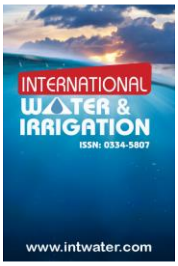


International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

Predicting Monthly Runoff: A BO-CNN-LSTM Approach for Enhanced Water Resource Management

Li Jie[1], Sun Ying Na[1], Cao Yue[2]

[1]School of Hydraulic and Electric Engineering, Heilongjiang University, Harbin 150080, China

[2]Heilongjiang Provincial Institute of Water Resources and Hydropower Survey, Design and Research Institute, Harbin 150080, China

Funding Program: National Social Science Foundation of China (20BS042)

Author Resume: Li Jie (2000-), Male, Jiuquan, Gansu, Master, major research interests in hydrology and water resources.

E-mail: 1751473960@qq.com

Corresponding author: Yingna Sun (1976-), female , Fujin, Heilongjiang, Associate Professor, Ph.D., research interests in hydrological water resources.

E-mail: 26787238@qq.com

Third author: Cao Yue(1980-), man, A native of Huanan, Heilongjiang, Senior engineer, His research interests are hydrology and water resources engineering.

E-mail:617177783@qq.com

Abstract: Monthly runoff prediction is critical for sustainable water resource management, flood control, and drought preparedness. Accurate forecasts assist decision-making and management of water resources, which are critical for human and environmental demands. Previous prediction algorithms frequently lack accuracy due to human-defined hyperparameters. Autonomous hyperparameter tuning is required for improved predictions. Objectives: To improve the accuracy of monthly runoff predictions, this research proposes a BO-CNN-LSTM model that integrates Bayesian optimization (BO) with a convolutional neural network-long short-term memory neural network (CNN-LSTM). Methods/Analysis: The BO algorithm is utilized to enhance the CNN-LSTM model's hyperparameters, which helps to overcome the limits of human-set parameters. Model training and validation were carried out using monthly runoff data from 1956 to 2019 from Lanxi Station in the Hulan River Basin. The model's prediction performance was assessed using root mean square error (RMSE), mean absolute error (MAE), fitting coefficient (R^2), and mean relative error (MAPE), and compared with other previous models. Findings: The results show that the BO-CNN-LSTM model outperforms conventional models in prediction accuracy and reduction of errors. The BO method efficiently enhances hyperparameters, resulting in higher performance metrics. This model gives more precise and dependable monthly runoff forecasts, which are critical for handling the Hulan River Basin's water resources. Novelty/Improvement: By combining BO and CNN-LSTM, the proposed model solves the constraints of previous models, improving prediction accuracy and reducing errors. This innovative technique presents a promising novel approach for monthly runoff prediction, with possible applications in larger hydrological prediction.

Keywords-Hulan River Basin; Bayesian optimization; CNN－LSTM model; runoff projections； Parameter optimization

1.Introduction

Monthly runoff prediction is an important part of hydrological modeling and water resource management [1]. Monthly runoff refers to the amount of water that flows into a river or stream over the course of a month. Accurate prediction of this runoff is critical for efficient water management, which includes reservoir operation, flood control, and drought planning [2]. Particularly, comprehending and forecasting runoff in locations like the Hulan River Basin is especially important because it influences local water resources and agricultural operations [3].

The Hulan River Basin, situated in northeastern China, is a key hydrological basin that contributes to the regional water cycle while also supporting varied ecosystems and human operations [4]. Effective runoff forecasting aids in the efficient management of water resources, the mitigation of flood hazards, and the provision of a consistent supply of water for agricultural and household purposes. By presenting timely and exact forecasts, water managers could make informed decisions to balance water utilization, defend against catastrophic events, and preserve ecological health.

Predicting monthly runoff entails examining historical runoff data and using different methods of modeling to anticipate future runoff amounts. Conventional prediction models, such as statistical approaches and simple machine learning techniques, frequently struggle to capture the intricate temporal patterns and correlations found in runoff data [5]. These models depend on human-defined hyperparameters, which might introduce biases and restrictions that hinder prediction accuracy. There is a pressing need for models that could independently tune hyperparameters are critical for improving the dependability of these predictions.

Several ways have been investigated to improve the accuracy of runoff forecasts. Conventional techniques like linear regression and time series analytics have been used, however, they frequently have low accuracy because of their incapability to deal with complicated, non-linear connections in the data. To solve these constraints, sophisticated methods have been developed, such as artificial neural networks and hybrid models. However many of these models continue to suffer issues like overfitting, computational ineffectiveness, and inferior performance caused by manually tweaked hyperparameters [6].

To address these constraints, this research offers a new BO-CNN-LSTM model that combines Bayesian optimization (BO) with convolutional neural network-long short-term memory neural network (CNN-LSTM). The BO method is used to tune hyperparameters independently, hence avoiding the difficulties related to manual tuning. The CNN component is intended to extract geographical information from runoff data, while the LSTM component captures temporal dependencies, making the model ideal for time series prediction problems. This incorporation attempts to improve the accuracy and dependability of monthly runoff predictions by combining the advantages of both CNNs and LSTMs, along with adjusted hyperparameters.

