Research on Optimal Selection of Runoff Prediction Models Based on Coupled Machine Learning Methods

Xing Wei
1035027480@qq.com
Chongqing Three Gorges University
Mengen CHEN
Chongqing Three Gorges University
Yulin ZHOU
Chongqing Three Gorges University
Jianhua ZOU
Chongqing Three Gorges University
Libo Ran
Chongqing Three Gorges University
Ruibo SHI
Chongqing Three Gorges University

Article

Keywords: Runoff Prediction, Three Gorges Reservoir Area, Variational Mode Decomposition, Sparrow Search Algorithm, Machine Learning

Posted Date: August 13th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4754786/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on December 30th, 2024. See the published version at https://doi.org/10.1038/s41598-024-83695-8.

scientific reports

OPEN

Research on Optimal Selection of Runoff Prediction Models Based on Coupled Machine Learning Methods

Xing WEI [1*], Mengen CHEN [1], Yulin ZHOU [1], Jianhua ZOU [1], Libo Ran [1], Ruibo SHI [1]

[1]School of Civil Engineering, Chongqing Three Gorges University, Chongqing 404100, China
[*]Corresponding author: weixing@sanxiau.edu.cn

Abstract: Runoff fluctuations under the influence of climate change and human activities present a significant challenge and valuable application in constructing high-accuracy runoff prediction models. This study aims to address this challenge by taking the Wanzhou station in the Three Gorges Reservoir area as a case study to optimize various prediction models. The study first selects artificial neural network (ANN) and support vector machine (SVM) as the base models. Then, it evaluates and selects from three time-series decomposition methods: Time-Varying Filter-based Empirical Mode Decomposition (TVF-EMD), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), and Variational Mode Decomposition (VMD). Subsequently, these decomposition methods are coupled with optimization algorithms, including Whale Optimization Algorithm (WOA), Grasshopper Optimization Algorithm (GOA), and Sparrow Search Algorithm (SSA), to construct various hybrid prediction models. The results indicate that: (1) Among the single prediction models, the Long Short-Term Memory (LSTM) model outperforms the Backpropagation Neural Network (BP) and SVM in terms of prediction accuracy; (2) The hybrid models show superior accuracy compared to the individual models, with the VMD-LSTM model outperforming the CEEMDAN-LSTM and TVF-EMD-LSTM models; (3) Among the coupled machine learning prediction models, the VMD-SSA-LSTM model achieves the highest accuracy. Employing a "decomposition-reconstruction" strategy combined with robust optimization algorithms enhances the performance of machine learning prediction models, thereby significantly improving the runoff prediction capabilities in watershed hydrological models.

Keywords: Runoff Prediction; Three Gorges Reservoir Area; Variational Mode Decomposition; Sparrow Search Algorithm; Machine Learning

Introduction

Runoff prediction is crucial in flood control, drought resistance, water resource development, and watershed management, contributing significantly to economic benefits and serving a wide array of applications. In recent years, the combined effects of global climate change and intense human activities have led to frequent extreme weather events and significant alterations in underlying surface geomorphology[1]. These changes exacerbate the nonlinearity and non-stationarity of watershed runoff sequences, presenting new challenges for accurate runoff forecasting[2]. The current focus of research in improving runoff prediction accuracy lies in optimizing predictive models and integrating multiple methodologies to enhance model performance, addressing issues such as insufficient prediction precision for practical needs and lack of model stability[3].

Runoff prediction models primarily fall into two categories: process-driven models based on physical hydrological frameworks, and data-driven models primarily utilizing machine learning[4]. Physical hydrological models require comprehensive calibration data and a certain level of expertise and experience. While they can achieve high-precision short-term runoff predictions in small watersheds, their generalizability for medium to long-term forecasts is limited due to the incomplete understanding of the underlying physical mechanisms of hydrological processes[5]. In contrast, machine learning models are adept at handling highly nonlinear and non-stationary time series, making them widely used in runoff prediction, with popular models including Artificial Neural Networks (ANN)[6], Support Vector Machines (SVM)[7], and Adaptive Neuro-Fuzzy Inference Systems (ANFIS)[8].

As leading machine learning models, ANN and SVM exhibit robust learning capabilities and are extensively applied in the field of runoff prediction, capable of addressing complex problems[9]. However, their performance is significantly influenced by model parameters. Given that runoff processes inherently exhibit trends, periodicity,

 natureportfolio

and high non-stationarity, these models may struggle to fully capture all characteristics of hydrological processes. Therefore, analyzing the temporal scale features of runoff and optimizing the models are essential for enhancing prediction accuracy[10]. Previous researchers have primarily improved hydrological forecasting accuracy through two approaches: (1) Data preprocessing methods decompose highly non-stationary runoff sequences into a series of relatively stable subsequences, simplifying the modeling process for each subsequence; (2) Integrating robust intelligent optimization algorithms to determine the optimal parameter combinations, thus enhancing the performance and predictive capacity of hydrological models[11].

While these methods aim to improve prediction accuracy, the precision obtained by coupling different methods with different machine learning models varies. Therefore, for specific meteorological data and historical runoff data of the study area, selecting high-accuracy, multi-method coupled machine learning models is beneficial for making more scientific analyses, assessments, and decisions regarding future runoff predictions in the research basin[12]. For example, Zhang et al.[13] developed a VMD-SSA-BiLSTM model by integrating Variational Mode Decomposition (VMD) for signal decomposition and preprocessing with the Sparrow Search Algorithm (SSA) for parameter optimization of the Bidirectional Long Short-Term Memory Neural Network (BiLSTM) model, which enhanced the smoothness of monthly runoff sequences and improved point prediction accuracy. WEN et al.[14] used VMD as a data preprocessing technique and coupled it with the emerging Sine Cosine Algorithm (SCA) and Extreme Learning Machine (ELM), demonstrating superior performance compared to single prediction models. Guo et al.[15] compared a single Least Squares Support Vector Machine (LSSVM) model, a CEEMDAN-LSSVM model, and a combined CEEMDAN-SVM–GM(1,1) model, finding that the combination model CEEMDAN-SVM–GM(1,1) exhibited higher accuracy.

In summary, this study aims to combine different preprocessing methods and optimization algorithms with machine learning models, compare their prediction results, and select a reliable, efficient, and practical multi-method hybrid machine learning model for predicting highly irregular, complex nonlinear, and multi-scale variable monthly runoff sequences. This approach offers a new perspective for improving runoff prediction accuracy under changing environmental conditions. The study will use the monthly runoff forecast at the Wanzhou hydrological station, a key control section in the heart of the Three Gorges Reservoir area, as a case study to validate the effectiveness and feasibility of the selected model.

Research Methodology

Machine Learning Models

With the advent of the big data era, machine learning has become increasingly effective in processing a nd transforming information from large datasets, learning from experience to make inferences and decisi ons. In this study, we utilize three machine learning models for monthly runoff prediction in the waters hed: BP, LSTM, and SVM.

