the CNN-LSTM model for simulation, and output the runoff simulation results for each IMF component.

Step 3: Overlay the simulated runoff results of each IMF component and reconstruct them into a complete runoff sequence to obtain the final daily runoff simulation results.

3.5. Model Evaluation Indicators

This article evaluates the simulation accuracy of the model using Nash Sutcliffe efficiency (NSE), coefficient of determination (R^2), and root mean square error (RMSE). The range of NSE values is (- ∞, 1], and the closer the NSE value is to 1, the higher the credibility of the model. The range of R^2 values is [0.1], and the closer R^2 is to 1, the higher the degree of fit between the predicted and measured values. RMSE can reflect the magnitude of the deviation between predicted and measured values, with a range of [0,+∞). The closer it is to 0, the smaller the error between the model's predicted and measured values.

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

ISSN: 2688-8645

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \overline{y_i})(\hat{y}_i - \overline{\hat{y}_i})}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}_i})^2}} \right)^2 \quad (4)$$

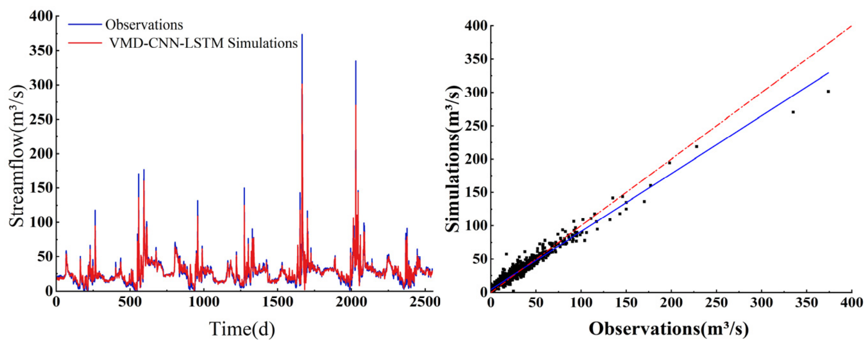
In the formula, y_i and $\hat{y}_i$ represent measured and simulated values, respectively, and n represents the total number of samples.

4. Results and Analysis

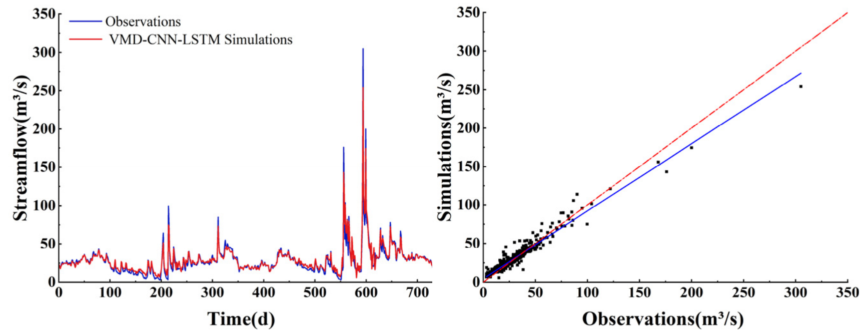
4.1. Simulation Results Analysis of VND-CNN-LSTM Model

Table 1. Evaluation Indicators for Training and Testing Sets of Each Model

Model	Training period			Testing period		
	R²	NSE	RMSE	R²	NSE	RMSE
LSTM	0.839	0.828	8.082	0.823	0.801	9.148
CNN-LSTM	0.883	0.872	6.952	0.863	0.840	8.195
VMD-CNN-LSTM	0.955	0.945	4.562	0.946	0.938	5.083