The primary contributions of this paper are as follows: the introduction of the BO-CNN-LSTM model to enhance monthly runoff prediction accuracy by autonomous hyperparameter tuning; the demonstration of the model's better performance compared to previous techniques utilizing extensive experiments with monthly runoff data from Lanxi Station in the Hulan River Basin; and a thorough assessment of the model's prediction accuracy utilizing metrics like root mean square error (RMSE), mean absolute error (MAE), fitting coefficient (R^2), and mean relative error (MAPE). These contributions highlight the usefulness of the BO-CNN-LSTM model to alleviate the constraints of conventional prediction approaches.

This paper aims to improve the accuracy of monthly runoff forecasts by introducing a unique model that solves the inadequacies of previous techniques. This study has implications for several aspects of water resource management, such as reservoir operations, flood forecasts, and water supply planning. The proposed model provides a more strong and dependable technique for forecasting monthly runoff, offering useful insights for enhancing water management tactics.

The rest of the paper is structured as follows: Section 2 discusses related works, including previous models and their constraints. Section 3 describes the study methodologies and assessment indicators, containing descriptions of convolutional neural networks (CNNs), long short-term memory networks (LSTMs), Bayesian optimization (BO), the BO-CNN-LSTM model, and the evaluation metrics employed. Section 4 describes example applications, such as the experimental surroundings, a summary of the research area, and data sources and processing techniques. Section 5 provides and examines the findings, with a concentration on parameterization and prediction accuracy assessment. At last, Section 6 summarizes the paper and discusses future research directions.

International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

2.Objectives

Monthly runoff prediction has been a major emphasis in hydrological research, with a variety of approaches implemented to improve prediction accuracy. Singh et al. [7] used an incorporated statistical-machine learning strategy for runoff forecast, including models like multiple linear regression (MLR), multiple adaptive regression splines (MARS), support vector machines (SVM), and random forests.

Their analysis found that RF beat other models in terms of accuracy for daily runoff predictions. Likewise, Li et al. [8] investigated many machine-learning models for predicting monthly runoff in the upper Yangtze River. Their technique included several global circulation indicators and surface meteorological indexes, indicating that the LSTM_Copula model was better in forecasting accuracy when compared to univariate models.Sibtain et al. [9] developed an alternative strategy, introducing a three-stage hybrid model that combines support vector machines (SVM) with signal decomposition methods, CEEMDAN and VMD. This model successfully tackled the nonlinear dynamics of runoff time series, outperforming various existing hybrids and independent models. Anaraki et al. [10] used several machine learning algorithms to estimate rainfall runoff in the Wadi Ouahrane Basin, Algeria. Their study showed the efficiency of the KNN-GTO model, which merged the K-nearest neighbor with a gorilla troop optimizer, attaining greater performance than previous models.Swagatika et al. [11] created a hybrid deep learning model called FT-LSTM to enhance runoff forecasting accuracy in the Brahmani River basin. Their model outperformed previous deep learning algorithms in terms of capturing temporal correlations and fitting runoff data. Dehbalaei et al. [12] also proposed a linear-nonlinear hybrid special model, which combines a multivariate linear stochastic model and a multilayer nonlinear machine learning model. This model predicted runoff more accurately than classic linear and nonlinear models. Clark et al. [13] compared deep learning models such as LSTM to conceptual models for monthly runoff forecasts across Australia. Their results indicated that LSTM models usually matched or exceeded the effectiveness of conceptual models, particularly in large catchments.Li et al. [14] studied medium and long-term runoff prediction approaches utilizing vast meteorological data and machine learning techniques. They investigated several feature selection techniques and ensemble learning models and found that Extreme Gradient Boosting (XGB) provided superior predicting skills than Random Forest (RF).Despite progress in these models, numerous restrictions remain. Several current models, including those stated above, depend largely on manual hyperparameter tuning, which can result in inferior efficiency and accuracy. Additionally, conventional models frequently struggle with the non-linear features and temporal relationships found in runoff data. There is a gap in the thorough incorporation of spatial feature extraction and temporal modeling inside a unified framework.

To tackle these challenges, the proposed BO-CNN-LSTM model is proposed. This model incorporates Bayesian optimization (BO) to automatically tune hyperparameters, increasing accuracy and decreasing biases related to manual tuning. The convolutional neural network (CNN) component extracts spatial features from runoff data when the long short-term memory (LSTM) network collects temporal dependencies. By merging these aspects with BO, the BO-CNN-LSTM model intends to improve prediction accuracy and also address existing approach shortcomings. This method provides a more robust strategy for modeling intricate runoff dynamics, creating a useful improvement in the area of hydrological prediction.