BP Neural Network (Backpropagation Neural Network): The primary characteristics of the BP neural network involve forward propagation of signals and backward propagation of errors. It adjusts the weights and thresholds of the entire network through backpropagation to minimize the sum of squared errors[16]. Previous applications in hydrological forecasting research have demonstrated that a BP neural network with a single hidden layer can meet the accuracy requirements for predictions. Therefore, a three-layer BP neural network model, comprising an input layer, a hidden layer, and an output layer, is sufficient to predict changes in hydrological elements accurately. In this study, we construct a three-layer BP neural network model for runoff prediction.

LSTM Neural Network (Long Short-Term Memory Network): The LSTM neural network is an improved model of the Recurrent Neural Network (RNN). It addresses the issues of gradient vanishing and gradient explosion that can occur in simple RNNs by incorporating specially designed memory cells capable of selective memory. These memory cells can choose to retain important information[17]. The LSTM network introduces three unique gates: the input gate, the forget gate, and the output gate, allowing the model to selectively protect and forget information, thus efficiently capturing dependencies in time series data.

SVM (Support Vector Machine): The SVM model introduces kernel functions to map non-linearly separable data in the original low-dimensional space into a higher-dimensional feature space where it becomes linearly separable. SVM seeks to minimize overall risk by finding an optimal trade-off between "structural risk" and "empirical risk" using a penalty factor[18]. To avoid the risk of "overfitting," SVM employs slack variables to create a "soft margin." This approach effectively addresses practical issues such as the curse of dimensionality and small sample sizes that are common in traditional machine learning. Consequently, SVM is widely used in the field of hydrological forecasting.

When using the above machine learning models for runoff prediction, it is crucial to ensure that the models converge and accurately capture the characteristics of the data. This is achieved by normalizing and denormalizing the input and output factors before training the models. The normalization process is carried out using the following formula:

$$\bar{x}_i = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}} \tag{1}$$

Where x_{imax} and x_{imin} are the maximum and minimum values in the iii-th input dataset, respectively, $\bar{x}_i$ is the normalized value for the iii-th input data point.

For denormalization, the formula used is:

$$y_i = \frac{\bar{y}_i - y_{imin}}{y_{imax} - y_{imin}} * 0.998 + 0.001 \tag{2}$$

Where y_{imax} and y_{imin} are the maximum and minimum values in the iii-th output dataset, $\bar{y}_i$ is the normalized value for the iii-th output data point, y_i is the denormalized value for the iii-th output data point after the denormalization process.

Decomposition Techniques

The complexity of the original sequence affects the capability of machine learning. The simpler the sequence, the easier it is to model using machine learning. The "decomposition-reconstruction" prediction model established based on preprocessing methods breaks down complex runoff sequences into simpler parts, resulting in better data fitting capabilities and relatively more accurate predictions[19].

CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) is a novel signal decomposition algorithm. This method resolves the modal mixing problem in Empirical Mode Decomposition (EEMD) by adding a finite number of adaptive Gaussian white noise steps during each decomposition stage, thereby enhancing the accuracy and stability of the decomposition[20].

TVF-EMD (Time-Varying Filter-based Empirical Mode Decomposition) improves traditional EMD by introducing time-varying filters. This method effectively addresses the issues of incomplete or excessive decomposition that may occur in conventional methods. The core of this approach lies in using time-varying filtering techniques to complete the sifting process, adaptively designing local cutoff frequencies by fully leveraging instantaneous amplitude and frequency information[21]. In summary, TVF-EMD is a fully adaptive signal decomposition method that enhances frequency separation performance and maintains stability at low sampling rates. It also exhibits robustness against noise interference, making it suitable for analyzing both linear and non-stationary signals.

VMD (Variational Mode Decomposition) is a new adaptive decomposition method based on Wiener filtering and Hilbert transform. Essentially, it applies the classic Wiener filter to multiple adaptive segment waves of the signal. By selecting modes and their central frequencies, VMD decomposes a complex signal into several mode components while maintaining each component as monotonic as possible, making the decomposed signal easier to interpret and analyze[22]. Moreover, VMD has strong anti-interference capabilities, effectively overcoming the influence of high-frequency noise and balancing the errors between time-domain signals and their corresponding frequency-domain signals.

Swarm Intelligence Algorithms

Given the high complexity and nonlinearity of runoff prediction problems, swarm intelligence algorithms are widely used due to their global optimization capabilities, strong adaptability, robustness, and high interpretability[23]. By integrating robust optimization algorithms, we can better determine the optimal parameter combinations and enhance the predictive performance of hydrological models. This study employs three swarm intelligence algorithms to couple with machine learning models and selects the combination model with the highest prediction accuracy.

Whale Optimization Algorithm (WOA) is a meta-heuristic optimization algorithm inspired by the hunting behavior of humpback whales. The main mechanism includes encircling prey, bubble-net feeding, and searching for prey. In WOA, each whale's position represents a potential solution. By continuously updating the positions of whales in the solution space, the algorithm ultimately finds the global optimal solution[24]. The position update mechanism is calculated by creating a spiral equation to compute the distance between the individual and the best position, using the formula:

$$X(t+1) = x'(t) + e^{bl} \times \cos(2\pi l) \times D_i \tag{3}$$

$$D_i = |x'(t) - x(t)| \tag{4}$$

Where $x'(t)$ is the current best foraging position, b is a constant used to adjust the shape of the spiral, l is a random number in the range [-1, 1], D_i is the distance between the current individual and the best foraging position, $x(t)$ is the current position of the individual.

Grasshopper Optimization Algorithm (GOA) was proposed by Shahrzad Saremi et al. in 2017[25]. It is a swarm intelligence optimization algorithm inspired by the foraging behavior of grasshopper swarms. GOA is known for its fast convergence speed and high search efficiency. The algorithm divides the search space into repulsive, comfortable, and attractive zones based on the repulsive and attractive forces among grasshoppers. It adjusts the effect of these forces according to the distance between grasshoppers and uses an abstraction function to find the optimal solution. The implementation steps of GOA are as follows:

Step 1 : Initialize the population size, population positions, maximum number of iterations, and control parameters.

Step 2 : Calculate the fitness value of each grasshopper and rank them in ascending order.

Step 3 : Update the control parameters.

Step 4 : Update the positions of the grasshoppers and calculate their fitness values.

Step 5 : Check if the termination condition is met. If yes, output the results; otherwise, repeat steps 2-4.

Sparrow Search Algorithm (SSA) was developed by Xue et al. in 2020[26], inspired by the foraging and anti-predation behaviors of sparrows. In SSA, the sparrow population is divided into three roles: discoverers, joiners, and scouts. Discoverers and joiners dynamically switch roles where discoverers provide the direction and area for foraging, and joiners follow the signals from the discoverers to forage. Scouts use anti-predation strategies to provide warnings, helping the population avoid local optima. The interaction among these roles facilitates the optimization search. The mathematical model of SSA is as follows:

Assuming the number of sparrows in the population is \(n \), the population is represented as:

$$G = \begin{pmatrix} x_{11} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \tag{5}$$

Where G represents the entire population, and d denotes the dimensionality of the sparrow population.