(a)Training period



(b)Testing period

Figure 2. Comparison between predicted and measured values of VMD-CNN-LSTM

According to Table 1, the VMD-CNN-LSTM model has an R^2 of 0.955, NSE of 0.945, and RMSE of 4.562 m^3/s during the training period; During the testing period, R^2 was 0.946, NSE was 0.938, and RMSE was 5.083 m^3/s. The results showed that VMD-CNN-LSTM achieved high simulation accuracy during both the training and testing periods, and the overall simulation effect was good. From the R^2 and NSE values during the training period, it can be seen that the model can effectively capture the changing patterns between data during the training period, and the simulation results are relatively accurate; The results of the testing period were slightly lower than those of the training period, but the R^2 and NSE values remained at a high level, indicating that VMD-CNN-LSTM has strong generalization ability in the Wuding River Basin and can achieve simulation tasks on different datasets. The increase in RMSE value reflects a slight increase in simulation error during the testing period, but the overall error is within an acceptable range. From the comparison of various evaluation indicators, it can be seen that the

ISSN: 2688-8645

performance during the testing period has decreased compared to the training period, indicating that the VMD-CNN-LSTM model has certain shortcomings in its application in the Wuding River Basin.

The comparison between the daily runoff simulation values and the measured values of the VMD-CNN-LSTM model is shown in Figure 2. It can be seen from the figure that the simulation results of the model are highly consistent with the trend of the measured values during both the training and testing periods, and the overall fitting results are good,. This indicates that the VMD-CNN-LSTM model has strong simulation ability in simulating daily runoff. However, especially at the peak, the model exhibits varying degrees of underestimation. The reason for this phenomenon may be the failure to consider the impact of vegetation changes and human activities on hydrological processes within the study area[16].

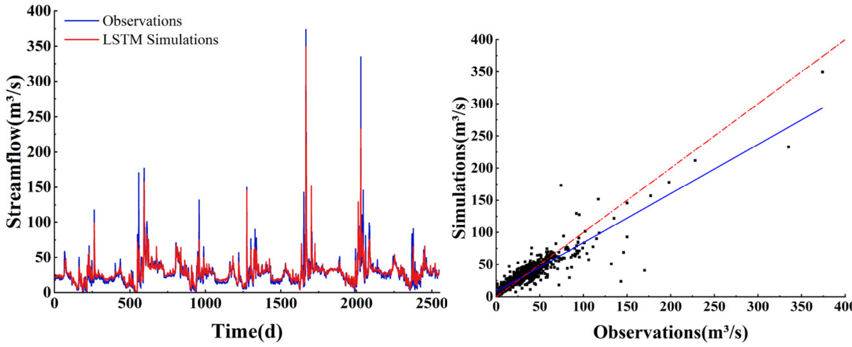
4.2. Comparative Analysis of Model Simulation Accuracy

In order to further verify the effectiveness of the VMD-CNN-LSTM model simulation results used in this paper, a comparative analysis was conducted with LSTM and CNN-LSTM models with the same input to evaluate the performance of the simulation. According to Table 1, the NSE of all three models during the testing period was greater than 0.8, indicating that all three models have high simulation accuracy and can accurately simulate the test data. Further analysis of the evaluation indicators of each model shows that the VMD-CNN-LSTM model outperforms the LSTM and CNN-LSTM models in all evaluation indicators.

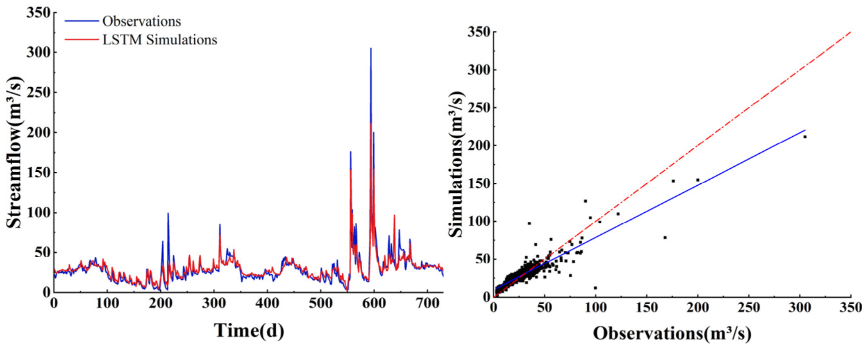
The R^2 value during the testing period of VMD-CNN-LSTM is 0.946, which is 0.123 and 0.083 higher than that of LSTM and CNN-LSTM, respectively. This indicates that the fitting effect of VMD-CNN-LSTM model is better than the other two models, demonstrating better simulation ability. In addition, the NSE value of the VMD-CNN-LSTM model during the testing period was 0.938, which was 0.137 and 0.098 higher than that of LSTM and CNN-LSTM, respectively, further verifying its high accuracy in daily runoff simulation. The higher the NSE value, the higher the agreement between the predicted and observed values of the model. Therefore, the excellent performance of VMD-CNN-LSTM in this indicator indicates that it has better prediction accuracy and stronger simulation ability. In terms of evaluation metric RMSE, the RMSE value of the VMD-CNN-LSTM model during the testing period was 5.083 m^3/s, significantly lower than the 9.148 m^3/s and 8.195 m^3/s of LSTM and CNN-LSTM, further indicating that the VMD-CNN-LSTM model has smaller simulation errors and more accurate simulation results.