3.Research methods and evaluation indicators

3.1 CNN

CNN models have demonstrated excellent feature extraction capabilities when processing time series data. Through convolutional operations, CNNs can effectively capture key local patterns and features in runoff data [15]. For runoff data, these local features may include important information such as seasonal variations, periodic fluctuations, peaks, and troughs. The sliding action of the convolutional kernel over different time windows enables the model to accurately extract these local features, which in turn leads to a deeper understanding of the changing patterns of the data. In the traditional CNN structure, the model consists of an input layer, a convolutional

International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

layer, a pooling layer, a fully connected layer, and an output layer alternately. Among these, the convolutional layer plays a central role, and the convolutional kernel inside it focuses on extracting the intrinsic features of the data. The working of the convolution kernel can be described by the following equation:

$$c_i = f(w \otimes x_{i:i+h-1} + b) \quad (1)$$

where: denotes the convolution operation; denotes the convolution operation i from i+h-1 data; denotes the feature data obtained by the above convolution; W denotes the convolution kernel parameters; b denotes the bias matrix; h denotes the height of the convolution kernel.

To reduce the computation and extract more effective features from the raw data, CNN uses the Inception structure. The structure is shown in Fig. 1. The Inception structure captures features at different scales by applying 1x1, 3x3, and 5x5 convolutional kernels in parallel.1x1 convolution is used to reduce the dimensionality and reduce the amount of computation, while 3x3 and 5x5 convolution are used to capture both local and broader global features. This multi-scale feature extraction method can comprehensively capture the features of runoff data. In addition, the parameter-sharing nature of the convolution operation helps to reduce computational and storage requirements and improve model efficiency.

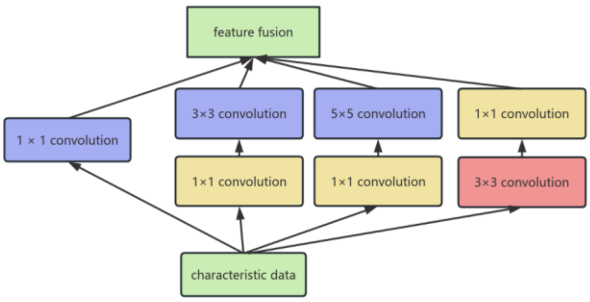


Figure 1. Inception structure diagram

3.2 LSTM

LSTM neural network is a variant of recurrent neural network used to process sequential data, which can better capture and process long-term dependencies by introducing gating mechanisms of input gates, forgetting gates, and output gates. This makes LSTM perform well in processing long sequence data and tasks that require memorization of long-term information. The LSTM structure is shown in Fig.

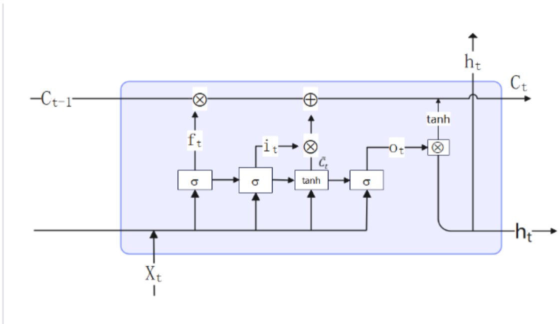


Fig.2. Structure diagram of a recurrent unit of LSTM network

where it and together determine the update of information in the memory cell, while determining which information in the hidden state will be output. The specific formula is as follows:

$$f_t = \sigma(w_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

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

$$\tilde{c_t} = \tanh(h_{t-1} \cdot w_c + x_t \cdot w_c + b_c) \quad (4)$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c_t} \quad (5)$$

$$o_t = \sigma(h_{t-1} \cdot w_0 + x_t \cdot w_0 + b_0) \quad (6)$$

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

Where:w_f、w_i、w_c、w_0 is the weight matrix;b_f、b_i、b_c、b_0 is the bias vector; h_{t-1}、h_t denotes the input of the previous cell and the output of this cell;c_{t-1}、$c_t$denotes the information state; $\tilde{c_t}$is the information state through the input gate; and σis the sigmoid function.

3.3 BO

Bayesian optimization (BO), based on Bayes' theorem see Eq. (8), is an iterative optimization method for global optimization problems. It models the objective function by building a probabilistic model and continuously uses this model to guide the optimization process to find the optimal solution. The basic idea is to consider the objective function as a black-box function whose exact form and properties are unknown. By constantly selecting new sample points for evaluation, Bayesian optimization attempts to find the global optimal solution with a limited number of samples by inferring the overall shape of the objective function from the existing sample points and the evaluation results of the objective function, and predicting which sample points may be the most promising.$\varepsilon_t$is the fitting error;

$$P(f|D) = \frac{P(D|f)P(f)}{P(D)} \quad (8)$$

Eq. $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_t, y_t)\}$

Eq.$f = f(x)$is the distribution of the unknown objective function;D denotes the set of observed parameters and observations;x_tIndicates the observed parameter;$y_t = f(x_t) + \varepsilon_t$Indicates an observed value;

The core parts of the Bayesian optimization algorithm are the probabilistic agent model and the collection function. In this paper, the Gaussian process is selected as the probabilistic agent model of the Bayesian optimization algorithm, and the EI function is selected as the acquisition function of Bayesian optimization. The principle and execution steps of its Bayesian optimization refer to reference [16].