The fitness value of a sparrow is given by:

$$F_X = \begin{pmatrix} f\left[x_{1,1}x_{1,2} \dots x_{1,d}\right] \\ \vdots \\ f\left[x_{n,1}x_{n,2} \dots x_{n,d}\right] \end{pmatrix} \tag{6}$$

Where F_X represents the fitness values of all sparrows, and f denotes the fitness value of an individual sparrow.

The position changes for the three different roles of sparrows are determined as follows, with the position change for discoverers given by:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \times exp\left(-\dfrac{i}{\alpha t_{max}}\right), R' < \delta_{st} \\ x + Q \times L, R' \geq \delta_{st} \end{cases} \tag{7}$$

Where $X_{i,j}^{t+1}$ and $X_{i,j}^{t}$ are the positions of the i-th sparrow in the j-th dimension in the $t+1$ and t iterations, α is a random number in the range (0,1), t_{max} is the maximum number of iterations, R' is the warning threshold, Q is a random number following a normal distribution, L is a $1 \times d$ matrix with all elements equal to 1.

The position change for joiners is given by:

$$X_{i,j}^{t+1} = \begin{cases} Q \times exp\left(\dfrac{X_w - X_{i,j}^{t}}{i^2}\right), i > n/2 \\ X_p^{t+1} + \left|X_{i,j}^{t} - X_p^{t+1}\right| \times A^{+} \times L, i \leq \dfrac{n}{2} \end{cases} \tag{8}$$

Where $X_{i,j}^{t+1}$ is the best position for the discoverer in the $t+1$ iteration, X_w is the current worst global position, $A^{+} = A^{T}(AA^{T})^{-1}$, A is a $1 \times d$ matrix with all elements being 1 or -1.

The position change for scouts is given by:

$$X_{i,j}^{t+1} = \begin{cases} X_b^{t} + \beta \times \left|X_{i,j}^{t} - X_b^{t}\right|, f_i > f_b \\ X_{i,j}^{t} + K \times \left(\dfrac{\left|X_{i,j}^{t} - X_w^{t}\right|}{(f_i - f_w) + \varepsilon}\right), f_i = f_b \end{cases} \tag{9}$$

Where X_b^{t} is the global best position after the t-th iteration, β is the step size control parameter, K is a random number in the range [-1,1], f_i is the current fitness value of the scout, f_b and f_w are the current best and worst fitness values, ε is a constant ensuring the denominator is not zero.

Evaluation Metrics

To comprehensively evaluate the prediction performance of the nine models, the following metrics are selected: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency (NSE), Mean Absolute Percentage Error (MAPE), and Correlation Coefficient (R). These metrics quantify the effectiveness of the models' predictions. The technical workflow is illustrated in Figure 1. The evaluation criteria are calculated using the following formulas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - y_i| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - y_i)^2} \tag{11}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Y_i - y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_1)^2} \tag{12}$$

$$MAPE = \frac{100\%}{n}\sum_{i}^{n}\left|\frac{Y_i - y_i}{Y_i}\right| \tag{13}$$

$$R = \frac{\sum_{i=1}^{n}[(Y_i - \overline{Y}_1)(y_i - \bar{y}_1)]}{\sqrt{\sum_{i=1}^{n}[(Y_i - \overline{Y}_1)^2(y_i - \bar{y}_1)^2]}} \tag{14}$$

Where Y_i is the i-th observed true sample, y_i is the i-th predicted test value, $\overline{Y}_1$ is the mean of all observed values, $\bar{y}_1$ is the mean of all predicted values.

Case Study Analysis

Overview of the Study Area

The study area is the Three Gorges Reservoir region, with the Wanzhou Hydrological Station located at the "heart" of this area selected as the research site. The Wanzhou Hydrological Station, situated in Wanzhou District, Chongqing, is a central reporting station for flood information. It covers a catchment area of approximately 970,000 km² and serves as a first-class hydrological station for monitoring the lower reaches of the main stream in the upper Yangtze River. It is a key national hydrological station and collects essential data for the construction, operation, and regulation of the Three Gorges Dam project. The research data include historical observations from the Wanzhou Hydrological Station (monthly runoff data from January 1960 to December 2019) and meteorological data from four weather stations, including the Wanzhou Meteorological Station. The Three Gorges Reservoir region and station information are shown in Figure 2.

Trend and Periodicity Analysis of Runoff Series

The Mann-Kendall trend test[27]and wavelet analysis[28]were employed to analyze the trend and periodicity of the runoff series at the research station, as illustrated in Figures 3 and 4.

From the Mann-Kendall trend test results of the monthly runoff series shown in Figure 3, it can be observed that there is a decreasing trend in the monthly runoff series over the past 60 years. The Z-value obtained from the Mann-Kendall trend test is 0.5009 ($Z < 1.28$), indicating that the monthly runoff series at the research station does not pass the significance test. Generally, due to the dual influences of human activities and climate change, river runoff series over extended periods tend to exhibit certain changes. Based on the runoff characteristics and simulation requirements, data are divided into training and testing sets according to the 75% and 25% split principle. Thus, the monthly runoff data for the first 45 years (1960-2004) are used as the training set, and the data for the last 15 years (2005-2019) are used as the testing set.

Using the nearly 60-year runoff series in the Three Gorges Reservoir area, the annual runoff wavelet coefficient contour map and wavelet variance diagram are produced after wavelet transformation, as shown in Figure 4. The annual runoff in the Three Gorges Reservoir area exhibits multi-scale time-frequency variation characteristics and multiple primary periodic changes, indicating a pattern of alternating wet and dry periods. The wavelet variance of the annual runoff series reveals two significant peaks, corresponding to primary periods of approximately 5 and 19 years. The first (largest) peak occurs on a 19-year time scale, indicating the strongest periodic oscillation around this interval. The main periods for monthly runoff are 9 months and 18 months, with the maximum wavelet variance occurring at 18 months. Considering the lag effects of climate factors on the runoff at the research station and the periodicity of monthly runoff, the study includes lags of 1 to 20 months as different delay periods for model input selection. Suitable lag periods are determined through lag testing.

Selection of Single Machine Learning Models for Runoff Prediction

Using monthly runoff series data from the Wanzhou Hydrological Station spanning 720 months from 1960 to 2019, and meteorological data from four weather stations, including the Wanzhou Meteorological Station, runoff is selected as the prediction target. Rainfall, temperature, and preceding runoff are used as input features for the machine learning models. Considering the monthly runoff cycle analysis from the previous sections, lag periods of 1 to 20 months are tested using RMSE as the reference target, as shown in Figure 5. The results indicate that a 15-month lag period is optimal. Consequently, all subsequent model inputs use a 15-month lag.