Comparing the line graphs of simulated and measured values during the training and testing periods in Figure 2-4, it can be intuitively seen that the VMD-CNN-LSTM model has a higher degree of fit between the simulated and measured values than the LSTM and CNN-LSTM models. During the training period, when the runoff was less than 100 m^3/s, the simulation results of the VMD-CNN-LSTM model among the three models were closer to the 1:1 line, reflecting its good simulation ability at low flow rates. As the measured values increase, the simulation results of the three models have varying degrees of deviation from the measured values, but the VMD-CNN-LSTM model has the smallest degree of dispersion in the simulation results, indicating that the model also has a relatively stable simulation ability when simulating high flow rates; The simulation results of LSTM and CNN-LSTM models exhibit greater fluctuations at higher flow rates, leading to a decrease in the accuracy of the simulation results. During the testing period, the VMD-CNN-LSTM model showed strong predictive ability, and the fitting degree between simulated and measured values was closer to the 1:1 line, indicating that the simulation accuracy of this model during the testing period was higher than that of the LSTM and CNN-LSTM models.

ISSN: 2688-8645

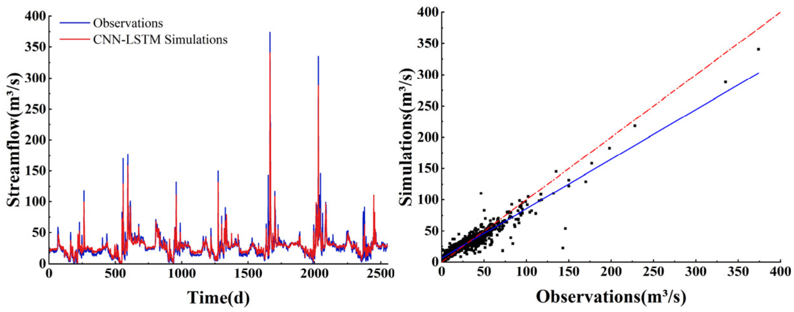


(a)Training period

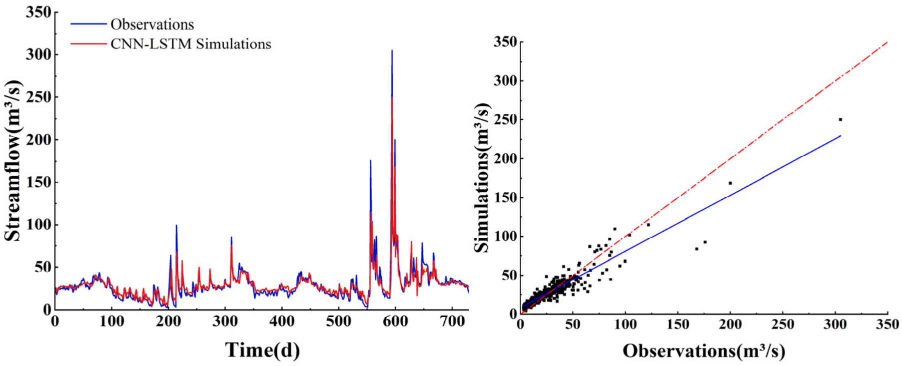


(b)Testing period

Figure 3. Comparison between LSTM predicted values and measured values



(a)Training period



(b)Testing period

Figure 4. Comparison between predicted and measured values of CNN-LSTM

5. Conclusion

(1) The VMD-CNN-LSTM model achieved superior accuracy in Wuding River Basin's daily runoff simulation compared to standalone LSTM and CNN-LSTM models, with training/ validation R^2 values of 0.955/0.946 and Nash-Sutcliffe efficiency coefficients of 0.945/0.938. It

demonstrated R^2 improvements of 0.123/0.083 and NSE gains of 0.137/0.098 over baseline models in respective phases, coupled with the lowest RMSE among all models, confirming its optimal performance.

(2) The integrated model exhibits enhanced generalization capacity, particularly in peak runoff simulations where it significantly reduces prediction errors compared to benchmark models. This validates its reliability in hydrological data processing and its technical advantage in capturing complex runoff dynamics.

References