3.4 BO－CNN－LSTM

The model of Bayesian optimization convolutional neural network combined with long and short-term memory neural network uses a Bayesian optimization algorithm to automatically search for the optimal model parameter configuration first, and then construct the BO-CNN-LSTM model based on the searched optimal parameter configuration, which consists of a convolutional neural network (CNN) model to extract the original runoff sequences in the temporal features, and then input the feature data into the long and short-term memory neural network with strong prediction fitting ability for prediction. The prediction flow of its combined prediction model is shown in Fig.3.

The steps of the BO-CNN-LSTM model to predict the runoff of the Hulan River are as follows:

Step 1: Obtain historical runoff data from 1956-2019. Preprocess the data by handling missing values through mean imputation, removing outliers using the interquartile range (IQR) method. Then divide this preprocessed

International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

data chronologically into a training set and test set, and linearly normalized. Specifically, earlier data was used for training and later data for testing, establishing a clear distinction between the training and validation phases.

Step 2: Put the training set and test set into the CNN model for feature extraction.

Step 3: Input the extracted feature data into the LSTM model for prediction.

Step 4: Using the objective function as a Bayesian optimization process, it's calculated using the measured and predicted values.

Step 5: Determine whether the maximum number of iterations is reached, if not, build a probabilistic agent model based on the maximum value and continue the loop until the maximum number of iterations is reached.

Step 6: If the maximum number of iterations are completed, return the optimal hyperparameter combination. The Bayesian Optimization algorithm optimizes specific hyperparameters, which include:

- Learning rate (α): The range of [0.0001, 0.01].
- Batch Size: The range of [16, 128].
- Number of Epochs: The range of [50, 500].
- Number of Filters in Each CNN Layer: The range of [32, 256]
- Kernel Size for Convolutional Layers: The range of [2, 5]
- Dropout Rate: The range of [0.1, 0.5]
- Number of LSTM units: The range of [100, 500].

Step 7: Use the optimal hyperparameter combination obtained by the Bayesian optimization algorithm to establish a prediction model.

Step 8: Input the feature data extracted in step (2) into the most effective LSTM model for prediction.

Step 9: Calculate the performance index to measure the model prediction performance.

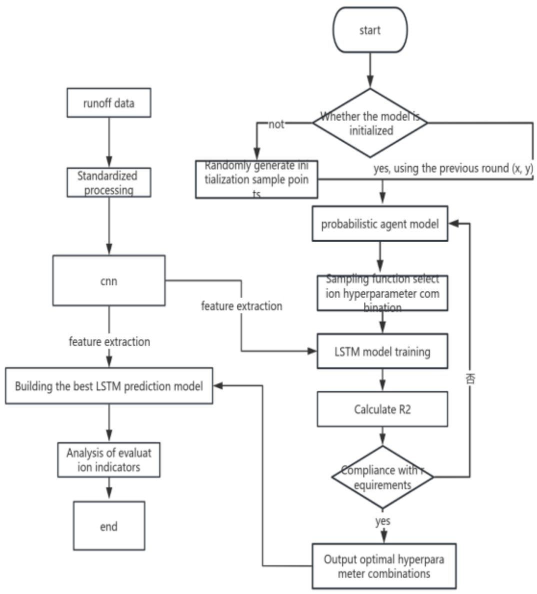


Fig.3 BO—CNN—LSTM model running flow chart

3.5 Evaluation indicators

In this paper, four metrics are selected to evaluate the model performance, namely, mean absolute error (MAE), mean square error (RMSE), correlation coefficient (R2), and mean relative error (MAPE), which are given by the formula:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{N}\left|Q_i^0 - Q_i^s\right| \quad (9)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(Q_i^0 - Q_i^s)^2} \quad (10)$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left[(Q_i^0-\overline{Q^0})(Q_i^s-\overline{Q_i^s})\right]}{\sqrt{\sum_{i=1}^{n}\left[(Q_i^0-\overline{Q^0})^2(Q_i^s-\overline{Q_i^s})^2\right]}}\right)^2 \quad (11)$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|Q_i^s-Q_i^0|}{Q_i^0} \quad (12)$$

Where: Q_i^0 is the measured runoff volume; Q_i^s is the predicted runoff volume; $\overline{Q^0}$ is the average value of the measured runoff volume; $\overline{Q_i^s}$ is the average value of the predicted runoff volume; N is the total number of runoff data;

4.Example applications

4.1 Experimental Environment Construction

In this experiment, the environment of python3.11.0, pytorch2.0.1+cu117 was used, the Bayesian optimization was performed by calling the optuna library of python, and the CNN-LSTM neural network model construction was performed by calling the torch library of python.

4.2 Overview of the study area

Hulan River originates from Taiping Gully between three big mountains and Taiping Ridge in the northeast of Tieli City at the western foot of Xiaoxinganling, and flows southwestward through Tieli City, Qing'an County, Beilin District of Suihua City, Wangkui County, Qinggang County, etc., and then turns to flow southward in the northern part of Lanxi County, and then turns to flow southeastward in the southern part of Lanxi County, and then finally merges into the Songhua River to the south of Hulan District of Harbin City. Hulan River has a total length of 506km, a watershed area of 31424, and an average annual runoff of 4.44 billion. The scope and location of the watershed are shown in the figure. The inter-annual variation of hydrological elements in the region is large, and the variation is especially significant in the downstream area. As the main control hydrological station, the downstream of Hulan River Basin, Lanxi Hydrological Station has a direct influence on the social and economic development of the Hulan River Basin and the protection of the ecological environment. Therefore, the monthly runoff prediction study based on the measured data of Lanxi Hydrological Station in the lower reaches of the Hulan River has good representativeness and practical significance.