Three prediction models—BP, LSTM, and SVM—are established and compared for optimization. The results are summarized in Table 1. Table 1 uses five evaluation metrics to assess the three models. The prediction accuracies of the BP, LSTM, and SVM models are relatively similar. The Nash-Sutcliffe Efficiency (NSE) for the LSTM and SVM models on the test set are 0.72513 and 0.7585, respectively, both greater than 0.7, achieving a B-grade runoff prediction level (NSE in the range [0.7, 0.9] indicates B-grade prediction accuracy). The BP model has an NSE of 0.67733, indicating a C-grade runoff prediction level (NSE in the range [0.5, 0.7] indicates C-grade prediction accuracy). Overall, the prediction accuracies of the LSTM and SVM models at the research site are higher than those of the BP model. Considering the LSTM model's sensitivity to time, better memory capabilities, and unique gating mechanism, it is more suitable for handling runoff prediction problems with strong time-lag, nonlinearity, and long time series[29]. Therefore, the LSTM model is selected for coupling with different methods for runoff prediction.

Runoff Prediction Optimization with "Decomposition Techniques" Coupled Machine Learning Models

Hybrid prediction models are established by coupling different decomposition methods with the LSTM model. The process involves decomposing the runoff series into multiple sub-series using different decomposition methods, modeling each sub-series individually, and then summing the predicted values of all sub-series to obtain the final predicted runoff series.

Given the better prediction results of the LSTM model, further hybrid models CEEMDAN-LSTM, TVF-EMD-LSTM, and VMD-LSTM are established and compared for optimization. The prediction results are shown in Figure 6. As seen in Figure 6, the LSTM models coupled with the three decomposition techniques outperform the standalone LSTM model in both training and test sets. The NSE for the VMD-LSTM model is 0.93676 (training set) and 0.84016 (test set), higher than those of the CEEMDAN-LSTM and TVF-EMD-LSTM models, representing improvements of 11.4% and 15.86%, respectively, compared to the standalone LSTM model.

To enable the LSTM model to better recognize runoff patterns and reduce prediction difficulty, decomposition techniques are used to decompose the runoff data for the study area. Figure 7 shows the time-domain decomposition diagram of the runoff series based on VMD. During the VMD decomposition process, the selection of the number of modes k and the penalty factor α significantly affects the decomposition results. A small k may lead to modal overlap and insufficient decomposition precision, failing to reduce the complexity of the original series effectively. A large k may result in some modes having the same decomposition frequency, causing over-decomposition. The selection of k is determined by observing the central frequency; when the central frequency selected at k_i is very close to those at k_{i+1} and k_{i+2}, it indicates possible frequency mixing, suggesting k_i as the optimal number of decomposition modes[30]. Based on this theory, k=6 is chosen. Additionally, through trial and error, the penalty factor α is set to 50. The time-domain decomposition diagram of the runoff series based on VMD is shown in Figure 7.

Optimization of Hybrid Machine Learning Models for Runoff Prediction

Beyond data preprocessing strategies, integrating robust optimization algorithms can also enhance the prediction accuracy of standalone models. To validate the effectiveness of different optimization algorithms coupled with machine learning models, three monthly runoff prediction models—VMD-WOA-LSTM, VMD-GOA-LSTM, and VMD-SSA-LSTM—are established and compared for optimization.

Table 2 presents the runoff prediction results for the training and test sets of four models: VMD-LSTM, VMD-WOA-LSTM, VMD-GOA-LSTM, and VMD-SSA-LSTM. Using five evaluation metrics to assess these models, the accuracy of VMD-WOA-LSTM, VMD-GOA-LSTM, and VMD-SSA-LSTM in both training and test sets is higher than that of VMD-LSTM, indicating that integrating different swarm intelligence optimization algorithms can improve model prediction accuracy. Additionally, the NSE for VMD-SSA-LSTM and VMD-WOA-LSTM models on the test set are 0.9488 and 0.92823, respectively, both greater than 0.9, achieving an A-grade prediction level (NSE $\geqslant$ 0.9 indicates A-grade prediction accuracy)[31]. Overall, the statistical indicators of VMD-SSA-LSTM in both training and test sets outperform those of VMD-WOA-LSTM and VMD-GOA-LSTM, suggesting that the integrated SSA algorithm has higher convergence accuracy and robustness compared to WOA and GOA. Moreover, VMD-SSA-LSTM shows better adaptability in runoff prediction at the Wanzhou station in the Three Gorges Reservoir area, thus delivering the best overall forecasting performance.

Figure 8 compares the predicted and observed runoff data for the training and test sets of the three hybrid models—VMD-WOA-LSTM, VMD-GOA-LSTM, and VMD-SSA-LSTM. The fitting curves and correlation coefficient R between the observed and simulated values indicate that all three models achieve good predictive performance, with VMD-SSA-LSTM slightly outperforming VMD-WOA-LSTM and VMD-GOA-LSTM. This demonstrates that the decomposition-reconstruction strategy and integration of robust swarm intelligence optimization algorithms effectively capture the nonlinear relationships in the original monthly runoff data and determine the optimal parameter combinations, thereby improving the predictive accuracy of single machine learning models and achieving superior runoff forecasting results.

Discussion

This study focuses on the Wanzhou station, located at the "heart" of the Three Gorges Reservoir area. Over the past 60 years, the monthly runoff series at this station has shown no significant increasing or decreasing trends, as confirmed by the Mann-Kendall trend test, indicating relative stability. Given the challenges posed by runoff prediction with strong time-lag effects, nonlinearity, and long time series, the LSTM model demonstrated superior prediction accuracy at this research site. This success can be attributed to the LSTM model's sensitivity to temporal data, enhanced memory capabilities, and unique gating mechanisms.

Furthermore, the application of a "decomposition-reconstruction" strategy combined with robust optimization algorithms in machine learning models has significantly improved the prediction accuracy for the runoff at the research site. As shown in Table 3, the VMD-SSA-LSTM model outperforms both the VMD-LSTM and standalone LSTM models across all five evaluation metrics. The VMD-SSA-LSTM model shows substantial improvements over the LSTM model, with the Nash-Sutcliffe Efficiency (NSE) increasing from 0.72513 to 0.984, upgrading the prediction accuracy level from B-grade to A-grade[32].

Additionally, the "decomposition-reconstruction" strategy proves advantageous for enhancing the accuracy of individual prediction models. However, this method often encounters boundary effects, which can negatively impact the quality of time-series modeling and overall predictive performance. Future research should focus on improving extension algorithms to reduce the boundary effect in decomposition preprocessing techniques.

Regarding the integration of robust swarm intelligence optimization algorithms, while these methods have rapidly developed robust theoretical frameworks and are widely applied across various fields due to their robustness, self-organization, and flexibility, they still suffer from slow convergence speeds and a tendency to get trapped in local optima. Thus, further research and discussion are required on how to enhance these algorithms to achieve more reasonable and efficient practical modeling[33].