[1] Ma K, He D, Liu S, et al. Novel time-lag informed deep learning framework for enhanced streamflow prediction and flood early warning in large-scale catchments[J]. Journal of Hydrology, 2024, 631: 130841.

[2] BAO Lina,TANG Deshan,HU Xiaobo 1 ,CHU Shi-ji 1. Runoff Prediction Model Based on Wavelet Decomposition and Arima Error Correction: Research and Application [J]. Journal of Changjiang River Scientific Research Institute,2018,35(12):18-21,33.

[3] Lei Xiaohui,Wang Hao,Liao Weihong. et al. Advances in hydro-meteorological forecast under changing environment [J]. Journal of Hydraulic Engineering,2018.49(1):9-18.

[4] Wen Xiaohu, Feng Qi, Deo R C, et al. Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems[J]. Journal of Hydrology,2019,570: 167-184.

[5] Deng Chao,Chen Chunyu. et al. Catchment runoff simulation by coupling data assimilation andmachine learning methods [J]. Advances in Water Science,2023,34(06):839-849.

[6] Deng H, Chen W, Huang G. Deep insight into daily runoff forecasting based on a CNN-LSTM model[J]. Natural Hazards, 2022, 113(3): 1675-1696.

[7] Yuan R, Cai S, Liao W, et al. Daily runoff forecasting using ensemble empirical mode decomposition and long short-term memory[J]. Frontiers in Earth Science, 2021, 9: 621780.

[8] Yao Z, Wang Z, Wang D, et al. An ensemble CNN-LSTM and GRU adaptive weighting model based improved sparrow search algorithm for predicting runoff using historical meteorological and runoff data as input[J]. Journal of Hydrology, 2023, 625: 129977.

[9] Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, 62(3), 531-544.

[10] HUANG NE., LONG SR., WU MLC., et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society. Mathematical, physical and engineering sciences,1998,454(1971):903-995.

[11] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE,1998, 86(11), 2278-2324.

[12] Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation,1997, 9(8), 1735-1780.

[13] Bharti B, Pandey A, Tripathi S K, et al. Modelling of runoff and sediment yield using ANN, LS-SVR, REPTree and M5 models[J]. Hydrology Research, 2017, 48(6): 1489-1507.

[14] Deng H, Chen W, Huang G. Deep insight into daily runoff forecasting based on a CNN-LSTM model[J]. Natural Hazards, 2022, 113(3): 1675-1696.

[15] Li Wenjia, Wu Lili, Wen Xiaohu, et al. Runoff simulation study based on LSTM-Seq2seq modeloptimized by Attention mechanism [J]. Journal of Glaciology and Geocryology,2023,1-13.

[16] Yang S, Kang T, Bu J, et al. Evaluating the impacts of climate change and vegetation restoration on the hydrological cycle over the Loess Plateau, China[J]. Water, 2019, 11(11): 2241.