International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

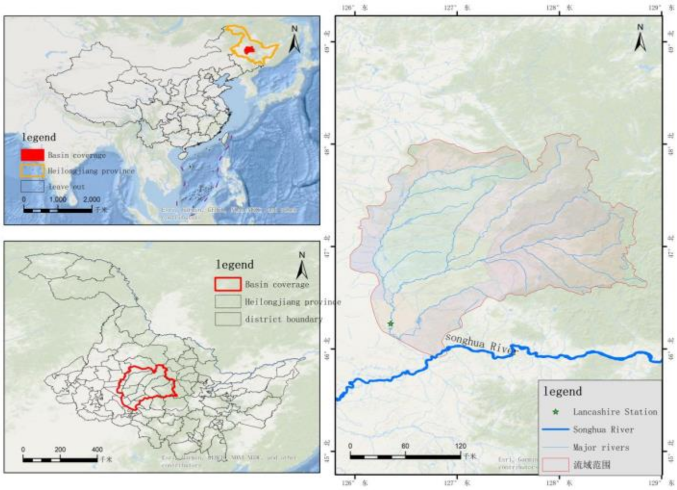


Fig.4 Watershed extent map

4.3 Data sources and processing

The monthly runoff data in this paper were provided by the Heilongjiang Provincial Institute of Water Resources and Electric Power Survey and Design, whose time range is January 1956-December 2019 for 768 months. According to the interannual variation rule, the first 12 months were selected as inputs to the model to predict the 13th month. The runoff data of 768 months from 1956 to 2019 were divided into a training set and a test set, whose first 54 years of data were used as model training, and the last 10 years of data were used as model validation.

5.Results and analysis

5.1 Parameterization

In this paper, the Bayesian optimization parameters are selected to optimize seven hyperparameters of the LSTM model, which will be passed as inputs to the objective function objective(trial), which uses the hyperparameters to configure the architecture of the model and the training process, and returns a performance metric (R^2) as the value of the objective function. The number of iterations is set to 200.

The LSTM model uses a two-hidden-layer structure with an equal number of neurons in both hidden layers, RMSE is selected for the loss function, Adam is selected for the optimizer, and the momentum parameter is set to (0.9, 0.999). The remaining parameter sizes of the LSTM model are derived from Bayesian optimization. The range of hyperparameters is determined empirically or experimentally, the learning rate is set to a logarithmic space from 10-4 to 10-1, the Dropout ratio is a floating point number between 0.1 and 1, the number of neurons is 10-55, the number of neural network layers is 1-8, the batch size is 20-250, the number of output channels of convolutional layer is 1-24, and the number of iterations is 1000-8000. Table 1 shows the optimization of the hyperparameters results after the optimization.

Table 1 Hyperparameter optimization results

hyper parameterization	Optimized value
learning_rate	0.00012893402557556514
drop_rate	0.43261164357221416
hidden_size	15
num_layers	7
batch_size	156
conv_output	21
epochs	3981

5.2 Evaluation of Prediction Accuracy

To verify the prediction effect and accuracy of the BO-CNN-LSTM model proposed in this paper, CNN-LSTM, LSTM, and GA-BP models are constructed at the same time, and the model accuracy is selected by the four evaluation indexes selected above, respectively. Evaluation, to compare and validate under the same conditions, all use standardized processed data input models, CNN-LSTM and LSTM models determine the hyperparameters of the model through manual experience, the structure of the CNN module and parameter settings in its CNN-LSTM are the same as those of the BO-CNN -LSTM model is the same, the LSTM module is set with two hidden layers and 500 neurons per layer, and the GA-BP model optimizes the weights and thresholds of the BP model by genetic algorithm. Its hidden layer activation function is set to sigmoid, output layer activation function is set to purelin, and 5 hidden layer neurons. All models were trained 1000 times with a learning rate of 0.001. The results of the different models in the test set are shown in Fig. 5, and the results of the evaluation indexes calculated for each model are shown in Table 2. From the comparison of the measured runoff and the predicted runoff in the test set in Fig. 5, it can be seen that the accuracy of the test set of the BO-CNN-LSTM model is higher than that of the CNN -LSTM with high accuracy, close to the fluctuation trend of the original monthly runoff time series, and similar to the predicted peak and minimum values, and similar to the fluctuation trend of the measured runoff time series.