In recent years, global hydrological processes have exhibited significant changes due to human activities and climate change. The nonlinearity, non-stationarity, and uncertainty of runoff sequences have become increasingly prominent, posing greater challenges for future runoff prediction[34]. Therefore, the focus of the next research phase will be on how to deeply explore the changing characteristics of runoff sequences under varying environmental conditions, enhance the learning and generalization capabilities of prediction models, and improve the accuracy of runoff predictions.

Conclusion

In this study, three machine learning models—BP, LSTM, and SVM—were employed to predict runoff at the Wanzhou station, located in the "heart" of the Three Gorges Reservoir area. The results indicate that the LSTM model provides the highest prediction accuracy, followed by SVM, with BP performing the worst. This suggests that the LSTM model's ability to capture trends and temporal correlations in sequence data makes it more effective for runoff prediction in regions with strong trend characteristics like the Three Gorges Reservoir area.

Using runoff decomposition techniques enhances the representation of essential characteristics such as periodicity and sequence properties in runoff data, which in turn improves prediction accuracy. Consequently, the three coupled models based on the "decomposition-reconstruction" strategy outperform the standalone models in terms of prediction accuracy. Among the coupled models, the monthly runoff prediction accuracy, from highest to lowest, is as follows: VMD-LSTM > CEEMDAN-LSTM > TVF-EMD-LSTM.

Among the nine runoff prediction models constructed in this study, the VMD-SSA-LSTM and VMD-GOA-LSTM models exhibit the highest prediction accuracy, with NSE values of 0.948 and 0.92823 on the test set, respectively, both reaching A-grade prediction accuracy levels and surpassing the other models. Overall, the VMD-SSA-LSTM model demonstrates superior adaptability and generalization performance, making it the best overall forecasting model.

In the runoff prediction for the Wanzhou station in the Three Gorges Reservoir area, the combination of the "decomposition-reconstruction" strategy and the integration of robust optimization algorithms successfully enhanced the prediction accuracy of individual runoff prediction models. However, there are still some limitations that need to be addressed in future research phases. The "decomposition-reconstruction" strategy suffers from boundary effects, which can affect the quality of time-series modeling and limit improvements in prediction accuracy. Further research could focus on improving extension algorithms to mitigate endpoint effects in decomposition preprocessing techniques. Additionally, robust swarm intelligence optimization algorithms often have slow convergence speeds and a tendency to get trapped in local optima. Improving these algorithms to achieve more reasonable and efficient practical modeling remains a key area for future development.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

References

1. Mosavi, A. Ozturk, P. Chau, K.-w. Flood Prediction Using Machine Learning Models: Literature Review. *Water*. **10**, 1536 (2018).
2. Zhang, J, Wang, D, Wang, Y. et al. Runoff Prediction Under Extreme Precipitation and Corresponding Meteorological *Conditions. Water Resources Management. 37, 3377–3394 (2023).*
3. Yang, H. & Li, W. Data decomposition, seasonal adjustment method and machine learning combined for runoff prediction: A case study. *Water Resources Management*. **37**, 557–581 (2023).
4. Devia G K, Ganasri B P, Dwarakish G S. A review on hydrological models. *Aquatic Procedia*. **4**: 1001–1007 (2015).
5. Niu W J, Feng Z K, Chen Y B, et al. Annual streamflow time series prediction using extreme learning machine based on gravitational searchalgorithm and variational mode decomposition. *Journal of hydrologic engineering*. **25**: 1943–5584 (2020).
6. Yang, X., Zhou, J., Zhang, Q. et al. Evaluation and Interpretation of Runoff Forecasting Models Based on Hybrid Deep Neural Networks. *Water Resour Manage*. **38**, 1987–2013 (2024).
7. Sharma, B., Goel, N.K. Streamflow prediction using support vector regression machine learning model for Tehri Dam. *Applied Water Science.* **14**, 99 (2024).
8. Tabbussum, R., Dar, A.Q. Performance evaluation of artificial intelligence paradigms—artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction. *Environmental Science Pollution Research*. **28**, 25265–25282 (2021).
9. Mohammadi, B. A review on the applications of machine learning for runoff modeling. Sustain. *Water Resources Management*. **7**, 98 (2021).
10. Zhou YL, Mu ZX, Peng L. et al. Mid-long term hydrological forecast of snowmelt runoff in western Tianshan Mountains based on mutual information and neural network[J]. *Journal of Yangtze River Scientific Research Institute*. **35**,17-21 (2018).
11. Yang, H., Li, W. Data Decomposition, Seasonal Adjustment Method and Machine Learning Combined for Runoff Prediction: A Case Study. *Water Resour Manage*. **37**, 557–581 (2023).
12. Liu, S., Qin, H., Liu, G. et al. Runoff Forecasting of Machine Learning Model Based on Selective Ensemble. *Water Resour Manage.* **37**, 4459–4473 (2023).
13. Zhang, X., Wang, X., Li, H. et al. Monthly runoff prediction based on a coupled VMD-SSA-BiLSTM model. *Scientific Reports.* **13**, 13149 (2023).
14. Wen J N, Zhong K F, Wen F Y & Jun Z Short term streamflow time series predictiomodel by machine learning tool based on data preprocessing technique and swarm intelligence algorithm, *Hydrological Sciences Journal*. **65**, 2590-2603, (2020).
15. Guo, S., Wen, Y., Zhang, X. et al. Runoff prediction of lower Yellow River based on CEEMDAN–LSSVM–GM(1,1) model. *Scientific Reports.* **13**, 1511 (2023).
16. He, S., Sang, X., Yin, J. et al. Short-term Runoff Prediction Optimization Method Based on BGRU-BP and BLSTM-BP Neural Networks. *Water Resour Manage*. **37**, 747–768 (2023).
17. Gao, L., Ha, D., Ma, L. et al. The prediction model of water level in front of the check gate of the LSTM neural network based on AIW-CLPSO. *Journal of Combinatorial Optimization*. **47**, 5 (2024).
18. Sharma, B., Goel, N.K. Streamflow prediction using support vector regression machine learning model for Tehri Dam. *Applied Water Science*. **14**, 99 (2024).
19. Yang, X., Chen, Z. & Qin, M. Monthly Runoff Prediction Via Mode Decomposition-Recombination Technique. *Water Resour Manage*. **38**, 269–286 (2024).
20. Katipoğlu, O.M. Integration of extreme learning machines with CEEMDAN and VMD techniques in the prediction of the multiscalar standardized runoff index and standardized precipitation evapotranspiration index. *Natural Hazards*. **120**, 825–849 (2024).
21. Jamei, M., Ali, M., Malik, A. et al. Forecasting Daily Flood Water Level Using Hybrid Advanced Machine Learning Based Time-Varying Filtered Empirical Mode Decomposition Approach. *Water Resour Manage*. **36**, 4637–4676 (2022).
22. Hu, H., Zhang, J. & Li, T. A Novel Hybrid Decompose-Ensemble Strategy with a VMD-BPNN Approach for Daily Streamflow Estimating. *Water Resour Manage*. **35**, 5119–5138 (2021).
23. Katipoğlu, O.M., Sarıgöl, M. Improving the accuracy of rainfall-runoff relationship estimation using signal processing techniques, bio-inspired swarm intelligence and artificial intelligence algorithms. *Earth Science Informatics*. **16**, 3125–3141 (2023).
24. Wang, WC., Wang, B., Chau, Kw. et al. Monthly runoff time series interval prediction based on WOA-VMD-LSTM using non-parametric kernel density estimation. *Earth Science Informatics*. 2373–2389 (2023).
25. Shahrzad Saremi, Seyedali Mirjalili, Andrew Lewis. Grass hopper optimisation algorithm: Theory and application. *Advances in Engineering Software* , **105**, 30-47 (2017).
26. XUE J K, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science & Control Engineering*. **8**, 22-34 (2020).
27. Alashan, S. Non-monotonic trend analysis using Mann–Kendall with self-quantiles. *Theoretical Applied Climatology*. **155**, 901–910 (2024).
28. Liang, H. Wavelet Analysis. In: Jaeger, D., Jung, R. *Encyclopedia of Computational Neuroscience* (2022).
29. Vo, T.Q., Kim, SH., Nguyen, D.H. et al. LSTM-CM: a hybrid approach for natural drought prediction based on deep learning and climate models. *Stochastic Environmental Research and Risk Assessment*. **37**, 2035–2051 (2023).
30. Zhang, X., Yin, Q., Liu, F. et al. Comparative study of rainfall prediction based on different decomposition methods of VMD. *Scientific Reports.* **13**, 20127 (2023).
31. Standard for hydrological information and hydrological forecasting: GB/T 22482—2008 [S]. Beijing: China Standard Press, 2008. (in Chinese)
32. Xiong Y, Zhou JZ, Sun N. et al. Monthly runoff forecasting based on adaptive variational mode decomposition and long short-term memory network[J]. *Journal of Hydraulic Engineering*. **54**, 172-183+198 (2023).
33. Wang, GY., Cheng, DD., Xia, DY. et al. Swarm Intelligence Research: From Bio-inspired Single-population Swarm Intelligence to Human-machine Hybrid Swarm Intelligence. *Machine Intelligence Research*. **20**, 121–144 (2023).
34. Reddy, N., Chitrahalli Lingaraju, M.K., Kotagi Girisha, S., Khatib, M. Runoff Modelling for the Watershed Using Hydrological Model Swat Under Changing Environment. In: Wan, Z. *Water Resources Management and Water Pollution Control, Environmental Science and Engineering* (2024).

Acknowledgments

The research was sponsored by the Natural Science Foundation of Chongqing Municipality

(No.CSTB2022NSCQ-MSX1392), the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202201224, KJQN202201210, KJQN202301209), the Science Technology Planning Project of WanZhou District, Chongqing (No. wzstc20230313), and Chongqing Engineering Research Center of Disaster Prevention & Control for Banks and Structures in Three Gorges Reservoir Area（No. SXAPGC23YB08）.

Author contributions statement

Author Contributions: Xing Wei, as the corresponding author, curated the data and prepared the original draft. Mengen Chen was responsible for the conceptualization and methodology of the study. Yulin Zhou contributed to the visualization and investigation of the study's findings. Jianhua Zou and Libo Ran provided overall supervision and project administration. Ruibo Shi conducted validation and contributed to writing, reviewing, and editing of the manuscript. All authors agree to submit this manuscript and affirm that it has not been published or submitted to any other journal.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.H.

Reprints and permissions information is available at www.nature.com/reprints.

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

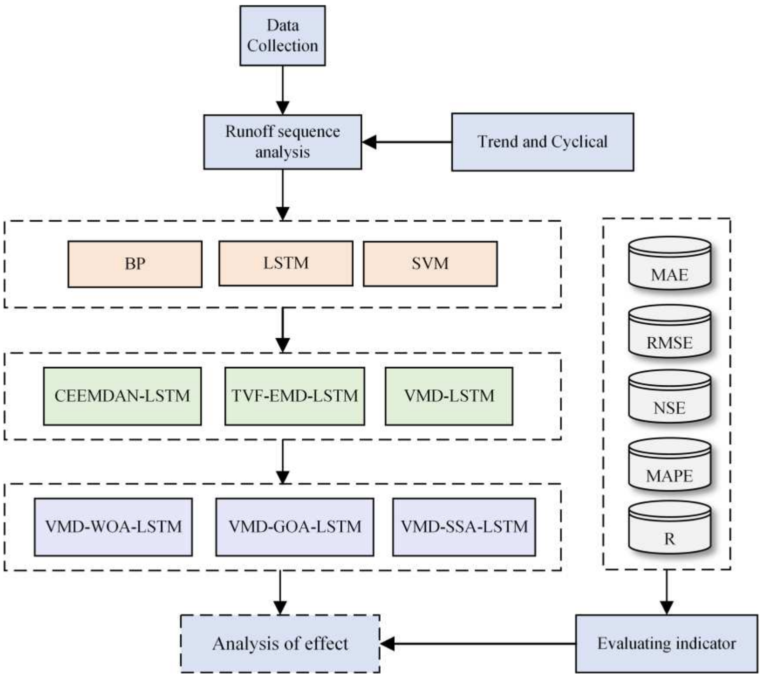

Fig 1. Technology roadmap.

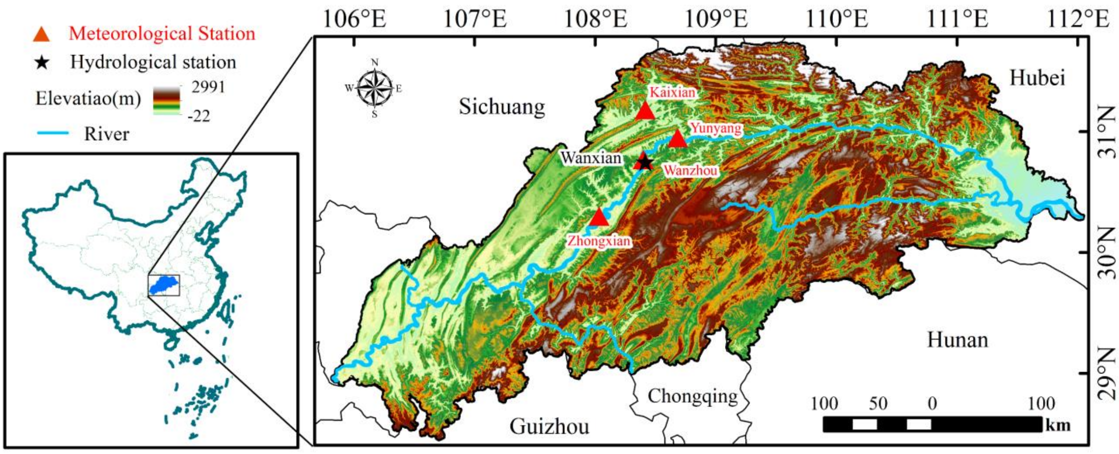

Fig 2. Three gorges reservoir area and wanxian hydrological station in china.