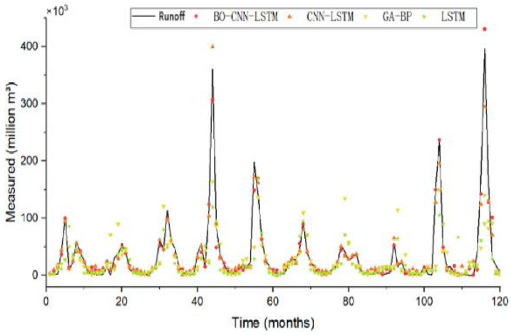


Fig.5 Comparison of measured and predicted runoff in the test set

Table 2 Calculation results of evaluation indexes of each model

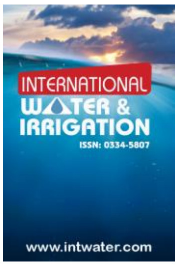
mould	point	MAE 10^8m^3	RMSE 10^8m^3	R^2	MAPE
BO－CNN－LSTM	training set	0.32	0.96	0.9823	12.74%
	test set	0.4	1.09	0.97	13.86%
CNN－LSTM	training set	0.73	1.27	0.9236	15.24%
	test set	0.998	1.8	0.916	16.44%
GA－BP	training set	1.54	1.25	0.8623	23.60%
	test set	1.72	1.83	0.8471	25.32%
LSTM	training set	1.43	0.52	0.8762	18.33%
	test set	1.67	0.73	0.85	21.62%

International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

Analyzing Table 2, the MAE, RMSE, R2, and MAPE of the training set of the BO-CNN-LSTM model showed significant improvement in the training period accuracy compared to that of the CNN-LSTM model by 0.41 × 108 m^3, 0.31 × 108 m^3, 0.0587, 2.5%, and the MAE, RMSE, R2, and MAPE of the test set of the BO-CNN-LSTM model have improved the accuracy compared to that of the CNN-LSTM model, respectively, by 0.598×108 m^3, 0.71×108 m^3, 0.054, 2.58%, and the prediction accuracy of BO-CNN-LSTM model is slightly better than that of CNN-LSTM model. The single LSTM model and the BP neural network (GA-BP) model optimized based on the genetic algorithm deviated more from the measured values, and their validation period MAPE was even 21.62% and 25.32%. The above results show that based on Bayesian inference and Gaussian process regression, the optimal hyper-parameter combinations can be found quickly within a limited number of evaluations, avoiding the problem of poor model accuracy due to the human subjective-energetic setting of hyper-parameters in traditional methods such as grid search.

The runoff at Lanxi station was small in January, February, and March of each decade, and then the weather turned warmer, snow and ice melted and precipitation increased from April onwards, and the runoff peaked in July, August, and September, after which the temperature kept decreasing and precipitation plummeted, and then fell back to the low value in December. Figure 6 reflects the number of different models that satisfy the forecast accuracy (MAPE ≤ 20%) in each month in the monthly runoff prediction in the validation period of 10 years, from which it can be seen that the different models have poor prediction results in the runoff prediction of January, February and March, and the prediction pass rate of the BO-CNN-LSTM model in January, February and March is 63.3%, CNN-LSTM prediction pass rate 60%, LSTM prediction pass rate 63.3% and GA-BP prediction pass rate 56.7%. In the runoff prediction of different models in July, August, and September, the BO-CNN-LSTM model achieves a pass rate of 93.3%, which is much higher than the 73.3% of the CNN-LSTM model, the 66.67% of the LSTM model and the 60% of the GA-BP model. BP model's 60%. The reason for the above results may be that the Lanxi station is in the winter icing period in January, February, and March, and the CNN model cannot effectively extract the spatial features related to the icing period. While in the BO-CNN-LSTM model, the BO parameter update will be more inclined to the fitting of peaks.

The BO-CNN-LSTM model in which the CNN uses inception structure, operates multiple convolutional kernels of different sizes in parallel and then splices their outputs, which increases the nonlinear capability of the model to the point of extracting effective time series features, thus increasing the predictive capability of the model. The above-analyzed results verify the high accuracy and superiority of the model proposed in this study. Therefore, it can be well applied to runoff prediction in the Hulan River Basin compared to the CNN-LSTM model.



International Water and Irrigation
Volume 43 Issue 2 (2024)
ISSN: 0334-5807
Pages 34 - 46