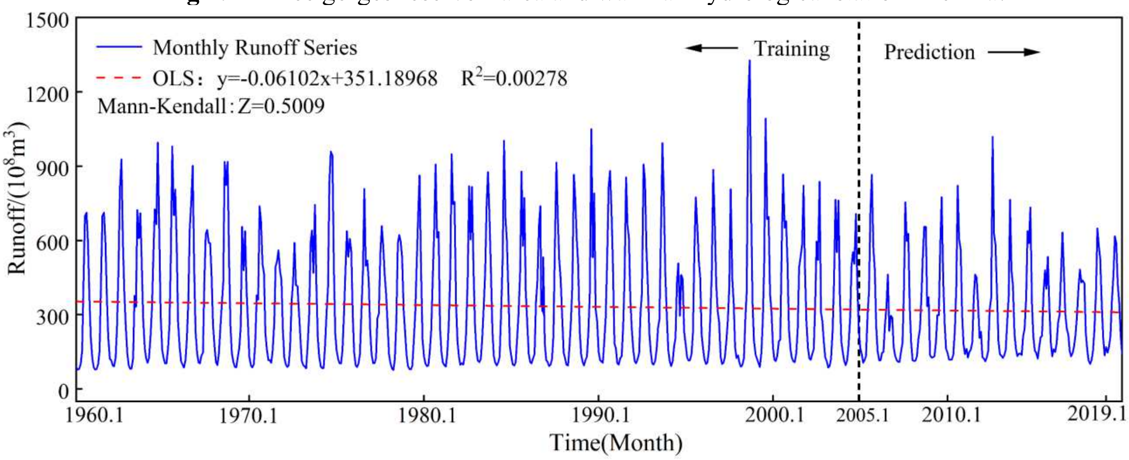

Fig 3. Mann-Kendall trend test of monthly runoff series.

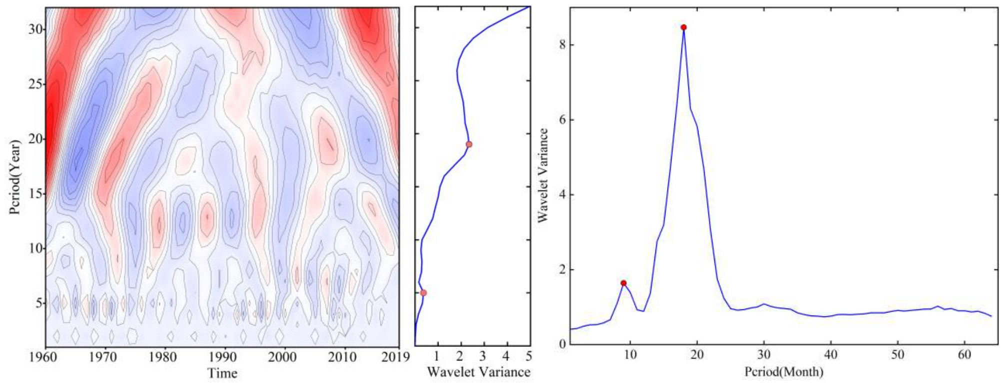

Fig 4. Wavelet analysis diagram of runoff series.

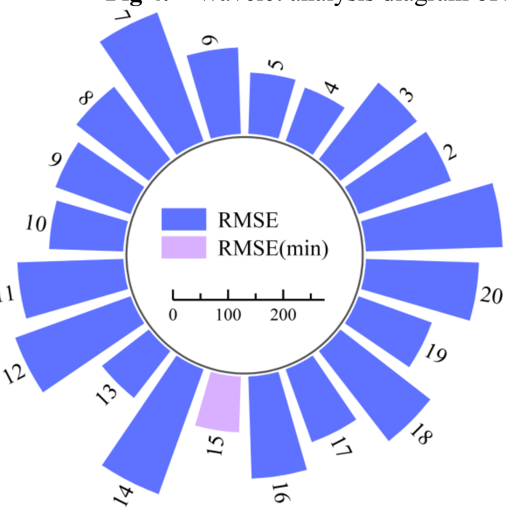

Fig 5. Delay test results.

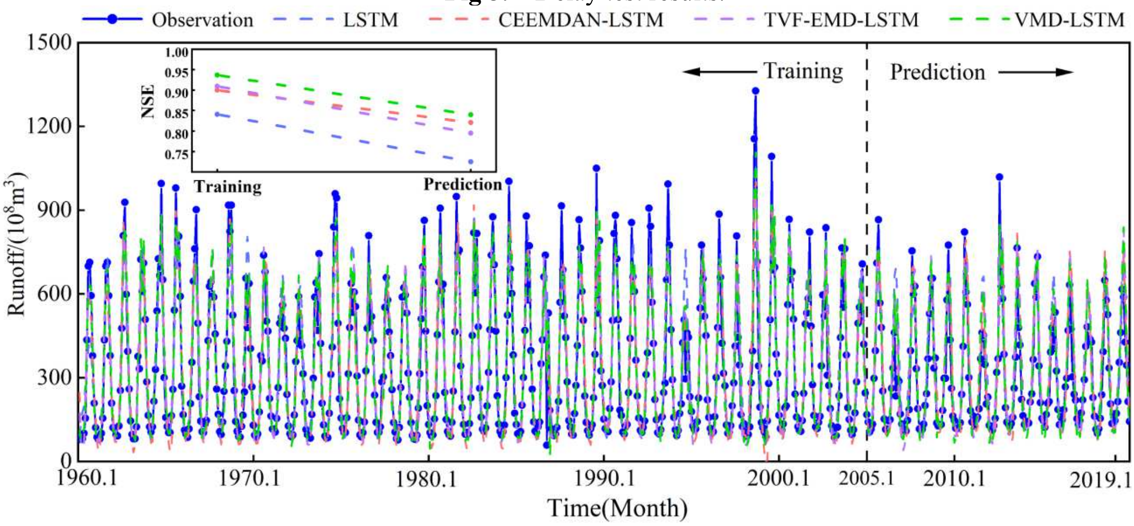

Fig 6. Decomposition technology coupled with machine learning prediction results.

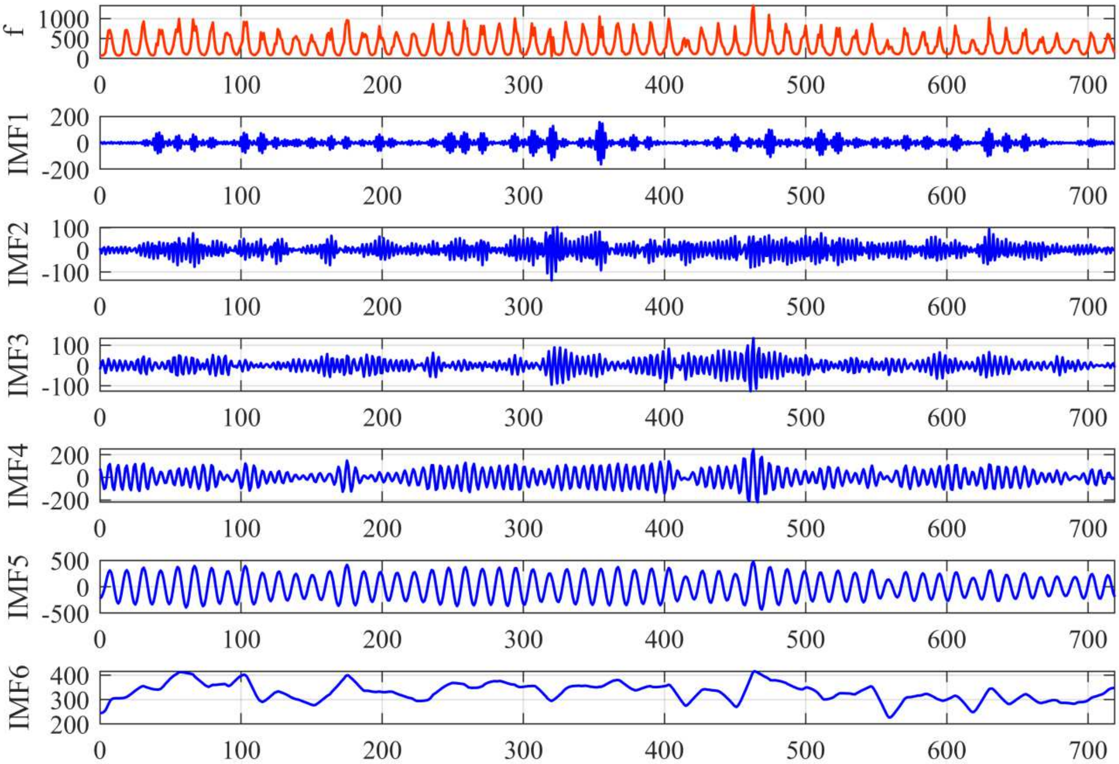

Fig 7. Time domain diagram of runoff sequence decomposition based on VMD.

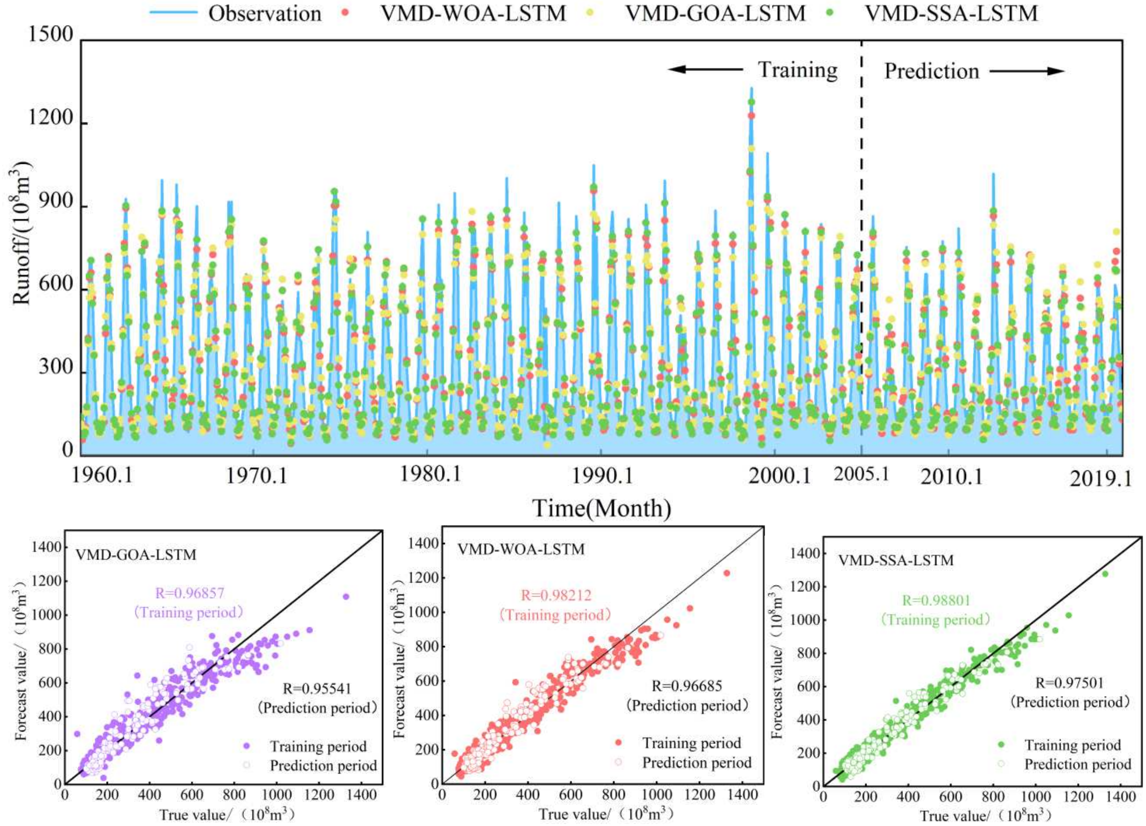

Fig 8. Runoff forecast results.

Table 1. The prediction results of a single prediction model.

	BP		LSTM		SVM	
	Training period	Prediction period	Training period	Prediction period	Training period	Prediction period
MAE	65.9511	70.5446	63.3973	66.3536	71.6384	69.6235
RMSE	105.0917	108.2022	101.3286	99.8668	90.3345	93.6078
NSE	0.82885	0.67733	0.84089	0.72513	0.87354	0.7585
MAPE	21.7201%	24.231%	20.3532%	22.7985%	30.5452%	27.359%
R	0.9109	0.87678	0.91701	0.88152	0.93524	0.89283

Table 2. Runoff forecast results.

Models	Phase	MAE	RMSE	NSE	MAPE	R
VMD-LSTM	Training	42.9235	63.8822	0.93676	14.7763%	0.96787
	Prediction	55.1466	76.155	0.84016	20.6369%	0.9444
VMD-WOA-LSTM	Training	34.6368	48.5316	0.93629	12.6545%	0.98212
	Prediction	37.7831	51.0313	**0.92823**	14.7229%	0.96685
VMD-GOA-LSTM	Training	42.8079	64.1212	0.93629	14.1958%	0.96857
	Prediction	43.2756	61.6045	0.8954	15.4499%	0.95541
VMD-SSA-LSTM	Training	29.181	39.6666	0.97562	11.9244%	0.98801
	Prediction	32.6549	43.4356	**0.948**	12.9048%	0.97509

Table 3. The prediction results of different methods coupling model in the test set.

Models	MAE	RMSE	NSE	R	MAPE
LSTM	66.3536	99.8668	0.72513	0.88125	22.7985%
VMD-LSTM	55.1466	76.155	0.84106	0.9444	20.6369%
VMD-SSA-LSTM	**32.6549**	**43.4356**	**0.948**	**0.97509**	**12.9048%**

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterialoriginaldata.xlsx