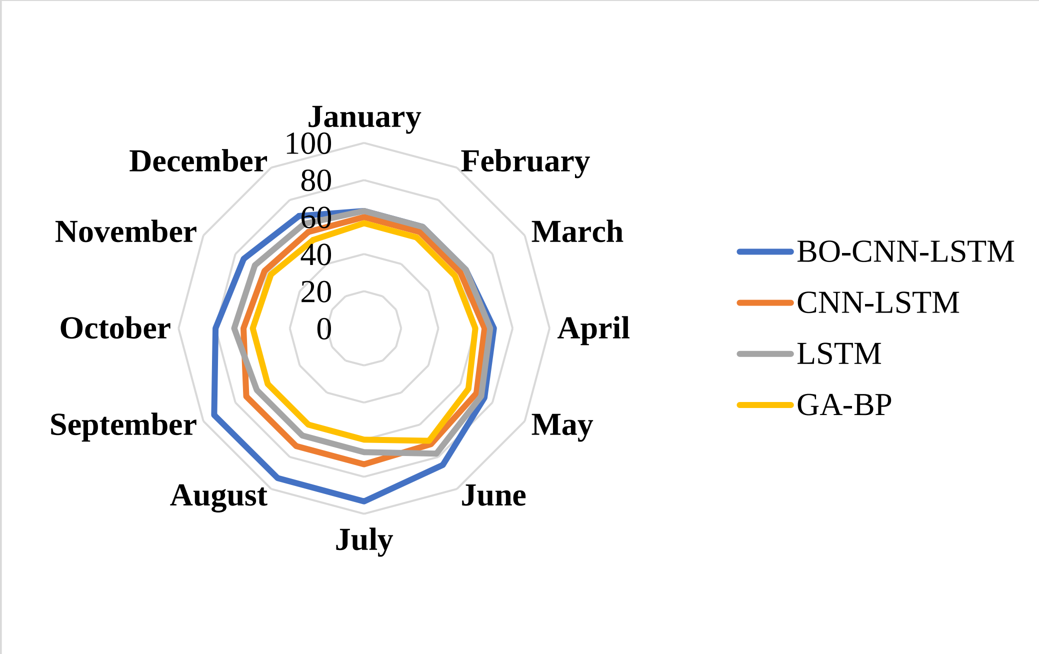


Fig.6 Forecast qualified quantity for different models in each month

The proposed BO-CNN-LSTM model achieves the best outcomes by combining BO and the hybrid CNN-LSTM architecture. Bayesian Optimization efficiently tunes hyperparameters, resulting in higher model performance without substantial manual tuning, which is frequently error-prone and inefficient. The CNN component's inception structure enables the extraction of diverse and complicated features by processing many convolutional operations in parallel, boosting the model's capability to grasp nuanced patterns in the input. This multi-kernel technique, paired with the temporal abilities of LSTM, yields a model that excels at detecting both spatial and temporal correlations in runoff data. The robust performance throughout multiple months, notably during high runoff seasons, demonstrates its capacity to generalize and adapt to varied seasonal variations, making it more trustworthy and accurate than standard models like CNN-LSTM, LSTM, and GA-BP.

The proposed BO-CNN-LSTM model surpasses previous CNN-LSTM, LSTM, and GA-BP models owing to its enhanced hyperparameter optimization via Bayesian Optimization, which provides the best configuration for model accuracy and effectiveness. The inception structure within the CNN layers allows the extraction of diverse and intricate spatial properties, considerably increasing the model's capability to catch complicated patterns in runoff data. This, paired with the LSTM's capacity to model temporal dependencies, results in a more extensive and accurate prediction ability. In comparison, the CNN-LSTM and LSTM models depend on manually defined hyperparameters and do not achieve a similar level of spatial feature extraction, whereas the GA-BP model suffers from the data's nonlinear and sequential structure. The superior architecture of the BO-CNN-LSTM leads to higher accuracy, particularly during peak and low runoff periods, as proven by its improved performance metrics across several evaluation indices.

6.Conclusions and outlook

1. the BO-CNN-LSTM combined model automatically finds the network hyperparameters through Bayesian optimization, which makes up for the problem of unstable model effect and is difficult to optimize due to the traditional manual setting of hyperparameters. the BO algorithm quickly finds the optimal solution based on the finite samples, which greatly improves the model's accuracy and performance in the task of predicting monthly runoff.

2. By comparing with the CNN-LSTM model in the time series prediction of monthly runoff in Hulan River Basin, the MAE and RMSE of the BO-CNN-LSTM model in the validation period decreased by $0.59\times108\ m^3$, $1.01\times108\ m^3$, and the R2 increased by 0.0446, and the R2 increased by 0.0446, and the R2 increased by 0.0446, and the R2 increased by 0.0446. 0.0446, verifying that it better learns and fits the nonlinear correlation of the data. The model fully utilizes CNN to extract multiscale features and LSTM to capture temporal dependence, which shows the superiority of this kind of complex temporal prediction.

3. The BO-CNN-LSTM model fully integrates the advantages of deep learning and adaptive algorithms by learning the input spatial structure through the Inception module, memorizing the sequence trend by LSTM, and finding the optimal model automatically by Bayesian optimization. It applies the important algorithmic ideas of hydrological prediction tasks to provide an efficient and generalized method for runoff prediction.

Refrences

[1] Ditthakit, P., Pinthong, S., Salaeh, N., Weekaew, J., Tran, T. T., & Pham, Q. B. (2023). Comparative study of machine learning methods and GR2M model for monthly runoff prediction. Ain Shams Engineering Journal, 14(4), 101941.

[2] Samantaray, S., Das, S. S., Sahoo, A., & Satapathy, D. P. (2022). Monthly runoff prediction at the Baitarani river basin by support vector machine based on Salp swarm algorithm. Ain Shams Engineering Journal, 13(5), 101732.

[3] Su, Q., Dai, C., Zhang, Z., Zhang, S., Li, R., & Qi, P. (2023). Runoff Simulation and Climate Change Analysis in Hulan River Basin Based on SWAT Model. Water, 15(15), 2845.

[4] Liu, Y., Li, H., Cui, G., & Cao, Y. (2020). Water quality attribution and simulation of non-point source pollution load flux in the Hulan River basin. Scientific Reports, 10(1), 3012.

[5] Wang, W. C., Cheng, Q., Chau, K. W., Hu, H., Zang, H. F., & Xu, D. M. (2023). An enhanced monthly runoff time series prediction using extreme learning machine optimized by Salp swarm algorithm based on time-varying filtering based empirical mode decomposition. Journal of Hydrology, 620, 129460.

[6] Yan, L., Lei, Q., Jiang, C., Yan, P., Ren, Z., Liu, B., & Liu, Z. (2022). Climate-informed monthly runoff prediction model using machine learning and feature importance analysis. Frontiers in Environmental Science, 10, 1049840.

[7] Singh, A. K., Kumar, P., Ali, R., Al-Ansari, N., Vishwakarma, D. K., Kushwaha, K. S., ... & Heddam, S. (2022). An integrated statistical-machine learning approach for runoff prediction. Sustainability, 14(13), 8209.

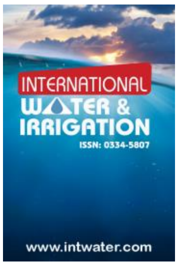
[8] Li, X., Zhang, L., Zeng, S., Tang, Z., Liu, L., Zhang, Q., ... & Hua, X. (2022). Predicting monthly runoff of the upper Yangtze River based on multiple machine learning models. Sustainability, 14(18), 11149.

[9] Sibtain, M., Li, X., Nabi, G., Azam, M. I., & Bashir, H. (2020). Development of a Three-Stage Hybrid Model by Utilizing a Two-Stage Signal Decomposition Methodology and Machine Learning Approach to Predict Monthly Runoff at Swat River Basin, Pakistan. Discrete Dynamics in Nature and Society, 2020(1), 7345676.

[10] Anaraki, M. V., Achite, M., Farzin, S., Elshaboury, N., Al-Ansari, N., & Elkhrachy, I. (2023). Modeling of monthly rainfall-runoff using various machine learning techniques in Wadi Ouahrane Basin, Algeria. Water, 15(20), 3576.

[11] Swagatika, S., Paul, J. C., Sahoo, B. B., Gupta, S. K., & Singh, P. K. (2024). Improving the forecasting accuracy of monthly runoff time series of the Brahmani River in India using a hybrid deep learning model. Journal of Water and Climate Change, 15(1), 139-156.

[12] Dehbalaei, F. N., Azari, A., & Akhtari, A. A. (2023). Development of a linear–nonlinear hybrid special model to predict monthly runoff in a catchment area and evaluate its performance with novel machine learning methods. Applied Water Science, 13(5), 118.



[13] Clark, S. R., Lerat, J., Perraud, J. M., & Fitch, P. (2024). Deep learning for monthly rainfall-runoff modeling: a large-sample comparison with conceptual models across Australia. Hydrology and Earth System Sciences, 28(5), 1191-1213.

[14] Li, Y., Wei, J., Wang, D., Li, B., Huang, H., Xu, B., & Xu, Y. (2021). A medium and long-term runoff forecast method based on massive meteorological data and machine learning algorithms. Water, 13(9), 1308.

[15] Zhao C. Capsule Network with Inception Structure for Image Classification[C]//Proceedings of the 4th International Conference on Advanced Information Science and System. 2022: 1-5.

[16] ZHOU Yanming, SHI Peng, QU Simin, et al. Optimization and application of Xin'anjiang model parameters based on Bayesian optimization algorithm[J]. Hydropower Energy Science,2019,37(5):6-9,127.

[17] XU Dongmei, LIAO Andong, WANG Wenchuan. Monthly runoff prediction based on VMD-EEMD-CNN-LSTM hybrid model[J]. Water Resources Planning and Design,2023,(07):57-63.

[18] XU Dongmei, WANG Yiyang, WANG Wenchuan. Annual runoff prediction based on Bayesian optimization algorithm for long and short-term memory neural network model[J]. Hydropower Energy Science,2022,40(12):42-46.

[19] ZHU Feilin, CHEN Jiabi, ZHANG Mi, XU Xiangrong, ZHONG Pingan. An integrated approach to medium- and long-term runoff prediction based on multi-model fusion[J]. Hydropower Generation,1-10.

[20] ZHAO Kun,FU Haiyan,LI Wei et al. Progress of watershed hydrological modeling[J]. Modern Agricultural Science and Technology,2009(23):267-270.

[21] WANG Xiu-Jie, QIAO Hong-Fei, ZENG Yong-Hong, et al. A combined model for daily runoff prediction based on TVF-EMD, GRA, and LightGBM[J]. Water Resources Conservation,2023,39(05):135-142+151.

[22] Dang Chiheng, Zhang Hongbo, Chen Keyu et al. Application of long and short-term memory neural networks in precipitation-runoff simulation in seasonal snowmelt basins[J]. Journal of North China University of Water Resources and Hydropower (Natural Science Edition),2020,41(05):10-18+33.